

eman ta zabal zazu



Universidad  
del País Vasco

Euskal Herriko  
Unibertsitatea

# Hierarchical modelling of patient-reported outcomes data based on the beta-binomial distribution

Author:

**Josu Najera-Zuloaga**

Supervisor(s):

**Dae-Jin Lee**

**Inmaculada Arostegui**

A thesis submitted for the degree of Philosophy Doctor in  
Mathematics & Statistics

November 2017





# Hierarchical modelling of patient-reported outcomes data based on the beta-binomial distribution

Author:  
**Josu Najera-Zuloaga**

Supervisor(s):  
**Dae-Jin Lee**  
**Inmaculada Arostegui**

A thesis submitted for the degree of Philosophy Doctor in  
Mathematics & Statistics

November 2017

This research was supported by the Basque Government through the BERC 360 2014-2017 and the Department of Education, Language Policy and Culture of the Basque Government IT-620-13 programs, by the Spanish Ministry of Economy and Competitiveness MINECO and FEDER: BCAM Severo Ochoa excellence accreditation SEV-2013-0323, MTM2013-40941-P and MTM2016-74931-P, and by grants from the Instituto de Salud Carlos III and by the European Regional Development Funds (RD12/0001/0001) - through the thematic networks - REDISSEC (Red de Investigación en Servicios de Salud en Enfermedades Crónicas).



*Nire eredu izan diren eta direnei,  
Osaba Aitorreri.*

Agurtu ditut lagunak,  
agurtu dut nire osaba,  
gazte nerabe aldiak  
agurtuak ditut jada.  
Txikitako maitasuna,  
gurasoen besarkada,  
izan zirena zirenez  
gaurkoan garena gara.  
Denak alde egiten digu  
baina agurra ezin da laga,  
izan ere bizitzaren  
susterren oinarria da.  
Gaur beste garai bat doa  
nire oroiminetara,  
baina amaiera oro  
hasiera ere bada.



---

## ACKNOWLEDGEMENTS

First of all, I would like to express my most sincere thanks to my supervisors, Dr. Inmaculada Arostegui and Dr. Dae-Jin Lee, for their constant support and encouragement, without which this thesis would have never been possible. Inma, zu izan zara lan hau argitaratu egin izanaren errudun handienetariko bat, hasiera batetik nigan sinistuz, behar nuen aukera eskainiz eta urteotan edukitako traba guztiak aholku zein lanaren bitartez apalduz, eskerrik asko bihotzez. Dae-Jin, gracias por todo el tiempo invertido en mí, por las lecciones, no sólo de estadística, y por la paciencia que has tenido. Nos conocimos ‘casualmente’ al principio de este camino que se ha hecho llamar ciencia, empezaste siendo mi director y terminas como un buen amigo.

Gracias también al Prof. Vicente Núnñez-Antón por la atención ofrecida y por facilitarme la oportunidad de realizar estancias dirigidas por investigadores de alto prestigio. Likewise, I would like to thank Prof. Yudi Pawitan from Karolinska Institutet in Stockholm and Prof. Jianxin Pan from Manchester University for the given research opportunity and the received support while I stayed with them. Asimismo, me gustaría agradecer al Dr. Cristobal Esteban del Hospital Galdakao el haberme facilitado los datos que han servido de motivación para esta tesis.

Nire eskerrik beroenak nire arreba zientifiko nagusiari, Irantzu Barriori, eskainitako laguntza eta aholku zintzoengatik eta elkar partekatutako barre, algara eta elkarrizketengatik. Quisiera agradecer al staff de BCAM la oportunidad que se me ha otorgado, destacando especialmente a Ainara, Miguel y los Enekos. Agradecer también a la gente de mi despacho la ayuda ofrecida, specially to Gabriela Capo for her constant happiness and her contribution to my spoken English. Agradecer



también a mi hermano chileno, Sebastián Zamorano, el apoyo y empuje ofrecidos en los momentos no tan buenos. Finalmente, cómo no, no quisiera olvidarme de mi 'kuadrila' de BCAM, Fabio Pizzichillo, Mario Fernández-Pendás, Gorka Kobeaga, Antsa Ratsimanetrimanana, Amaia Iparragirre, Julia Kroos, Biagio Cassano y Simone Rusconi por hacer que mi paso por BCAM haya sido una grandísima experiencia.

Nire eskerrik beroenak egunerokotasunean, momentu on zein txarretan, nire ondoan eduki izan ditudan lagun guztiei, zazpi RocknRollei, Renteri, Gorkari eta abar luze bati; eskertzekoa da bizitzan horrelako lagun zintzoekin topo egin izana. Nire eskerrik goxo eta sakonenak nire familia guztiari. Zuek izan zarete txikitatik hezitu eta pertsona heldu bilakatu nauzuenak eta, gaur egun, lan honen eraikuntza posible izan bada, zuek guztiek urteetan zehar egindako lanarengatik da. Aitite eta amamak eskertu nahiko nituzke lehenengoz, zuek izan zarete nire eredurik zuzenena, beti edozertan laguntzeko prestutasuna eta maitasuna erakutsiz. Izekoak, osabak, amabitxi, aitabitxi, lehengusin eta lehengusuak, eskerrik asko egunerokotasunean bizitzak eskaintzen dituen eta, honen moduko proiektuak ahalbidetzen dituzten, momentu goxoetan hor egoteagatik. Nire eskerrik zintzoenak ere Zolloko familiari, hasiera batetatik zuen parte sentiarazteagatik eta hiru urte hauetan eskainitako erraztasunengatik. Eskerrik asko bereziki nire etxeko hirukoteari, ama eta aitari nire euskarri, hauspo eta eredu etengabea izateagatik, eta Gorkari, zuri, isilpean bada ere, nigatik azaleratzen duzun guztiarengatik. Azkenik, eskerrik asko Ibone Larizgoitziari, gorabeheraz beteriko bidai honetan indarra, adorea eta ausardia zer diren etengabe irakasten dizkidan bidelagunari.

Eskerrik asko,  
Muchas Gracias,  
Thank you.

---

# CONTENTS

<b>Summary</b>	<b>xxi</b>
<b>Laburpena</b>	<b>xxvii</b>
<b>1 Introduction to patient-reported outcomes</b>	<b>1</b>
1.1 Patient-reported outcomes . . . . .	1
1.2 Measuring instruments: Questionnaires . . . . .	4
1.2.1 Short Form-36 Health Survey . . . . .	5
1.2.2 St. George’s Respiratory Questionnaire . . . . .	6
1.2.3 Mini Mental Score Examination . . . . .	7
1.3 Motivating data . . . . .	8
1.3.1 COPD Study . . . . .	8
1.3.2 Paquid Research Programme . . . . .	11
1.4 Distributional features . . . . .	13
1.4.1 The exponential family . . . . .	13
1.4.2 The beta-binomial distribution . . . . .	25
1.4.3 Distributional fit to the datasets . . . . .	29
1.5 Objectives of the thesis . . . . .	38
<b>2 Cross-sectional analysis: a beta-binomial regression approach</b>	<b>41</b>
2.1 Introduction . . . . .	41
2.1.1 Generalised linear models . . . . .	42
2.1.2 Beta-binomial regression background . . . . .	44

2.2	Beta-binomial regression approaches . . . . .	46
2.2.1	Marginal approach . . . . .	46
2.2.2	Conditional approach . . . . .	52
2.3	Application to the Short Form-36 . . . . .	57
2.4	Simulation study . . . . .	63
2.4.1	Scenarios set up . . . . .	63
2.4.2	Results . . . . .	65
2.5	Conclusions and discussion . . . . .	70
<b>3</b>	<b>Longitudinal analysis: a beta-binomial mixed-effects model approach</b>	<b>75</b>
3.1	Introduction . . . . .	76
3.1.1	Introduction to longitudinal models . . . . .	76
3.1.2	Methodologies for analysing longitudinal data . . . . .	77
3.2	Mixed-effects models in the literature . . . . .	80
3.2.1	Fixed or random effects? . . . . .	83
3.2.2	Linear mixed-effects model . . . . .	85
3.2.3	Generalised linear mixed-effects model . . . . .	90
3.3	Beta-binomial mixed-effects model . . . . .	94
3.3.1	Model definition . . . . .	95
3.3.2	Estimation . . . . .	96
3.3.3	Inference . . . . .	108
3.3.4	Latent variable interpretation . . . . .	110
3.3.5	Similar approaches in the literature . . . . .	113
3.3.6	Comparison to other approaches . . . . .	118
3.4	Simulation study . . . . .	119
3.5	Application to real data . . . . .	124
3.5.1	Mini Mental Score Examination . . . . .	124
3.5.2	St. George's Respiratory Questionnaire . . . . .	127
3.6	Conclusions and discussion . . . . .	134
<b>4</b>	<b>Multivariate analysis</b>	<b>139</b>
4.1	Introduction . . . . .	139
4.2	Background . . . . .	141
4.3	Beta-binomial mixed-effects model approach . . . . .	143
4.3.1	Shared random effects approach . . . . .	143

---

4.3.2	Correlated random effects approach . . . . .	148
4.4	Application to COPD Study . . . . .	149
4.4.1	Cross-sectional analysis . . . . .	149
4.4.2	Longitudinal analysis . . . . .	154
4.5	Discussion and future work . . . . .	156
<b>5</b>	<b>Software development</b>	<b>159</b>
5.1	R-packages for the beta-binomial distribution . . . . .	160
5.2	PR0reg R-package . . . . .	161
5.2.1	BB: The beta-binomial distribution . . . . .	162
5.2.2	BBest: Estimation of the parameters of a beta-binomial dis- tribution . . . . .	165
5.2.3	BBreg: Fit a marginal beta-binomial regression model . . . . .	168
5.2.4	BBmm: Fit a beta-binomial mixed-effects regression model . . . . .	171
5.2.5	Additional functions . . . . .	177
<b>6</b>	<b>Conclusions and further work</b>	<b>179</b>
<b>Appendices</b>		
<b>Appendix A</b>	<b>Recoding process of the Short Form-36 Health Survey</b>	<b>189</b>
<b>Appendix B</b>	<b>Iterative estimation methodologies</b>	<b>191</b>
B.1	Newton-Raphson procedure . . . . .	191
B.2	Iterative weighted least squares algorithm . . . . .	192
<b>Appendix C</b>	<b>Laplace approximation</b>	<b>197</b>
<b>Appendix D</b>	<b>Matrix differentiation</b>	<b>199</b>
D.1	Properties and definitions . . . . .	199
D.2	Matrix differentiation of the proposed models . . . . .	203
<b>References</b>		<b>207</b>



---

## SUMMARY

Nowadays there is growing interest in patient-centered healthcare system. A patient-reported outcome (PRO) is any report on the status of a patient's health condition, quality of life, or functional status associated with health care or treatment that comes directly from the patient, without interpretation of the patient's response by a clinician or anyone else. PROs are increasingly used as primary outcome measures in observational and experimental studies as they inform clinicians and researchers about the health-status of patients and generate data to facilitate improved care. In fact, numerous studies have recommended that objective indicators combined with PROs would be considered a more comprehensive form of outcome evaluation. PROs are usually obtained using rating scale questionnaires consisting of questions or items, grouped into one or more subscales, often called dimensions or domains. There exist some disease-specific questionnaires as well as generic questionnaires that even can be applied to healthy subjects. Traditionally, PROs are calculated by assigning rank scores to the patient's item responses and summing the scores across a group of items and creating overall scores by dimensions that are usually rescaled. Therefore, PROs have an integer and bounded nature which typically accumulate values in one or both edges of the score scale, leading to U, J or inverse J-shaped distributions which the usual exponential family distributions are not able to fit properly.

In order to overcome the poor distributional fit provided by the exponential family members, the beta-binomial distribution has been proposed in the literature for analysing PROs, leading to adequate distributional fits. The beta-binomial distribution is defined as a mixture between a binomial and a beta distribution; or in

other words, as a binomial distribution which probability parameter is random and follows a beta distribution,

$$Y|\theta \sim \text{Bin}(m, \theta) \quad \text{where} \quad \theta \sim \text{Beta}(\alpha_1, \alpha_2),$$

where  $\alpha_1, \alpha_2 > 0$ . Conditional and marginal mean and variances of the beta-binomial distribution can be therefore derived as

$$\begin{aligned} \mathbb{E}[Y] &= m \frac{\alpha_1}{\alpha_1 + \alpha_2}, & \text{Var}[Y] &= m \frac{\alpha_1}{\alpha_1 + \alpha_2} \left[ 1 + (m-1) \frac{1}{\alpha_1 + \alpha_2 + 1} \right], \\ \mathbb{E}[Y|\theta] &= m\theta, & \text{Var}[Y|\theta] &= m\theta(1-\theta), \end{aligned}$$

where the marginal density function is given by

$$\begin{aligned} f(y) &= \int_0^1 f(y|\theta) f(\theta) d\theta = \binom{m}{y} \int_0^1 \theta^y (1-\theta)^{m-y} \frac{\theta^{\alpha_1-1} (1-\theta)^{\alpha_2-1}}{\text{B}(\alpha_1, \alpha_2)} d\theta \\ &= \binom{m}{y} \frac{\Gamma(\alpha_1 + y)}{\Gamma(\alpha_1)} \frac{\Gamma(\alpha_2 + m - y)}{\Gamma(\alpha_2)} \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1 + \alpha_2 + m)}. \end{aligned}$$

It is straightforward to notice the complexity of the beta-binomial density function, which does not belong to the exponential family.

When the aim is to assess the relationship of different covariates on PROs, regression analysis is the most useful framework for statistical modelling. Despite of the appropriate distributional characteristics of the beta-binomial distribution for PRO analysis, in a regression framework, the fact that beta-binomial distribution does not belong to the exponential family makes inappropriate the use of classical regression models, such as generalised linear models (GLMs) or generalised linear mixed-effects models (GLMMs).

Researchers at the Respiratory Service at Galdakao Hospital in Spain designed a longitudinal study where the health-status of patients with Chronic Obstructive Pulmonary Disease (COPD) was repeatedly measured. COPD is a pulmonary disease, it is one of the major causes of mortality and it is associated with high levels of disability. In fact, according to estimates from the World Health Organization, by 2020 it shall become the third most frequent cause of death. The objective of the COPD Study was to measure the health-status and evolution of COPD patients who were followed for up to five years. Two different PRO questionnaires were used to assess the health-status of the patients: a generic, the Short Form-36 Health Survey, and a pulmonary disease-specific, the St. George's Respiratory Questionnaire. Ad-

ditionally, a set of selected time-dependent and independent variables were recoded in the study and considered as covariates for the model development.

The main goal of this thesis is to propose a beta-binomial regression approach that could deal with longitudinal and hierarchical data in PROs framework. This main objective is split into five specific goals: (i) a review and comparison of existing beta-binomial regression approaches; (ii) the development of a methodology for the analysis of multilevel or hierarchical beta-binomial data, in particular, longitudinal data; (iii) the proposal of a multivariate regression model based on the beta-binomial distribution for analysing jointly the dimensions provided by PRO questionnaires; (iv) the implementation of the proposed methodology as Open Source Software to the scientific community and (v) the application of the proposed methodologies to COPD data providing clinically relevant interpretations.

The first beta-binomial regression approach is developed in a cross-sectional or independent data context, and it establishes the base for more complex models. Based on the conditional or marginal interpretation of the beta-binomial distribution, two different regression approaches have been proposed in the literature. On the one hand, the marginal approach, denoted as *BBreg*, consists of a parametrization of the beta-binomial distribution assuming that  $\alpha_1 = p/\phi$  and  $\alpha_2 = (1-p)/\phi$ , where  $0 < p < 1$  and  $\phi > 0$ . That way, the marginal expectation of the beta-binomial distribution is defined as  $\mathbb{E}[Y] = mp$  and  $p$  can be interpreted as a probability parameter, while  $\text{Var}[Y] = mp(1-p)[1 + (m-1)\phi/(1+\phi)]$  determinates  $\phi$  as the dispersion parameter. Therefore, similar to the logistic regression, a logit link function can be applied to the probability parameter connecting it with the given covariates through a linear predictor,

$$Y_i \sim \text{BB}(m_i, p_i, \phi),$$

$$\text{logit} \left( \frac{p_i}{1-p_i} \right) = \mathbf{x}'_i \boldsymbol{\beta},$$

where  $m_i$  is the maximum number of score (or summed binary variables) for the  $i$ th observation,  $\boldsymbol{\beta}$  are the regression coefficients and  $\mathbf{x}_i$  is the  $i$ th row of a full rank matrix composed by the given covariates,  $i = 1, \dots, n$ . Estimation of the parameters is done via maximum likelihood. On the other hand, the conditional approach denoted as *BBhglm* is based on the hierarchical generalised linear models (HGLMs), where not necessarily Gaussian random effects are included in a linear predictor of a GLM. Consequently, the model assumes that conditional on some



beta distributed random effects the outcomes follow a binomial distribution. Hence, the BBglm applies a logistic regression in the conditioned expectation including the beta random effects in the linear predictor.

$$Y_i|u_i \sim \text{Bin}(m_i, p_i), \quad \text{where } u_i \sim \text{Beta}(1/\lambda, 1/\lambda),$$

$$\text{logit} \left( \frac{p_i}{1 - p_i} \right) = \mathbf{x}'_i \boldsymbol{\beta} + \text{logit} \left( \frac{u_i}{1 - u_i} \right),$$

where  $\boldsymbol{\beta}$  are the regression coefficients of the fixed effects,  $u_i$  are the random effects,  $\mathbf{x}_i$  is the  $i$ th row of a full rank matrix composed by the given covariates and  $\lambda$  is the positive parameter of the distribution of the random effects,  $i = 1, \dots, n$ . In this thesis, we show through a real data application and a simulation study that when the objective is to measure the effect of the covariates in PROs, the marginal approach offers more adequate results regarding the statistical significance of the effect and moreover, a convenient interpretation in terms of odds-ratios.

The BBreg model assumes independence among observations, and hence, some extension is required in order to apply it to correlated or multilevel data, the longitudinal COPD Study for instance. In this thesis, we extend the BBreg approach to the inclusion of random effects in the linear predictor in a mixed-effects regression context,

$$Y_i|\mathbf{u} \sim \text{BB}(m_i, p_i, \phi), \quad \text{where } \mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{D}),$$

$$\text{logit} \left( \frac{p_i}{1 - p_i} \right) = \mathbf{x}'_i \boldsymbol{\beta} + \mathbf{z}'_i \mathbf{u},$$

where  $\boldsymbol{\beta}$  and  $\mathbf{x}_i$  have been defined in the previous formulae,  $\mathbf{u}$  are the random effects,  $\mathbf{z}_i$  is the  $i$ th row of the design matrix  $\mathbf{Z}$  that determines the correlation structure of the data and  $\mathbf{D}$  is the variance-covariance matrix of the random effects,  $i = 1, \dots, n$ . This way, the random effects will account for the correlation that may exist among the observations. The model is denoted as *BBmm*. Due to the conditional beta-binomial distribution, GLMM estimation theory is not directly applicable, and therefore, we develop a specific estimation and inference process. The marginal likelihood of the model is approximated by Laplace approximation considering the joint likelihood and an adjusted term. Nevertheless, we show that the adjusted term does not carry any information about the fixed effects and consequently, the approximation can be done by means of the joint likelihood. A delta method based procedure is proposed for the estimation of the fixed effects and the prediction of the

random effects. However, in order to estimate the dispersion parameters (i.e.,  $\phi$  from the beta-binomial distribution and the variance of the random effects), a penalisation of the likelihood is performed to avoid the bias due to the successive estimation of fixed and random effects in an iterative algorithm. We compare our proposal with similar approaches in the literature, such as generalised additive models for location, scale and shape (GAMLSS). The main difference between GAMLSS and BBmm approaches lies in the penalisation of the likelihood used in BBmm to estimate the dispersion parameters. We show through a simulation study that the penalisation does not only improve the estimation of the dispersion parameters in terms of the reduction of the bias, but also in the estimates of the rest of the parameters in the model.

We apply the BBmm approach to the COPD longitudinal study where clinically relevant and valid results about the evolution of patients with COPD are obtained. However, either by BBreg or BBmm approaches, each dimension provided by PRO questionnaires is analysed separately. In general, a PRO may be interpreted as a multivariate outcome constructed by responses of the same patients that may have some correlation and hence, we extend the BBmm model to a multivariate approach. In this thesis, we develop a mixed-effects regression approach, called the shared random effects model, for the joint analysis of all the dimensions in the same PRO questionnaires. The model assumes that each individual has the same random effect across all the dimensions. The implementation of this model under the mixed-effects beta-binomial framework proposed in this thesis is straightforward and it also has the advantage that the complexity of the model is not increased by the number of PRO dimensions. We apply this model in both cross-sectional and longitudinal COPD data, leading to clinically and statistically relevant results compared to the individual analysis of each PRO dimension.

We also implement all the methodology proposed in this thesis in an R-package called `PROreg` which is available at CRAN. The package includes the regression models developed in this thesis and other functionalities.

The research of this thesis leads the way for some future work that must be considered. For instance, we show that the proposed shared random effects model has some limitations when dealing with longitudinal multivariate data and hence, the use of different but correlated random effects structures would be of interest for further research. However, the correlated random effects approach leads to Laplace's approximation problems in the marginal likelihood when more than two dimensions are jointly analysed. Techniques for dealing with the mentioned limitation are discussed

in the thesis within the beta-binomial framework. Additionally, several proposals are left as future work: (i) joint modelling of mean and variance components in BBmm approach, (ii) joint analysis of survival and longitudinal PROs, (iii) testing of the variance components in the proposed models; and (iv) inclusion of non-linear covariate effects in BBmm approach.

In conclusion, in this thesis we develop several regression approaches based on the beta-binomial distribution for the analysis of PROs. Moreover, although the illustration of the models is only performed in PRO context, they can also be useful in any U, J or inverse J-shaped discrete and bounded data (due to overdispersion). In fact, this type of data can appear in different contexts and research fields, such as Finance (e.g. the estimation of the probability of a claim in an insurance product), or Biology (e.g. presence/absence of species or bacteria in agricultural experiments). The implementation of the methodology through the `PROreg` R-package provides a useful tool that clinical researchers can use for statistical modelling in a wide variety of studies in order to draw conclusions from PROs.

---

## LABURPENA

Gaur egun, gero eta ohikoagoa da pazientearen inguruan oinarritutako osasun sistema, zeinak pazientetik eratorritako behaketak (PEBk) darabiltzan. PEB pazientearengandik zuzenean, inolako medikuren bitartekaritza zein interpretazio gabe, pazientearen osasun egoeraren, bizi kalitatearen edo gaixotasun zein tratamendurekin lotutako egoera funtzionalaren edozein neurketa da. PEBk oinarritzko behaketa modura erabiltzen hasi diren ikerketen kopurua etengabe ari da hasten, izan ere, PEBek pazienteen osasun egoeraren informazioa isladatzen diete ikertzaila zein medikuei eta gaixoen zaintza hobe dezaketen datuak eskaini. Are gehiago, badira, informazio osoagoa lortu nahiean, indikadore objektiboak PEBekin ustartzearen garrantzia nabarmentzen duten hainbat ikerketa. Orokorrean, PEBk puntuazio eskaladun inkesten bitartez lortzen dira, zenbait galderaz osatua daudenak, galdera horiek taldeka edo dimentsioka multzokatuta egonik. Gaixotasun zehatzentzat garatutako inkestak zein inkesta orokorrak, populazio osasuntsuan ere aplikatu daitezkenak, existitzen dira. Normalean, PEBk pazienteak ihardetsi beharreko galderen erantzun posibleak puntuatuz eta puntuazio horiek taldeka batuz kalkulatu dira. Ondorioz, paziente bakoitzak neurtu nahi den PEBren puntuazio zehatz bat lortzen du multzokatze irizpide izan den dimensio bakoitzean. Azkenik, lortutako puntuazioen eskala eraldatzen da normalean. Beraz, PEBek izaera oso eta bornatua daukate eta, orokorrean, balioak puntuazio eskalaren alde batetan zein bietan multzokatzen dituzte, familia exponentzialeko banaketek ondo dohitzerik ez duten U, J edo alderantzizko J itxurak aurkeztuz.

Familia exponentzialeko kideek daukaten dohikuntza gabezia gainditze aldera, banaketa beta-binomiala proposatua izan da literaturan PEB analizatzeko, dohikuntza

egokiak lortuz. Banaketa beta-binomiala, izenak adierazi bezala, banaketa binomial eta beta banaketa baten arteko nahastura bat da, edo baliokidea dena, zorizko beta banaketa darraien probabilitate parametrodun binomiala da,

$$Y|\theta \sim \text{Bin}(m, \theta) \quad \text{non} \quad \theta \sim \text{Beta}(\alpha_1, \alpha_2)$$

non  $\alpha_1, \alpha_2 > 0$ . Banaketa beta-binomialaren itxaropen baldintzatu eta marginalak hurrenez-hurren, hurrengo formulak emanak dira,

$$\begin{aligned} \mathbb{E}[Y] &= m \frac{\alpha_1}{\alpha_1 + \alpha_2}, & \text{Var}[Y] &= m \frac{\alpha_1}{\alpha_1 + \alpha_2} \left[ 1 + (m-1) \frac{1}{\alpha_1 + \alpha_2 + 1} \right], \\ \mathbb{E}[Y|\theta] &= m\theta, & \text{Var}[Y|\theta] &= m\theta(1-\theta), \end{aligned}$$

non dentsitate funtzio marginala

$$\begin{aligned} f(y) &= \int_0^1 f(y|\theta) f(\theta) d\theta = \binom{m}{y} \int_0^1 \theta^y (1-\theta)^{m-y} \frac{\theta^{\alpha_1-1} (1-\theta)^{\alpha_2-1}}{\text{B}(\alpha_1, \alpha_2)} d\theta \\ &= \binom{m}{y} \frac{\Gamma(\alpha_1 + y)}{\Gamma(\alpha_1)} \frac{\Gamma(\alpha_2 + m - y)}{\Gamma(\alpha_2)} \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1 + \alpha_2 + m)} \end{aligned}$$

den. Nabaria da banaketa beta-binomialaren dentsitate funtzioaren konplexutasunaz ohartzea, eta bide batez, familia exponentzialaren parte ez dela konturatzea.

Helburua zenbait koaldagaiek PEBetan duten eragina aztertzea eta neurtzea deanean, erregresio ereduak eredugintza estatistikoaren herramintarik erabilienetakoa dira. PEBk dohitzerako orduan banaketa beta-binomialak emaitza oso onak lortzen baditu ere, familia exponentzialaren parte ez izateak murriztu egin du banaketa horretan oinarritutako erregresioen erabilera. Izan ere, aplikazio praktikoan gehienetan erabiltzen diren erregresio lineal orokortuek (ELOek) zein erregresio lineal orokortu mixtoek (ELOMek), menpeko aldagaiaren banaketa familia exponentzialaren parte izatea suposatzen dute ezinbestean.

Galdakaoko Ospitaleko arnaskatze zerbitzuko ikertzaileek birikietako burtxadura kroniko gaixotasuna (BBKG) zuten pazienteen osasun egoeraren denboran zeharreko neurketak gauzatu zituzten. BBKG birikietako gaixotasun nahiko larria da, munduko heriotz tasa handienetakoa duena. Bestalde, ezintasun fisiko zein mental handiekin lotzen den gaixotasuna da. Osasunaren Mundu Erakundearen iritziz, 2020. urterako gaixotasunek eragindako hirugarren heriotza kausa bilakatuko da. Ikerketa honen helburua BBKG pairatzen duten pazienteen osasun egoera aztertzea da eta, bost urtetan zehar aztertutak eta neurtutak izan diren pazienteen osasun gara-

pena neurtzea. Pazienteen osasun egoera neurtzeko bi inkesta ezberdin erabili ziren: batetik, inkesta generiko bat, Short Form-36 Osasun Inkesta, eta bestetik, birikitako gaixotasunak aztertzeko inkesta espezifiko bat, St George Arnaskatze Inkesta. Osasunaren egoeraren neurketa hauekin batera, denboran zehar aldakorrak zein estatikoak diren beste zenbait aldagai ere neurtu zituzten, erregresio eredugintzan koaldagai modura erabiliko direnak.

Tesi honen helburu nagusia denboran zeharreko zein korrelaturiko PEB datuak aztertzeko banaketa beta-binomialan oinarritutako erregresio eredu bat garatzea da. Aldi berean, helburu nagusi hau beste bost azpi-helburutan banatzen da: (i) literaturan ageri diren banaketa beta-binomialan oinarritutako erregresioen berrikuste eta komparaketa; (ii) banaketa beta-binomialetik eratorritako datu korrelatu zein hierarkikoen, bereziki denporazko datuen, azterketa ahalbidetuko duen ereduak garatzea; (iii) PEBen inkestek eragiten dituzten dimentsio guztiak amankomunean aztertuko dituen eta banaketa beta-binomialan oinarrituko den eredu multidimentsionala proposatzea; (iv) garaturiko eredu guztiak software askean implementatzea zientzia eta ikerketa komunitateari baliagarria izan dakizkion; eta, (v) BBKG pairatzen duten pazienteen ikerketan garaturiko ereduak aplikatzea eta klinikoki esanguratsuak diren emaitzak ondorioztatzea.

Banaketa beta-binomialan oinarritutako lehen erregresio ereduak neurketen arteko askatasuna beharrezkoa du eta hurrengo eredu konplexuagoak garatzeko giltzarria izango da. Banaketa beta-binomialaren ikuspegi marginal edo baldintzatuaren arabera erregresio eredu ezberdinak gara daitezke. Alde batetik, eredu marginala, *BBreg* deritzoguna, banaketa beta-binomialaren parametroen eraldaketa batean oinarritzen da,  $\alpha_1 = p/\phi$  eta  $\alpha_2 = (1-p)/\phi$  non  $0 < p < 1$  eta  $\phi > 0$ . Eraldaketa horretan oinarriturik, banaketaren itxaropena marginala  $\mathbb{E}[Y] = mp$  moduan definitzen da eta ondorioz,  $p$  probabilitate parametro modura interpreta daiteke; era berean,  $\text{Var}[Y] = mp(1-p)[1 + (m-1)\phi/(1+\phi)]$ ,  $\phi$  dispersio parametro moduan zehaztuz. Ondorioz, erregresio logistikoarekin egin antzera, logit funtzio lokailua aplikatuz dakiok probabilitate parametroari eta ereduaren erabili nahi ditugun koaldagaiekin erlazionatu herrengo ekuazioen bitartez,

$$Y_i \sim \text{BB}(m_i, p_i, \phi),$$

$$\text{logit} \left( \frac{p_i}{1-p_i} \right) = \mathbf{x}'_i \boldsymbol{\beta},$$

non  $m_i$   $i$ . neurketak lor dezaken puntuazio kopuru altuena den (edo batutako al-

gai binario kopurua),  $\beta$  erregresioaren koefizienteak eta  $\mathbf{x}_i$  koaldagaiek definitzen duten hein osoko  $\mathbf{X}$  matrizearen  $i$ .garren errenkada den,  $i = 1, \dots, n$ . Parametroen estimazioa egiantz handieneko prozedura erabiliz gauzatzen da. Beste alde batetik, *BBhglm* deritzogun eredu baldintzatua erregresio linear okortu hierarkikoetan (ELOH) oinarritzen da non Gaussiarrak ez diren zorizko efektuak ere gaineratu daitezkeen ELO baten auresale linealean. Ereduak suposatzen du behaketek beta banaketa darraiten zorizko efektu batzuek baldintzatutako banaketa binomial bat darraitela. Ondorioz aldagaiaren itxaropen baldintzatuan erregresio logistiko bat ezarri eta zorizko beta efektuak gaineratzen ditu auresale linealean,

$$Y_i|u_i \sim \text{Bin}(m_i, p_i), \quad \text{non } u_i \sim \text{Beta}(1/\lambda, 1/\lambda),$$

$$\text{logit} \left( \frac{p_i}{1 - p_i} \right) = \mathbf{x}'_i \boldsymbol{\beta} + \text{logit} \left( \frac{u_i}{1 - u_i} \right),$$

non  $\beta$  efektu finkoak,  $u_i$  zorizko efektuak,  $\mathbf{x}_i$  koaldagaiek osatutako hein osoko matrizearen  $i$ . errenkada eta  $\lambda$  zorizko efektuen banaketaren parametro positiboa diren,  $i = 1, \dots, n$ . Helburua koaldagaiek PEBn duten eragina aztertzea denean, tesi honetan erakusten dugu, aplikazio erreal zein simulazio azterketa baten bitartez, eredu marginalak emaitza hobekia lortzen dituela estimazioaren adierazgarritasun estatistikoari dagokionez eta, are gehiago, oso interpretazio txukuna eskaintzen duela odds-ratioen bitartez.

BBreg ereduak behaketen arteko askatasuna bere egiten du eta ondorioz, hedapen bat beharrezkoa da datu korrelatuak aztertu nahi badira, denboran zeharreko datuak kasu. Tesi honetan BBreg eredu zorizko efektuen alorretik hedatzen dugu, banaketa normala darraiten zorizko efektuak ereduaren auresale linealean txertatuz,

$$Y_i|\mathbf{u} \sim \text{BB}(m_i, p_i, \phi), \quad \text{where } \mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{D}),$$

$$\text{logit} \left( \frac{p_i}{1 - p_i} \right) = \mathbf{x}'_i \boldsymbol{\beta} + \mathbf{z}'_i \mathbf{u},$$

non  $\mathbf{u}$  zorizko efektuak,  $\mathbf{z}_i$  datuen korrelazio egitura zehazten duen  $\mathbf{Z}$  diseinu matrizearen  $i$ . errenkada eta  $\mathbf{D}$  zorizko efektuen bariantza-kobariantza matrizea diren,  $i = 1, \dots, n$ . Hortaz, zorizko efektuek behaketen arteko korrelazioa berenganatuko dute. Ereduri honi *BBmm* deritzogu. Baldintzatutako banaketa beta-binomiala dela eta, ELOMen estimazio teoria ez da kasu honetan aplikagarria eta ondorioz, estimazio eta inferentzia prozedura zehatza garatu dugu. Ereduaren egiantz marginala Laplacen hurbilketa erabiliz kalkulatu da egiantz bateratuaren eta penalizazio gai

baten bitartez. Hala ere, tesi honetan penalizazio gaiak efektu finkoen inguruko informaziorik eskaintzen ez duela erakusten dugu, beraz, zorizko efektu zein efektu finkoen estimazioa egiantz bateratua erabiliz egin daitekeela ondorioztatu. Ereduko efektuen estimazioa gauzatzeko delta metodoan oinarritutako prozedura bat garatzen da. Dena den, ereduko dispersio parametroak (banaketa-beta binomialaren dispersio parametroa  $\phi$  eta zorizko efektuen bariantza) estimatzerako orduan, aurrez gauzaturiko zorizko zein efektu finkoen estimazioak sor dezakeen alborapena ekiditeko, egiantz batertuari penalizazio irizpide bat ezartzen zaio. Gure eredu proposamena literaturan aurki daitezkeen antzeko ereduak alderatzen dugu, lokalizazio, eskala eta itxurazko eredu batukor orokortuekin (LEIEBO) bereziki. Proposatutako BBmm eta LIEEBO ereduaren arteko ezberdintasunak nabariena dispersio parametroak estimatzeko orduan egiantz bateratua gauzatzen den penalizazioan datza. Simulazio azterketa baten bitartez penalizazioak dispersio parametroen estimazioaren alborapena murrizteaz gain, gainerako parametro guztien estimazioaren alborapena ere murrizten duela erakusten dugu.

BBmm ereduaren BBKG pairatzen duten pazienteen luzerazko ikerketan aplikatu dugu, klinikoki esanguratsuak eta baliagarriak diren emaitzak lortuz. Hala ere, orain arteko azterketa erreala guztietan, BBreg zein BBmm ereduak erabiliz, PEBk osatzen dituzten dimentsioak banan-banan analizatu ditugu, haien artean egon litekeen korrelazioa kontuan hartu gabe. Izan ere, paziente berberak beteriko inkesten erantzunetatik sortzen dira dimentsioak, beraz, paziente berberak dimentsio ezberdinetan lortutako puntazioen artean korrelazioa egon daitekeela pentsatzeak logikoa dirudi. Hori dela eta, luzerazko datuetarako proposaturiko BBmm ereduaren analisi multidimensionalak gauzatzeko garatu dugu. Tesi honetan, eredu unidimentsionaletik multidimentsionalerako pausua eredu mixtoak erabiliz garatzen da, partekaturiko zorizko efektuen ikuspuntutik. Ereduak dimentsio guztiek zorizko efektu amankomuna partekatzen dutela suposatzen du eta modu horretan, korrelazio egitura bat gaineratzen dio ereduari. Eredu honen inplementazioa, estimazio zein inferentzia aldetik, zuzena da aurretik garaturiko BBmm ereduaren oinarrituz. Azterketa multidimentsionalak gauzatzeko litearturan zenbait eredu existitzen diren arren, partekaturiko zorizko efektuen ereduaren aukerategia, nagusiki, dimentsioak gehitu ahala ereduaren konplexutasuna handitzen ez delako egin dugu, eta beraz, garaturiko egiantz hurbilketa guztiek egokiak izaten jarraitzen dute. Ereduaren BBKG duten pazienteen datu basean aplikatu dugu, paziente bakoitzaren lehen neurketa bakarrik zein denbora zeharreko neurketa guztiak kontsideratuz. Datu errealetarako ereduaren aplikazioak emaitza kliniko zein estatistiko baliagarriak eskaintzen ditu.



Besteak beste, komeni da aipatzea, garaturiko banaketa beta-binomialean oinarrituriko eredu guztiak R software estatistikoa, **PROreg** paketearen bitartez, inplementatu ditugula, dagoeneko CRAN-en erabilgarri dagoena. Paketeak, aipaturiko ereduez gain, PEBk aztertzerako orduan baliagarriak izan daitezkeen beste zenbait funtzio ere gaineratzen ditu.

Tesi honetan zehar garaturiko ikerketak etorkizunean aztertuko beharreko zenbait alorren azalratzea eragin du ezinbestean. Adibidez, partekaturiko zorizko efektuen ereduak, aplikagarritasunari dagokionean, erreztasun ugari eskaintzen baditu ere, luzerazko datuen azterketa multidimentsionala garatzeko orduan emaitzen egokitasuna kolokan jarri dezakeen mugak ere badituela erakusten dugu. Mugak gaintze aldera, irtenbide bat zorizko efektu ezberdin baina korrelatuak dimentsio bakoitzean aplikatzea da. Hala ere, zorizko efektu korrelatuen eredu honek egiantz marginalaren integralaren dimentsioa nabarmen handitzen du PEBk osatzen ditzuten eta aztergai diren dimentsio kopurua handituz gero. Integralaren dimentsioa areagotzeak hurbilketa tekniken egokitasun eza dakar, BBmm ereduan garaturiko Laplacen hurbilketarena adibidez. Hala eta guztiz ere, hurbilketen ezegokitasun hau gaintutako luketen teknika ezberdinen inguruko eztabaida bat ere garatzen da tesian. Bestalde, zenbait proposamen ere etorkizunean garatzeko aukera moduan uzten dira: (i) BBmm ereduaren itxaropen eta bariantza parametroen modelizazio bateratua, (ii) PEBen luzerazko eta biziraupeneko analisi bateratua, (iii) proposaturiko ereduen bariantza parametroen estimazioaren inferentzia, eta (iv) koaldagaien efektu ez linealen gainerapena BBmm ereduan.

Azkenik, borobiltze aldera, tesi honetan PEBk aztertzekeo banaketa beta-binomialean oinarrituriko zenbait erregresio eredu garatu ditugu. Hala ere, ereduen ilustrazioa PEBen kontextuan irudikatu badugu ere, U, J edo alderantzizko J itxura duten datu oso eta bornatuetan ere aplikagarriak izan daitezke. Izan ere, datu mota hau testuinguru zein aztergai diren arlo askotan ageri da, Finantzak (aseguru etxeetan bezeroen produktu eskaera ezberdinen probabilitatea neurtzeko adibidez) eta Biologia (laboraltza experimentuetan bakterio edo espezie zehatz baten presentzia/absentzia aztertzekeo adibidez) hurrenez-hurren. Are gehiago, garaturiko metodologia **PROreg** R-paketean inplementatua egoteak herreminta oso erabilgarria eta zuzena eskaintzen die kliniko zein ikertzaileei beta-binomialean oinarritutako ereduak aplikatu ahal izateko eta bide batez, emaitza erabilgarri zein adierazgarriak ondorioztatzekeo euren ikergaiaren inguruan.

---

---

# CHAPTER 1

---

## INTRODUCTION TO PATIENT-REPORTED OUTCOMES

*“Deus ez da inorena, bizitza bera ere ez”*

---

Xabier Lete, 1944 – 2010

### 1.1 Patient-reported outcomes

Patient-reported outcome (PRO) measurements are increasingly used as primary outcome measures in observational and experimental studies, which play an important role in health care and understanding health outcomes. A PRO is any report on the status of a patient’s health condition that comes directly from the patient, without interpretation of the patient’s response by a clinician or anyone else (US Department of Health and Human Services, 2012). PROs are used to subjectively assess outcomes, such as pain, quality of life or satisfaction with care, that are difficult or impossible to measure physically without a patient’s evaluation and whose key questions require patient’s input on the impact of a disease or treatment. Consequently, PROs inform clinicians and researchers about issues associated with health-status that are most important to patients and their families (Dobrozsi and Panepinto, 2015) and generate data to facilitate improved care of the patient. In addition, PROs have gradually become an important element and a crucial source for monitoring disease conditions or assessing the effectiveness of treatment, especially in some health problems such as subjective discomfort and psychological distress

(Chang, 2007). Therefore, the U.S. Food and Drug Administration (FDA) has recommended that objective indicators combined with PROs would be considered a more comprehensive form of outcome evaluation since 2006 (Speight and Barendse, 2010).

PROs are usually obtained using rating scale questionnaires, which are made up of questions, called *items*, grouped into one or more subscales, often called *dimensions* or *domains*. The patient is presented with a series of related items and is asked to respond with ordered categories that represent magnitude estimates of his or her personal experience related to the content of the items. The term instrument is used to refer to the collection of items and response categories. Traditionally, PRO instruments have been scored by assigning rank scores to the patient's item responses, summing the scores across a group of items and creating overall scores by dimensions that are usually rescaled.

PROs provide composite scores, one by dimension, from a series of questions around a central concept to quantify the level of distress or impairment caused by a patient's symptoms, disease, or treatment. A variety of PROs has been developed and validated to assess symptoms (such as nausea and vomiting, insomnia, constipation, and pain), patient physical, social, and emotional function, and more complex constructs such as the impact of health on peer relationships. There are generic and disease-specific questionnaires that can be used to measure these outcomes, as well as available measures for the use in pediatric and adult patients.

Therefore, PROs are important supplements to traditional medicine as they offer a new insight of the health-status of patients. Indeed, some decades ago the World Health Organization (WHO) defined the health as a complete physical, mental and social welfare, and not only as a disability or lack of disease. In this context a new concept appeared, the Health-Related Quality of Life (HRQoL). The measurement of HRQoL provides information about the disease and its impact on the patient in a standardised, comparable and objective way (Goldsmith, 1972). Indeed, HRQoL is a general PRO which is being increasingly used as an outcome in clinical trials, effectiveness research, and research on quality of care (Wilson and Cleary, 1995). Factors that have facilitated this increased usage include the accumulating evidence that measures of HRQoL are valid and reliable (McDowell and Newell, 1987), the publication of several large clinical trials showing that these outcome measures are responsive to important clinical changes (Croog et al., 1986) and the successful development and testing of shorter instruments that are easier to understand and administer (Hunt et al., 1985). Because these measures describe or characterise

what the patient has experienced as the result of medical care, they are useful and important supplements to traditional physiological or biological measures of health-status.

In fact, measuring HRQoL or, PROs in general, can help determine the burden of preventable disease, injuries, and disabilities, and it can provide valuable new insights into the relationships between HRQoL and risk factors. Therefore, to study the relationship of HRQoL with patient and disease characteristics has become one of the primary aims of many PROs studies. In this framework, regression models are necessary in order to assess the effect that clinical and socio-demographic variables have on the HRQoL of individuals.

There exist some well-designed studies that emphasise the utility of PROs (Au et al., 2010) especially for patients with chronic diseases. For instance, patients with cancer commonly experience symptoms that lead to impaired physical function, emotional function and social function resulting in decreased health-status (Fisch et al., 2012). However, patient symptoms often go undetected during typical clinic interviews, and therefore, clinicians may underestimate the impact of the symptoms from interview alone (Buckner et al., 2014). The impact of cancer and cancer treatment on patient health-status can be systematically quantified by PROs (Montazeri, 2008). The use of PROs in these populations has identified symptoms and impairments in patient function. Moreover, cancer patients with better symptom management and better PRO scores live longer with less distress (Montazeri, 2009). PROs generate data that can facilitate better care regardless of diagnosis or prognosis for patients with cancer or other chronic illnesses.

Once the nature of the PROs has been defined, and the validity of different studies has been illustrated, this chapter will continue as follows. In the next section, we introduce some PRO instruments in the form of questionnaires, we describe the health aspect they assess and introduce some specific characteristics such as the number of items or number of health dimensions they provide. In Section 1.3 we describe two PRO studies carried out for measuring different aspects of the health-status for different disease patients. This data will be used to validate the proposed methodology throughout this thesis. In Section 1.4 we describe the statistical features of PROs which make commonly used modelling techniques inappropriate in this framework. We focus our attention on the distributional fit of the PROs provided in each dataset. Finally, in Section 1.5 we describe the structure and objectives of the thesis.

## 1.2 Measuring instruments: Questionnaires

PROs are measured through instruments to assess for instance, the health-status or HRQoL of patients. Different measuring instruments have been developed in the literature, most of them in the form of questionnaires, which decompose the health aspect they evaluate in different dimensions. Some questionnaires can be provided to patients with different diseases or healthy subjects, and they are considered *generic* questionnaires. On the contrary, others are designed for subjects with specific diseases, the so-called *specific* questionnaires. Each survey contains a different number of dimensions and decomposes each dimension in a different score scale. In Section 1.3 we are going to present two different PRO studies which contain measurements provided by patients with different diseases. In the first study patients with Chronic Obstructive Pulmonary Disease (COPD) were recruited and their health-status was measured by means of two different questionnaires, one generic and the other pulmonary diseases specific. The second study is focused on the assessment of dementia where a specific questionnaire was used to measure the cognitive status of patients.

First, we describe the Short Form-36 (SF-36) Health Survey, a generic questionnaire which is broadly used for different populations. It provides measurements of the HRQoL of patients in eight dimensions. The scores provided by the SF-36 in COPD patients will be analysed in Chapter 2 and Chapter 4. Second, we describe the St. George's Respiratory Questionnaire (SGRQ), a specific questionnaire for chronic airflow limitation diseases. It measures the HRQoL and the impact of the disease in the patients and it provides three different dimensions. The same as the SF-36, it has also been applied to patients with COPD. Therefore, the HRQoL of patients with COPD has been assessed by means of two different PRO questionnaires, one generic and the other disease-specific. The information provided by the SGRQ will be analysed in Chapter 3 and Chapter 4. Finally, we introduce the Mini Mental State Examination (MMSE) which was developed for measuring the cognitive status of individuals. It only provides one dimension that covers seven cognitive aspects. Outcomes provided by the MMSE will be analysed in Chapter 3.

It is worth mentioning that while the SF-36 is a generic questionnaire, the MMSE and SGRQ were developed for specific diseases. In fact, each questionnaire analyses different aspects of the health-status of patients and provides different measures. Therefore, we consider that the analysis of the cited questionnaires by different methodologies will provide a global view of PROs and it will furnish the reader with different techniques to deal with any other PRO.

### 1.2.1 Short Form-36 Health Survey

The SF-36 Health Survey was developed within the Medical Outcomes Study (Ware et al., 1993) and it is one of the most widely used generic instruments. It measures generic HRQoL concepts, and it provides an objective way to measure HRQoL from the patients' point of view by scoring standardised responses to standardised questions. The validity and reliability of this instrument have been broadly tested (Stansfeld et al., 1997).

The SF-36 questionnaire has 36 items, with different answer options. It was constructed to represent eight health dimensions, which are *physical functioning* (PF), *role physical* (RP), *bodily pain* (BP), *general health* (GH), *vitality* (VT), *social functioning* (SF), *role emotional* (RE) and *mental health* (MH). Each item is assigned to a unique health dimension. Each of the eight multi-item dimensions contains two to ten items. The first four dimensions are mainly physical, whereas the last four measure mental aspects of HRQoL. The standardised scoring system is thoroughly described by the original authors (Ware et al., 1993). Briefly, each score is calculated with an algorithm based on the original items assigned to this dimension. For each dimension, the answers to the items are first recoded and then added in a weighted sum fashion. The resulting raw scores are then transformed to standardised scale scores from 0 to 100, where a higher score indicates a better health-status.

In order to provide a better understanding of the construction of the SF-36 Health Survey dimensions, we show in Table 1.1 the number of items related to each dimension and the number of possible values each dimension can obtain.

Table 1.1: Number of items related to the construction of each SF-36 dimension and the number of values each dimension can obtain.

Dimension	No. of items	No. of possible values
<i>Physical functioning</i>	10	21
<i>Role physical</i>	4	5
<i>Bodily pain</i>	2	27
<i>General health</i>	5	39
<i>Vitality</i>	4	21
<i>Social functioning</i>	2	9
<i>Role emotional</i>	3	4
<i>Mental health</i>	5	26

Consequently, the SF-36 Health Survey generates a profile of HRQoL outcomes on eight health dimensions. Additionally, the SF-36 includes an item related to health transition, which is not used in the scoring of the eight health dimensions. The authors of the SF-36 Health Survey also provide normative scores for each health dimension. Each SF-36 score is first standardised using the mean and standard deviations obtained from the general population and then transformed to a norm-based (mean=50, standard deviation = 10) scoring (Ware et al., 1993). Two summary measures, one physical and one mental, can be created from the eight main domains. These two summary scores are generated using the physical and mental factor score coefficients from the general population, and they are also transformed to norm-based scoring (Ware et al., 1994). However, we have used neither norm-based scores nor summary scores in this work.

Finally, it is worth mentioning that the SF-36 was rated in 2002 by the British Medical Journal (Garratt et al., 2002) as the most frequently used PRO of generic health in the scientific publications.

### 1.2.2 St. George's Respiratory Questionnaire

The SGRQ was designed to quantify the impact of chronic airflow limitation on health and perceived well-being (quality of life), and to be sufficiently sensitive to respond to changes in disease activity (Jones et al., 1991). The SGRQ can provide a psychosocial impact profile of these patients that cannot be identified by the tests of lung function. Clinically, it has been shown to be a valuable tool in quantifying the impact of chronic obstructive airways diseases on the symptom, functional measures and well-being (Doll et al., 2003; Peruzza et al., 2003) and in evaluating the effectiveness of health care (Singh et al., 2001). Compared to other chronic airflow limitation specific questionnaires, it offers standardised measures. In fact, the absence of suitable sensitive and standardised questionnaires of health and well-being limits the application and value of quality of life measurements in respiratory medicine, as if the questionnaire is not standardised between patients, it is difficult to compare different studies and different study populations (Jones et al., 1992).

The SGRQ consists of 50 items. Principal-component analysis of the responses to these items supported the partition of the questionnaire into three sections or dimensions. The first, *symptoms*, contains items concerned with the level of symptomatology, including frequency of cough, sputum production, wheeze, breathlessness, and the duration and frequency of attacks of breathlessness or wheeze. The

second, *activity*, is concerned with physical activities that either cause or are limited by breathlessness. Finally, the last dimension corresponds to the *impacts* and it covers such as employment, being in control health, panic, stigmatisation, the need for medication and its side effects, and expectations for health and disturbance of daily life. Items specifically relating to the anxiety and depression were not included in the SGRQ, since many established measures exist for this area of health. Each of the three dimensions of the questionnaire is scored separately in the range 0 to 100, zero score indicating no impairment of life quality. Additionally, a summary score utilising responses to all items is constructed, which is defined as the *total* SGRQ score. This score also ranges from 0 to 100, where a lower score means a better health-status. More details about the construction of the scores can be found at Quirk and Jones (1990).

### 1.2.3 Mini Mental Score Examination

The MMSE is a questionnaire that measures the cognitive mental status of patients (Folstein et al., 1975). The objective of the survey is to offer a brief screening tool to provide a quantitative assessment of cognitive impairment and to record cognitive changes over time. The MMSE consists of 11 simple questions grouped into 7 cognitive domains: *orientation to time*, *orientation to place*, *registration of three words*, *attention and calculation*, *recall of three words*, *language* and *visual construction*. It only requires 5-10 minutes to the administer, which makes it very practical to use serially and routinely. In fact, it has become one of the most popular psychometric tests used to quantify global cognitive functioning and cognitive change in population-based longitudinal studies. It is only based on the cognitive aspects of mental functions of patients, avoiding questions about mood, abnormal mental experiences and the form of thinking.

The MMSE is divided into two parts, the first one covers orientation, memory, and attention, and the second part covers the ability to follow verbal and written commands. The maximum score number an individual can obtain is 21 in the first part and 9 in the second, therefore, a possible score of 30 is used to provide a picture of an individual's present cognitive performance based on direct observation of completion of test items, where a higher value means a better cognitive status.



## 1.3 Motivating data

### 1.3.1 COPD Study

Chronic obstructive pulmonary disease is one of the major causes of mortality worldwide and it is associated with high level of disability (Pauwels and Rabe, 2004). Some well-designed studies have found a measured prevalence of COPD in Europe between 4% and 10% of adults, and it is expected to increase over the next years (Halbert et al., 2003; Buist et al., 2008). According to estimates from the World Health Organization (WHO), by 2020 it shall become the third most frequent cause of death, following coronary and cerebrovascular diseases (Murray and López, 1997). COPD is a respiratory system disease with irreversible damage of pulmonary and bronchial tubes, which represents the state of chronic airflow limitation (Jones and Higenbottam, 2007). It not only causes physiological discomfort but also has a psychosocial influence on individuals. The clinical assessment of COPD often involves measurement of lung function parameters (e.g. FEV1) and exacerbation level of a patient to evaluate the disease progress and the therapeutic effect (Cosio and Agustí, 2010). However, the overall impact of COPD on individuals is multi-faceted and not entirely reflected by these clinical parameters. For this reason, it is now realised that no single measure can adequately reflect the nature or severity of COPD and it often needs to be supplemented by other indicators from a patients perspective, such as those related to PROs or HRQOL. To date, evaluation of the treatment effect has emphasised the improvement of the quality of life rather than the small gains in survival rate or physiological indicators (Wiklund, 2004).

Researchers at the Respiratory Service at Galdakao Hospital in Spain designed the COPD Study, a longitudinal study whose main goal was to measure the health-status and evolution of patients being treated for COPD. Patients were recruited at five outpatient respiratory clinics affiliated with the hospital and consecutively included in the study for one year, starting in January 2003. Patients were eligible for the study if they had been diagnosed with COPD for at least six months and they had been receiving medical care at one of the hospital respiratory outpatient facilities for at least six months. Their COPD had to be stable for six weeks before enrolment. Patients were followed for up to five years. Two main outcome measurements were collected: (i) Generic HRQoL was measured using version 1.2 of the SF-36 Health Survey (see Section 1.2.1), which corresponds to the version 1.4 of the Spanish version. (ii) Respiratory specific health-status was measured with the SGRQ (see Section 1.2.2). In addition, a set of selected time independent and time dependent

variables recoded in the study and considered as covariates for the models were socio-demographic variables such as gender and age at entry in the cohort, together with forced expiratory volume in one second in percentile (FEV1%), body mass index (BMI), dyspnea (measured with the modified scale of the Medical Research Council, (Mahler et al., 2009)), the 6-minute walking tests (American Thoracic Society, 2002) and presence of anxiety and depression measured by the Hospital Anxiety and Depression (HAD) scale (Zigmond and Snaith, 1983) among others.

Esteban et al. (2016) divided the individuals participating in the COPD Study in some clusters where four subtypes were identified. They conclude that subtypes A, B, and C, had marked respiratory profiles with a continuum in severity of several variables, while the fourth, subtype D, had a more systemic profile with intermediate respiratory disease severity. Subtype A was associated with less dyspnea, better HRQoL and lower comorbidity, and subtype C with the most severe dyspnea, and poorer pulmonary function and quality of life, while subtype B was between subtypes A and C. Subtype D had higher rates of hospitalization the previous year and comorbidities.

Table 1.2 shows a socio-demographic and clinical summary of the collected time independent and time dependent exploratory variables. Both discrete and continuous variables are analysed for the different time points. On the one hand, for the discrete exploratory variables, the number of individuals in each level and the proportion are shown. On the other hand, for the continuous covariates, the mean and the standard deviation for each time point are displayed. Additionally, Table 1.2 shows the number of individuals that remain in the cohort for different time points.

In terms of the descriptive analysis of the response measurements in the COPD Study, Table 1.3 shows the mean and standard deviation of the original standardised scores of the SF-36 and the row scores of the SGRQ. In general terms, it can be appreciated that the passing of time affects each dimension of both questionnaires differently. Regarding SF-36 dimensions, Table 1.3 shows that while in some dimensions the mean and standard deviation of the original standardised scores change considerably as the time goes by (e.g. *role physical*), there are some others that hardly change (e.g. *physical functioning*). Moreover, it can be appreciated that patients with COPD get the worst results in average in physical dimensions such as *physical functioning*, *role physical* or *general health*. In addition, it is worth noticing that the largest variability occurs in *role physical* and *role emotional* dimensions. In fact, these are the dimensions with the lowest number of possible values (see Table 1.1), and hence, the standardisation to the 0-100 scale scatters more the scores than

Table 1.2: Descriptive analysis of the covariates in the COPD Study.

	Time framework			
	Baseline	1-Year	2-Years	5-Years
No. Individuals	$n = 543$	$n = 480$	$n = 425$	$n = 324$
<b>Discrete variables</b>	$n$ (%)	$n$ (%)	$n$ (%)	$n$ (%)
Sex*				
<i>Male</i>	522 (96.13)	459 (95.62)	405 (95.29)	308 (95.06)
<i>Female</i>	21 (3.86)	21 (4.38)	20 (4.71)	16 (4.94)
Cluster*				
<i>A</i>	164 (30.20)	157 (32.70)	148 (34.82)	137 (42.28)
<i>B</i>	195 (35.91)	177 (36.87)	155 (36.47)	114 (35.18)
<i>C</i>	89 (16.39)	71 (14.79)	60 (14.12)	39 (12.04)
<i>D</i>	95 (17.50)	75 (15.64)	62 (14.59)	34 (10.50)
Anxiety				
<i>No</i>	459 (84.51)	409 (85.21)	368 (86.59)	265 (81.79)
<i>Yes</i>	84 (15.47)	71 (14.79)	57 (13.41)	59 (18.21)
Depression				
<i>No</i>	506 (93.19)	439 (91.46)	389 (91.53)	299 (92.28)
<i>Yes</i>	37 (6.81)	41 (8.54)	36 (8.47)	25 (7.72)
Dyspnea				
<i>1</i>	69 (12.71)	85 (17.71)	75 (17.65)	57 (17.59)
<i>2</i>	264 (48.62)	248 (51.67)	188 (44.24)	134 (41.36)
<i>3</i>	166 (30.57)	127 (26.46)	142 (33.41)	100 (30.86)
<i>4-5</i>	44 (8.10)	20 (4.17)	23 (5.41)	33 (10.19)
<b>Continuous variables</b>	Mean (SD)			
Age at baseline*	68.32 (8.32)	67.61 (8.36)	67.42 (8.29)	66.24 (8.36)
FEV1%	55.00 (13.31)	55.21 (16.05)	57.87 (14.66)	54.27 (14.81)
BMI	28.28 (4.43)	28.33 (5.24)	28.10 (4.44)	27.64 (4.79)
Walking Test	408.89 (92.43)	420.56 (117.55)	412.92 (115.28)	397.36 (123.00)

SD: Standard Deviation; BMI: Body Mass Index; FEV1%: Forced Expiratory Volume in one second in percentile. Symbol \* stands for time independent covariates.

in the other dimensions, causing original standardised scores with higher variability. In terms of the results provided by the SGRQ, apparently, as time goes by there is no much change neither in the mean nor the standard deviation of the dimensions for the survivors.

Table 1.3: Descriptive analysis of PROs provided by the SF-36 and the SGRQ in the COPD Study.

	Time framework			
	Baseline	1-Year	2-Year	5-Year
No. Individuals	$n = 543$	$n = 480$	$n = 425$	$n = 324$
Dimensions	Mean (SD)			
<b>Short Form-36</b>				
<i>Physical functioning</i>	57.76 (24.38)	58.17 (24.95)	57.79 (24.68)	56.46 (24.96)
<i>Role physical</i>	65.61 (38.92)	60.99 (39.93)	62.65 (40.01)	55.48 (41.16)
<i>Bodily pain</i>	71.09 (29.26)	67.74 (30.33)	69.25 (29.96)	68.58 (29.00)
<i>General health</i>	44.67 (21.93)	43.36 (23.32)	42.28 (22.47)	41.80 (20.92)
<i>Vitality</i>	59.36 (24.96)	58.27 (24.00)	59.64 (23.42)	57.58 (23.88)
<i>Social functioning</i>	81.58 (24.46)	79.92 (25.89)	82.18 (24.14)	77.89 (26.13)
<i>Role emotional</i>	80.17 (35.91)	73.96 (39.42)	76.71 (37.95)	70.37 (41.14)
<i>Mental health</i>	73.42 (22.92)	73.17 (21.86)	73.39 (22.16)	71.63 (23.11)
<b>St. George</b>				
<i>Symptoms</i>	44.54 (22.18)	42.48 (22.36)	43.40 (23.25)	44.06 (23.38)
<i>Activity</i>	48.69 (24.94)	45.90 (24.97)	46.89 (24.74)	47.37 (25.35)
<i>Impacts</i>	32.05 (20.89)	30.36 (21.21)	30.23 (20.32)	30.39 (20.89)

SD: Standard Deviation.

### 1.3.2 Paquid Research Programme

It is well known that increasing longevity and declining fertility rates are shifting the age distributions of populations toward older age groups in many parts of the world, including Europe, the United States of America and, in fact, most industrialised nations (Anderson et al., 2000). Improved sanitation, medical technology, and healthcare services, as well as increased individual wealth, have all contributed to rising life expectancy (WHO, World Health Statistics report, 2017). According to the United Nations demographics indicator, the relative population of individuals aged 65 and above will increase rapidly in industrialised countries by an average of 16.8 percent between 2000 and 2020 (Anderson et al., 2000).

As the population grows older, age-related diseases such as dementia will increase, and issues such as providing proper health care and disease treatment will

come to the forefront. The resulting financial and personal costs might devastate the world's economic and healthcare systems, in addition to burdening many families worldwide. Changes in public policies must be implemented to accommodate financial security, healthcare provision and living arrangements (Chan, 2001).

Dementia is a cognitive disorder that affects the brain and results in failing memory and personality changes (Martin, 2009). In 2010 there were an estimated 35.6 million people with dementia worldwide. This number will nearly double every 20 years, resulting in an estimated 65.7 million in 2030 and 115.4 million in 2050. Much of this increase will occur in developing countries. At present, 58 percent of people with dementia reside in developing countries; by 2050, this figure will rise to 71 percent. By 2050, individuals aged 60 years and over will account for 22 percent of the world's population, with four-fifths living in Asia, Latin America and Africa (Ferri et al., 2005). The incidence of mental and neurological illnesses is high in Europe too, with nearly 165 million people (38 percent of the population) suffering from disorders such as depression, anxiety, insomnia or dementia each year (Wittchen et al., 2011).

The Paquid research program was designed to study the incidence of dementia and Alzheimer's disease in elderly people in South-Western France (Letenneur et al., 1994). Subjects were randomly selected from the electoral rolls of 37 parishes in Gironde and followed-up over a maximum period of 20 years. Three criteria had to be met for subjects to be included in the study: (i) to be more than 65 years by 31 December 1987; (ii) to be living at home at the time of the initial data collection phase; and (iii) to give their informed consent to participate in the study. The selection procedure led to the inclusion of 4050 elderly subjects living at home, which, finally, 2792 agreed to participate in the program. Intellectual functioning was examined through a series of psychometric tests which were the most sensitive for following a cognitive decline in elderly individuals. The battery test included the MMSE questionnaire introduced in Section 1.2.3. Additionally, socio-demographical variables of the individuals were measured, such as depressive symptomatology and subjective health measures. More detailed information about the study and the sample can be found in the original reference (Letenneur et al., 1994).

In this thesis, we have considered a subsample of the Paquid research programme. The data are publicly available in `lcmm` R-package (Proust-Lima et al., 2017). The data consists of 2050 observations over 498 subjects and includes dementia information variables, such as dependency level, depressive symptomatology, dementia status and age at dementia diagnosis and also time independent socio-demographic

variables such as educational level and age at entry in the cohort.

Table 1.4 shows a descriptive analysis of the outcome and the exploratory covariates available in the subsample of the Paquid Study. For discrete covariates, we show the number of patients and the relative percentage, while for continuous the mean and standard deviation are displayed.

Several insights can be carried out from Table 1.4. For instance, it can be appreciated that the number of individuals in the cohort decreases notoriously as the study goes on. Moreover, as time goes by, the dementia status of patients worsens considerably and, hence, while the percentage of no dependency in the initial point was 25.30%, in the end, it decreases to 3.80%. However, as regards to the MMSE score, apparently, there is not any evolution over time, as means and standard deviations remain quite constant for all the time points.

## 1.4 Distributional features

In this section, we present the statistical challenges there exist when trying to fit PROs. First, we present the exponential family as a very well-known class of models which include most of the distributions used in practice. However, some special characteristics that PROs usually present make the fit by exponential family distributions inadequate. Therefore, we will define the beta-binomial distribution, which has been illustrated in the literature to get satisfactory distributional fits of PROs (Arostegui et al., 2007). Finally, we fit some exponential family distributions together with the beta-binomial distribution to the PROs provided in the two datasets presented in Section 1.3.

### 1.4.1 The exponential family

The exponential family is a very wide class of models that includes most of the commonly used distributions in practice. A general  $p$ -parameter exponential family depends on the parameter vector  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)$  and its log-density is of the form

$$\log f(x|\boldsymbol{\theta}) = \sum_{i=1}^p \eta_i(\boldsymbol{\theta})T_i(x) - A(\boldsymbol{\theta}) + c(x),$$

for known functions  $A(\boldsymbol{\theta})$  and  $c(x)$ , and  $\eta_i(\boldsymbol{\theta})$  and  $T_i(x)$  for each  $i$ ,  $i = 1, \dots, p$ . The parameters  $\eta_i$ 's are called the *natural parameters* of the family, and  $T_i$ 's the *natural statistics*.

Table 1.4: Descriptive analysis of the covariates available in the subsample of the Paquid research programme.

	Time framework								
	Baseline	1	2	3	4	5	6	7	8
No. Individuals	<i>n</i> = 498	<i>n</i> = 405	<i>n</i> = 329	<i>n</i> = 254	<i>n</i> = 203	<i>n</i> = 150	<i>n</i> = 110	<i>n</i> = 75	<i>n</i> = 26
<b>Discrete variables</b>	<i>n</i> (%)								
Educational level*									
<i>No</i>	145 (29.12)	113 (27.90)	85 (25.84)	63 (24.80)	46 (22.66)	36 (24.00)	26 (23.64)	16 (21.33)	5 (19.23)
<i>Yes</i>	353 (70.88)	292 (72.10)	244 (74.16)	191 (75.20)	157 (77.34)	114 (76.00)	84 (76.36)	59 (78.67)	21 (80.77)
Dependency level									
<i>No</i>	126 (25.30)	117 (28.90)	97 (29.50)	67 (26.40)	39 (19.20)	28 (18.60)	2 (1.80)	5 (6.60)	1 (3.80)
<i>Mild</i>	243 (48.80)	156 (38.50)	114 (34.65)	89 (35.00)	76 (37.40)	55 (36.60)	49 (44.55)	24 (32.00)	3 (11.50)
<i>Moderate</i>	115 (23.10)	114 (28.10)	94 (28.60)	79 (31.00)	74 (36.50)	56 (37.40)	49 (44.55)	36 (49.00)	20 (76.90)
<i>Severe</i>	14 (2.80)	18 (4.50)	24 (7.25)	19 (7.60)	14 (6.90)	11 (7.40)	10 (9.20)	10 (13.40)	2 (7.80)
Dementia diagnosis									
<i>No</i>	370 (74.30)	283 (69.90)	225 (68.40)	171 (67.30)	133 (65.50)	95 (63.30)	68 (61.80)	48 (64.00)	21 (80.80)
<i>Yes</i>	128 (25.70)	122 (30.10)	104 (31.60)	83 (32.60)	70 (34.50)	55 (36.60)	42 (38.20)	27 (36.00)	5 (19.20)
<b>Continuous variables</b>	Mean (SD)								
Age at entry*	74.23 (6.39)	73.63 (6.06)	73.24 (5.84)	72.41 (5.19)	71.98 (4.88)	71.42 (4.54)	71.17 (4.32)	70.33 (3.82)	69.67 (3.04)
Age at dementia	83.79 (6.93)	84.80 (6.46)	85.97 (5.77)	86.64 (5.34)	87.44 (4.89)	88.24 (4.55)	88.78 (4.37)	89.05 (3.87)	89.47 (4.01)
Age	75.88 (6.47)	77.52 (6.22)	79.43 (6.09)	80.97 (5.47)	83.04 (5.07)	85.08 (4.65)	86.93 (4.38)	88.50 (3.71)	90.04 (3.12)
<b>Outcome variable</b>	Mean (SD)								
MMSE	26.97 (2.61)	26.70 (3.05)	26.50 (3.47)	26.74 (3.37)	26.32 (3.97)	25.68 (4.85)	25.65 (4.99)	25.67 (3.57)	26.00 (2.93)

SD: Standard Deviation; Age at dementia: Age at dementia diagnosis for patients diagnosed with dementia and last contact for patients free of dementia. Symbol \* stands for time independent covariates.

The exponential family includes both discrete and continuous random variables such as the normal, binomial, Poisson or gamma distributions. However, although it covers a wide range of distributions, not all the models are included in the exponential family, the Cauchy and the  $t$ -distribution for instance.

**Example 1.1.** Consider the commonly used normal distribution with parameter vector  $\theta = (\mu, \sigma^2)$ , we have that the density function is defined as

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}.$$

If we rewrite the density function in an exponential form as

$$\log f(x|\mu, \sigma^2) = \frac{\mu x}{\sigma^2} - \frac{x^2}{2\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{1}{2} \log(2\pi\sigma^2),$$

we realise that the normal distribution defines a two-parameter exponential family distribution with natural parameters  $\eta_1 = \mu/\sigma^2$  and  $\eta_2 = -1/(2\sigma^2)$ , and natural statistics  $T_1(x) = x$  and  $T_2(x) = x^2$ .

□

**Example 1.2.** For the Poisson model with mean  $\lambda$  we have that the log-density function is defined as

$$\log f(x|\lambda) = x \log \lambda - \lambda - \log x!.$$

Hence, it is straightforward to prove that the Poisson distribution is a one-parameter exponential family member.

□

The joint distribution of an independent and identically distributed (iid) sample from an exponential family also belongs to the exponential family. The previous statement is very useful when we are developing a model for a group of observations that are assumed to be drawn from the same distribution. In addition, for exponential family distributions of the form

$$\log f(x|\theta) = \theta x - A(\theta) + c(x), \tag{1.1}$$



it is shown that

$$\begin{aligned}\mu &= \mathbb{E}[X] = A'(\theta) \\ \text{Var}[X] &= A''(\theta) = \frac{\partial}{\partial \theta} \mathbb{E}[X] = A'' [A'^{-1}(\mu)] = v(\mu),\end{aligned}\tag{1.2}$$

where  $v(\cdot)$  is the so-called variance function for the mean  $\mu$  (see McCullagh and Nelder (1989) for further details). Therefore,  $A(\theta)$  implies a certain relationship between the expectation and the variance of the random variable. The displayed relationship between the mean and the variance is very useful in modelling approaches, as the distribution family is completely defined by knowing the first two moments.

For statistical modelling, it is often adequate to consider a two-parameter model known as the exponential dispersion model (Jørgensen, 1987). Based on an observation  $x$  the log-likelihood of the scalar parameters  $\theta$  and  $\phi$  is of the form

$$\log L(\theta, \phi) = \frac{x\theta - A(\theta)}{\phi} + c(x, \phi),\tag{1.3}$$

where  $A(\theta)$  and  $c(x, \phi)$  are assumed known functions. In this form the parameter  $\theta$  is called the *canonical parameter* and  $\phi$  the *dispersion parameter*. Since  $A(\theta)$  and  $c(x, \phi)$  can be anything there are infinitely many submodels in the dispersion exponential family, though the density must satisfy

$$\sum_x \exp \left\{ \frac{x\theta - A(\theta)}{\phi} + c(x, \phi) \right\} = 1,$$

which forces a certain relationship between  $A(\theta)$  and  $c(x, \phi)$ .

Compared to the more rigid exponential family defined in Equation (1.1), in this case, the variance is not closely defined by the mean as

$$\text{Var}[X] = \phi A''(\theta) = \phi \frac{\partial}{\partial \theta} \mathbb{E}[X] = \phi v(\mu).$$

This is indeed, the biggest advantage of the exponential dispersion model over the more rigid model in Equation (1.1).

In practise, while  $A(\theta)$  is explicitly given,  $c(x, \phi)$  is left implicit in the model. However, this is not a problem as far as the maximum likelihood estimation (MLE) of  $\theta$  is concerned, since the score equation,

$$S(\theta) = \frac{\partial}{\partial \theta} \log L(\theta, \phi) = \frac{x - A'(\theta)}{\phi},$$

does not involve the term  $c(x, \phi)$ . However, it does not allow the performance of a likelihood based estimation of  $\phi$  or a full-likelihood inference of both parameters. One option is to estimate  $\phi$  with the method of moments, although there exist some other techniques. In this work, we are going to develop a likelihood approximation approach, usually called as *quasi-likelihood* (Wedderburn, 1974), which in general is easy to implement.

First, we need to define some regularity conditions that the density function of the random variable must satisfy in order to develop the quasi-likelihood theory.

**Definition 1.1.** *Assume that we have  $X_i, i = 1, \dots, n$ , iid variables distributed with density  $f(x|\theta)$  and let define  $\hat{\theta}$  and  $\theta_0$  as the MLE of the parameter vector  $\theta$  and the true but unknown parameter value respectively. Then if,*

1.  $\theta_0 \in \text{Int}(\Omega)$ , where  $\Omega$  is the parameter space
2. the true but unknown parameter value  $\theta_0$  is identified, i.e.

$$\theta_0 = \arg \max_{\theta \in \Omega} \mathbb{E} \log f(X_i|\theta)$$

3. the log-likelihood function

$$l(\theta|x_1, \dots, x_n) = \sum_{i=1}^n \log f(x_i|\theta)$$

is continuous in  $\theta$

4.  $\mathbb{E} \log f(X_1, \dots, X_n|\theta)$  exists
5. the log-likelihood function is such that  $\frac{1}{n}l(\theta|x_1, \dots, x_n)$  converges almost surely to  $\mathbb{E} \log f(X_i|\theta)$  uniformly in  $\theta \in \Omega$ , i.e.

$$\sup_{\theta \in \Omega} \left| \frac{1}{n}l(\theta|x_1, \dots, x_n) - \mathbb{E} \log f(X_i|\theta) \right| < \epsilon \text{ almost surely for some } \epsilon > 0$$

6. the log-likelihood function is twice continuously differentiable in a neighbourhood of  $\theta_0$
7. integration and differential operators are interchangeable

8. the expected Fisher information matrix

$$\mathcal{I}(\boldsymbol{\theta}_0) = \mathbb{E} \left( -\frac{\partial^2 \log f(X_1, \dots, X_n | \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right)$$

exists and it is non-singular

then, we say that regularity conditions are satisfy.

It is worth mentioning that all the members of the exponential family and commonly used distributions satisfy the regularity conditional defined in Definition 1.1 (Pawitan, 2001).

On the one hand, at fixed  $\phi$ , the MLE of  $\theta$ , denoted as  $\hat{\theta}$ , for the observation  $x$  is the solution of the score equation,

$$S(\hat{\theta}) = \frac{\partial}{\partial \theta} \log L(\theta | x, \phi) \Big|_{\theta=\hat{\theta}} = \frac{x - A'(\hat{\theta})}{\phi} = 0,$$

which implies that  $A'(\hat{\theta}) = x$ .

On the other hand, under regularity conditions, it can be proved that the MLE of  $\theta$  follows a normal distribution,

$$\hat{\theta} \sim \mathcal{N}(\theta_0, \mathcal{I}(\theta_0)^{-1}),$$

where in dispersion exponential families it is equivalently to (Pawitan, 2001)

$$\hat{\theta} \sim \mathcal{N}(\theta, I(\hat{\theta})^{-1}),$$

being  $I(\cdot)$  the observed Fisher information matrix, defined as

$$I(\hat{\theta}) = -\frac{\partial^2}{\partial \theta^2} \log L(\theta | x, \phi) \Big|_{\theta=\hat{\theta}} = \frac{A''(\hat{\theta})}{\phi}.$$

Consequently, we have that the asymptotic distribution of the MLE of  $\theta$  is defined as

$$f(\hat{\theta}) \approx (2\pi)^{-1/2} I(\hat{\theta})^{1/2} \exp \left\{ -\frac{I(\hat{\theta})}{2} (\hat{\theta} - \theta)^2 \right\}. \quad (1.4)$$

Moreover, if the likelihood for the  $x$  contribution is regular, i.e. the log-likelihood is well approximated by a quadratic function, we can apply a second order Taylor

series around the MLE,

$$\log L(\theta) \approx \log L(\hat{\theta}) + S(\hat{\theta})(\theta - \hat{\theta}) - \frac{1}{2}I(\hat{\theta})(\hat{\theta} - \theta)^2, \quad (1.5)$$

and get the following approximation of the log-ratio of the likelihoods,

$$\log \frac{L(\theta)}{L(\hat{\theta})} \approx -\frac{1}{2}I(\hat{\theta})(\hat{\theta} - \theta)^2. \quad (1.6)$$

Given Equation (1.4) and Equation (1.6) we obtain that the density function of the MLE can be approximated by,

$$f(\hat{\theta}) \approx (2\pi)^{-1/2} I(\hat{\theta})^{1/2} \frac{L(\theta)}{L(\hat{\theta})}.$$

As we only have one observation, the estimation of the mean of the distribution,  $\hat{\mu}$ , is equal to the observed unique outcome  $x$ . Additionally, from exponential families (see Equation (1.2)) we have that  $\mu = \mathbb{E}[X] = A'(\theta)$ , and hence, that  $\hat{\mu} = A'(\hat{\theta})$ .

Therefore, we can conclude that the approximated density function for the single observation  $x$  is defined as

$$f(x) = f(\hat{\mu}) = f(\hat{\theta}) \left| \frac{\partial \hat{\theta}}{\partial \hat{\mu}} \right| = f(\hat{\theta}) A''(\hat{\theta})^{-1} \approx (2\pi \phi v(x))^{-1/2} \frac{L(\theta)}{L(\hat{\theta})} \quad (1.7)$$

where  $v(x)$  is the variance function evaluated in the observation  $x$  and  $\hat{\theta}$  is the MLE of the canonical parameter for the unique observation. Notice that the term  $c(x, \phi)$  cancels out in the likelihood ratio term, so we end up with something simpler.

The last term in Equation (1.7) is directly related to the *deviance function* which is defined as

$$D(x, \theta) = 2 \log \frac{L(\hat{\theta}, \phi = 1|x)}{L(\theta, \phi = 1|x)} = 2 \left[ x\hat{\theta} - x\theta - A(\hat{\theta}) + A(\theta) \right], \quad (1.8)$$

where  $\hat{\theta}$  is the MLE for the single observation  $x$ . If we had  $x_1, \dots, x_n$  an iid sample, the total deviance of the model would be defined as

$$D = \sum_{i=1}^n D(x_i, \theta), \quad (1.9)$$

where for fixed  $\phi$  the MLE of  $\theta$  is the parameter that minimises the total deviance (Jørgensen, 1997). In fact, although the deviance is useful and necessary to de-

velop the quasi-likelihood theory, it is commonly used as a goodness-of-fit criteria to compare nested models.

Finally, we get the expression of the approximation of the log-likelihood of an exponential dispersion family distribution for a single observation  $x$  as

$$\log L(\theta, \phi|x) \approx -\frac{1}{2} \log(2\pi\phi v(x)) - \frac{1}{2\phi} D(x, \theta). \quad (1.10)$$

**Example 1.3** In this example, we consider the binomial dispersion model or, simply, the binomial distribution with dispersion parameter. First of all, we rewrite the binomial distribution in an exponential family way. Once the canonical parameter  $\theta$  and the  $A(\cdot)$  function of the binomial distribution are identified, we try to extend it to the dispersion model approach and we approximate the likelihood function with the theory developed above.

Let us assume that the random variable  $X$  follows a binomial distribution with parameters  $m$  and  $p$ . Then, the density function is defined as

$$f(X = x|p) = \binom{m}{x} p^x (1-p)^{m-x}.$$

Notice that we do not condition the density function on the maximum score number  $m$  as it is usually given as known and, instead, focus our attention on the probability parameter  $p$ .

Equivalently, we can rewrite the log-density of the binomial distribution in an exponential family way

$$\log f(x|p) = x \log \frac{p}{1-p} + m \log(1-p) + \log \binom{m}{x},$$

where we define  $\theta = \log[p/(1-p)]$  as the canonical parameter of the distribution and  $A(\theta) = -m \log(1-p) = m \log(1 + e^\theta)$ .

Notice that the relationships defined in Equation (1.2) are satisfied as follows,

$$\begin{aligned} A'(\theta) &= \frac{\partial}{\partial \theta} m \log(1 + e^\theta) = m \frac{e^\theta}{1 + e^\theta} = mp = \mathbb{E}[X], \\ A''(\theta) &= \frac{\partial}{\partial \theta} A'(\theta) = \frac{\partial}{\partial \theta} m \frac{e^\theta}{1 + e^\theta} = m \frac{e^{-\theta}}{(1 + e^\theta)^2} = mp(1-p) = \text{Var}[X]. \end{aligned}$$

For that canonical parameter and  $A(\cdot)$  function we can define the log-likelihood

of the dispersion model as

$$\log L(\theta, \phi|x) = \frac{x\theta - m \log(1 + e^\theta)}{\phi} + c(x, \phi), \quad (1.11)$$

where  $\theta = \log [p/(1 - p)]$ .

As it was already mentioned, the model defined in Equation (1.11) is enough if we want to estimate the canonical parameter or, equivalently, the probability parameter  $p$ . However, it allows neither a full likelihood inference of  $\theta$  (or  $p$ ) nor a likelihood based estimation of the dispersion parameter  $\phi$ . Nevertheless, we can approximate the log-likelihood of the binomial dispersion model with the quasi-likelihood approximation defined in Equation (1.10).

First, we know that for a single observation the estimation of the probability parameter is defined as the ratio between the number of successes and the number of trials, i.e.  $\hat{p} = x/m$ , hence, we define the deviance of the model as

$$D(x, \theta) = 2 \left[ x \log \frac{x/m}{1 - x/m} - x \log \frac{p}{1 - p} + m \log \left( 1 - \frac{x}{m} \right) - m \log(1 - p) \right].$$

Consequently, using Equation (1.10), the approximated log-likelihood of the binomial dispersion model or binomial distribution with dispersion parameter is defined as

$$\log L(\theta, \phi|x) \approx -\frac{1}{2} \log \left( 2\pi\phi \frac{x}{m} \left( 1 - \frac{x}{m} \right) \right) - \frac{1}{2\phi} D(x, \theta).$$

The above formula allows the full-likelihood based estimation of all the parameters involving the binomial distribution with dispersion parameter. Additionally, it also allow for an inference procedure of the parameters.

□

It has been mentioned that the exponential dispersion model relaxes the relationship between the mean and the variance, however, in some cases, this accommodation is not enough for the correct distributional fit of the variable. In fact, most of the PROs include some characteristics that make the fit by exponential family distributions inefficient (Arostegui et al., 2007). For instance, we consider the SF-36 dimensions provided in the COPD Study (see Section 1.3.1) as an illustration of the inefficient fit of PROs by exponential family distributions.

Many of PROs studies, in particular, HRQoL studies, perform statistical analysis assuming that the dimensions provided by the SF-36 follow a normal distribution (Pal et al., 2017; Sánchez-García et al., 2017). Figure 1.1 shows the distribution of the original standardised SF-36 dimensions of the COPD Study in a cross-sectional setting, in this case corresponding to the baseline measurements. It evidences that, on the one hand, the dimensions have different density shapes and, on the other hand, the weak distributional fit of the normal distribution. In fact, it can be appreciated that most of the SF-36 dimensions do not show bell-shaped or Gaussian densities, and instead, they follow skewed distributions which accumulate values in one or two edges of the distribution scale. Additionally, it is worth mentioning that while the normal distribution is defined on the real line  $\mathbb{R}$ , the original standardised SF-36 dimensions are bounded to  $[0,100]$ . Moreover, the raw dimensions are constructed by summing up some item scores and they can only reach a finite number of values (see Table 1.1). Therefore, although the raw scores are standardised to the 0-100 scale, they still maintain the integer feature as it can be appreciated in the *role emotional* histogram in Figure 1.1, which only can take values equal to 0, 33.33, 66.66 or 100. This assumption does not match with the continuity of the normal distribution and, hence, it makes the fit of the original standardised SF-36 dimensions by the normal distribution quite senseless.

The COPD Study is a specific study that measures specific PROs in an specific population. However, it has been illustrated in the literature that many PROs measured in different populations share the previously cited characteristics (Izem et al., 2014; Arostegui and Núñez-Antón, 2008). Indeed, the PROs are usually bounded and tend to accumulate values in one or both sides of the range, which makes them far from Gaussian or bell-shaped distributions. Additionally, we must also mention the integer feature of PROs due to the way they are constructed. Therefore, the number of distributions available in the exponential family that may match with the cited characteristics reduces considerably. In fact, only two distributions in the exponential family can be adapted to the features of the PROs, although with some limitations: (i) the beta distribution; and (ii) the binomial distribution.

The beta distribution is a continuous two parameter exponential family model which is bounded in the open interval  $(0,1)$ . The density function of the beta distribution is given by

$$f(x|\alpha_1, \alpha_2) = \frac{x^{\alpha_1-1}(1-x)^{\alpha_2-1}}{B(\alpha_1, \alpha_2)}, \quad (1.12)$$

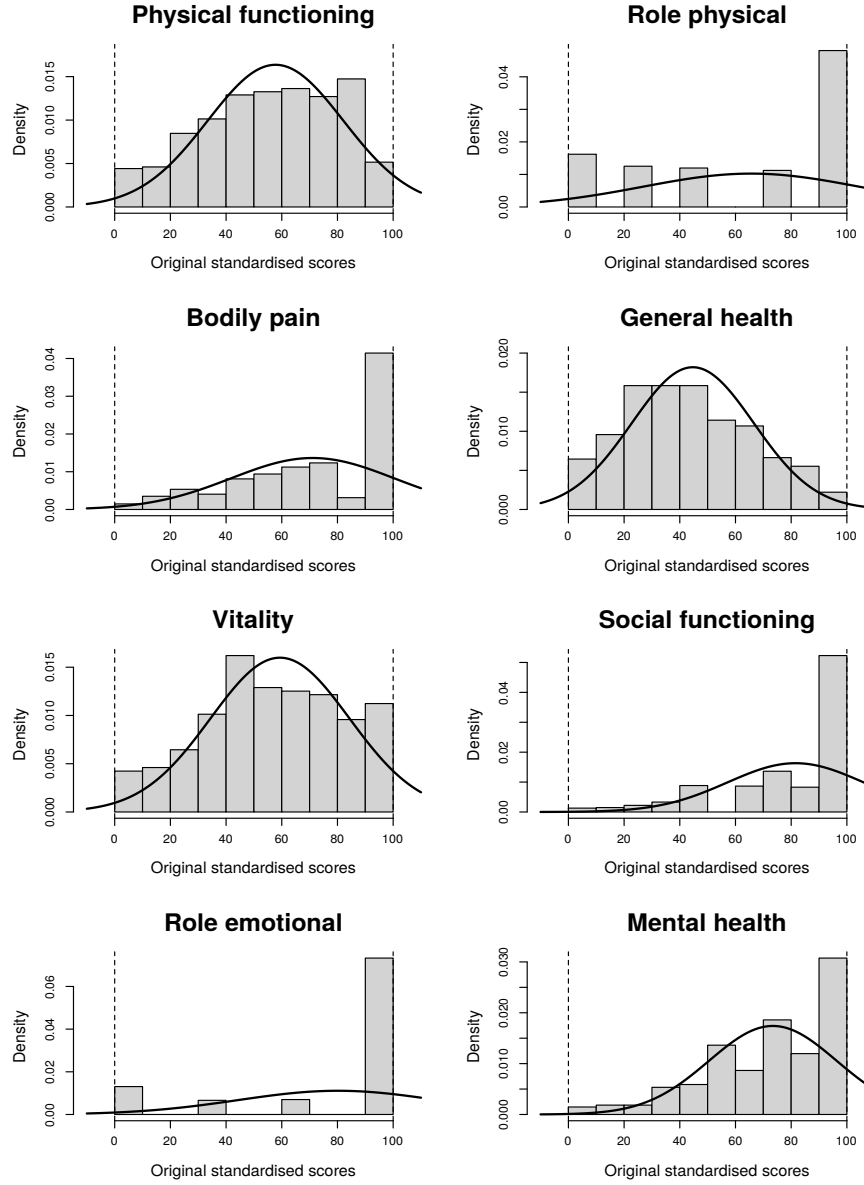


Figure 1.1: Histograms of the original standardized SF-36 scores in cross sectional COPD data. The black lines correspond to the normal distributional fit.

where  $\alpha_1$  and  $\alpha_2$  are the positive parameters of the distribution and  $B(\alpha_1, \alpha_2)$  is the beta function defined as

$$B(\alpha_1, \alpha_2) = \int_0^1 t^{\alpha_1-1} (1-t)^{\alpha_2-1} dt. \quad (1.13)$$



Figure 1.2 shows the different shape the beta distribution can reach depending on the values of the parameters. It can be appreciated that U, J or inverse J-shapes can be obtained. Therefore, due to its flexibility, it has been proposed by several authors in the literature to model PROs (Basu and Manca, 2012; Hunger et al., 2011). Regarding the distribution domain, PROs must be rescaled to the  $(0,1)$  interval before they are analysed by the beta distribution. This lead to some floor and ceiling problems as, for instance, if an individual reaches the maximum score number, the rescale process will link the score value with 1, but the beta distribution is not defined in the closed interval. Different techniques have been described in the literature to overcome this problem (Smithson and Verkuilen, 2006; Verkuilen and Smithson, 2012). However, the beta distribution has some other limitations that, from our point of view, do not match with the nature of PROs. First, we have mentioned that, although PRO scores are assumed continuous, they have an integer nature and, hence, they do not fit with the continuity of the beta distribution. Additionally, the rescale to the  $(0,1)$  interval can lead to some loss information because the value of the maximum score number is ignored. In fact, we need to keep in mind that PROs are mainly constructed using questionnaires, therefore, the more questions are performed (the more items are summed up), the better is the outcome in terms of variability.

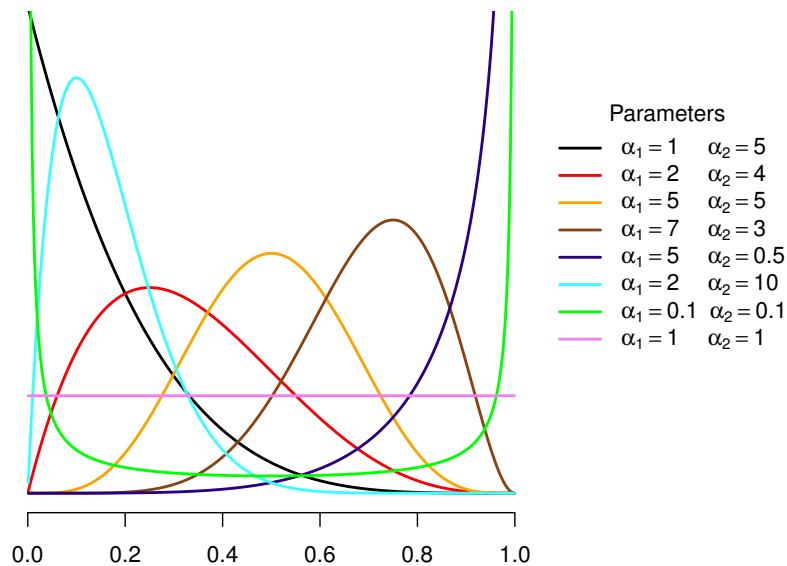


Figure 1.2: Different shapes of the beta distribution for different values of the parameters.

The other distribution from the exponential family that could suit with the nature of PROs is the well-known binomial distribution. It is a bounded integer distribution and, conversely to the beta distribution, it takes into account the maximum score number to calculate the variability of the estimates. The problem of the binomial distribution is the rigidity of its density function that, in general terms, displays a Gaussian form. Even if a binomial dispersion model is considered, the dispersion parameter  $\phi$  only would scale the distribution, but would not change its shape. Consequently, U, J or inverse J-shaped forms will not be correctly fitted by the binomial distribution (Najera-Zuloaga et al., 2017). In Section 1.4.3 we will fit PROs provided in the studies introduced in Section 1.3 by the binomial and binomial with dispersion parameter distributions. It will be shown that the binomial distribution does not offer a good fit to the data.

We have mentioned in the previous paragraphs the advantages and disadvantages of the beta and binomial distributions. On the one hand, we have illustrated the flexibility of the beta distribution but, we have shown its weakness when fitting the nature of PROs. On the other hand, we have explained that the binomial distribution suits to the features of PROs, but offers very poor fittings due to its rigidity. Therefore, we could think of a mixture of both distributions, where the nature of PROs will be preserved by the binomial distribution, but assume that the probability parameter follows a beta distribution for accommodating its flexibility in the model. In fact, the mixture of the binomial and beta distributions is called the *beta-binomial* distribution and it has already been proposed in the literature for fitting PROs (Arostegui et al., 2007).

### 1.4.2 The beta-binomial distribution

The beta-binomial distribution is defined as a mixture of the binomial and beta distributions. It consists of a finite sum of Bernoulli variables whose probability parameter is random and follows a beta distribution.

For instance, consider that we have some binary random variables  $Y_j, j = 1, \dots, m$  where  $m \in \mathbb{N}$ . Assume that conditional on a random variable  $u$  the binary variables are iid from a Bernoulli distribution with parameter  $u$ . Additionally, consider that the random variable  $u$  follows a beta distribution with parameters  $\alpha_1$  and  $\alpha_2$ . Namely, we have that for  $j = 1, \dots, m$

$$Y_j|u \sim \text{Ber}(u) \quad \text{and} \quad u \sim \text{Beta}(\alpha_1, \alpha_2). \quad (1.14)$$

The density function of the beta distribution has been introduced in Equation (1.12) where the first and second order moments are defined as

$$\mathbb{E}[u] = \psi \quad \text{and} \quad \mathbb{V}\text{ar}[u] = \psi(1 - \psi) \frac{\phi}{1 + \phi}, \quad (1.15)$$

being  $\psi = \alpha_1/(\alpha_1 + \alpha_2)$  and  $\phi = 1/(\alpha_1 + \alpha_2)$ , and hence,  $0 < \psi < 1$  and  $\phi > 0$ .

On the contrary, for a binary outcome  $x$  the density function of the Bernoulli distribution with probability parameter  $u$  is defined as

$$f(x|u) = u^x(1 - u)^{1-x},$$

where the first and second order moments are given by

$$\mathbb{E}[X] = u \quad \text{and} \quad \mathbb{V}\text{ar}[X] = u(1 - u). \quad (1.16)$$

Therefore, based on the first and second order moments of the beta and Bernoulli distributions defined in Equation (1.15) and Equation (1.16) respectively, we can conclude with the marginal moments of the mixed distribution as

$$\begin{aligned} \mathbb{E}[Y_j] &= \mathbb{E}[\mathbb{E}[Y_j|u]] = \mathbb{E}[u] = \psi, \\ \mathbb{V}\text{ar}[Y_j] &= \mathbb{V}\text{ar}[\mathbb{E}[Y_j|u]] + \mathbb{E}[\mathbb{V}\text{ar}[Y_j|u]] = \mathbb{V}\text{ar}[u] + \mathbb{E}[u(1 - u)] \\ &= \mathbb{V}\text{ar}[u] + \mathbb{E}[u] - \mathbb{E}[u^2]. \end{aligned}$$

We know from the definition of the variance function that for a random variable  $u$  we have that

$$\mathbb{V}\text{ar}[u] = \mathbb{E}[u^2] - \mathbb{E}[u]^2,$$

and, hence, we have that the marginal variance for each binary variable is defined as

$$\mathbb{V}\text{ar}[Y_j] = \mathbb{V}\text{ar}[u] + \mathbb{E}[u] - \left( \mathbb{V}\text{ar}[u] + \mathbb{E}[u]^2 \right) = \psi(1 - \psi).$$

Notice that the marginal expectation and variance for each binary outcome corresponds to the Bernoulli distribution with probability parameter  $\psi$ . However, due to the fact that all the binary variables  $j = 1, \dots, m$  are conditioned on the same random variable, we have that the covariance and correlation within observations

are defined as

$$\begin{aligned}\mathbb{Cov}[Y_j, Y_k] &= \mathbb{Cov}[\mathbb{E}[Y_j|u], \mathbb{E}[Y_k|u]] + \mathbb{E}[\mathbb{Cov}[Y_j, Y_k|u]] \\ &= \mathbb{Cov}[u, u] = \text{Var}[u] = \psi(1-\psi) \frac{\phi}{1+\phi}, \\ \text{Corr}[Y_j, Y_k] &= \frac{\mathbb{Cov}[Y_j, Y_k]}{\sqrt{\text{Var}[Y_j]} \sqrt{\text{Var}[Y_k]}} = \frac{\phi}{1+\phi},\end{aligned}\tag{1.17}$$

$\forall j, k = 1, \dots, m$  and  $j \neq k$ . Therefore, Equation (1.17) determines the parameter  $\phi$  as the *dispersion* or *correlation* parameter of the distribution.

If we sum up all the random variables  $Y_j$ ,  $j = 1, \dots, m$ , we define a new variable as

$$Y = \sum_{j=1}^m Y_j,\tag{1.18}$$

which follows the so called beta-binomial distribution.

The marginal density function of the beta-binomial distribution is defined as

$$\begin{aligned}f(y) &= \int_0^1 f(y|u) f(u) du = \int_0^1 f\left(\sum_{i=1}^m y_i | u\right) f(u) du \\ &= \int_0^1 \binom{m}{\sum_{i=1}^m y_i} \prod_{i=1}^m f(y_i | u) f(u) du \\ &= \binom{m}{y} \int_0^1 u^y (1-u)^{m-y} \frac{u^{\alpha_1-1} (1-u)^{\alpha_2-1}}{\text{B}(\alpha_1, \alpha_2)} du \\ &= \binom{m}{y} \int_0^1 \frac{u^{\alpha_1+y-1} (1-u)^{\alpha_2+m-y-1}}{\text{B}(\alpha_1, \alpha_2)} \\ &= \binom{m}{y} \frac{\text{B}(\alpha_1 + y, \alpha_2 + m - y)}{\text{B}(\alpha_1, \alpha_2)}.\end{aligned}$$

At this point, we can use the following property of the beta function,

$$\text{B}(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)},$$

where  $\Gamma(\cdot)$  is the gamma function defined as

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt.\tag{1.19}$$

Therefore, the marginal density function of the beta-binomial distribution is given

by

$$f(y) = \binom{m}{y} \frac{\Gamma(\alpha_1 + y)}{\Gamma(\alpha_1)} \frac{\Gamma(\alpha_2 + m - y)}{\Gamma(\alpha_2)} \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1 + \alpha_2 + m)},$$

or equivalently, using that  $\Gamma(x + 1) = x!$ ,  $\forall x \in \mathbb{N}$ , we obtain

$$\begin{aligned} f(y) &= \binom{m}{y} \frac{\prod_{k=0}^{y-1} (\alpha_1 + k) \prod_{k=0}^{m-y-1} (\alpha_2 + k)}{\prod_{k=0}^{m-1} (\alpha_1 + \alpha_2 + k)} \\ &= \binom{m}{y} \frac{\prod_{k=0}^{y-1} (\psi + k\phi) \prod_{k=0}^{m-y-1} (1 - \psi + k\phi)}{\prod_{k=0}^{m-1} (1 + k\phi)}, \end{aligned} \quad (1.20)$$

where if  $y = 0$ , then  $\prod_{k=0}^{y-1} \log(\phi + k\phi) = 1$  or equivalently,  $\prod_{k=0}^{y-1} (\alpha_1 + k) = 1$ ; and if  $y = m$ , then  $\prod_{k=0}^{m-y-1} \log(1 - \psi + k\phi) = 1$ , or equivalently,  $\prod_{k=0}^{m-1} (\alpha_1 + \alpha_2 + k) = 1$ .

Additionally, the marginal first and second order moments of the beta-binomial distribution are defined as

$$\begin{aligned} \mathbb{E}[Y] &= \mathbb{E} \left[ \sum_{i=1}^m Y_i \right] = \sum_{i=1}^m \mathbb{E}[Y_i] = m\psi, \\ \text{Var}[Y] &= \text{Var} \left[ \sum_{i=1}^m Y_i \right] = \sum_{i=1}^m \text{Var}[Y_i] + \sum_{j \neq k} \text{Cov}[Y_j, Y_k] \\ &= \sum_{i=1}^m \psi(1 - \psi) + \sum_{j \neq k} \psi(1 - \psi) \frac{\phi}{1 + \phi} \\ &= m\psi(1 - \psi) + m(m - 1)\psi(1 - \psi) \frac{\phi}{1 + \phi} \\ &= m\psi(1 - \psi) \left[ 1 + (m - 1) \frac{\phi}{1 + \phi} \right]. \end{aligned} \quad (1.21)$$

Notice, that the marginal moments defined in Equation (1.21) correspond to the binomial distribution with probability parameter  $\psi$  and a number of trials  $m$ , except for the variance that includes an additional multiplicative term. In fact, the binomial distribution is constructed following a similar procedure, however, the summed up Bernoulli variables are assumed independent. This is, indeed, the reason of the multiplicative term in the variance of the beta-binomial distribution. The bigger

is the correlation between the binary variables ( $\phi/(1 + \phi) \gg 0$ ), the larger is the difference between the variance of the binomial and the beta-binomial distributions. Finally, it is worth mentioning that when the dispersion or correlation parameter  $\phi$  is zero, the beta-binomial distribution reduces to the binomial distribution.

Alternatively, there exists a more direct way of defining a beta-binomial distribution.

**Definition 1.2.** *The random variable  $Y$  follows a beta-binomial distribution, if conditioned on the beta distributed random variable  $u$ , it is drawn from a binomial distribution with probability parameter  $u$ , i.e.*

$$Y \sim BB(m, \psi, \phi) \quad \text{if} \quad Y|u \sim \text{Bin}(m, u) \quad \text{and} \quad u \sim \text{Beta}(\alpha_1, \alpha_2)$$

where  $\psi = \alpha_1/(\alpha_1 + \alpha_2)$  and  $\phi = 1/(\alpha_1 + \alpha_2)$ .

There exist different packages in R to fit a beta-binomial distribution, such as `rmutil`, `TailRank`, `emdbook`, `VGAM` and `gamlss`. Moreover, we have developed our own functions available in `PROreg` R-package (see Chapter 5 further details).

In terms of the distributional shape of the beta-binomial distribution, Figure 1.3 shows the different forms it can reach for different parameter values and a fixed number of summed binary variables, or in a binomial framework, a maximum score number equal to  $m = 10$ . It can be appreciated that the beta-binomial distribution preserves the characteristics of the binomial distribution that suit with the nature of PROs (discrete and bounded). Nevertheless, compared to the binomial distribution, Figure 1.3 displays the flexibility that the introduction of the beta distribution offers. In fact, if we compare the shapes of the beta-binomial distributions with the shapes of the beta distributions in Figure 1.2 for the same parameters, we realise that the beta-binomial distribution is somehow a ‘discretised’ version of the beta distribution in the  $0$ - $m$  scale where both maximum and minimum values can be reached.

### 1.4.3 Distributional fit to the datasets

Once the beta-binomial distribution has been introduced and its suitability for the accommodation of the nature of PROs has been exposed, we are going to estimate, as an illustration, the distributional fit of the PROs provided in the studies we have described in Section 1.3.

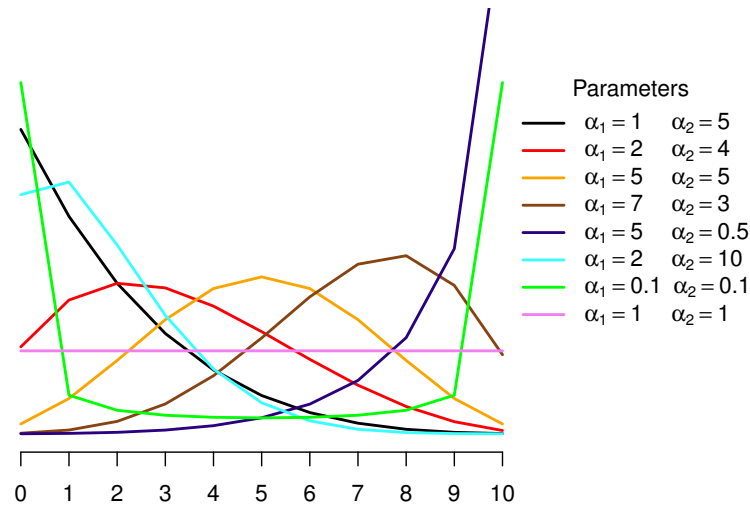


Figure 1.3: Different shapes of the beta-binomial distribution for different values of the parameters and a fixed number of summed binary observations  $m = 10$ .

### Short Form-36 Health Survey

In Section 1.4.2 we have introduced the beta-binomial distribution showing its flexibility and shape variety which matches with most of the shapes the SF-36 dimensions show in the COPD Study (see Figure 1.1). In fact, we have already mentioned that the beta-binomial distribution has been proposed in the literature for fitting PROs (Arostegui et al., 2007), in particular, HRQoL dimensions provided by the SF-36. However, due to the performed standardisation of the raw scores to the  $[0, 100]$  interval, recoding of the scores to a binomial form is necessary in order to fit the beta-binomial distribution to the SF-36 Health Survey scores.

Arostegui et al. (2013) proposed and evaluated a method of recoding continuous and bounded scores, such as HRQoL scores, to a binomial form. The method was mainly based on the possible number of values each dimension can obtain, which, as it has been explained in Section 1.2.1 and showed in Table 1.1, it comes from the number of items related to the construction of the score in each dimension. Indeed, the methodology transforms the sorted scale of possible values of each dimension score to an ordinal scale from 0 to  $m$ , i.e. to a discrete and bounded scale. The real interval  $[0, 100]$ , which is the range of the original standardised scores, is divided into some subintervals, and then, each subinterval is linked to the value it corresponds

to in the  $0 - m$  order scale, where  $m + 1$  is the number of intervals. Consequently, score values within each subinterval are recoded with the value the sub-interval was linked to in the  $0 - m$  scale. The way the subintervals are constructed is the main contribution of Arostegui et al. (2013) and the subdivision of the  $0 - 100$  scale for each dimension is available in the Appendix of the mentioned work. However, for the sake of clarity, we give more detailed information about the recoding process in Appendix A.

Conversely to Figure 1.1 where the original standardised scores are displayed, Figure 1.4 shows the distribution of the eight recoded SF-36 scores in patients with COPD at baseline. Additionally, it shows the fit by the beta-binomial distribution, together with the binomial and binomial dispersion models. Although represented in different scales, similar shapes can be observed for the original standardised scores (Figure 1.1) and the recoded scores (Figure 1.4).

Figure 1.4 illustrates that the distributions of the recoded SF-36 scores are, generally, very skewed, accumulating values at the boundaries. It can be appreciated that, as it was mentioned in Section 1.4.1, due to the characteristics of PROs, in particular HRQoL, both the binomial and the dispersion binomial distributions offer a poor fit in most of the recoded scores (e.g. *role physical*, *bodily pain*, *social functioning*, *role emotional* and *mental health*). In fact, as it was mentioned before, the binomial distribution with dispersion parameter only scales the binomial density function, but it does not alter its shape. Therefore, if we had performed a regression model based on this distribution, results would not be reliable as the binomial distribution does not reflect a good fit to the data. Figure 1.4 also shows that the scores have different shapes (e.g. bell, U or J-shaped), due to the fact that in some dimensions people tend to answer more or less extreme than in others. Consequently, there is an individual within variability in each dimension, that as it can be appreciated, the beta-binomial distribution is able to accommodate.

Figure 1.5 shows descriptively the distribution of the scores of the eight SF-36 dimensions provided by patients with COPD based on different categorical variables, such as gender, dyspnea, anxiety, and depression. It allows a description of the effect of each categorical characteristic in the HRQoL of patients. Each axis of the radar chart corresponds to a recoded SF-36 dimension. The scales have been standardised to the interval defined by the length of the axis and divided into three cut points (25%, 50% and 75%) for a better visualisation of the mean values. In Figure 1.5b we can appreciate the influence of the dyspnea in the different scores, where lower levels of dyspnea are associated with higher health-status in all the dimensions. However,



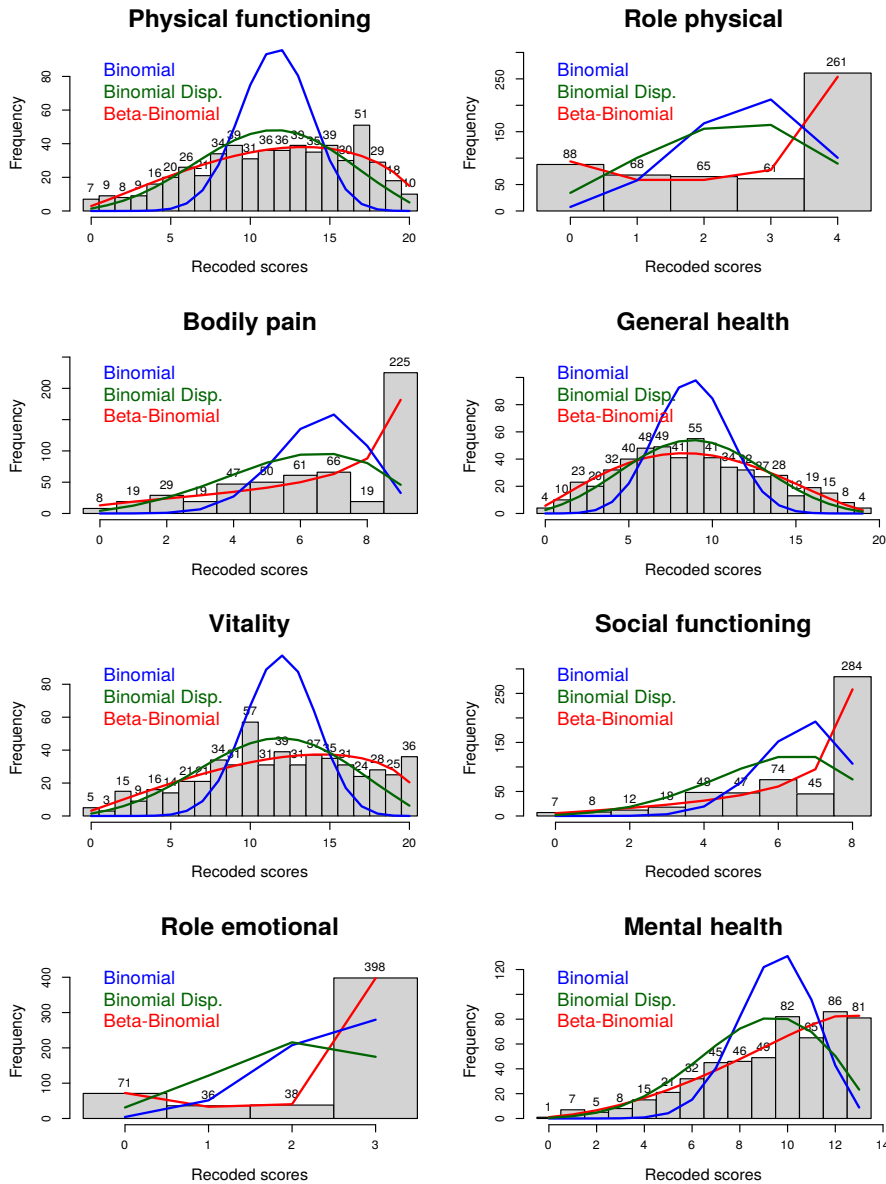
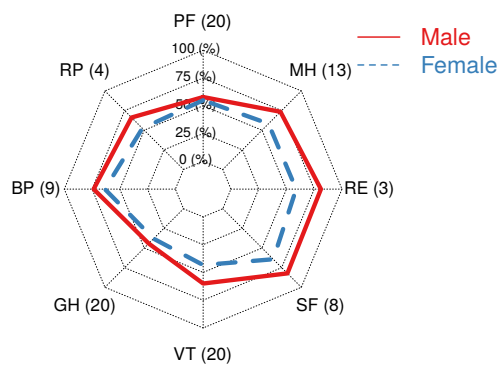


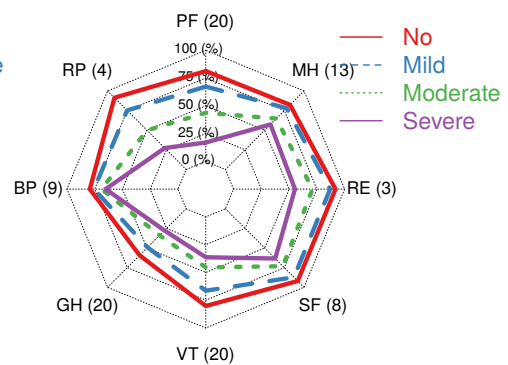
Figure 1.4: Histograms of the recorded SF-36 scores in cross-sectional COPD data. The blue and the green lines correspond to the binomial and binomial dispersion fits respectively. The red line corresponds to the fit by the beta-binomial distribution. Frequencies are shown at the top of each bar

it can be shown that the mean effect of different dyspnea levels is not equal in all the dimensions, as the effect in *physical functioning* is higher than in *bodily pain*,

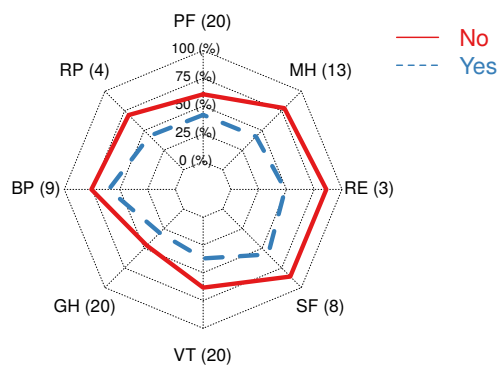
differentiating category effect between physical and mental dimensions. Figure 1.5a also shows that the mean perception of HRQoL is better in males than in females in all the dimension, being the mental dimensions where the difference is higher. On the other hand, Figures 1.5c-1.5d show that, as expected, the anxiety or depression status worsens in average the health-status of COPD patients in all the dimensions, the anxiety especially in *role emotional* and the depression in *vitality*.



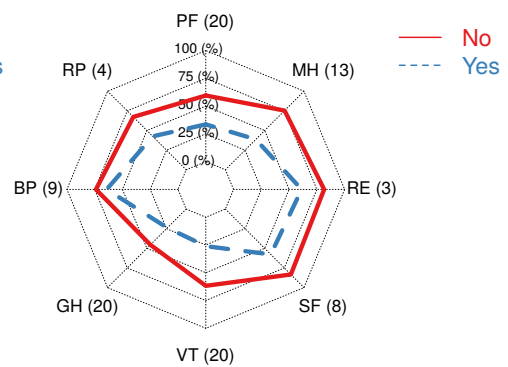
(a) Distribution of the health-status by gender



(b) Distribution of the health-status by dyspnea



(c) Distribution of the health-status by anxiety



(d) Distribution of the health-status by depression

Figure 1.5: Distribution of the health-status or HRQoL in COPD patients at baseline based on some categorical variables. Numbers between parenthesis are the maximum values of the recoded SF-36 scores in each dimension.

## St. George's Respiratory Questionnaire

The SGRQ has been introduced in Section 1.2.2, where it was mentioned that it provides three different domains which cover different aspects of the health-status of the patients with respiratory diseases. In fact, a principal component analysis was performed on the 50 items of the SGRQ, leading to the decomposition of the outcome in the following three dimensions: *Symptoms*, *Activity* and *Impacts*. The resulting three scores were then standardised to the 0 – 100 scale. Therefore, similar to the dimensions provided by the SF-36, a recoding process of the SGRQ domains might be carried out before the beta-binomial distribution is applied.

PROs are used to discriminate differences between patients and evaluate changes within patients. Unlike most health-status questionnaires, there is evidence that the SGRQ behaves similarly when used to make comparisons between patients or detect changes within patients (Jones, 2005). Consequently, the same estimation of the minimum clinically important difference can be used for both applications. Clinical thresholds are used most commonly to judge whether a treatment has a clinically worthwhile effect, or whether it is superior to another treatment. Jones (2002) showed that the estimate of the SGRQ is consistently around four units, regardless of the method of estimation and the number of the subjects contributing to the estimate. Therefore, treatments that produce an improvement of the order of 4 units have found wide acceptance once in use, so it seems reasonable to expect any new treatment proposed for COPD to produce an advantage over placebo that is not significantly inferior to a 4 unit difference (Jones, 2005).

We can develop a recoding process of the SGRQ dimensions based on the idea that a 4 points change in the 0 – 100 scale can be considered as a clinically significant change. In fact, we can assume that a change in the health-status only ‘exists’ if 4 units are exceeded in the score and, hence, divide the 0 – 100 scale into 4 length subintervals recoding the value of each score with the value of the subinterval it belongs. Table 1.5 shows the recoding of the SGRQ scores based on that criteria.

Consequently, after the recoding has been applied, we get three recoded SGRQ scores which are integer and take values from 0 to 24, where a higher point, as in the original scale, means a worse health-status. Figure 1.6 displays the histograms of the distributions of the original scores (on the left-hand side) and recoded scores (on the right-hand side) of the SGRQ three dimensions for the patients at baseline. As it can be appreciated, the recoding process maintains the distributional features of the scores, which show similar shapes on both sides of the figure. Additionally, Figure

Table 1.5: Recoding of the SGRQ scores for the three dimensions.

Interval		Recoded	Interval		Recoded
[0, 4)	→	0	[52, 56)	→	13
[4, 8)	→	1	[56, 60)	→	14
[8, 12)	→	2	[60, 64)	→	15
[12, 16)	→	3	[64, 68)	→	16
[16, 20)	→	4	[68, 72)	→	17
[20, 24)	→	5	[72, 76)	→	18
[24, 28)	→	6	[76, 80)	→	19
[28, 32)	→	7	[80, 84)	→	20
[32, 36)	→	8	[84, 88)	→	21
[36, 40)	→	9	[88, 92)	→	22
[40, 44)	→	10	[92, 96)	→	23
[44, 48)	→	11	[96, 100]	→	24
[48, 52)	→	12			

1.6 displays the distributional fit of the scores (*symptoms*, *activity* and *impacts* domains) by the normal, binomial, binomial with dispersion parameter and beta-binomial distributions.

First of all, it is worth mentioning that the normal and binomial with dispersion parameter distributions reach very similar fit shapes conditional on the different scale and characteristics of the data, which it is assumed continuous in the first one and integer in the second one. However, it is evidenced that the normal distribution is not bounded in the 0 – 100 range, which leaves distribution tails out of the scale, especially in the *impacts* domain. As regards to the fit in the recoded scores, we can state that the three domains behave quite different and, consequently, different results are obtained from the fit of the different distributions in each dimension. First, the *symptoms* shows a slightly bell-shaped (almost flat) distribution, where both the binomial with dispersion parameter and the beta-binomial distributions display similar results. Second, the *activity* only reaches values in some determined scores, which complicates the correct fit by any distribution. However, compared to both binomial distributions, the beta-binomial offers a flatter distribution and, therefore, accumulates less error in the zero or low frequency scores. Finally, the *impacts* scores are accumulated on the left-hand side of the scale, meaning that there is not much impact of the disease in COPD patients. Similar to the recoded *mental health* SF-36 dimension in Figure 1.4, the bell shape of the density function

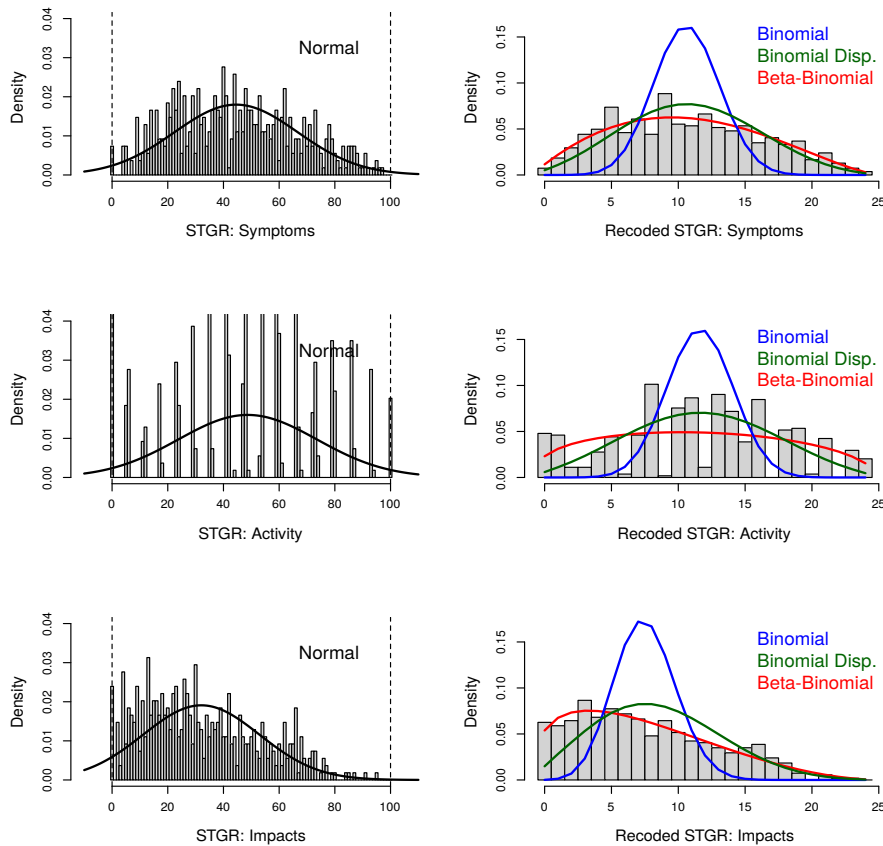


Figure 1.6: Histograms of the SGRQ scores in the COPD Study at baseline. In the left-hand side, the original scores are displayed, while on the right-hand side the recoded scores are shown. The figure offers the fit to the data by the normal (left), binomial, dispersion binomial and beta-binomial (right) distributions.

of the binomial distributions does not allow an appropriate fit to the data, even if a dispersion parameter is included. In this domain, it can be easily appreciated that the beta-binomial distribution offers the most accurate fit. Summarizing, in one recoded dimension the binomial distribution with dispersion parameter and the beta-binomial distribution offer very similar results, however, in the other two the beta-binomial is the most adequate, especially in the *Impacts* dimension. Therefore, we propose the use of the beta-binomial distribution as a unified way of analysing the SGRQ scores in the COPD Study.

### Mini Mental State Examination

In the following lines, we are going to perform the distributional fit to the MMSE scores in the Paquid Research Programme subsample introduced in Section 1.3.2. The MMSE offers tools for measuring the cognitive status of patients being evaluated and, conversely to the SF-36, it only provides a unique score. Moreover, the provided score is an integer in the 0 – 30 scale, and hence, no recoding process is required before the beta-binomial distribution is applied.

The Paquid research programme was carried out in a longitudinal framework, and a maximum of 9 measurements were performed for each individual in the cohort. Therefore, in order to show and compare the adequacy of the beta-binomial distribution in the longitudinal study, we perform different distributional fits on each of the time points. However, for the sake of clarity, as scores behave similarly in each time point, Figure 1.7 only displays the distribution of the first four-time point measurements. In addition, it shows the fit by the beta-binomial and binomial distributions with and without dispersion parameter.

The main conclusion we obtain from Figure 1.7 is that observations tend to accumulate on the right-hand side of the score scale, meaning that patients have a good mental status in general. Consequently, it is easy to appreciate that any of the first four measurements do not follow a Gaussian or bell-shape distribution. In fact, that is one of the reasons for the poor performance of both binomial distributions, as they assume a bell-shaped distribution. On the contrary, the beta-binomial distribution does not display any bell-shaped form, which makes its adjustment more accurate taking into account the characteristics of the data. Indeed, Figure 1.7 shows that, in low frequency scores, the beta-binomial distribution gets much better fit than the binomial distributions. Although, in high frequency scores the fit is not as good as the previous ones.

We have already shown in Section 1.4.2 that the beta-binomial corresponds to the binomial distribution when the dispersion parameter  $\phi$  is equal to zero, meaning that there is no extra variability or overdispersion. However, Figure 1.7 displays quite different shapes for each distribution and, therefore, we can conclude that there exists overdispersion in MMSE scores for the different time points. Therefore, we propose the use of the beta-binomial distribution to fit MMSE scores.

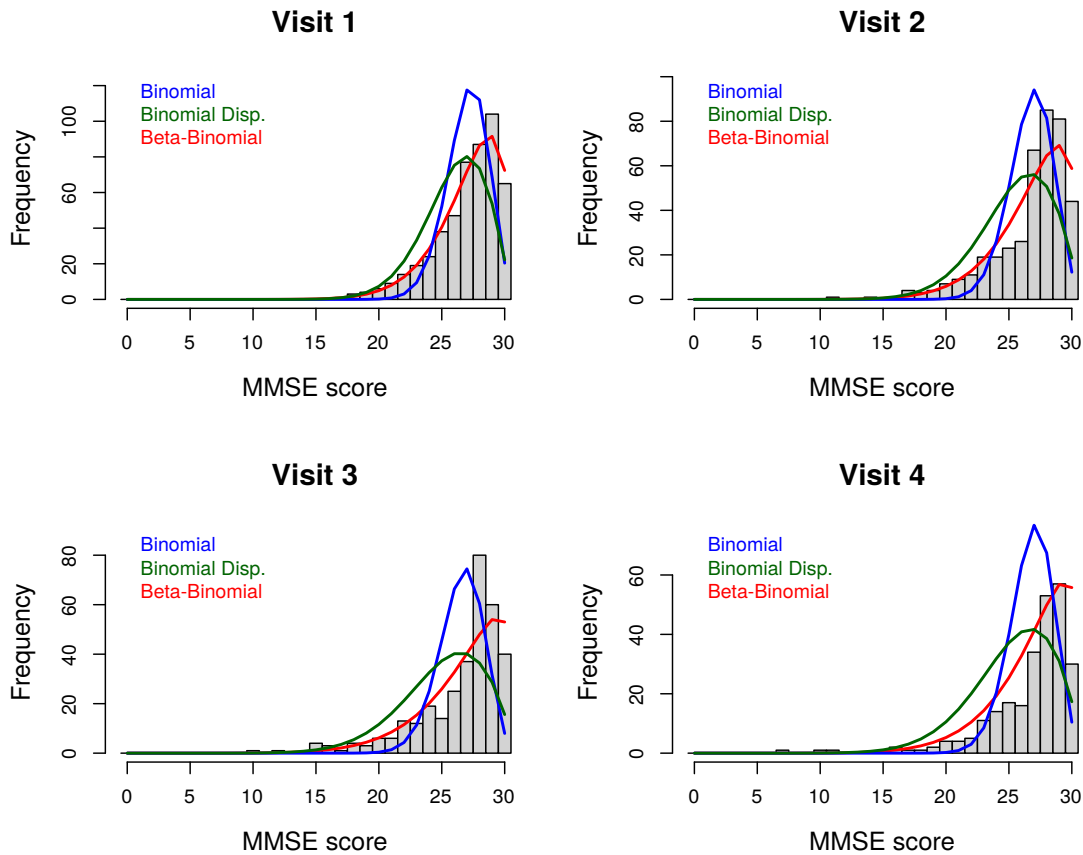


Figure 1.7: Histograms of the MMSE score at the first four visits in Paquid subsample data. The blue and the green lines correspond to the binomial and dispersion binomial models respectively. The red line corresponds to the fit by the beta-binomial distribution.

## 1.5 Objectives of the thesis

This thesis is focused on the development of regression approaches based on the beta-binomial distribution for the appropriate analysis of PROs in different scenarios. In general terms, we can break the objectives of this thesis down in five main goals. First, in order to detect and quantify the effect of the covariates on PROs in cross-sectional studies, we will propose a regression model based on the beta-binomial distribution and we will define a method to make inference in this context. However, PRO studies are usually carried out over time where patients are repeatedly measured leading to non-independent data. Therefore, our second ob-

---

jective is to extend the cross-sectional beta-binomial regression model to multilevel data framework, where not only longitudinal studies, but also any kind of correlated data could be analysed. Due to the fact that that PROs often provide different dimensions referring to several health aspects, the third objective is to develop a regression model based on the beta-binomial distribution for the joint analysis of all the dimensions, i.e. the third objective is to develop a multivariate beta-binomial regression model. The theoretical development of statistical regression models may be useless unless a practical tool is provided. Hence, the fourth goal is the implementation of the regression models as Open Source Software that could be easily used in clinical research. Finally, the fifth and last objective is the application of the proposed methodology to COPD data in order to get clinically valid and relevant results involving the health-status of COPD patients.





---

---

## CHAPTER 2

---


# CROSS-SECTIONAL ANALYSIS: A BETA-BINOMIAL REGRESSION APPROACH


*“The limits of my language means the limits of my world”*

---

Ludwig Wittgenstein, 1889 – 1951

*The work developed in this chapter has already been accepted in the Statistical Methods in Medical Research journal and partially presented in the Conference - 31st International Workshop on Statistical Modelling Conference.*

 *Najera-Zuloaga, J., Lee, D.-J. and Arostegui, I. (2017). Comparison of beta-binomial regression model approaches to analyze health-related quality of life data. Statistical Methods in Medical Research (in press)*

 *31st International Workshop on Statistical Modelling. Comparison of beta-binomial regression approaches to analyze health-related quality of life data. Najera-Zuloaga, J., Lee, D.-J. and Arostegui, I. Proceedings volume II. Rennes, July 4 - 8, 2016.*

### 2.1 Introduction

One of the most important goals of PRO studies is the measurement of the effect that some specific observed variables, such as life habits, disease's characteristics or socio-demographical properties, have on the life status of the patients. A variety

of statistical methods are available to deal with the mentioned objective, however, due to PROs features referred to in Section 1.4, not all the methodologies would be appropriate.

Typically, regression models have been developed in the literature to estimate the relationship between an outcome variable and some given exploratory variables. In the classical linear models (LMs) the random vector  $\mathbf{Y} = (Y_1, \dots, Y_n)$ , referred to as the outcome, is connected to the given covariates by means of a linear regression

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where  $\mathbf{X}$  is a full rank matrix which each column corresponds to the measurements of each exploratory variable,  $\boldsymbol{\beta}$  are the regression coefficients and  $\boldsymbol{\epsilon}$  is the error vector. Usually it is assumed that the error term is a vector of iid variables normally distributed with zero mean and the same variance. This model specification sets up some conditions on the outcome variable  $\mathbf{Y}$ , such as continuity, which is not met in all the real examples. For instance, one may be interested in the relationship between some covariates and dichotomous or count outcomes, however, the distributional assumption of LMs may fail in this situations. The extension of classical LMs to cover non-normal responses is one of the largest success of the statistical inference. McCullagh and Nelder (1989) developed the well-known generalised linear models (GLMs) which are based on exponential family distributions, covering a huge range of real situations.

### 2.1.1 Generalised linear models

Due to the fact that GLMs are a well-known regression methodology (McCullagh and Nelder, 1989), in this section, we will shortly introduce and describe them. Unlike the classical LMs, GLMs do not assume normality of the outcome, and instead, restrict their assumption to the exponential family distributions. The exponential family, which has been introduced in Section 1.4.1, covers a wide variety of distributions, and hence, it makes GLMs useful in a huge range of studies.

In general terms, GLM methodology applies a monotonic function defined in the real line, which is commonly called the *link function*, to the expectation of the outcomes, and then, connects it with the given covariates by a linear predictor. For instance, assume that  $\mathbf{y} = (y_1, \dots, y_n)'$  are the observed  $n$  independent outcomes

which follow an exponential family distribution of the following form

$$f(y_i|\theta_i, \phi) = \exp \left\{ \frac{y_i\theta_i - A(\theta_i)}{\phi} + c(y_i, \phi) \right\},$$

where  $\theta_i$  is the canonical parameter of the distribution  $i = 1, \dots, n$  and  $A(\cdot)$  and  $c(\cdot)$  are known functions. As it was mentioned, GLMs connect the expectation of the outcomes with the given covariates by means of a link function  $h(\cdot)$ ,

$$h(\mu_i) = \mathbf{x}'_i\boldsymbol{\beta}$$

where  $\mu_i = \mathbb{E}[Y_i]$ ,  $\mathbf{x}_i$  is the  $i$ th row of a full rank matrix  $\mathbf{X}$  composed by the covariates and  $\boldsymbol{\beta}$  are the regression parameters. From exponential family theory (see Section 1.4.1), we know that  $\mu_i = A'(\theta_i)$ , and therefore, there exists an implied relationship,

$$A'(\theta_i) = h^{-1}(\mathbf{x}'_i\boldsymbol{\beta}),$$

between the canonical parameter and the regression parameters. The choice of  $h(\cdot)$  such that  $h(\mu_i) = \theta_i$ , or equivalently  $\theta_i = \mathbf{x}_i\boldsymbol{\beta}$ , is called *canonical link* function. The use of the canonical link facilitates the estimation and inference procedures, and consequently, it is usually selected in most of the applications, although any monotonic function could be used. Notice that the canonical link function depends on the canonical parameter and, hence, each exponential family distribution has a different canonical link function.

Therefore, GLM approach only needs two specifications for defining a model, the distribution from the exponential family and the link function. However, it was mentioned in Section 1.4.1 that, by fixing the first two moments the distribution family is completely defined due to the relationships presented in Equation (1.2). Consequently, a GLM could be specified by the definition of the link function and the first two moments of the distribution of the outcome.

Following exponential family distribution theory (see Section 1.4.1) the log-likelihood of a GLM is defined as

$$\log L(\boldsymbol{\beta}, \phi|\mathbf{y}) = \sum_{i=1}^n f(y_i|\boldsymbol{\beta}, \phi) = \sum_{i=1}^n \frac{y_i\theta_i - A(\theta_i)}{\phi} + c(y_i, \phi), \quad (2.1)$$

where  $A'(\theta_i) = h^{-1}(\mathbf{x}'_i\boldsymbol{\beta})$  or  $\theta_i = \mathbf{x}'_i\boldsymbol{\beta}$  if the canonical link function is used and the dispersion parameter is assumed constant for all the outcomes.

Regarding the estimation of the regression parameters  $\boldsymbol{\beta}$ , the Iterative Weighted Least Squares (IWLS) is one of the most widely used approaches (McCullagh and Nelder, 1989). It extends the Newton-Raphson algorithm to the GLM case, where an iterative estimating formula is obtained. More details about the construction of the IWLS procedure are provided in Appendix B.2.

However, in some distribution, such as the binomial, if the dispersion model defined in Equation (1.3) is considered, then we do not have an explicit formulation of the density function because  $c(\cdot)$  is unknown. In this situations, as well as in the distributional inference in Section 1.4.1, the estimation of the regression parameters  $\boldsymbol{\beta}$  remains equal as  $c(\cdot)$  do not depend on  $\boldsymbol{\beta}$ , however, it does depend on  $\phi$ . Therefore, we are not allowed to perform a likelihood-based estimation of  $\phi$ , neither a full inference of  $\boldsymbol{\beta}$  as the second derivatives of the likelihood depend on  $\phi$ .

In Chapter 1, Section 1.4.1, we have introduced the so-called quasi-likelihood approximation as a full inference procedure of the parameters in an exponential family dispersion model. The quasi-likelihood methodology is based on the normalisation of the density function of the outcomes through a quadratic approximation of the likelihood, which leads to a deviance approximation. Based on the quasi-likelihood theory, the MLE of  $\boldsymbol{\beta}$  is the value that minimises the total deviance of the model defined in Equation (1.9) (Wedderburn, 1974). Additionally, it offers an explicit formulation of the approximation of the log-likelihood of the model as

$$\log L(\boldsymbol{\beta}, \phi | \mathbf{y}) \approx \sum_{i=1}^n -\frac{1}{2} \log(2\pi\phi v(y_i)) - \frac{1}{2\phi} D(y_i, \theta_i), \quad (2.2)$$

where  $A'(\theta_i) = h^{-1}(\mathbf{x}'_i \boldsymbol{\beta})$ ,  $h(\cdot)$  is the link function,  $\mathbf{x}_i$  is the  $i$ th row of a full rank matrix  $\mathbf{X}$  composed by the covariates,  $v(\cdot)$  is the variance function of the specific exponential family member and  $D(\cdot)$  is the total deviance of the model defined in Equation (1.8).

The quasi-likelihood function defined in Equation (2.2) allows a full inference process in both regression and dispersion parameters of the model.

### 2.1.2 Beta-binomial regression background

The classical GLMs are a very general regression methodology where a wide variety of outcomes and covariates can be analysed. In fact, they do not restrict their distributional assumption to a unique distribution as LMs, and moreover, the possible use of different link functions enriches the models. However, although the exponen-

tial family includes most of the distributions that do occur in real studies, there are some exceptions. Indeed, as it was explained in Chapter 1, the observations' within-correlation and overdispersion that most PROs present make the fit by exponential family distributions inadequate.

The beta-binomial distribution was proposed in 2007 to fit the SF-36 Health Survey scores (Arostegui et al., 2007). The proposal was mainly motivated because of the ordinal feature that many of the SF-36 scores exhibit. Furthermore, they showed that the beta-binomial regression is a good option to detect significant predictors of SF-36 scores and they provided a nice interpretation of the effect of explanatory variables on HRQoL when SF-36 is used. The authors also compared results using multiple linear regression (MLR) and beta-binomial regression for real and simulated data, showing that performance of the beta-binomial approach was better or similar than the MLR approach in all the dimensions of the SF-36. Comparison of MLR and beta-binomial regression approaches was performed based on distributional assumptions.

In 2012, Arostegui et al. (2012) generalised the SF-36 results and presented eight methods of analysis of PROs under different assumptions that lead to different interpretation of the results. The methods were: MLR with least square and bootstrap estimations, tobit regression, ordinal logistic and probit regressions, beta-binomial regression, binomial-logistic-normal regression and coarsening. All methods were applied to scores obtained from two of the health dimensions of the SF-36 Health Survey. They showed that the beta-binomial regression approach renders satisfactory results in a broad number of situations, with a very convenient clinical interpretation of the results.

Therefore, the beta-binomial distribution has been proposed in the literature not only for fitting the distribution of PROs but also as the given distribution of the outcomes for performing regression models. However, there are two different ways of implementing a regression model based on the beta-binomial distribution in the literature. On the one hand, the marginal beta-binomial regression can be implemented, which applies a logistic regression in the probability parameter of a beta-binomial distribution (Forcina and Franconi, 1988). In this setting, estimation is done via maximum likelihood where the delta algorithm (Jørgensen, 1984) is applied. On the other hand, the binomial-beta model developed by Lee and Nelder (1996) can be used as a particular case of the Hierarchical Generalised Linear Models (HGLMs). In the conditional or HGLM approach beta distributed random effects are included in the linear predictor of a logistic model to accommodate the overdispersion and

correlation of PROs. Keeping in mind that there could exist some differences in the results due to average-specific (marginal) and subject-specific (conditional) approaches respectively, it would be reasonable if both methodologies may result in similar conclusions. Nevertheless, none of the existing literature in the analysis of PRO data has performed a comparison of both approaches in terms of adequacy and regression parameter interpretation context. Therefore, in order to clarify differences between both approaches and find out the optimal methodology in terms of the estimation of covariate effect, in this chapter we carry out a deep comparison study between both models in both real and simulated data.

The rest of the chapter is organised as follows. Section 2.2 presents a description of both methodological approaches to perform a regression model based on the beta-binomial distribution. The application to COPD Study is carried out in Section 2.3 where the defined models result on different parameter and standard deviation estimates, that lead to different conclusions regarding the effect of the covariates in the HRQoL of patients with COPD. Consequently, Section 2.4 is focused on a simulation study that provides comparisons of the approaches in controlled scenarios. Finally, in Section 2.5, we provide a brief discussion of the obtained results, as well as some general conclusions and recommendations.

## 2.2 Beta-binomial regression approaches

It was described in Section 1.4.2 that the beta-binomial consists of a mixture between a beta and a binomial distribution. Therefore, as it was explained, there exists two ways of defining the distribution model: (i) the marginal; and (ii) the conditional. In the marginal model we integrate out the beta distributed random variable, and get a closed-form equation for the density function. In the conditional approach we do not perform any integration and, instead, the model is left in a conditional form. Each of the model definitions leads to a different regression approach. This section is focused on the description of both methodologies where, apart from the definition, the estimation procedure is broadly discussed.

### 2.2.1 Marginal approach

The first approach, which we denote as *BBreg*, is based on a marginal regression model approach. *BBreg* assumes that the observed outcomes are drawn from a beta-binomial distribution and applies a logistic regression model in the marginal

expectation. Therefore, it does not consider the beta-binomial as a mixture, and instead, it is focused on the marginal distribution where beta effects are integrated out in the density function. BBreg model definition is similar to GLMs as a link function is applied to the expectation of the outcomes before linking it with a linear predictor consisting of the covariates. However, there is a crucial difference: the beta-binomial distribution does not belong to the exponential family. Consequently, GLM inference procedure cannot be directly applied as it is developed based on some properties that exclusively exponential family distributions satisfy.

Let  $\mathbf{Y} = (Y_1, \dots, Y_n)'$  be a set of iid random variables where each  $Y_i$  follows a beta-binomial distribution with parameters  $p_i$  and  $\phi$  and a maximum score number equal to  $m_i$ ,  $i = 1, \dots, n$ . Namely, we have that

$$Y_i \sim \text{BB}(m_i, p_i, \phi), \quad i = 1, \dots, n,$$

where

$$\mathbb{E}[Y_i] = m_i p_i \quad \text{and} \quad \text{Var}[Y_i] = m_i p_i (1 - p_i) \left[ 1 + (m_i - 1) \frac{\phi}{1 + \phi} \right]. \quad (2.3)$$

Notice that we assume that the dispersion parameter  $\phi$  is equal for all the random variables associated with each observation. In fact, in PRO context,  $\phi$  measures the dimension within correlation for the binary responses of the same individual. Therefore, it is logical to think that the correlation depends on the specific dimension and hence, that remains equal for all the patients.

Based on the binomial distribution, Equation (2.3) offers a straightforward interpretation of the beta-binomial distribution parameters. In fact, this can be very useful in the medical framework as it allows the interpretation of the regression parameters' effect based on the meaning of the modelled distribution parameter. On the one hand, as it can be appreciated, the mean corresponds to the binomial mean, and hence, similar to the binomial distribution,  $p_i$  can be interpreted as the probability parameter of obtaining one success or one point in PRO framework. On the other hand, the second order moment adds a multiplicative term to the binomial variance allowing for overdispersion. Indeed,  $\phi$  was defined as the correlation parameter between the binary outcomes in Section 1.4.2 and therefore, we can conclude that the overdispersion of the data is generated by the intraclass correlation of the summed up Bernoulli (binary) outcomes. Notice that when  $\phi = 0$ , the model corresponds to the binomial case. On the whole, the beta-binomial distribution can be inter-



preted as a binomial distribution where intraclass correlation and overdispersion are measured.

Due to the easy and useful interpretation of the distribution parameters, and in order to simplify the notation, from now on we will consider  $p_i$  the probability parameter of the beta-binomial distribution,  $\phi$  as the dispersion parameter and  $m_i$  as the maximum score number. The main objective of PRO studies is the measurement of the effect that some covariates may have on the expected PRO being analysed. In other words, studies are interested in modelling the mean as a function of the covariates. Therefore, we will only connect or relate the probability parameter as a function of the covariates, letting  $\phi$  as a constant.

Assume that we have  $\mathbf{y} = (y_1, \dots, y_n)'$  a set of observations. Following Forcina and Franconi (1988), we can connect the probability parameter of the beta-binomial distribution with a linear predictor consisting of some given covariates  $X_1, \dots, X_p$  by means of a logit link function as

$$\log \frac{p_i}{1 - p_i} = \eta_i = \mathbf{x}'_i \boldsymbol{\beta},$$

where  $\eta_i$  is the linear predictor that corresponds to the  $i$ th observation,  $\boldsymbol{\beta}$  is the  $(p + 1) \times 1$  vector of regression parameters and  $\mathbf{x}_i$  the  $i$ th row of a full rank design matrix  $\mathbf{X}$  composed by the given covariates,  $i = 1, \dots, n$ .

The log-likelihood of the described model is defined as

$$\begin{aligned} \log L(\boldsymbol{\beta}, \phi | \mathbf{y}) &= \sum_{i=1}^n \log f(y_i | \boldsymbol{\beta}, \phi) \\ &= \sum_{i=1}^n \log \left[ \binom{m_i}{y_i} \frac{\Gamma\left(\frac{1}{\phi}\right)}{\Gamma\left(\frac{1}{\phi} + m_i\right)} \frac{\Gamma\left(\frac{p_i}{\phi} + y_i\right)}{\Gamma\left(\frac{p_i}{\phi}\right)} \frac{\Gamma\left(\frac{1 - p_i}{\phi} + m_i - y_i\right)}{\Gamma\left(\frac{1 - p_i}{\phi}\right)} \right], \end{aligned}$$

where  $\boldsymbol{\beta}$  is included in the equation through the relationship given by

$$p_i = \frac{e^{\mathbf{x}'_i \boldsymbol{\beta}}}{1 + e^{\mathbf{x}'_i \boldsymbol{\beta}}}, \quad (2.4)$$

and  $\Gamma(\cdot)$  is the gamma function defined in Equation (1.19).

As done in Equation (1.20), if we use that  $\Gamma(x + 1) = x\Gamma(x)$  for all  $x \in \mathbb{R}$  and  $\Gamma(0) = 1$ , we obtain a new expression of the beta-binomial density function, and hence, of the likelihood, which is easier to manipulate. Therefore, the log-likelihood

of the marginal beta-binomial regression model can be simplified as

$$\log L(\boldsymbol{\beta}, \phi | \mathbf{y}) = \sum_{i=1}^n \left[ \log \binom{m_i}{y_i} + \sum_{k=0}^{y_i-1} \log(p_i + k\phi) + \sum_{k=0}^{m_i-y_i-1} \log(1 - p_i + k\phi) - \sum_{k=0}^{m_i} \log(1 + k\phi) \right], \quad (2.5)$$

where if  $y_i = 0$  then,  $\sum_{k=0}^{y_i-1} \log(p_i + k\phi) = 0$ ; and if  $y_i = m_i$  then,  $\sum_{k=0}^{m_i-y_i-1} \log(1 - p_i + k\phi) = 0$ .

In order to get the MLE of  $\boldsymbol{\beta}$  we perform the first order derivative of the log-likelihood, which leads to the next score equation

$$S(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} \log L(\boldsymbol{\beta}, \phi | \mathbf{y}) = \left[ \frac{\partial}{\partial \mathbf{p}} \log L(\boldsymbol{\beta}, \phi | \mathbf{y}) \right] \frac{\partial \mathbf{p}}{\partial \boldsymbol{\beta}} = \boldsymbol{\xi}' \mathbf{S} \mathbf{X} \quad (2.6)$$

where  $\mathbf{S} = \text{diag}(p_1(1 - p_1), \dots, p_n(1 - p_n))$  and  $\boldsymbol{\xi} = (\xi_1, \dots, \xi_n)'$  being

$$\xi_i = \sum_{k=0}^{y_i-1} \frac{1}{p_i + k\phi} - \sum_{k=0}^{m_i-y_i-1} \frac{1}{1 - p_i + k\phi}, \quad (2.7)$$

(see Appendix D.2, Equation (D.1) and Equation (D.2) for further details).

Given that Equation (2.6) has no explicit solution, iterative algorithms are used to solve Equation (2.6). The Newton-Raphson algorithm is one of the most used procedures for solving score equations. It consists of a quadratic approximation of the function to be solved which leads to an iterative estimation of the root based on the approximation through the ratio between the value of the function and the value of the derivative (see Appendix B.1). When dealing with estimation in regression models, we try to maximise the likelihood, or equivalently, solve the derivative of the log-likelihood function. Therefore, approximating ratio consists of the division between the first and second order derivatives of the log-likelihood function which may have a tendency to be unstable for many reasons. One reason is that the negative of the second derivative of the log-likelihood, or equivalently, the observed Fisher information  $I(\boldsymbol{\beta})$ , may be negative unless  $\boldsymbol{\beta}$  is already very close to the MLE  $\hat{\boldsymbol{\beta}}$ . In fact,  $I(\hat{\boldsymbol{\beta}})$  determines the sharpness of the peak in the likelihood function around its maximum, and consequently, it must be positive-defined. However, occasionally the  $I(\boldsymbol{\beta})$  term is also used where  $\boldsymbol{\beta}$  is arbitrary, but this can be inadequate as it may not be positive-defined. Moreover, apart from negativity, there could be additional

problems as  $I(\boldsymbol{\beta})$  can be singular and not invertible or it can have both negative and positive eigenvalues. In order to correct the estimation procedure, the Fisher's scoring algorithm is used. This method is defined as a Newton-Raphson algorithm where the matrix of second derivatives is replaced by its expectation,  $\mathcal{I}(\boldsymbol{\beta}) = \mathbb{E}[I(\boldsymbol{\beta})]$ , i.e., the Fisher expected information matrix. Indeed,  $\mathcal{I}(\boldsymbol{\beta})$  is always non-negative, and even strictly positive in regular cases.

In GLMs, due to exponential family properties, if a canonical link function is used, the Fisher observed information matrix is equivalent to the expected information matrix as  $\partial^2 \theta_i / \partial \boldsymbol{\beta} \partial \boldsymbol{\beta}' = 0$  (see Appendix B.2). Therefore, in these model approaches, the Newton-Raphson algorithm is accurate and stable. However, we have mentioned before that the beta-binomial distribution does not belong to the exponential family, and therefore, Newton-Raphson or GLM estimation procedure cannot be applied directly. Moreover, due to the complexity of the beta-binomial distribution,  $I(\boldsymbol{\beta})$  is very hard to calculate and approximation procedures must be used. Therefore, in this work, we have developed a modification of the Newton-Raphson estimating procedure based on the delta algorithm (Jørgensen, 1984).

Basically, the delta algorithm generalises Fisher's scoring method, and it is derived from a modification of the Newton-Raphson algorithm where the matrix of the second derivatives is replaced by an approximation of the form

$$-\frac{\partial^2 L}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} \approx T(\boldsymbol{\beta})' K(\boldsymbol{\mu}) T(\boldsymbol{\beta}), \quad (2.8)$$

where  $K(\boldsymbol{\mu})$  is a suitable  $n \times n$  symmetric positive-defined matrix called the weight matrix,  $T(\boldsymbol{\beta})$  is the  $n \times p$  matrix  $\partial \boldsymbol{\mu} / \partial \boldsymbol{\beta}$  and  $\boldsymbol{\mu}$  is an intermediate parameter vector of length  $n \times 1$ , being  $n$  the number of observations.

There are several options for the election of the weight matrix in the delta algorithm. In fact, we have that the negative second derivative of the log-likelihood is defined as

$$-\frac{\partial^2 L}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = T(\boldsymbol{\beta})' I(\boldsymbol{\mu}) T(\boldsymbol{\beta}) - \sum_{i=1}^n \frac{\partial^2 \mu_i}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} u(\mu_i), \quad (2.9)$$

where  $u(\boldsymbol{\mu}) = \partial L / \partial \boldsymbol{\mu}$ . Hence, we choose  $K(\boldsymbol{\mu})$  close to  $I(\boldsymbol{\mu})$ . But as the second term in Equation (2.9) is ignored in the approximation, this term must be small in absolute value if the delta algorithm is to have a reasonable rate of convergence. Jørgensen (1984) showed that the delta algorithm is expected to work well only for models which are nearly linear or fit the data well (Jørgensen, 1984).

For additive log-likelihood models,  $I(\boldsymbol{\mu})$  and  $\mathcal{I}(\boldsymbol{\mu})$  are diagonal, so we take  $K(\boldsymbol{\mu})$

diagonal and denote by  $K(\mu_i)$  the  $i$ th diagonal element. The election of  $K(\mu_i)$  equal to the expected information weight

$$\mathcal{I}(\mu_i) = \mathbb{E} \left[ -\frac{\partial^2 \log L}{\partial \mu_i^2} \right]$$

simplifies the delta algorithm to the Fisher's scoring method. Moreover, whereas other choices of  $K(\mu_i)$  depend on the particular parametrisation of  $\boldsymbol{\mu}$ , the expected information weight is invariant to reparametrisation of  $\boldsymbol{\mu}$  (Jørgensen, 1984). Additionally, as we have mentioned before, the expected weights are always positive for any regular model.

In the BBreg approach the intermediate parameter  $\boldsymbol{\mu}$  corresponds to the probability parameter  $\boldsymbol{p}$  and therefore, the second derivative of the log-likelihood with respect to the regression parameters can be approximated using expected information weights, as

$$\frac{\partial^2 \log L}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} \approx \mathbb{E} \left[ -\frac{\partial \boldsymbol{p}'}{\partial \boldsymbol{\beta}} \frac{\partial \log L}{\partial \boldsymbol{p} \boldsymbol{p}'} \frac{\partial \boldsymbol{p}}{\partial \boldsymbol{\beta}} \right] = \boldsymbol{X}' \boldsymbol{S} \mathbb{E}[\boldsymbol{V}] \boldsymbol{S} \boldsymbol{X} \quad (2.10)$$

where  $\boldsymbol{S}$  and  $\boldsymbol{\xi} = (\xi_1, \dots, \xi_n)'$  have been previously defined in Equation (2.7) and  $\boldsymbol{V} = \text{diag}(v_1, \dots, v_n)$  being

$$v_i = -\frac{\partial \xi_i}{\partial p_i} = \sum_{k=0}^{y_i-1} \frac{1}{(p_i + k\phi)^2} + \sum_{k=0}^{m_i-y_i-1} \frac{1}{(1-p_i + k\phi)^2}, \quad (2.11)$$

(see Appendix D.2, Equation (D.1) and Equation (D.3) for further details).

Consequently, the estimation equation for the MLE of  $\boldsymbol{\beta}$  is defined as

$$\begin{aligned} \hat{\boldsymbol{\beta}}^{(r+1)} &= \hat{\boldsymbol{\beta}}^{(r)} - \left\{ -\frac{\partial \boldsymbol{p}'}{\partial \boldsymbol{\beta}} \mathbb{E} \left[ \frac{\partial^2}{\partial \boldsymbol{p} \partial \boldsymbol{p}'} \log L(\boldsymbol{\beta}, \phi | \boldsymbol{y}) \right] \frac{\partial \boldsymbol{p}}{\partial \boldsymbol{\beta}} \right\}^{-1} \left[ \frac{\partial}{\partial \boldsymbol{\beta}} \log L(\boldsymbol{\beta}, \phi | \boldsymbol{y}) \right]' \\ &= \hat{\boldsymbol{\beta}}^{(r)} + \{(\boldsymbol{S} \boldsymbol{X})' \mathbb{E}[\boldsymbol{V}] (\boldsymbol{S} \boldsymbol{X})\}^{-1} \boldsymbol{X}' \boldsymbol{S} \boldsymbol{\xi} \\ &= \hat{\boldsymbol{\beta}}^{(r)} + \{\boldsymbol{X}' \boldsymbol{S} \mathbb{E}[\boldsymbol{V}] \boldsymbol{S} \boldsymbol{X}\}^{-1} \boldsymbol{X}' \boldsymbol{S} \boldsymbol{\xi}, \end{aligned}$$

and after some simple calculations it leads to

$$\hat{\boldsymbol{\beta}}^{(r+1)} = (\boldsymbol{X}' \boldsymbol{S} \mathbb{E}[\boldsymbol{V}] \boldsymbol{S} \boldsymbol{X})^{-1} \boldsymbol{X}' \boldsymbol{S} \mathbb{E}[\boldsymbol{V}] \boldsymbol{S} \boldsymbol{\nu},$$

where  $\boldsymbol{\nu} = \boldsymbol{X} \boldsymbol{\beta}^{(r)} + (\boldsymbol{S} \mathbb{E}[\boldsymbol{V}])^{-1} \boldsymbol{\xi}$ , and the previous matrices are evaluated at  $\hat{\boldsymbol{\beta}}^{(r)}$ .

However, due to the complexity of the beta-binomial density function and, espe-

cially, the  $v_i$  terms  $i = 1, \dots, n$ , the computation of  $\mathbb{E}[\mathbf{V}]$  is intractable. Therefore, it is necessary to replace  $\mathbb{E}[\mathbf{V}]$  with  $\mathbf{V}$  and use the observed weights instead. This adjustment will usually increase the rate of convergence, though the algorithm may be less stable if the starting points are far from the maximum (Jørgensen, 1984). However, Forcina and Franconi (1988) tested the algorithm and convergence was always obtained in two to four iterations.

Therefore, we get the MLE of  $\boldsymbol{\beta}$  by the iterative use of the following equation

$$\hat{\boldsymbol{\beta}}^{(r+1)} = (\mathbf{X}'\mathbf{S}\mathbf{V}\mathbf{S}\mathbf{X})^{-1}\mathbf{X}'\mathbf{S}\mathbf{V}\mathbf{S}\boldsymbol{\nu}, \quad (2.12)$$

where  $\boldsymbol{\nu} = \mathbf{X}\boldsymbol{\beta}^{(r)} + (\mathbf{S}\mathbf{V})^{-1}\boldsymbol{\xi}$ , and the previous matrices are evaluated at  $\hat{\boldsymbol{\beta}}^{(r)}$ .

The estimates of  $\boldsymbol{\beta}$  are functions of  $\phi$ . Hence, if we replace  $\boldsymbol{\beta}$  with  $\hat{\boldsymbol{\beta}}$  or, equivalently,  $\mathbf{p}$  with  $\hat{\mathbf{p}}$ , in the log-likelihood function, we obtain the profile log-likelihood with respect to  $\phi$  (see Appendix D.2, Equation (D.4)). This function is easy to compute, and simple numerical methods are available to locate the maximum  $\hat{\phi}$  (Forcina and Franconi, 1988).

### 2.2.2 Conditional approach

The second approach considers the beta-binomial distribution as a mixture and, unlike the BBreg approach, it models the probability parameter of the conditional binomial distribution. In fact, it performs a logistic regression in the probability parameter of the conditional binomial distribution but includes specific beta distributed random effects in the linear predictor to accommodate both the overdispersion and observations within correlation. Therefore, compared to the previous marginal approach, the conditional model is constructed in a mixed-effects framework.

Generalised linear mixed models (GLMMs) are a very widely used methodology in different frameworks, as they allow the inclusion, in addition to the usual fixed effects, one or more random effects in the linear predictor of GLMs (McCulloch and Searle, 2001). However, the distribution of the random effects is better decided by the properties of the data or the purposes of inference. Therefore, although the normal distribution is convenient for specifying correlations in the data, the use of other distributions for the random effects greatly enriches these class of models. Lee and Nelder (1996) extended GLMMs to hierarchical GLMs (HGLMs), in which the distribution of random components is extended to conjugates of arbitrary distributions from the exponential family.

On the one hand, the HGLMs are defined as, the response vector  $\mathbf{Y}$ , conditioned on some given random components  $\mathbf{u}$ , follows an exponential family distribution satisfying that

$$\mathbb{E}[\mathbf{Y}|\mathbf{u}] = \boldsymbol{\mu} \quad \text{and} \quad \mathbb{V}\text{ar}[\mathbf{Y}|\mathbf{u}] = \lambda V(\boldsymbol{\mu}),$$

and the linear predictor takes the form

$$\boldsymbol{\eta} = g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v},$$

where  $\mathbf{v} = l(\mathbf{u})$ , the random effects, are the transformation of the random components  $\mathbf{u}$  through the scale, and  $\boldsymbol{\beta}$  are the fixed effects. On the other hand, the random components  $\mathbf{u}$  follow a distribution conjugate to an exponential family of distributions with parameter  $\varphi$  where

$$\mathbb{E}[u_i] = \psi_c \quad \text{and} \quad \mathbb{V}\text{ar}[u_i] = \varphi V_c(\psi_c),$$

for  $i = 1, \dots, q$ , and  $q$  is the number of random effects.

In Bayesian probability theory, a prior distribution is considered conjugate to another distribution (likelihood) if the posterior distribution through the Bayes theorem belongs to the same family of distributions. For instance, we have shown that the beta-binomial distribution consists of a conditional response following a binomial distribution whose probability parameter is assumed random and drawn from a beta distribution. Indeed, we assume that the observation  $y$  is drawn from a beta-binomial distribution, if conditional on the random parameter  $u$ ,  $u \sim \text{Beta}(\alpha_1, \alpha_2)$ , it follows a binomial distribution with probability parameter  $u$ . The posterior distribution of the model is defined as

$$\begin{aligned} f(u|\alpha_1, \alpha_2, y) &= \frac{f(y|u)f(u|\alpha_1, \alpha_2)}{f(y|\alpha_1, \alpha_2)} \\ &= \frac{\binom{m}{y} u^y (1-u)^{m-y} u^{\alpha_1-1} (1-u)^{\alpha_2-1} \text{B}(\alpha_1, \alpha_2)^{-1}}{\int_0^1 \binom{m}{y} u^y (1-u)^{m-y} u^{\alpha_1-1} (1-u)^{\alpha_2-1} \text{B}(\alpha_1, \alpha_2)^{-1} du} \\ &= \frac{u^{y+\alpha_1-1} (1-u)^{m-y+\alpha_2-1}}{\int_0^1 u^{y+\alpha_1-1} (1-u)^{m-y+\alpha_2-1} du} \\ &= \frac{u^{y+\alpha_1-1} (1-u)^{m-y+\alpha_2-1}}{\text{B}(y+\alpha_1, m-y+\alpha_2)} \\ &\sim \text{Beta}(y+\alpha_1, m-y+\alpha_2), \end{aligned}$$

which follows a beta distribution of parameters  $y + \alpha_1$  and  $m - y + \alpha_2$ , being  $\text{B}(\cdot)$

the beta function defined in Equation (1.13). Therefore, we have shown that the binomial and beta are conjugate distributions. Hence, we can consider the beta-binomial model as a special case of HGLMs, and define it as

$$Y_i|u_i \sim \text{Bin}(m_i, p_i) \quad \text{and} \quad u_i \sim \text{Beta}(\alpha_1, \alpha_2),$$

where  $p_i$  is connected to  $u_i$  by a linear predictor  $i = 1, \dots, n$ , being  $n$  the number of observations.

In fact, when we construct a HGLM we must choose  $l(\cdot)$  the scale on which the random effects occur linearly in the linear predictor, that is called the *weak canonical scale* (Lee et al., 2006). This weak canonical scale allows the model to maintain invariance of inference with respect to equivalent modelling approaches. The linear predictor of the binomial-beta regression model is defined as

$$\eta_i = \text{logit}(p_i) = \mathbf{x}'_i \boldsymbol{\beta} + v_i, \quad (2.13)$$

where  $\mathbf{x}'_i$  is the  $i$ th row of a full rank design matrix  $\mathbf{X}$  composed by the given covariates,  $\boldsymbol{\beta}$  are the fixed effects and  $v_i = l(u_i) = \text{logit}(u_i)$  is the random effect attributed to observation  $i$ ,  $i = 1, \dots, n$ . Therefore, in the binomial-beta HGLM, the scale of the random effects corresponds to the logit transformation. Note that the model in Equation (2.13) is the binomial-beta HGLM, which we denote as *BBhglm*.

Constraints must be specified in either the random or fixed part of the model to maintain the structure of the model. Lee and Nelder (2001) proposed to impose constraints in the random part of the model, fixing  $\psi_c$ , the expectation of the random components  $u_i$   $i = 1, \dots, n$ , equal to the value that the scale transforms to zero. Namely, in *BBhglm* they imposed,

$$\mathbb{E}[u_i] = \psi_c = 1/2. \quad (2.14)$$

The restriction in Equation (2.14) leads to a strict relationship between the parameters of the distribution of the random effects,

$$\mathbb{E}[u_i] = \frac{\alpha_1}{\alpha_1 + \alpha_2} = \frac{1}{2} \implies \alpha_1 = \alpha_2,$$

and hence, it is assumed that the random components satisfy  $u_i \sim \text{Beta}(1/\alpha, 1/\alpha)$ , where  $\alpha > 0$  (Lee et al., 2006). This assumption fixes the dispersion parameter of the beta distribution as  $\varphi = \alpha/2$ .

In HGLMs, especially in the BBhglm, the computation of the marginal likelihood is not straightforward, and moreover, it is totally uninformative about the random effects  $\mathbf{v}$ . Consequently, Lee and Nelder (1996) proposed the so-called  $h$ -likelihood (or hierarchical-likelihood) as an approach to perform inference in HGLMs.

Let  $\mathbf{y} = (y_1, \dots, y_n)'$  be a set of the observed outcomes. The  $h$ -likelihood of any HGLM is defined by

$$h = h(\boldsymbol{\beta}, \mathbf{v}, \lambda, \alpha | \mathbf{y}) = \log f(\mathbf{y} | \boldsymbol{\beta}, \mathbf{v}, \lambda) + \log f(\mathbf{v} | \alpha), \quad (2.15)$$

where the first term is the conditional log-likelihood of the response and the second term is the log-likelihood corresponding to the random effects in the linear predictor. In the BBhglm the first term corresponds to the binomial density function, while the second term corresponds to the beta density function through the logit transformation. Consequently, the  $h$ -likelihood in BBhglm is given by

$$\begin{aligned} h(\boldsymbol{\beta}, \mathbf{u}, \alpha | \mathbf{y}) &= \log f(\mathbf{y} | \boldsymbol{\beta}, \mathbf{u}, \lambda) + \log \left[ f(l^{-1}(\mathbf{v}) | \alpha) \frac{\partial l^{-1}(\mathbf{v})}{\partial \mathbf{v}} \right] \\ &= \log f(\mathbf{y} | \boldsymbol{\beta}, \mathbf{u}, \lambda) + \log [f(\mathbf{u} | \alpha) l^{-1}(\mathbf{v}) (1 - l^{-1}(\mathbf{v}))] \\ &= \sum_{i=1}^n \frac{y_i \theta_i - m_i \log(1 + e^{\theta_i})}{\lambda} + c(y_i, \lambda) \\ &\quad + \log \left[ \frac{u_i^{1/\alpha-1} (1 - u_i)^{1/\alpha-1}}{B(1/\alpha, 1/\alpha)} u_i (1 - u_i) \right] \\ &= \sum_{i=1}^n \frac{y_i \log \frac{p_i}{1-p_i} + m_i \log(1 - p_i)}{\lambda} + \frac{1}{\alpha} \log(u_i (1 - u_i)) + d(y_i, \lambda, \alpha) \\ &= \sum_{i=1}^n \frac{y_i \log \frac{p_i}{1-p_i} + m_i \log(1 - p_i)}{\lambda} + \frac{\frac{1}{2} \log \frac{u_i}{1-u_i} + \log(1 - u_i)}{\alpha/2} \\ &\quad + d(y_i, \lambda, \alpha) \\ &= \sum_{i=1}^n \frac{y_i \log \frac{p_i}{1-p_i} + m_i \log(1 - p_i)}{\lambda} + \frac{\psi_c \log \frac{u_i}{1-u_i} + \log(1 - u_i)}{\varphi} \\ &\quad + d(y_i, \lambda, \alpha) \end{aligned}$$

where  $d(y_i, \lambda, \alpha) = c(y_i, \lambda) - \log B(1/\alpha, 1/\alpha)$ .

In fact, conjugacy allows to consider  $\log f(\mathbf{v} | \alpha)$  as the log-likelihood of quasi-data  $\boldsymbol{\psi} = \mathbf{1}_n \psi_c$ , where  $\mathbf{1}_n$  is a vector of 1s with length  $n$ , with quasi-fixed parameters  $u_i$ , satisfying the relationship  $\mathbb{E}[\psi_i] = u_i$  and  $\text{Var}[\psi_i] = \varphi V(u_i)$ ,  $i = 1, \dots, n$ . Therefore,



the estimation of HGLMs can be seen as an augmented GLM with the response variable  $\mathbf{y}_a = (\mathbf{y}', \boldsymbol{\psi}')'$ , where

$$\mathbb{E}[\mathbf{Y}] = \boldsymbol{\mu}, \quad \mathbb{E}[\boldsymbol{\psi}] = \mathbf{u},$$

$$\text{Var}[\mathbf{Y}] = \lambda V(\boldsymbol{\mu}), \quad \text{Var}[\boldsymbol{\psi}] = \varphi V(\mathbf{u}).$$

Moreover, the augmented linear predictor is defined as

$$\boldsymbol{\eta}_a = (\boldsymbol{\eta}', \boldsymbol{\eta}'_c)' = \mathbf{T}\boldsymbol{\omega},$$

where  $\boldsymbol{\eta} = g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v}$  is the linear predictor of the original model,  $\boldsymbol{\eta}_c = l(\mathbf{u}) = \mathbf{v}$  is the augmented linear predictor,  $\boldsymbol{\omega} = (\boldsymbol{\beta}', \mathbf{v}')'$  are fixed unknown parameters and quasi-parameters and  $\mathbf{T}$  is the augmented model matrix which is defined as

$$\mathbf{T} = \begin{pmatrix} \mathbf{X} & \mathbf{Z} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}.$$

If the dispersion parameters  $\lambda$  and  $\alpha$  are fixed, based on the augmented GLM approach, we can compute the IWLS algorithm to estimate  $\boldsymbol{\omega}$ , and consequently  $\boldsymbol{\beta}$  and  $\mathbf{v}$ . Therefore, the estimating equations are defined as

$$\mathbf{T}'\boldsymbol{\Sigma}_a\mathbf{T}\hat{\boldsymbol{\omega}} = \mathbf{T}'\boldsymbol{\Sigma}_a^{-1}\mathbf{z}_a$$

where, on the one hand, the augmented adjusted dependent variable  $\mathbf{z}_a = (\mathbf{z}', \mathbf{z}'_c)'$  consists of

$$z_i = \eta_i + (y_i - \mu_i)(\partial\eta_i/\partial\mu_i) \quad \text{and} \quad z_{ci} = v_i + (\psi_c - u_i)(\partial v_i/\partial u_i).$$

And on the other hand, the prior weight  $\mathcal{J}^{-1}$  is defined as

$$\mathcal{J} = \begin{bmatrix} \boldsymbol{\Lambda} & \mathbf{0} \\ \mathbf{0} & \mathbf{A} \end{bmatrix},$$

where  $\boldsymbol{\Lambda} = \text{diag}(\lambda_i)$  and  $\mathbf{A} = \text{diag}(\alpha_i)$ , and the weight function

$$\mathbf{W}_a = \begin{bmatrix} \mathbf{W} & \mathbf{0} \\ \mathbf{0} & \mathbf{W}_c \end{bmatrix}$$

is defined as

$$\mathbf{W} = (\partial\boldsymbol{\mu}/\partial\boldsymbol{\eta})^2 V(\boldsymbol{\mu})^{-1} \quad \text{and} \quad \mathbf{W}_c = (\partial\mathbf{u}/\partial\mathbf{v})^2 V(\mathbf{u})^{-1},$$

which leads to the overall weight function  $\boldsymbol{\Sigma}_a^{-1}$  defined as

$$\boldsymbol{\Sigma}_a = \mathcal{J}\mathbf{W}^{-1} = \begin{bmatrix} \boldsymbol{\Sigma} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_c \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Lambda}\mathbf{W}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}\mathbf{W}_c^{-1} \end{bmatrix}.$$

However, for the estimation of the dispersion parameters  $\lambda$  and  $\alpha$ , the MLE may be substantially biased owing to the estimation of  $\boldsymbol{\beta}$  (Lee and Nelder, 1996). Therefore, a modification of the  $h$ -likelihood must be applied in order to obtain a consistent estimation of the dispersion parameters. Let us consider the following function which consists of a likelihood function  $l$  and a parameter  $\alpha$ ,

$$p_\alpha(l) = \left\{ l - \frac{1}{2} \log \left[ \det \left( \frac{1}{2} \mathbf{D}(l, \alpha) / \pi \right) \right] \right\} \Big|_{\alpha=\hat{\alpha}},$$

where  $\mathbf{D}(l, \alpha) = -\partial^2 l / \partial \alpha^2$  and  $\hat{\alpha}$  solves  $\partial l / \partial \alpha = 0$ . The defined  $p_\alpha(l)$  is a profile likelihood that eliminates nuisance effects  $\alpha$  from  $l$ . In fact,  $\mathbf{D}(l, \alpha)$  is called the adjusted term for such elimination. On the one hand, Cox and Reid (1987) showed that for fixed effects  $\boldsymbol{\beta}$  the use of  $p_\beta(l)$  is equivalent to conditioning the likelihood on  $\hat{\boldsymbol{\beta}}$ . On the other hand, Lee and Nelder (2001) proved that the use of  $p_v(l)$  for the random effects is equivalent to integrating them out.

Therefore, for the estimation of the dispersion parameters, Lee and Nelder (1996) proposed to use the adjusted profile  $h$ -likelihood

$$p_\omega(h) = \left\{ h - \frac{1}{2} \log \left[ \det \left( \frac{1}{2} \mathbf{D}(h, \omega) / \pi \right) \right] \right\} \Big|_{\omega=\hat{\omega}},$$

where  $h$  is the  $h$ -likelihood defined in Equation (2.15),  $\mathbf{D}(h, \omega)$  corresponds to the Hessian matrix of the HGLM and  $\omega$  corresponds to the fixed parameters of the augmented model. Iterative use of  $\partial p_\omega(h) / \partial \lambda$  and  $\partial p_\omega(h) / \partial \varphi$  (or  $\partial p_\omega(h) / \partial \alpha$ ) leads to the maximum adjusted profile  $h$ -likelihood estimations of the dispersion parameters.

## 2.3 Application to the Short Form-36

In this section we will analyse data from the COPD Study in order to conclude with remarkable relationships between the HRQoL of patients and clinical and

socio-demographical variables listed in Section 1.3.1. It has been stated that regression models based on the beta-binomial distribution are appropriate candidates for analysing PRO data and, in particular, HRQoL provided by the SF-36 Health Survey (Arostegui et al., 2012). However, we have mentioned in the previous section that there are two different approaches, the marginal (BBreg) and the conditional (BBhglm), for implementing that type of model. Therefore, we are going to make use of both regression approaches to analyse data from the COPD Study and, that way, we will not only look for relationships between the HRQoL and covariates, but also check the performance of both methodologies in real data application.

In terms of statistical packages and software, on the one hand, we have implemented the BBreg approach in the R-package `PROreg` available at CRAN, <https://cran.r-project.org/web/packages/PROreg/index.html>. Further discussion about the specific function that performs the BBreg approach is provided in Chapter 5. On the other hand, the BBhglm approach is implemented in the `hg1m` R-package (Ronnegard et al., 2010) also available at CRAN.

The eight dimensions of the SF-36 Health Survey were the response variables and clinical and sociodemographic variables listed in Table 1.2 were considered as independent variables. Separated models were performed for each of the health dimension of the SF-36 and exclusively data from the first visit to the outpatient clinic was considered. For variables selection, we retained in the model those covariates whose influence in HRQoL was statistically significant ( $p\text{-value} < 0.05$ ) in at least one of the modelling approaches. For simplicity, clarity and brevity of exposition, we only show results for three of the eight health dimensions of the SF-36. The selected three dimensions (*physical functioning*, *mental health* and *role emotional*) illustrate different shapes of the distribution (see Figure 1.4) and a wide range of maximum number of scores ( $m$ ), from 4 to 20.

Tables 2.1-2.3 provide the results obtained from the analysis of the mentioned SF-36 dimensions in the COPD Study by both beta-binomial regression approaches. Estimates of the regression coefficients, their standard deviations and test of significance associated with the BBreg and the BBhglm modelling approaches for the selected three health dimensions of the SF-36 Health Survey are displayed. We also show the estimates in logarithmic scale of the dispersion parameter of each approach,  $\alpha$  for BBhglm and  $\phi$  for BBreg.

Real data application leads to several conclusions and interpretation. As regards to the effects of the covariates in the SF-36 dimensions, it can be appreciated that while in *physical functioning* dimension both algorithms lead to similar estimates,

Table 2.1: Effect of explanatory variables in the *physical functioning* dimension measured by both regression approaches based on the beta-binomial distribution.

Physical functioning	BBhglm			BBreg		
	$\hat{\beta}$	SD( $\hat{\beta}$ )	p-value	$\hat{\beta}$	SD( $\hat{\beta}$ )	p-value
Dyspnea						
<i>Mild</i>	-0.616	0.111	<0.001	-0.580	0.112	<0.001
<i>Moderate</i>	-1.339	0.122	<0.001	-1.281	0.120	<0.001
<i>Severe</i>	-2.317	0.178	<0.001	-2.207	0.176	<0.001
Depression						
<i>Yes</i>	-0.541	0.139	<0.001	-0.544	0.130	<0.001
Anxiety						
<i>Yes</i>	-0.416	0.096	<0.001	-0.404	0.090	<0.001
Sex						
<i>Female</i>	0.469	0.167	0.005	0.461	0.155	0.003
FEV1%	0.007	0.003	0.011	0.006	0.002	0.012
BMI	-0.019	0.007	0.011	-0.018	0.007	0.009
Age	0.013	0.004	0.002	0.012	0.004	0.002
Walking Test	0.004	10 <sup>-4</sup>	<0.001	0.004	10 <sup>-4</sup>	<0.001
log( $\alpha$ )	-2.656	0.084	—	—	—	—
log( $\phi$ )	—	—	—	-2.826	0.115	—

SD: Standard Deviation; BMI: Body Mass Index; FEV1%: Forced Expiratory Volume in one second in percentile.

in *mental health* and, especially, in *role emotional* dimension regression parameter estimates and statistical significances are completely different. For example, for *role emotional* dimension, on the one hand, the estimation of the coefficient corresponding to anxiety is  $-6.145$  in BBhglm approach and  $-1.649$  in BBreg, being both statistically significant in the model. On the other hand, the p-value corresponding to the estimate of moderate dyspnea is statistically significant in BBreg approach ( $< 0.001$ ), but not in BBhglm ( $0.434$ ).

Due to the fact that the logit link function is used in both methodologies, the interpretation of the regression coefficients  $\beta$  in both approaches is equivalent to the log odds-ratio in a binomial logistic regression model. For instance, the coefficient of depression in the *physical functioning* model for BBreg approach is  $-0.544$ , which means that based on this model the presence of depression increases by  $1/\exp(-0.544) = 1.72$  the odds of having a smaller *physical functioning* score.

Table 2.2: Effect of explanatory variables in the *mental health* dimension measured by both regression approaches based on the beta-binomial distribution.

<b>Mental health</b>	BBhglm			BBreg		
	$\hat{\beta}$	SD( $\hat{\beta}$ )	p-value	$\hat{\beta}$	SD( $\hat{\beta}$ )	p-value
Dyspnea						
<i>Mild</i>	-0.353	0.234	0.134*	-0.294	0.141	0.037
<i>Moderate</i>	-0.853	0.246	<0.001	-0.704	0.145	<0.001
<i>Severe</i>	-1.132	0.320	<0.001	-0.961	0.181	<0.001
Anxiety						
<i>Yes</i>	-1.480	0.204	<0.001	-1.290	0.108	<0.001
Depression						
<i>Yes</i>	-0.966	0.298	0.002	-0.853	0.157	<0.001
$\log(\alpha)$	-0.7647	0.069	—	—	—	—
$\log(\phi)$	—	—	—	-2.263	0.115	—

SD: Standard Deviation. Symbol \* stands for regression coefficients that are not statistically significant.

Table 2.3: Effect of explanatory variables in the *role emotional* dimension measured by both regression approaches based on the beta-binomial distribution.

<b>Role emotional</b>	BBhglm			BBreg		
	$\hat{\beta}$	SD( $\hat{\beta}$ )	p-value	$\hat{\beta}$	SD( $\hat{\beta}$ )	p-value
Anxiety						
<i>Yes</i>	-6.145	2.062	0.003	-1.649	0.226	<0.001
Dyspnea						
<i>Mild</i>	-2.600	5.229	0.619*	-0.614	0.418	0.142*
<i>Moderate</i>	-3.981	5.080	0.434*	-1.379	0.413	<0.001
<i>Severe</i>	-5.603	5.496	0.309*	-2.048	0.467	<0.001
$\log(\alpha)$	2.735	0.095	—	—	—	—
$\log(\phi)$	—	—	—	0.668	0.150	—

SD: Standard Deviation. Symbol \* stands for regression coefficients that are not statistically significant.

Therefore, at first sight, it seems that both regression approaches lead to completely different conclusions about the effect of the covariates in the HRQoL of pa-

tients with COPD. However, care must be required when comparing marginal and conditional models. In fact, although the interpretation of the parameters is made in the same way, it is worth noticing that they refer to different measurements. For instance, the BBreg approach should be interpreted in terms of a marginal response, and hence, conclusions should be taken in terms of population. Indeed, the linear predictor of BBreg approach is constructed based on the marginal expectation of the outcome variable,

$$\text{logit}(\mathbb{E}[Y_i]) = \text{logit}(p_i^m) = \mathbf{x}'_i \boldsymbol{\beta}.$$

On the contrary, the linear predictor of the BBhglm approach depends on the conditional mean,

$$\text{logit}(\mathbb{E}[Y_i|u_i]) = \text{logit}(p_i^c) = \mathbf{x}'_i \boldsymbol{\beta} + v_i.$$

We denote  $\mathbf{p}^m$  and  $\mathbf{p}^c$  to refer to the marginal and conditional means respectively, where  $\mathbb{E}[\mathbf{p}^c] = \mathbf{p}^m$ . Therefore, the conditional BBhglm approach describes individual responses, and consequently, interpretation of the parameters is done holding the value of the random effect (a particular value that corresponds to each individual).

Due to the fact that the logit and the expectation operator do not commute (i.e.  $\mathbb{E}[\text{logit}(p_i^c)] \neq \text{logit}(\mathbb{E}[p_i^c]) = \text{logit}(p_i^m)$ ), it has been shown that each approach is modelling a different measurement, and hence, we cannot compare them directly. However, there are still some features shown in the real data application that should be explained. First of all, as mentioned before, due to the model definition, we know that regression coefficients estimates through marginal and conditional models may differ. However, differences seem to be larger than expected. Moreover, we know that if a covariate does not affect the individuals, it has no effect on populations; however, the real data application does not show the same. In fact, it can be appreciated in Table 2.3 that the effect of the mild dyspnea is statistically significant in the marginal approach, but not in the conditional approach, which does not make sense with the previous statement. Furthermore, standard deviations of the estimates are completely different in both approaches, which could tell that one of the models is over or under estimating the variances.

Figure 2.1 shows the distribution of the analysed three SF-36 dimensions and the model-fit by the BBhglm approach. It can be appreciated the subject-specific feature of the approach, where the inclusion of a beta random effect per individual accommodates the dispersion of the fitted values. Therefore, it is shown that the distribution of the fitted values corresponds to the observed distribution of the re-

sponses, especially in *role emotional* dimensions, where results tended to be more misleading (see Table 2.3). Consequently, Figure 2.1 shows that, apparently, the *BBhglm* approach is correct, at least concerning fitted values, and that it is fitting the relationship between the HRQoL of the patients and covariates adequately.

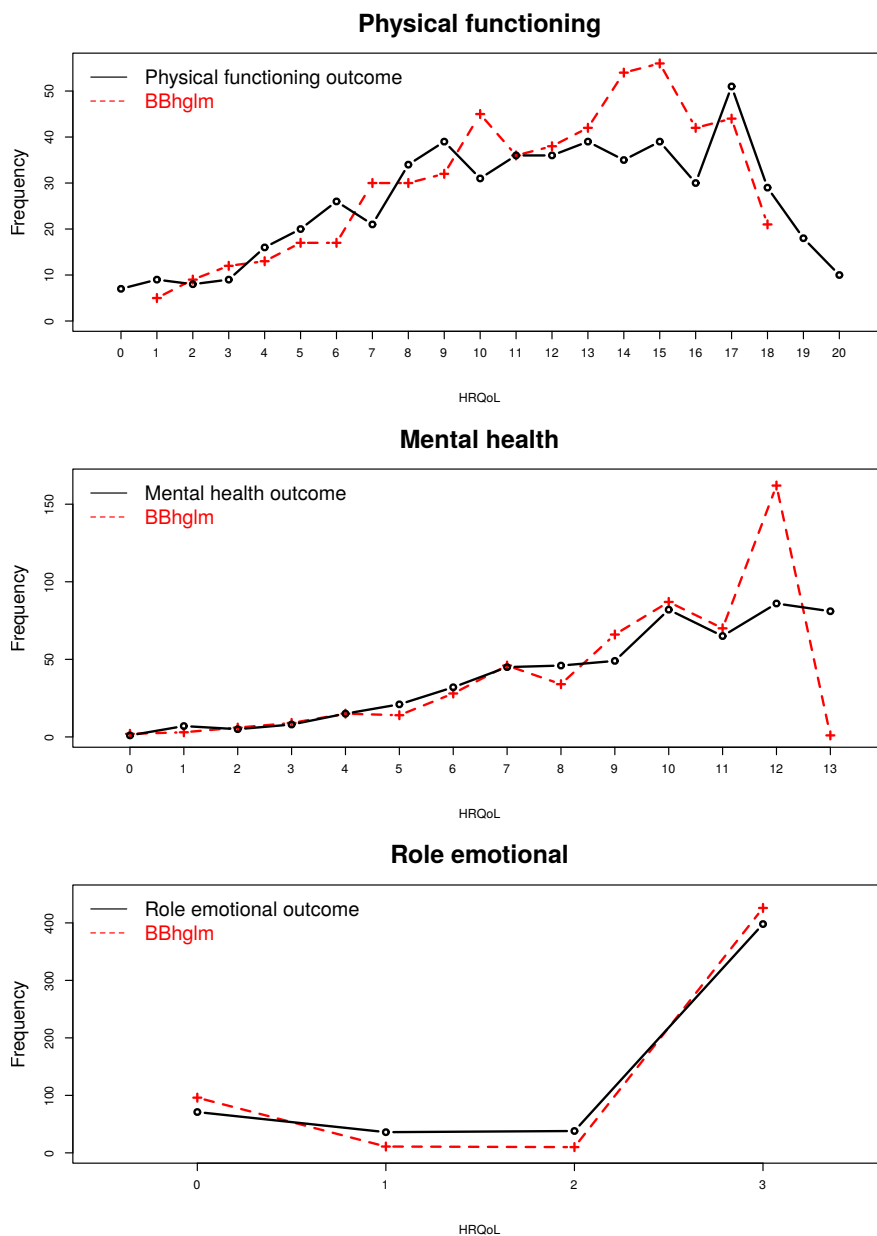


Figure 2.1: Observed distribution and fitted distribution by the *BBhglm* approach of the analysed SF-36 scores.

In general, we have explained that due to the model specification we cannot compare estimates from both approaches directly. However, from our point of view, there are some issues, such as statistical significance and over/under inflation of the variance, that must be addressed in order to conclude with the most appropriate approach to measure the effect of the covariates in PROs, for instance, the HRQoL of patients with COPD. It seems that as the dispersion parameter increases, both approaches conclude in more different results. Therefore, in the next section, we focus on the comparison of the two methodological approaches through a complete simulation study which is divided in different scenarios depending on the value of the dispersion parameter.

## 2.4 Simulation study

In this section we perform a simulation study to compare the adequacy in terms of parameters estimation, standard deviation and significance of both beta-binomial regression modelling approaches in PROs framework.

In terms of the software implementation, the R-packages described in Section 2.3 have been used in order to perform the simulation study, i.e. `PROreg` for BBreg approach and `hg1m` for BBhglm.

### 2.4.1 Scenarios set up

First of all, we set up some controlled scenarios defining specific values for the parameters of the model. Different scenarios correspond to different situations that can occur in real practise.

Given the recoded 8 health dimensions provided by the SF-36, we consider three groups based on the maximum score  $m$ , i.e.: *few* (3 and 4), *standard* (8, 9 and 10) and *large* (19 and 20) (see Figure 1.4). Consequently, in order to generalise the results, the simulation study has been also divided into three scenarios considering a maximum score of 4, 10 and 20. Finally, we have generated 500 random realizations of 100 observations of a dependent variable  $Y$  assuming a beta-binomial distribution with fixed probability and dispersion parameters.

In order to understand the behaviour of the methodologies in PROs framework, we are going to focus the simulation exercise on a regression approach with a single continuous covariate. The probability parameter of the beta-binomial distribution has been calculated as shown in Equation (2.4) for a fixed value of  $\beta_0$  and  $\beta_1$  equal



to 1 and  $-0.3$  respectively, and a fixed covariate  $X$  simulated assuming a normal distribution with mean 3 and standard deviation equal to 2.

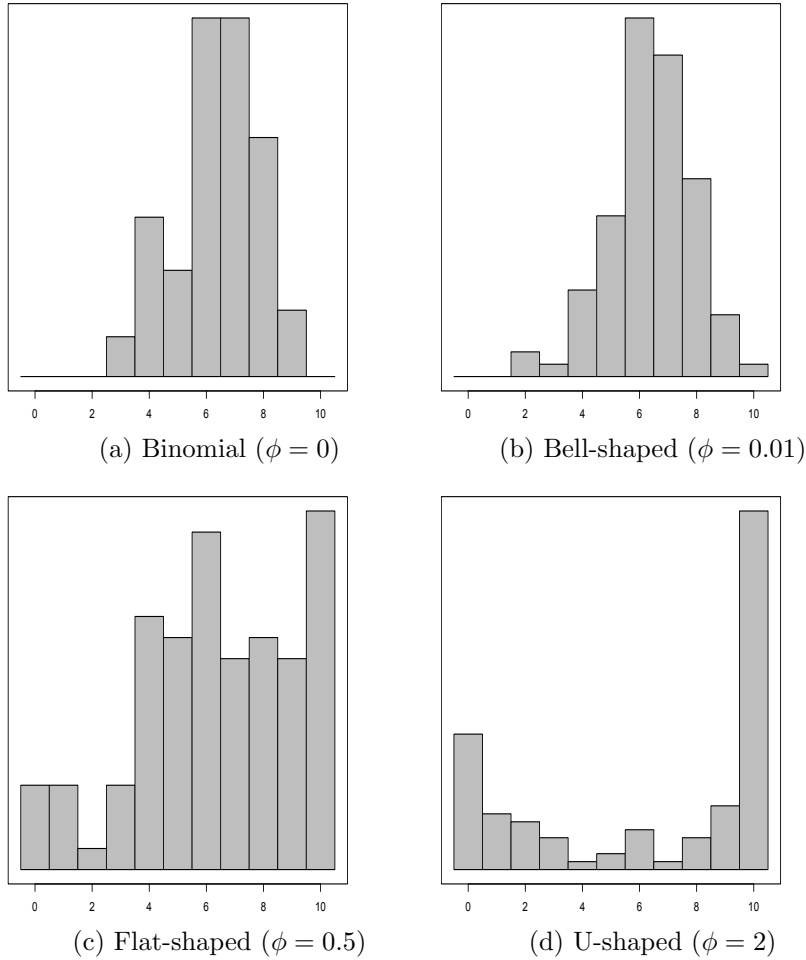


Figure 2.2: Distribution of the beta-binomial distribution for  $p = 0.5$ ,  $m = 10$  and different values of  $\phi$ .

The value of the dispersion parameter  $\phi$  defines different scenarios (the shape of the beta distribution), as for a fixed probability parameter the shape of the distribution changes considerably for different values of  $\phi$ . Values greater than 0.5 provides U-shaped distributions, values lower than 0.5 bell-shaped and a value equal to 0.5 flat-shaped. Figure 2.2 illustrates possible scenarios showing how the shape of the beta-binomial distribution changes for a fixed probability parameter equal to 0.5 considering the values  $\phi = 0.01$  (see Figure 2.2b),  $\phi = 0.5$  (see Figure 2.2c) and  $\phi = 2$

(see Figure 2.2d) for the dispersion parameter. If there is no overdispersion,  $\phi = 0$ , it corresponds to the usual binomial distribution (see Figure 2.2a). Hence, when the value of the dispersion parameter becomes greater the distribution is further from the mean value and the observations are accumulated at both extremes of the scale.

### 2.4.2 Results

As it was mentioned before, the main objective of most PROs analysis is focused on the estimation of the effect that some previously measured variables may have on the health-status of some population. Therefore, regression models are applied to the data where the interest lies in the estimation and interpretation of the regression coefficients  $\beta$ . It was detailed in Section 2.4.1 that, for the sake of clarity, the regression model performed in this simulation study only consists of a covariate, and hence, two regression parameters are included in the linear predictor: the intercept ( $\beta_0$ ) and the slope ( $\beta_1$ ). Consequently, although some details about the estimation of the intercept are provided, the objective of the simulation study lies in the correct measurement of the slope, the regression coefficient that multiplies the simulated covariate in the linear predictor.

In order to compare the performance of both beta-binomial regression approaches, the marginal (BBreg) and the conditional (BBhglm), Tables 2.4-2.6 include the empirical mean and expected mean square error (EMS) of the estimates of the intercept coefficient ( $\beta_0$ ). In addition, Tables 2.4-2.6 also show a deeper comparison analysis of the slope ( $\beta_1$ ) where, apart from the mean and EMS, the empirical standard deviation (ESD) and the average standard deviation (ASD) of the estimates are shown together with the coverage probability of the 95% Wald confidence intervals for the estimates of the slope ( $\beta_1 = -0.3$ ) and the percentage the simulated covariate effect is statistically significant in each model (PCSS). Moreover, Figures 2.3-2.5 show the boxplots of the estimates of the slope in the 500 simulations for both beta-binomial regression approaches.

The simulation study shows that similar to the real data application, differences in the results provided by both beta-binomial regression approaches depend on the scenarios. Indeed, results begin to differ as the dispersion parameter  $\phi$  and maximum score number  $m$  increase. For instance, Table 2.4 and Figure 2.3 show that when there is a low dispersion in the data ( $\phi = 0.01$ ) results provided by both approaches tend to be very similar. However, if a large dispersion is considered ( $\phi = 2$ ) results provided by both approaches are completely different (see Table 2.6 and Figure 2.5).

Table 2.4: Results of the simulation study for the bell-shaped distribution ( $\phi = 0.01$ ) for  $n = 100$  individuals and  $R = 500$  replicates.

		$\beta_0 = 1$		$\beta_1 = -0.3$					
	Model	Mean	EMS	Mean	ESD	ASD	EMS	CP	PCSS
$m = 4$	BBhglm	1.015	0.048	-0.305	0.061	0.065	0.004	97.8	100
	BBreg	1.009	0.046	-0.303	0.059	0.064	0.003	97.4	100
$m = 10$	BBhglm	0.997	0.019	-0.299	0.039	0.041	0.001	96.4	100
	BBreg	0.995	0.019	-0.298	0.038	0.041	0.001	96.2	100
$m = 20$	BBhglm	1.013	0.013	-0.303	0.031	0.030	0.001	93.8	100
	BBreg	1.010	0.013	-0.302	0.031	0.030	0.001	93.4	100

EMS: Expected Mean Square errors; ESD: Empirical Standard Deviation; ASD: Average Standard Deviation; CP: Coverage Probability of 95%; PCSS: Percentage the Covariate effect is Statistically Significant.

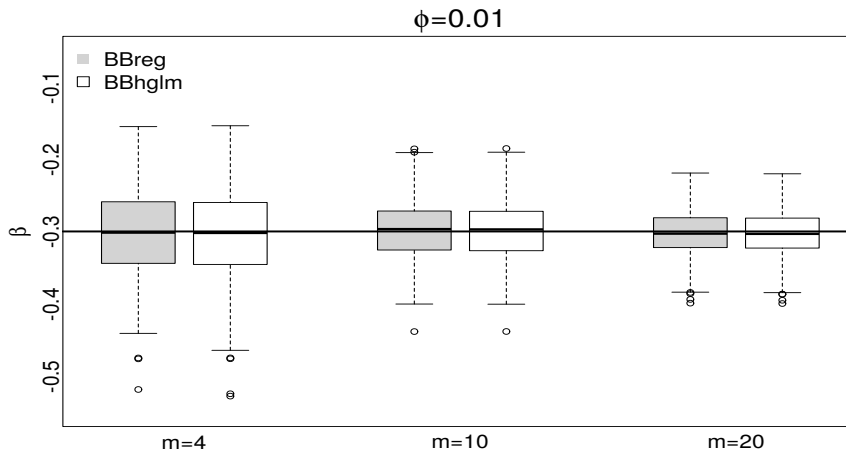


Figure 2.3: Boxplots of the slope estimates in the simulation study for the bell-shaped scenario ( $\phi = 0.01$ ). Simulations performed for  $n = 100$  individuals and  $R = 500$  replicates.

It has been mentioned before that, due to the fact that marginal and conditional approaches do not model the same quantity, they cannot be compared directly. Therefore, conclusions of the simulation study will not be obtained by comparing

Table 2.5: Results of the simulation study for the flat-shaped distribution ( $\phi = 0.5$ ) for  $n = 100$  individuals and  $R = 500$  replicates.

	Model	$\beta_0 = 1$		$\beta_1 = -0.3$					
		Mean	EMS	Mean	ESD	ASD	EMS	CP	PCSS
$m = 4$	BBhglm	1.286	0.346	-0.385	0.143	0.124	0.028	93.8	94.2
	BBreg	0.967	0.090	-0.291	0.082	0.086	0.007	96.8	96.0
$m = 10$	BBhglm	1.426	0.421	-0.427	0.131	0.125	0.033	90.0	96.0
	BBreg	0.957	0.072	-0.289	0.071	0.074	0.005	95.8	98.6
$m = 20$	BBhglm	1.589	0.649	-0.478	0.150	0.132	0.054	81.4	96.0
	BBreg	0.997	0.066	-0.303	0.070	0.070	0.005	95.8	99.4

EMS: Expected Mean Square errors; ESD: Empirical Standard Deviation; ASD: Average Standard Deviation; CP: Coverage Probability of 95%; PCSS: Percentage the Covariate effect is Statistically Significant.

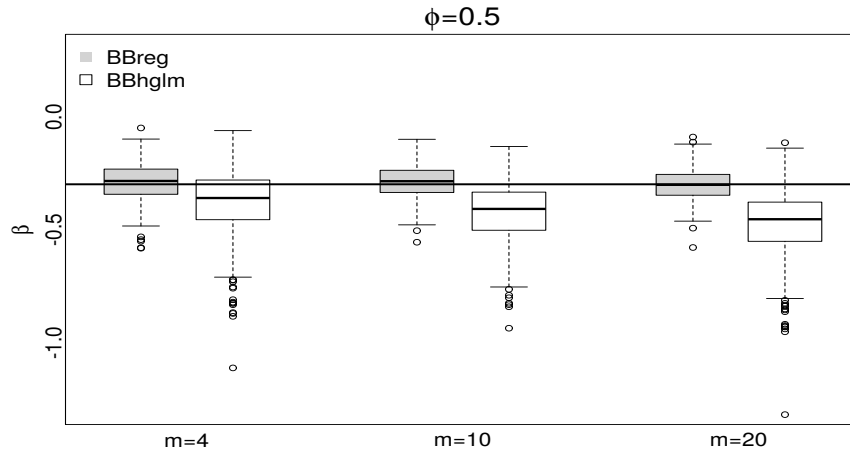


Figure 2.4: Boxplots of the slope estimates in the simulation study for the flat-shaped scenario ( $\phi = 0.5$ ). Simulations performed for  $n = 100$  individuals and  $R = 500$  replicates.

both approaches like with like, but instead, we will analyse results provided by each approach on its own. Finally, some conclusion about the adequacy of the approaches will be offered together with some recommendations.

Table 2.6: Results of the simulation study for the U-shaped distribution ( $\phi = 2$ ) for  $n = 100$  individuals and  $R = 500$  replicates.

		$\beta_0 = 1$		$\beta_1 = -0.3$					
	Model	Mean	EMS	Mean	ESD	ASD	EMS	CP	PCSS
$m = 4$	BBhglm	1.990	2.433	-0.610	0.347	0.270	0.217	91.4	61.0
	BBreg	0.798	0.134	-0.250	0.091	0.103	0.011	93.6	73.0
$m = 10$	BBhglm	2.719	6.002	-0.823	0.481	0.369	0.505	86.8	41.2
	BBreg	0.791	0.130	-0.248	0.085	0.096	0.010	93.8	77.0
$m = 20$	BBhglm	3.027	8.050	-0.920	0.539	0.413	0.675	78.2	24.2
	BBreg	0.777	0.134	-0.247	0.084	0.093	0.010	93.4	79.8

EMS: Expected Mean Square errors; ESD: Empirical Standard Deviation; ASD: Average Standard Deviation; CP: Coverage Probability of 95%; PCSS: Percentage the Covariate effect is Statistically Significant.

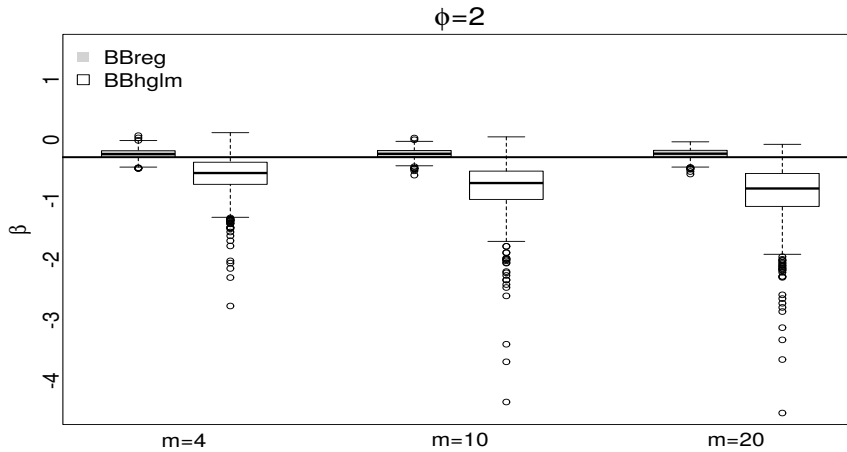


Figure 2.5: Boxplots of the slope estimates in the simulation study for the U-shaped scenario ( $\phi = 2$ ). Simulations performed for  $n = 100$  individuals and  $R = 500$  replicates.

On the one hand, as regards BBreg approach, several remarkable conclusions can be appreciated. First of all, regarding the bias of the estimates, it is shown that the EMS of the estimates remain very small in all the scenarios, especially

in the slope parameter where the highest value is equal to 0.011 (see Table 2.6). Therefore, we conclude that variability in the data does not affect the parameter estimation by BBreg approach, which maintains very close to the real value. In terms of the variance of the estimates, Tables 2.4-2.6 show that the ESD and ASD of the estimates are almost equal in all the scenarios, concluding that the algorithm is coherent with the estimation process. Additionally, it can be appreciated that as the maximum score number  $m$  increases the variance of the estimates (both empirical and average) decreases and, as the bias remains constant, the EMS is reduced. Indeed, this is a logical result as increasing the maximum score number  $m$  we are somehow increasing the sample size, and therefore, the estimator is more efficient. Finally, we have to remark that the interpretation of the regression parameters in BBreg approach is made in terms of population average, which can be very useful for clinical applications.

On the other hand, regarding BBhglm approach, results differ for several reasons. Firstly, we have to keep in mind that the interpretation of the bias does not make any sense in this case where the value being modelled was not the marginal expectation, but the logit of the conditional mean instead. In fact, the simulation study was performed in a marginal context, and hence, regression parameters in a conditional approach do not have the same meaning. Nevertheless, it can be appreciated in Figures 2.3-2.5 and Tables 2.4-2.6 that the estimates begin to distance from the marginal approach value as the dispersion parameter increases. Indeed, as noticed by Lee and Nelder (2004), in a mixed-effects regression context, different model approaches tend to be more similar as the heterogeneities of individuals are small. However, not only the bias but the standard deviation of the estimates (empirical and average) increase for larger values of the dispersion parameter  $\phi$ . Compared to BBreg approach, we realise that the BBhglm approach doubles the variances in large variability scenarios. Consequently, significance tests of the parameter estimates based on the estimated standard deviations tend to accept the null hypothesis, and therefore, conclude that there does not exist any relationship between the covariate and the response variable. In this sense, we find a contradiction between marginal and conditional models, because as it was stated by Senn in the comments to the paper by Lee and Nelder (2004): “After all, if the treatment cannot affect individuals, it has no effect on populations...”. However, Tables 2.5-2.6 show that the previous statement does not hold in the simulation study, because, for instance, in the last scenario ( $m = 20$  and  $\phi = 2$ ) the percentage the covariate effect is statistically significant in the model is equal to 79.8% in BBreg, but only

24.2% in BBhglm. In addition, it can be appreciated that the increment of the maximum number of scores  $m$  does not reduce the variability of the estimates in BBhglm approach, in fact, it increases, which in our opinion could be because the estimator is not efficient.

To sum up, we have shown that the BBreg approach performs adequately in terms of bias and variances of the estimations. Moreover, even in large variability scenarios, it is able to capture the effect of the covariate on the response variable. In contrast, BBhglm results cannot be interpreted based on the marginal mean because conditional approaches are considered, and instead, they should be interpreted on an individual level holding the same value of the predicted random effect. However, we have shown that when the variability in the data is large enough ( $\phi = 1$  or  $\phi = 2$ ), the BBhglm approach is not able to capture the effect of covariates that do affect the population average.

Finally, it is worth mentioning that while the BBreg approach converged in all the simulated scenarios, the BBhglm approach had several convergence problems, especially in the U-shaped scenario ( $\phi = 2$ ) where it got a convergence rate of 45.09%, 39.18% and 37.37% when  $m = 4$ ,  $m = 10$  and  $m = 20$  respectively.

## 2.5 Conclusions and discussion

It is known that differences in the behavior of regression coefficients in so-called marginal and conditional models are based on a failure to compare them like with like. In fact, it was shown by Lee and Nelder (2004) that these differences are mainly caused by the choice of unidentifiable constraints on the random effects. When we are defining a mixed-effect model, it may be reasonable to assume that an individual's unobserved trait ( $v_i$ ) follows a certain distribution. However, the center of this distribution cannot be identified as it is confounded with the intercept term. One solution is to fix the first moment of the random effects to some previously defined constraint. Nevertheless, the restriction of the expectation of the random effects affects the estimation procedure of the model (Lee and Nelder, 1996) and, in fact, the election of the constraints is crucial if the fixed effects in different models are going to be comparable in general (Lee and Nelder, 2004).

Some HGLMs, for instance, models based on a conditional Poisson distribution, display an easy decomposition of the model which allows the comparison of different random effects models, but even the marginal model. For example, assume two random effects models based on the Poisson distribution. On the one hand, the

normal-Poisson HGLM (Poisson GLMM), which we will denote as NP, is defined as

$$Y_i|v_i \sim \text{Poisson}(\mu_i^n) \quad \text{and} \quad v_i \sim \mathcal{N}(0, \lambda^n)$$

$$\eta_i^n = \log(\mu_i^n) = \mathbf{x}'_i \boldsymbol{\beta}^n + v_i.$$

On the other hand, the gamma-Poisson HGLM, denoted as GP, is defined as

$$Y_i|u_i \sim \text{Poisson}(\mu_i^g) \quad \text{and} \quad u_i \sim \text{Gamma}(1, \lambda^g)$$

$$\eta_i^g = \log(\mu_i^g) = \mathbf{x}'_i \boldsymbol{\beta} + \log u_i,$$

or equivalently,

$$\mu_i^g = \exp(\mathbf{x}'_i \boldsymbol{\beta}) u_i,$$

where  $\text{Gamma}(1, \lambda^g)$  denotes the gamma distribution with mean 1 and variance  $\lambda^g$ . Note that the superscript  $n$  makes reference to the NP model and  $g$  to the GP.

Each conditional model leads to a specific marginal model. The GP model leads to the marginal model

$$\log \mathbb{E}[Y_i] = \log \mathbb{E}[\mathbb{E}[Y_i|u_i]] = \log \mathbb{E}[\mu_i^g] = \mathbf{x}'_i \boldsymbol{\beta}^g$$

while the NP leads to

$$\mathbb{E}[\log(\mu_i^n)] = \mathbf{x}'_i \boldsymbol{\beta}^n.$$

These two models are different because the expectation does not commute with the log function, and hence, fixed coefficients cannot be directly compared. Indeed, the definition of the marginal mean depends on the scale on which the margins are formed. However, even the NP model offers an easy decomposition to the marginal mean as

$$\mathbb{E}[Y_i] = \mathbb{E}[\exp(\mathbf{x}'_i \boldsymbol{\beta}^n + v_i)] = \exp(\mathbf{x}'_i \boldsymbol{\beta}^n) \mathbb{E}[v_i^*],$$

where  $v_i^*$  follows the well known log-normal distribution. Indeed, the expectation of the log-normal distribution has an analytic known expression

$$\mathbb{E}[v_i^*] = \exp\left(\frac{\lambda^n}{2}\right).$$

Therefore, we have that

$$\mathbf{x}'_i \boldsymbol{\beta}^g = \mathbf{x}'_i \boldsymbol{\beta}^n + \frac{\lambda^n}{2},$$

and hence,  $\boldsymbol{\beta}^g$  and  $\boldsymbol{\beta}^n$  are comparable.



In the binomial-beta HGLM, however, the transformation of the linear predictor to the marginal first order is not so direct. In fact, in BBhglm approach we have that

$$\eta_i^b = \text{logit}(\mathbb{E}[Y_i|u_i]) = \text{logit}(p_i^b) = \mathbf{x}'_i \boldsymbol{\beta}^b + \log \frac{u_i}{1+u_i},$$

where  $\mathbb{E}[y_i|u_i] = p_i^b$  is the conditional mean and the superscript  $^b$  makes reference to the use of beta distributed random effects. If we try to calculate the marginal expectation of the outcome variable in BBhglm approach, we have that

$$\mathbb{E}[Y_i] = \mathbb{E}[\mathbb{E}[Y_i|u_i]] = \mathbb{E}[mp_i^b] = m\mathbb{E}[p_i^b],$$

where

$$p_i^b = \frac{u_i \exp(\mathbf{x}'_i \boldsymbol{\beta}^b)}{1 + u_i [\exp(\mathbf{x}'_i \boldsymbol{\beta}^b) - 1]}$$

which does not follow any known distribution, but specially, it does not have an analytic expectation. Therefore, we cannot compare BBreg and BBhglm regression coefficients like to like. Nevertheless, from a practical point of view, researchers working on PROs must be provided with a valid method of analysis for this kind of data.

Therefore, in this chapter, we have carried out a simulation study where the adequacy of BBreg and BBhglm approaches when analysing PRO data was examined. Among different conclusions, the most relevant one was that, in some situations, the conditional BBhglm approach is not able to identify covariates which had an statistically significant effect on the marginal mean of the outcome. However, it seems clear to think that if a covariate cannot affect individuals, it has no effect on populations. In fact, the previous result does not go with marginal and conditional models assumption, as if a variable has a statistically significant effect on the marginal mean, the effect on the conditional mean must be statistically significant as well (Lee and Nelder, 2004). Consequently, due to the fact that the main objective of most of the PROs studies is to find out relationships that could exist between the measured covariates and the response variables, we recommend the use of the marginal approach as a unified technique for analysing PRO data.

In conclusion, our results showed that when the goal of the study is to detect and interpret the effect of explanatory variables in the health-status of patients, the method of analysis must be cautiously selected. In fact, when there is large variability in the data, we showed that the BBhglm approach does not offer appropriate

---

results in terms of covariates statistical significance. Therefore, we recommend the use of BBreg approach as a unified technique for analysing PRO data and, in order to provide a useful tool to the researchers, it has been implemented by the authors in the `PROreg` R-package available in CRAN (further discussion about the `PROreg` R-package is compiled in Chapter 5).



---

---

## CHAPTER 3

---


# LONGITUDINAL ANALYSIS: A BETA-BINOMIAL MIXED-EFFECTS MODEL APPROACH


*“As far as the law of mathematics refer to reality, they are not certain; and as far as they are certain they do not refer to reality”*

---

Albert Einstein, 1878 – 1955

*The work developed in this chapter is being reviewed in Biometrical Journal and it has been partially presented in the Conference - 38th Annual Conference of the International Society for Clinical Biostatistics.*

 *Najera-Zuloaga, J., Lee, D.-J. and Arostegui, I. (2017). A beta-binomial mixed-effects model approach for analysing longitudinal patient reported outcomes. Biometrical Journal (under review)*

 *38th Annual Conference of the International Society for Clinical Biostatistics. Beta-binomial mixed-effects model for analyzing longitudinal binomial data with overdispersion. Najera-Zuloaga, J., Lee, D.-J. and Arostegui, I. Proceedings. Vigo, July 10 - 13, 2017.*

## 3.1 Introduction

### 3.1.1 Introduction to longitudinal models

Longitudinal studies play a key role in many frameworks such as health, social and behavioural sciences. They are indispensable to study the change in an outcome over time. The measurement of study participants repeatedly through the time allows the direct analysis of temporal changes within individuals and the factors that influence the change. Due to the fact that the study of the temporal change is so essential to almost every discipline, the number of longitudinal studies has increased in the past 30 to 40 years (Fitzmaurice et al., 2008).

A longitudinal study is defined as an investigation where participant outcomes and possibly treatments or exposures are collected at multiple follow-up times. Indeed, a longitudinal study generally contains multiple or repeated measurements on each subject in the research. Consequently, the repeated measurements are correlated within subjects and, thus, it requires special statistical techniques for valid analysis and inference. There is a wide variety of statistical challenges when analysing longitudinal data:

*Heterogeneity:* it is usual, specially in behavioural sciences, that there exist differences between individuals. The overall mean response in a sample drawn from a population does not provide any information about each individual deviation. Indeed, when the same subjects are repeatedly measured over time, their responses are multivariate and have a complex random-error structure that must be accommodated in the model. Moreover, each individual can be expected to have each own trend line, which deviates systematically from the overall mean trend line. Furthermore, in many studies personal features can be unobservable, leading to unexplained heterogeneity in the population. Consequently, the modelling of this unobserved heterogeneity is one of the main challenges of longitudinal models. For example, the modelling of the unobserved heterogeneity in terms of variance components (or random effects in mixed-effects models) is one way to accommodate the correlation between the repeated measurements over time and to better describe individual differences in the statistical characterization of the observed data.

*Time-dependent covariates:* longitudinal studies allow repeated measurements not only of the outcome variable, but also of the covariates. Consequently, time-dependent covariates can be included in the model. The inclusion of covariates

in the model that change over time can result in many complex analytic issues. The goal is to estimate the dynamic relationship between the variables over time, but notice that it is an individual within relationship, and may vary from individual to individual. Consequently, while our main goal is to determine if there exists a relationship between the variables in the whole population, one must be able to model the dynamic relationship within individuals.

*Missing data:* longitudinal studies tend to present missing data, as not all the subjects remain in the study until it finishes. Indeed, the basic problem is that even in randomized and well-controlled clinical trials, subjects available at the beginning of the study can be considerably different from subjects available at the end.

*Hierarchical structure:* in addition to the correlation structure produced by the repeated measurements of the same individual, the clustering of individuals within some grouping units (schools, hospitals, clinics, etc.) produces an additional correlation structure between the observations, that violates the independence assumption of traditional fixed-effects models.

There are still more challenges when analysing longitudinal data, however as it has been stated we could consider the cited as the most relevant. Moreover, as the complexity of the model increases, challenges to deal with the correlation structure of the defined model increase. The advantage of longitudinal models, however, is that all the available information from each subject can be used in the analysis, increasing both the statistical power and the ability to estimate subject-specific effects. Furthermore, the use of the full information decreases the bias due to arbitrary exclusion of subjects with incomplete response or the simple imputation of values for replacing missing responses.

### 3.1.2 Methodologies for analysing longitudinal data

During the last years, several modelling approaches have been developed by many authors in order to deal with the previous mentioned challenges associated with longitudinal models. In this section, we will enumerate the most widely used methodologies when analysing longitudinal data. We will also introduce them shortly, and we will define the advantages and disadvantages of each one.

*Reduction:* the reduction approach, referred as *derived variable* by some authors (Hedeker and Gibbons, 2006), is based on the reduction of the repeated mea-

surements into a summary variable. Indeed, once reduction has been performed, this approach is no longitudinal any more, since there is only one observation per subject. The main problem of this approach is that the uncertainty in the derived variable is proportional to the number of repeated measurements for which it was computed. In unbalanced cases each subject has a different number of measurements and hence different uncertainties. Consequently, the homoscedasticity assumption of the model is not ensured. Moreover, the reduction of the repeated measures decreases the number of observations and hence, there is a considerably loss of statistical power. Finally, the reduction of the outcome does not allow the inclusion of time-dependent variables in the model, as the temporal aspect of the data is removed.

*Analysis of variance:* the analysis of variance (ANOVA) for repeated measurements (Winer, 1971) is used to compare three or more group means where the participants are the same in each group. The model assumes compound symmetry which implies constant variances and covariances over time. This is an assumption that hardly will be held in longitudinal data for two different reasons. First, attrition, variances will increase over time because the number of people that response reduces. Second, it looks reasonable to assume that covariances for proximal measurements will be larger than covariances for distal measurements. The model allows a different trend line per subject, however, the trends only differ in the intercept, which implies that all the subjects behave equally over time. It looks more reasonable that subjects differ not only in the baseline, but also in the rate of change (slope) from the overall trend line.

*Multivariate analysis of variance:* the multivariate analysis of variance (MANOVA) was proposed for longitudinal data analysis by Bock (1985). The MANOVA model is simply an ANOVA model with several dependent variables, i.e., while the ANOVA model tests for differences in means between two or more groups, the MANOVA model tests differences in two or more vectors of means. This approach transforms the repeated measurements to orthogonal polynomial coefficients (e.g. constant, linear, quadratic growth rates), which are used as multivariate responses in the MANOVA model. The main disadvantage of this approach is that it does not deal with missing data, so all the subjects must have the same number of repeated measurements, which is very unlikely in practise.

*Mixed-effects models:* mixed-effects regression models are quite widely used in different frameworks, specially for the analysis of longitudinal data (Laird and Ware, 1982). For example, as it was introduced in Section 2.1.1 GLMMs are a general methodology that include random effects in the linear predictor of a GLM, which can easily accommodate the correlation structure of longitudinal data. We will develop these models in more detail in Section 3.2. Mixed-effects regression models include the term mixed-effects because they consists of a fixed component (regression coefficients) and a random component (random effects). Mixed-effects regression models are quite robust to missing data and irregularly spaced measurements, furthermore, they can easily deal with time-independent and time-dependent covariates.

*Generalised estimating equations:* The generalised estimating equations (GEEs) (Zeger and Liang, 1982) are a general alternative to mixed-effects models, which are computationally very convenient. GEE approach extends the classical GLMs (see Section 2.1.1) to the case of correlated data. They can be used to analyse a wide variety of outcomes and do not require complex numerical evaluation of the likelihood for nonlinear models. They model the overall mean relationship of the variables and the within-subject dependency separately. GEE models are also called marginal models, where the term marginal makes reference to the assumption that the mean response only depends on the covariates of interest and not on any random effects or previous responses.

Among the previously defined methodologies for analysing longitudinal data, the most widely used and appropriate include the mixed-effects regression approaches and GEE. The larger difference between these two approaches is that GEE models are based on quasi-likelihood estimation, and so the full likelihood of the data is not specified. Therefore, while GEE models are considered *partial-likelihood* methods, the mixed-effects models are considered *full-likelihood* methods as they use all the available data from each subject. The advantage of statistical models based on partial-likelihood is that they are computationally easier and generalise quite easily to different distribution forms of the repeated outcome variables. However, they are more restrictive in their assumption regarding missing data, limiting their applicability in some cases. Moreover, full-likelihood models provide subject-specific effects which are quite useful when analysing individual-within variability and when predicting future responses for a given subject or a group of subjects in hierarchical structures.



On the one hand, GEE approach calculates the marginal mean for each subject, even if some of those means have limited information due to subject drop out. Then standard errors are adjusted taking into account the correlation structure of the repeated measures over time and/or subject clustering. On the other hand, mixed effects regression approaches use all the available information to calculate subject-specific trends that would have been observed if the subjects had stayed until the end of the study. Hence, if future subject responses are related to previous measurements, both approaches can conclude quite different estimated mean responses at the end of the study. In fact, the main difference between both methodologies appears when the missing data are dependent on the previous observed responses for each subject. However, it is difficult to imagine that if the missing data for a given subject had been observed, the response would not have been related to previous measurements of the same subject. That is, GEE assume that the missing data are missing at random and do not depend on the previous measurements.

Therefore, in this thesis we will consider a mixed-effects approach as the most appropriate for the analysis of hierarchical or longitudinal PRO data. In fact, this chapter is based on the development of a mixed-effects model based on the beta-binomial distribution. To achieve that goal, in Section 3.2 we make a review of the existing literature describing the most used mixed-effects regression approaches. Then, in Section 3.3 we present the description of the model we propose, the development of an estimation and inference methodology and the comparison of its performance with similar approaches in the literature. In order to show the performance of our proposal and compare it with available methodology in the literature, a simulation study is carried out in Section 3.4. Finally, with the purpose of showing the applicability of the developed methodology, we apply it in both COPD Study and Paquid Research Programme described in Section 1.3.1 and Section 1.3.2 respectively. We finish the chapter providing some conclusions in Section 3.6.

## 3.2 Mixed-effects models in the literature

During the last years, many authors have described in the literature different approaches to deal with mixed-effects regression models in a longitudinal data framework: variance component models (Dempster et al., 1981), random-effects models (Laird and Ware, 1982), empirical Bayes models (Hui and Berger, 1983), random coefficient models (De Leeuw and Kreft, 1986), mixed models (Longford, 1987), random regression models (Gibbons et al., 1988), two-stage models (Bock, 1989),

multilevel models (Goldstein, 1995) and hierarchical linear models (Raubenbush and Bryk, 2002). In fact, all the previous mentioned approaches have a common characteristic, they accommodate the dependence of the repeated measurements by the inclusion of random effects per subject. These random subject effects describe subject-specific trends over time and measure the correlation structure of the data. In addition, they calculate the degree of subject variation that exists in the population of subjects. Apart from that, as it has been mentioned in Section 3.1, mixed-effects model are very useful for analysing longitudinal data, basically because they are very flexible with missing data, allowing the inclusion of subjects with incomplete data in the model. Consequently, compared to procedures where complete data is necessary, the use of all the available information increases the statistical power of the model and decreases the bias estimation, as subjects with complete data may not be a representative set of all the population. Additionally, studies where the follow-up times are not uniform can be incorporated in the model, as mixed-models assume the time as a continuous variable, and hence, subjects' responses do not have to be measured at the same time points.

During the last years many regression models have been proposed for analysing various types of data, where, in most cases, the underlying mean structure is still the linear model defined as

$$h(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta},$$

where  $h(\cdot)$  is defined as a monotonic link function,  $\mathbf{X}$  is a full rank matrix composed by the covariates and  $\boldsymbol{\beta}$  is a fixed parameter vector of length  $p + 1$ , being  $p$  typically small.

We have mentioned that the use of mixed-effects models arises when the independence assumption of the observed measurements fail. In the next section we will describe in more detail the scenarios where the essential independence assumption of the observations fail. For instance, assume that  $Y_{ij}$  is the random variable associated with  $j$ th observation of the  $i$ th subject. We will consider that the number of measurements change from subject to subject, so for each subject  $i$ ,  $i = 1, \dots, n$ , we denote the number of observations as  $t_i$ . Consequently, we can determine each  $Y_{ij}$  as

$$Y_{ij} = \mu + u_i + e_{ij}, \tag{3.1}$$

where  $\mu$  is the fixed overall mean parameter,  $u_i$  is the subject effect that determines the mean for subject  $i$ , and  $e_{ij}$  is the residual effect that accounts for the error of each  $j$  observation of the  $i$ th subject,  $i = 1, \dots, n$ ,  $j = 1, \dots, t_i$ .

The random effects are usually assumed Gaussian with mean zero and known variance structure, however, other types of distributions could be considered. For instance, as it was deeply discussed in Section 2.2.2, Lee and Nelder (1996) developed the HGLMs, where the random effects can follow any distribution conjugate to the distribution of the conditioned dependent variable. Nevertheless, in this section we will only focus on the usual mixed-effects models, where the random effects are assumed normal with mean zero.

Hence, we assume that the subject specific random effects  $u_i$  are iid  $u_i \sim \mathcal{N}(0, \sigma_u^2)$ , and that the residual effects  $e_{ij}$  are iid  $e_{ij} \sim \mathcal{N}(0, \sigma^2)$ , being independent random effects. Additionally, we assume that observations from the same subjects are correlated. Given the model in Equation (3.1), the covariance between two different measurements of the same subject is determined by

$$\begin{aligned} \mathbb{Cov}(Y_{ij}, Y_{ik}) &= \mathbb{Cov}(\mu + u_i + e_{ij}, \mu + u_i + e_{ik}) \\ &= \mathbb{Cov}(u_i + e_{ij}, u_i + e_{ik}) \\ &= \mathbb{Cov}(u_i, u_i) + \mathbb{Cov}(u_i, e_{ik}) + \mathbb{Cov}(e_{ij}, u_i) + \mathbb{Cov}(e_{ij}, e_{ik}) \\ &= \mathbb{Var}(u_i) = \sigma_u^2, \end{aligned}$$

for  $i = 1, \dots, n$  and  $j, k = 1, \dots, t_i$ , being  $k \neq j$ . Consequently, by the definition in Equation (3.1), we have determined a correlation structure between different observations for the same subject, which matches with the dependence assumption of the data. Hence, the correlation between different observations of the same subject is defined as,

$$\mathbb{Corr}(Y_{ij}, Y_{ik}) = \frac{\mathbb{Cov}(Y_{ij}, Y_{ik})}{\sqrt{\mathbb{Var}(Y_{ij})\mathbb{Var}(Y_{ik})}} = \frac{\sigma_u^2}{\sigma_u^2 + \sigma^2}.$$

In summary, we have shown that the inclusion of subject-specific random effects not only allows subjects to differ from the overall mean, but also accommodates the dependence structure of different observations of the same individual.

It is straightforward to define the model in Equation (3.1) as

$$\mathbf{Y}_i = \mu \mathbf{1}_{t_i} + \mathbf{v}_i,$$

where  $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{it_i})'$  is the vector of the measurements of individual  $i$ ,  $\mathbf{1}_{t_i}$  is the vector of 1s of length  $t_i$  and  $\mathbf{v}_i$  are the random effects, which follow a multivariate

normal distribution with mean 0 and variance-covariance matrix

$$\mathbf{D} = \sigma^2 \mathbf{I}_{t_i} + \sigma_u^2 \mathbf{J}_{t_i},$$

where  $\mathbf{I}_{t_i}$  is a  $t_i \times t_i$  identity matrix and  $\mathbf{J}_{t_i}$  is a  $t_i \times t_i$  matrix of 1s. Different variance-covariance matrices can be chosen depending on the dependence structure of the data, such as uncorrelated subjects, uncorrelated between and within subjects, uncorrelated between and autocorrelated within subjects, correlated between but not within subjects, etc (McCulloch and Searle, 2001).

In the next section, we will discuss on when a covariate effect should be considered as fixed or as random. Then, we will describe the most commonly used mixed-effects model approaches: the linear mixed model (LMM), where the dependent variable follows a normal distribution, and, the already mentioned GLMM for dependent variables following any distribution from the exponential family.

### 3.2.1 Fixed or random effects?

When fitting a statistical model, assumptions must be considered in order to the model be appropriate. When we are interested in applying a regression model to a dataset, one of the most important decisions we have to make is if a covariate that we want to introduce in the model is going to be considered as fixed or, conversely, it is going to be assumed as a random sample of a random distribution. We have already mentioned that in longitudinal studies, due to heterogeneity, i.e. the lack of independence in the repeated measurements, a random effect is usually considered for each subject. However, apart from the correlation between repeated responses of the same subject, there could be other reasons why an additional covariate should be assumed as random. As it was pointed out in the previous section, subjects could be clustered in grouping units (schools, clinics, hospitals, etc.), and consequently, the data could present a correlation structure due to hierarchies. We present two examples in order to understand when a subject clustering factor should be considered as fixed or random.

#### **Example 3.1: Weight loss by pills**

Suppose we have a clinical trial in which five pills are administered to individuals trying to reduce their weight. Assume that  $Y_{ij}$  is the random variable associated with the weight reduction for  $j$ th person receiving the pill  $i$ . We

can consider that

$$\mathbb{E}[Y_{ij}] = \mu_i = \mu + \alpha_i, \quad (3.2)$$

where  $\mu$  is a general mean and  $\alpha_i$  is the effect on the weight reduction due to pill  $i$ ,  $i = 1 \dots, 5$ .

In this modelling of the expected value of  $Y_{ij}$ , the  $\mu_i$ 's (and  $\alpha_i$ 's) are fixed effects because the five pills used in the clinical trial are the only pills being studied. They are the only five being used, and in using them there is no thought for any other pill. This is the concept of fixed effects. We consider the pills being used and no others, and so the effects are called *fixed effects*, and the model, *fixed effects model*.

□

### Example 3.2: Nutritional centres

Assume that we choose a pill, anyone, for losing weight from Example 3.1 and develop a clinical trial where the pill is administered to patients from 7 random nutritional centres in the Basque Country. The model for the random variable  $Y_{ij}$ , which represents the weight loss for patient  $j$  at the  $i$ th nutritional centre, would be

$$\mathbb{E}[Y_{ij}] = \mu + u_i \quad (3.3)$$

with  $i = 1, \dots, 7$  for the 7 nutritional centres. Notice that it is reasonable to think of those centres as a random sample of centres from some distribution of nutritional centres, maybe all the centres in the Basque Country.

Notice that Equation (3.3) is essentially the same algebraically as Equation (3.2), except for having  $u_i$  in place of  $\alpha_i$ . However, the underlying model assumptions are different. On the one hand, in Equation (3.2) the effect of pill  $i$  in the loss weight,  $\alpha_i$ , is considered a fixed effect as it was already decided that it was a pill of interest in the study. On the other hand, in Equation (3.3) each  $u_i$  is the effect of weight loss of patients being observed in nutritional centre  $i$ , but  $i$  centre is not a pre-selected centre, it is just one centre from among all the nutritional centres in the Basque Country that has been numbered in the trial as centre  $i$ . Consequently, nutritional centres in the study have been randomly selected with the object of treating them as a representation of all the nutritional centres in the Basque Country. Hence, inference and results from the study can be made about all the population. In fact, the main feature

of random effects is that they can be used as the basis for making inferences about populations from which they were selected. Therefore,  $u_i$  is called a *random effect* and the model in Equation (3.3) is considered a *mixed-effects model*, as it combines fixed-effects,  $\mu$ , and random-effects,  $u_i$ . As result, the model in Equation (3.3) allows the inference about the variance of the random effects, which measures the variation among nutritional centres. Furthermore, they allow the prediction of centres that are likely to have the best weight reduction.

□

We have shown two examples where the decision of whether certain effects are fixed or random was quite obvious, but there are some cases that it is not. The context the data was collected and the approach that is given to a covariate in the model are determinant so as to decide if the covariate should be treated as fixed or random. The main question when deciding the behaviour of a factor in the model is if it is reasonable to consider the levels of the factor as a random sample from a population which have a distribution. If the answer is yes, then the effects should be considered as random, otherwise, if it is not, then they are fixed.

### 3.2.2 Linear mixed-effects model

In this subsection we will study the LMMs whose definition allows the extension to non-Gaussian models. Moreover, theoretical results are 'clean' in the sense that estimation through approximated likelihood matches with the estimation through the marginal likelihood. Basically, the LMMs are the generalization of the LMs to the inclusion of random effects. The standard LMM specifies that

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}, \quad (3.4)$$

where  $\mathbf{Y}$  is a  $n \times 1$  vector of the random dependent variable,  $\mathbf{X}$  and  $\mathbf{Z}$  are  $n \times p$  and  $n \times q$  known design matrices,  $\boldsymbol{\beta}$  is a  $p \times 1$  vector of the fixed effects,  $\mathbf{u}$  is a  $q \times 1$  vector of the random effects and  $\mathbf{e}$  corresponds to the  $n \times 1$  vector of error terms. The model assumes that  $\mathbf{u}$  and  $\mathbf{e}$  are independent and normally distributed.

Without loss of generality we can assume that the expectation of the random effects is  $\mathbf{0}$  as, if it were otherwise,  $\mathbb{E}[\mathbf{u}] = \boldsymbol{\tau}$ , we could rewrite Equation (3.4) as  $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\tau} + \mathbf{Z}(\mathbf{u} - \boldsymbol{\tau})$ . Defining  $\mathbf{X}^* = [\mathbf{X} \ \mathbf{Z}]$ ,  $\boldsymbol{\beta}^* = (\boldsymbol{\beta}', \boldsymbol{\tau}')'$  and  $\mathbf{u}^* = (\mathbf{u} - \boldsymbol{\tau})$ ,

we obtain an equivalent formulation of the model as  $\mathbf{Y} = \mathbf{X}^*\boldsymbol{\beta}^* + \mathbf{Z}\mathbf{u}^* + \mathbf{e}$  where  $\mathbb{E}[\mathbf{u}^*] = \mathbf{0}$ .

Although the elements of  $\mathbf{u}$  are random variables, it is useful to define the model conditional on their unobservable but realised values. This specification of the model allows an immediate extension to non-normal models. We assume that conditional on the unobservable random effects  $\mathbf{u}$ , the dependent outcome variable is normally distributed with expectation equal to

$$\mathbb{E}[\mathbf{Y}|\mathbf{u}] = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}, \quad (3.5)$$

where  $\mathbb{E}[\mathbf{Y}|\mathbf{u}]$  means  $\mathbb{E}[\mathbf{Y}|\mathbf{U} = \mathbf{u}]$ , being  $\mathbf{U}$  the random variable and  $\mathbf{u}$  the realizations of the random variable.

Let's assume that the variance-covariance matrices of the random effects  $\mathbf{D}$  and error term  $\boldsymbol{\Sigma}$  are parametrized by an unknown vector of variance components  $\boldsymbol{\theta}$ . Hence, the conditional variance of the measured outcome variable is defined as  $\text{Var}[\mathbf{Y}|\mathbf{u}] = \boldsymbol{\Sigma}$ . Therefore, we assume that conditional on the unobservable random effects  $\mathbf{u}$ , the outcome variable is drawn from the following distribution

$$\mathbf{Y}|\mathbf{u} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}, \boldsymbol{\Sigma}), \quad (3.6)$$

where  $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{D})$  and  $\boldsymbol{\Sigma}$  is a matrix parametrized by the unknown parameter vector  $\boldsymbol{\theta}$ . The advantage of this new definition in Equation (3.5) is that we do not have to specify an error term, which does not make sense in non-linear models. Following we present the parameter estimation process for LMMs which is very useful for future developments of more complicated models.

First, from Equation (3.6) we conclude that the marginal expectation and variance of the outcome variable are defined as

$$\begin{aligned} \mathbb{E}[\mathbf{Y}] &= \mathbb{E}[\mathbb{E}[\mathbf{Y}|\mathbf{u}]] = \mathbb{E}[\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}] = \mathbf{X}\boldsymbol{\beta} \\ \text{Var}[\mathbf{Y}] &= \text{Var}[\mathbb{E}[\mathbf{Y}|\mathbf{u}]] + \mathbb{E}[\text{Var}[\mathbf{Y}|\mathbf{u}]] \\ &= \text{Var}[\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}] + \mathbb{E}[\boldsymbol{\Sigma}] \\ &= \mathbf{Z}\mathbf{D}\mathbf{Z}' + \boldsymbol{\Sigma} = \mathbf{V}. \end{aligned} \quad (3.7)$$

Notice, that the fixed effects enter only through the mean, while the random effects design and variance-covariance matrices enter only through the marginal variance of the outcome variable.

Based on the first and second order moments defined in Equation (3.7), the marginal log-likelihood of the model is defined as

$$\begin{aligned}\log L(\boldsymbol{\beta}, \boldsymbol{\theta} | \mathbf{y}) &= \log f(\mathbf{y} | \boldsymbol{\beta}, \boldsymbol{\theta}) \\ &= -\frac{1}{2} \log |\mathbf{V}| - \frac{1}{2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}),\end{aligned}\tag{3.8}$$

where  $\boldsymbol{\theta}$  enters in the equation through the marginal variance  $\mathbf{V}$  and  $\mathbf{y} = (y_1, \dots, y_n)'$  is the observed vector for the outcome variables.

For fixed  $\boldsymbol{\theta}$ , if we derive the marginal log-likelihood with respect to  $\boldsymbol{\beta}$ , we obtain the score equations for the fixed effects of the model as

$$S(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} \log L(\boldsymbol{\beta}, \boldsymbol{\theta} | \mathbf{y}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' \mathbf{V}^{-1} \mathbf{X}.\tag{3.9}$$

Matching Equation (3.9) equal to zero, we reach the well-known generalised or weighted least-squares formula, from which we get the maximum likelihood estimation of the fixed effects in the model,

$$(\mathbf{X}' \mathbf{V}^{-1} \mathbf{X}) \boldsymbol{\beta} = \mathbf{X}' \mathbf{V}^{-1} \mathbf{y}.\tag{3.10}$$

We obtain the standard errors of the estimates of the fixed effects by the negative of the inverse of the Fisher information matrix, which in LMMs is defined as

$$I(\hat{\boldsymbol{\beta}}) = \mathbf{X}' \mathbf{V}^{-1} \mathbf{X}.$$

Sometimes, specially when observations are grouped by a specific characteristic, the interest of the analysis is not only to estimate the fixed coefficients, but also to predict the realized values of the random effects. Note that assumptions concerning random and fixed effects are completely different. While the fixed effects are considered constant values, the random effects are considered effects that come from a larger population of effects. Consequently, the way of dealing with the two kind of effects should not be the same. Hence, to emphasize this distinctions, we would say that fixed effects are estimated, but we would use the prediction for the obtained realized values of the random effects. The problem is that, in these situations, the marginal log-likelihood function in Equation (3.8) does not offer any information concerning the random effects.

The log-likelihood of all the parameters in the model is based on the joint density



of the dependent outcome variable and random effects, i.e.

$$L(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{u}|\mathbf{y}) = f_{y|u}(\mathbf{y}|\mathbf{u})f_u(\mathbf{u}),$$

where the first term is the conditional distribution of the outcome variable, which from Equation (3.6) we know that follows a normal distribution, and the second term corresponds to the distribution of the random effects, which by model assumption is considered normal. Hence, we can express the joint log-likelihood of the parameters as,

$$\begin{aligned} \log L(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{u}|\mathbf{y}) &= -\frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})'\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}) \\ &\quad - \frac{1}{2}\log|\mathbf{D}| - \frac{1}{2}\mathbf{u}'\mathbf{D}^{-1}\mathbf{u}. \end{aligned} \quad (3.11)$$

Again, for fixed  $\boldsymbol{\theta}$ , given the fixed effects parameters  $\boldsymbol{\beta}$ , we take the derivative of the log-likelihood with respect to the random effects and obtain the score equation for  $\mathbf{u}$  as

$$S(\mathbf{u}) = \frac{\partial}{\partial \mathbf{u}} \log L(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{u}|\mathbf{y}) = \mathbf{Z}'\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}) - \mathbf{D}^{-1}\mathbf{u}.$$

Setting it to zero, the prediction of the random effects is the solution of

$$(\mathbf{Z}'\boldsymbol{\Sigma}^{-1}\mathbf{Z} + \mathbf{D}^{-1})\mathbf{u} = \mathbf{Z}'\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \quad (3.12)$$

Equation (3.12) offers the so-called best linear unbiased predictor (BLUP) of the random effects, best in the sense of minimized mean square error of prediction (Robinson, 1991).

Notice that the maximum likelihood estimates of the fixed effects in Equation (3.10) and the BLUP of the random effects in Equation (3.12), are the joint maximiser of the joint log-likelihood in Equation (3.11) (Pawitan, 2001). This provides a useful technique when dealing with non-normal models, where the computation of the marginal likelihood is not so direct.

Combining both score equations for  $\boldsymbol{\beta}$  and  $\mathbf{u}$  from the joint log-likelihood in Equation (3.11), we have that

$$\begin{pmatrix} \mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X} & \mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{Z} \\ \mathbf{Z}'\boldsymbol{\Sigma}^{-1}\mathbf{X} & \mathbf{Z}'\boldsymbol{\Sigma}^{-1}\mathbf{Z} + \mathbf{D}^{-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \mathbf{u} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{y} \\ \mathbf{Z}'\boldsymbol{\Sigma}^{-1}\mathbf{y} \end{pmatrix}. \quad (3.13)$$

The advantage of this formulation is that we can estimate fixed effects and predict random effects without having to compute the marginal variance  $\mathbf{V}$  or its inverse. The estimating procedure, which is called Jacobi or Gauss-Seidel method in linear algebra, also known as the iterative backfitting algorithm in statistics, is fully explained in Breiman and Friedman (1985).

Until now, we have considered the variance parameters fixed in the estimation of the fixed and random effects in the model. However, following, we develop the estimation algorithm of the variance parameter  $\boldsymbol{\theta}$ . From the marginal log-likelihood in Equation (3.8) we can derive the profile log-likelihood of the parameter vector  $\boldsymbol{\theta}$  as

$$\log L(\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\beta}) = -\frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})' \mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \quad (3.14)$$

where  $\hat{\boldsymbol{\beta}}$  is the maximum likelihood estimation of the fixed effects obtained from either the marginal score equation in Equation (3.10) or the joint estimation procedure in Equation (3.13). The disadvantage of this formulation of the profile log-likelihood for  $\boldsymbol{\theta}$  is that we have to compute the marginal variance  $\mathbf{V}$ , but specially its inverse. However, an alternative formulation has been developed in the literature which is based on the decomposition of all the terms involving the variance-covariance matrix  $\mathbf{V}$ . By Woodbury formula in Property D.2 and a partitioned matrix result in Appendix D.1, the profile log-likelihood in Equation (3.14) is reduced to

$$\begin{aligned} \log L(\boldsymbol{\theta}) &= -\frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{Z}\hat{\mathbf{u}})' \boldsymbol{\Sigma}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{Z}\hat{\mathbf{u}}) \\ &\quad - \frac{1}{2}\log|\mathbf{D}| - \frac{1}{2}\hat{\mathbf{u}}' \mathbf{D}^{-1}\hat{\mathbf{u}} - \frac{1}{2}\log|\mathbf{Z}'\boldsymbol{\Sigma}^{-1}\mathbf{Z} + \mathbf{D}^{-1}| \\ &= \log L(\hat{\boldsymbol{\beta}}, \boldsymbol{\theta}, \hat{\mathbf{u}}) - \frac{1}{2}\log|\mathbf{Z}'\boldsymbol{\Sigma}^{-1}\mathbf{Z} + \mathbf{D}^{-1}|, \end{aligned} \quad (3.15)$$

where  $\boldsymbol{\theta}$  enters in the function through  $\boldsymbol{\Sigma}$ ,  $\mathbf{D}$ ,  $\hat{\boldsymbol{\beta}}$  and  $\hat{\mathbf{u}}$  (see Pawitan (2001) for more details). Again,  $\hat{\boldsymbol{\beta}}$  and  $\hat{\mathbf{u}}$  are the maximum likelihood joint estimates in Equation (3.13). It is important to note that, the profile log-likelihood of the variance parameters in Equation (3.15), also called *modified profile log-likelihood*, is equal to the joint-likelihood function in Equation (3.11) plus the additional term

$$-\frac{1}{2}\log|\mathbf{Z}'\boldsymbol{\Sigma}^{-1}\mathbf{Z} + \mathbf{D}^{-1}|,$$

where  $\mathbf{Z}'\boldsymbol{\Sigma}^{-1}\mathbf{Z} + \mathbf{D}^{-1}$  corresponds with the Fisher information matrix of the random effects, i.e. the negative of the second derivative of the joint log-likelihood in

Equation (3.11) with respect to  $\mathbf{u}$ . A full inference procedure can be carried out for the dispersion components  $\boldsymbol{\theta}$  through the modified profile log-likelihood. Furthermore, the modified profile log-likelihood is much easier to manipulate than likelihood functions involving the term  $\mathbf{V}$  or its inverse.

Generalisation and extension to several random effects is straightforward, as if the model were defined as

$$\mathbb{E}[\mathbf{Y}|\mathbf{u}_1, \mathbf{u}_2] = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\mathbf{u}_1 + \mathbf{Z}_2\mathbf{u}_2,$$

we could define the random effects as  $\mathbf{u} = (\mathbf{u}'_1, \mathbf{u}'_2)'$  and the random effects design matrix as  $\mathbf{Z} = [\mathbf{Z}_1 \ \mathbf{Z}_2]$ . Hence, the previous model will reduce to

$$\mathbb{E}[\mathbf{Y}|\mathbf{u}] = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}$$

and all the previous theory would be now applicable.

### 3.2.3 Generalised linear mixed-effects model

The use of a random structure is not only restricted to linear models, as we have explained in the previous section. Sometimes, it is useful to incorporate random effects into non-linear models to accommodate correlated data, or to consider levels of a factor as selected from a population of levels (hierarchies) in order to make inference to that population. GLMMs extend the previously developed LMMs to non-linear models and the classical GLMs (see Section 2.1.1) to the inclusion of random effects (McCulloch and Searle, 2001).

The GLMM model, like the LMM model, assumes that the outcome response vector  $\mathbf{Y}$  consists of conditionally independent elements, i.e. given a vector of random effects  $\mathbf{u}$ , the dependent outcomes  $Y_1, \dots, Y_n$  are conditionally independent, each with a distribution density from the exponential family,

$$\begin{aligned} Y_i|\mathbf{u} &\sim f_{y_i|\mathbf{u}}, \text{ indep. } i = 1, \dots, n \\ f_{y_i|\mathbf{u}} &= \exp\{[y_i\psi_i - A(\psi_i)]/\phi - c(y_i, \phi)\}. \end{aligned} \tag{3.16}$$

where  $y_i$  is the observation that corresponds to the random variable  $Y_i$ .

From the exponential family theory we know that the conditional expectation of the outcomes, which we denote  $\mu_i$ , is related to  $\psi_i$  via the identity  $\mu_i = \partial A(\psi_i)/\partial \psi_i$ . Indeed, it is the transformation of this mean what we want to model as a linear

function of both fixed and random effects,

$$\begin{aligned}\mathbb{E}[Y_i|\mathbf{u}] &= \mu_i, \\ h(\mu_i) &= \mathbf{x}'_i\boldsymbol{\beta} + \mathbf{z}'_i\mathbf{u},\end{aligned}$$

where  $h(\cdot)$  is a known link function,  $\mathbf{x}'_i$  is the  $i$ th row of the full rank model matrix  $\mathbf{X}$  for the fixed effects,  $\boldsymbol{\beta}$  are coefficients of the fixed effects,  $\mathbf{z}'_i$  is the  $i$ th row of the model matrix  $\mathbf{Z}$  for the random effects and  $\mathbf{u}$  are the random effects. We complete the model specification assigning a distribution to the random effects, which is usually assumed normal with mean  $\mathbf{0}$  and variance-covariance matrix  $\mathbf{D}$ , i.e.  $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{D})$ . Finally, we assume that the variance-covariance matrix of the random effects  $\mathbf{D}$  is specified by a vector of parameters  $\boldsymbol{\lambda}$ .

Let be  $\boldsymbol{\theta}$  defined as a unknown vector of variance components of the model which consists of the dispersion parameter of the exponential family distribution,  $\phi$ , and the parameters needed for specifying the variance-covariance matrix of the random effect,  $\boldsymbol{\lambda}$ .

In LMMs, we have shown in Equation (3.7) that the marginal expectation of the response variable is modelled in the same way as in LMs, i.e.  $\mathbb{E}[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta}$ . Consequently, as it has been mentioned, in LMMs the estimation of the fixed effects by the marginal model matches with the conditional model estimation (joint estimation). However, due to the fact that the model specification in GLMMs is made conditional on the value of  $\mathbf{u}$ , and not in a marginal way, some aspects must be taken into account. In this case, the marginal expectation of  $Y_i$  is defined as

$$\mathbb{E}[Y_i] = \mathbb{E}[\mathbb{E}[Y_i|\mathbf{u}]] = \mathbb{E}[\mu_i] = \mathbb{E}[h^{-1}(\mathbf{x}'_i\boldsymbol{\beta} + \mathbf{z}'_i\mathbf{u})],$$

which, in general, cannot be simplified due to the non-linear function  $h^{-1}(\cdot)$ . Moreover, the marginal variance of  $Y_i$  is computed as

$$\begin{aligned}\text{Var}[Y_i] &= \text{Var}[\mathbb{E}[Y_i|\mathbf{u}]] + \mathbb{E}[\text{Var}[Y_i|\mathbf{u}]] \\ &= \text{Var}[\mu_i] + \mathbb{E}[\phi v(\mu_i)] \\ &= \text{Var}[h^{-1}(\mathbf{x}'_i\boldsymbol{\beta} + \mathbf{z}'_i\mathbf{u})] + \mathbb{E}[\phi v(h^{-1}(\mathbf{x}'_i\boldsymbol{\beta} + \mathbf{z}'_i\mathbf{u}))],\end{aligned}$$

where  $\text{Var}[Y_i|u] = \phi v(\mu_i)$  is the variance of an exponential family distribution (see Section 1.4.1). Again, the marginal variance cannot be simplified due to the non-linear function  $h^{-1}(\cdot)$ . Similar to LMMs, the introduction of random effects in the linear predictor of a GLMM also defines a correlation structure between observations

which have any random effect in common. Namely, for  $i, j = 1, \dots, n$  being  $i \neq j$ , we assume that

$$\begin{aligned} \text{Cov}(Y_i, Y_j) &= \text{Cov}(\mathbb{E}[Y_i|\mathbf{u}], \mathbb{E}[Y_j|\mathbf{u}]) + \mathbb{E}[\text{Cov}(Y_i, Y_j|\mathbf{u})] \\ &= \text{Cov}(\mu_i, \mu_j) + \mathbb{E}[0] \\ &= \text{Cov}(h^{-1}(\mathbf{x}'_i\boldsymbol{\beta} + \mathbf{z}'_i\mathbf{u}), h^{-1}(\mathbf{x}'_j\boldsymbol{\beta} + \mathbf{z}'_j\mathbf{u})), \end{aligned}$$

which cannot be implicitly solved due to non-linear terms in the formula.

Based on GLMMs specification in Equation (3.16), we can derive the marginal likelihood of the model as

$$L(\boldsymbol{\beta}, \boldsymbol{\theta}|\mathbf{y}) = \int_{\mathbb{R}^q} \prod_{i=1}^n f_{y_i|u}(y_i|\mathbf{u}, \boldsymbol{\theta}) f_u(\mathbf{u}|\boldsymbol{\theta}) d\mathbf{u} \quad (3.17)$$

where  $f_{y_i|u}(y_i|\mathbf{u}, \boldsymbol{\theta})$  is the exponential family distribution of the conditional outcomes,  $f_u(\mathbf{u}|\boldsymbol{\theta})$  is the normal density function and  $y_i$  is the observation of the outcome variable  $Y_i$ . Notice that the integration is over the  $q$ -dimensional distribution of  $\mathbf{u}$ .

While in LMMs the marginal likelihood in Equation (3.8) is relatively simple, the general case is notoriously difficult. In general, there is no closed form solution for integrating the random effects out. Hence, numerical methods, such as Gauss-Hermite quadrature (McCulloch and Searle, 2001), must be applied. However, although the numerical approximation of the marginal likelihood works relatively well in some cases (e.g. independent random effects), for more complicated structures (e.g. correlated random effects) the exact approach is no longer tractable.

The basis of the likelihood approximation in GLMMs is the extended or joint likelihood approach (Pawitan, 2001), which it is defined as

$$\log L(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{u}) = \log f_{\mathbf{y}|u}(\mathbf{y}|\mathbf{u}, \boldsymbol{\theta}) + \log f_u(\mathbf{u}|\boldsymbol{\theta}), \quad (3.18)$$

where, as before,  $\log f_{\mathbf{y}|u}(\mathbf{y}|\mathbf{u}, \boldsymbol{\theta})$  corresponds to the exponential family distribution of the conditional outcome vector and  $\log f_u(\mathbf{u}|\boldsymbol{\theta})$  is the normal density function of the random effects. This approximated log-likelihood is generally much easier to evaluate and optimize than the marginal likelihood presented in Equation (3.17). For fixed  $\boldsymbol{\theta}$ , unlike in LMMs, in GLMMs estimations through the extended likelihood are not exactly equal to those through the marginal likelihood. However, Lee and Nelder (1996) showed that under fairly general conditions the two estimates are

asymptotically close.

For a fixed value of  $\boldsymbol{\theta}$ , similar to GLMs, we can use a quadratic approximation of the conditional log-density function  $\log f_{y|u}(\mathbf{y}|\mathbf{u})$  to normalize the log-likelihood form and derive an iterative weighted least squares algorithm for performing the estimation (see Appendix B.2). Hence, given some initial values of the fixed and random effects,  $\boldsymbol{\beta}^0$  and  $\mathbf{u}^0$ , the conditional log-likelihood of the outcomes in Equation (3.18) can be approximated by

$$-\frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}(\mathbf{y}^w - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})' \boldsymbol{\Sigma}^{-1} (\mathbf{y}^w - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}),$$

where  $\mathbf{y}^w$  is defined as the working vector with elements

$$y_i^w = \mathbf{x}'_i \boldsymbol{\beta}^0 + \mathbf{z}'_i \mathbf{u}^0 + \frac{\partial h}{\partial \mu_i}(y_i - \mu_i^0) \quad \text{where} \quad \mu_i^0 = h^{-1}(\mathbf{x}'_i \boldsymbol{\beta}^0 + \mathbf{z}'_i \mathbf{u}^0), \quad (3.19)$$

and  $\boldsymbol{\Sigma}$  is a diagonal matrix of the variance of the working vector defined as

$$\Sigma_{ii} = \left( \frac{\partial h}{\partial \mu_i} \right)^2 \phi v_i(\mu_i^0), \quad (3.20)$$

where  $\phi v_i(\mu_i^0)$  is the conditional variance of  $Y_i$  given  $\mathbf{u}$ ,  $i = 1, \dots, n$ . Notice that the derivatives  $\partial h / \partial \mu_i$  are also evaluated at the current values of  $\boldsymbol{\beta}$  and  $\mathbf{u}$ .

Hence, we can derive an approximation of the extended log-likelihood of GLMMs in Equation (3.18) as

$$\begin{aligned} \log L(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{u}) \approx & -\frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}(\mathbf{y}^w - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})' \boldsymbol{\Sigma}^{-1} (\mathbf{y}^w - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}) \\ & - \frac{1}{2}\log|\mathbf{D}| - \frac{1}{2}\mathbf{u}' \mathbf{D}^{-1} \mathbf{u}, \end{aligned} \quad (3.21)$$

which leads to the usual log-likelihood of a LMM shown in Equation (3.11). Consequently, based on LMM's joint estimation development, we can conclude the following estimation equations,

$$\begin{pmatrix} \mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X} & \mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{Z} \\ \mathbf{Z}'\boldsymbol{\Sigma}^{-1}\mathbf{X} & \mathbf{Z}'\boldsymbol{\Sigma}^{-1}\mathbf{Z} + \mathbf{D}^{-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \mathbf{u} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{y}^w \\ \mathbf{Z}'\boldsymbol{\Sigma}^{-1}\mathbf{y}^w \end{pmatrix},$$

where, in this case,  $\mathbf{y}^w$  is the working vector defined in Equation (3.19) through the normalization of the conditional log-likelihood of the responses and  $\boldsymbol{\Sigma}$  is the variance matrix of the working vector defined in Equation (3.19), which needs to be

recomputed in each iteration of the estimation procedure. Therefore, notice that the computation of the estimates of the fixed and random effects in GLMMs requires repeated applications of norm-based formulas.

We have shown that even in LMMs the computation of the profile log-likelihood of the variance components  $\boldsymbol{\theta}$  cannot be done directly from the joint or extended log-likelihood owing to the bias of the estimates of fixed and random effects. It is shown in Equation (3.15) that the profile log-likelihood of the variance components is equivalent to a modification of the joint log-likelihood. The approximation methods of variance components estimation in GLMMs can be performed in different ways (Pawitan, 2001), however one of the most widely used technique is to estimate  $\boldsymbol{\theta}$  by maximising a modified profile log-likelihood similarly as in the normal case. Breslow and Clayton (1993) developed a restricted likelihood approach for performing estimation of variance components in GLMMs based on a heuristic justification in terms of Laplace's integral approximation. The method consists of a penalisation of the extended likelihood defined in Equation (3.21) as

$$\log L(\boldsymbol{\theta}) = \log L(\hat{\boldsymbol{\beta}}, \boldsymbol{\theta}, \hat{\boldsymbol{u}}) - \frac{1}{2} \log |\mathbf{Z}' \boldsymbol{\Sigma}^{-1} \mathbf{Z} + \mathbf{D}^{-1}|,$$

where  $\hat{\boldsymbol{\beta}}$  and  $\hat{\boldsymbol{u}}$  are the computed joint estimations for fixed  $\boldsymbol{\theta}$ . Notice that the formula corresponds to the penalisation of the joint likelihood in LMMs. However there are some differences. For instance, in the normal case,  $\boldsymbol{\Sigma}$  is typically a simple function of a variance component, but in GLMMs it is also a function of  $\boldsymbol{\mu}$ , and, consequently, of  $\boldsymbol{\beta}$  and  $\boldsymbol{u}$ . Therefore, since  $\boldsymbol{\mu}$  is unknown it is convenient to compute  $\boldsymbol{\Sigma}$  using  $\hat{\boldsymbol{\beta}}$  and  $\hat{\boldsymbol{u}}$ . It has been shown in many examples that the estimation procedure provides solutions that are close to the exact marginal estimates (Pawitan, 2001).

### 3.3 Beta-binomial mixed-effects model

The objective of this section is to extend the BBreg approach presented in Chapter 2 allowing for the inclusion of random effects in the linear predictor. As it has been stated though the thesis, the beta-binomial distribution does not belong to the exponential family and, consequently, GLMM (McCulloch and Searle, 2001), or even, HGLM (Lee and Nelder, 1996), theory cannot be directly applied in this framework.

In the next sections, we will define a mixed-effects regression model based on the beta-binomial distribution. Additionally, we will develop an estimation and

inference procedure for all the parameters involved in the model. However, not all the parameters in the model share the same characteristics, and hence, care must be taken when estimating them, considering different approaches. Moreover, we will provide a goodness-of-fit test based on the deviance. Finally, we will present some approaches that fit similar models in the literature and we will compare them with our proposal.

### 3.3.1 Model definition

Let  $\mathbf{Y} = (Y_1, \dots, Y_n)'$  be a set of random outcome variables of length  $n$ , which conditioned on the random effects  $\mathbf{u}$ , are assumed to be iid drawn from a beta-binomial distribution. To complete the model specification, as in LMMs and GLMMs, we assume that the random effects  $\mathbf{u}$  follow a multivariate normal distribution with mean  $\mathbf{0}$  and variance-covariance matrix  $\mathbf{D}$ , which depends on a vector of parameters  $\boldsymbol{\lambda}$  and it is non-singular. Summarizing, we assume that we have

$$Y_i | \mathbf{u} \sim \text{BB}(m_i, p_i, \phi), \quad i = 1, \dots, n, \quad (3.22)$$

where  $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{D})$ . In the same way as in GLMMs, we define the parameter vector  $\boldsymbol{\theta} = (\phi, \boldsymbol{\lambda}')'$  consisting of all the dispersion or variance components of the model, i.e. the correlation or dispersion parameter of the conditional beta-binomial distribution  $\phi$  and the vector of all variance parameters of the random effects  $\boldsymbol{\lambda}$ .

Consider that the response outcomes  $\mathbf{Y}$  depend on a set of given covariates  $X_1, \dots, X_k$ . Hence, following BBreg approach, we connect the probability parameter of the beta-binomial distribution with the observed covariates and random effects, assumed fixed when conditioning, by means of a logistic link function, i.e.

$$\eta_i = \text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \mathbf{x}_i' \boldsymbol{\beta} + \mathbf{z}_i' \mathbf{u}, \quad i = 1, \dots, n, \quad (3.23)$$

where  $\eta_i$  is the linear predictor of the model,  $p_i$  is the probability parameter of the conditional beta-binomial distribution,  $\boldsymbol{\beta}$  are the fixed effects,  $\mathbf{x}_i$  is the  $i$ th row of the full rank model matrix  $\mathbf{X}$  composed by the given covariates,  $\mathbf{u}$  are the random effects and  $\mathbf{z}_i$  is the  $i$ th row of the model matrix for the random effects. Notice that, as in all mixed-effects models, the correlation structure of the observed data, such as longitudinal repeated measures or hierarchies, is defined by the model matrix of the random effects  $\mathbf{Z}$ . We will denote the presented beta-binomial mixed-effects model as *BBmm*.



### 3.3.2 Estimation

Let assume that  $\mathbf{y} = (y_1, \dots, y_n)'$  is the vector of the  $n$  observations of the outcome variables. The marginal likelihood in BBmm approach is defined as

$$L(\boldsymbol{\beta}, \boldsymbol{\theta}|\mathbf{y}) = \int_{\mathbb{R}^q} \prod_{i=1}^n f_{y|u}(y_i|\boldsymbol{\beta}, \phi, \mathbf{u}) f_u(\mathbf{u}|\boldsymbol{\lambda}) d\mathbf{u} \quad (3.24)$$

where  $q$  is the number of levels or components of the random effects,  $f_{y|u}(y_i|\boldsymbol{\beta}, \phi, \mathbf{u})$  is the beta-binomial density function defined in Equation (1.20) and  $f_u(\mathbf{u}|\boldsymbol{\lambda})$  is the distribution of the random effects which is assumed normal. Similar to GLMMs, the marginal likelihood cannot be evaluated in closed form, and moreover, due to the complexity of the beta-binomial distribution, numerical computation is almost intractable. Therefore, approximation procedures must be developed to perform the inference in the model.

Equivalently, we can consider the marginal likelihood of the model in Equation (3.24) in an exponential form as

$$L(\boldsymbol{\beta}, \boldsymbol{\theta}|\mathbf{y}) = \int_{\mathbb{R}^q} \exp \left\{ \sum_{i=1}^n \log f_{y|u}(y_i|\boldsymbol{\beta}, \phi, \mathbf{u}) + \log f_u(\mathbf{u}|\boldsymbol{\lambda}) \right\} d\mathbf{u}.$$

Thus, considering that the summation of twice differentiable regular functions is a twice differentiable regular function, we can apply the Laplace's method for the integral approximation of the marginal likelihood of the model (see Appendix C for further details). Consequently, ignoring multiplicative constant terms, we have that

$$\log L(\boldsymbol{\beta}, \boldsymbol{\theta}|\mathbf{y}) \approx l(\boldsymbol{\beta}, \boldsymbol{\theta}|\mathbf{y}, \tilde{\mathbf{u}}) = h(\boldsymbol{\beta}, \boldsymbol{\theta}|\mathbf{y}, \tilde{\mathbf{u}}) - \frac{1}{2} \log |\mathbf{M}| \quad (3.25)$$

where,

$$\begin{aligned} h(\boldsymbol{\beta}, \boldsymbol{\theta}|\mathbf{y}, \mathbf{u}) &= \sum_{i=1}^n \log f_{y|u}(y_i|\boldsymbol{\beta}, \phi, \mathbf{u}) + \log f_u(\mathbf{u}|\boldsymbol{\lambda}) \\ &= \sum_{i=1}^n \left[ \sum_{k=0}^{y_i-1} \log(p_i + k\phi) + \sum_{k=0}^{m_i-y_i-1} \log(1 - p_i + k\phi) - \sum_{k=0}^{m_i-1} \log(1 + k\phi) \right] \\ &\quad - \frac{1}{2} \log |\mathbf{D}| - \frac{1}{2} \mathbf{u}' \mathbf{D}^{-1} \mathbf{u} \end{aligned} \quad (3.26)$$

is the joint log-likelihood of the model,  $\tilde{\mathbf{u}}$  is the solution of  $\partial h / \partial \mathbf{u} = 0$  and  $\mathbf{M}$  is

the adjusted term defined as

$$\mathbf{M} = \frac{\partial^2 h}{\partial \mathbf{u} \partial \mathbf{u}'} \Big|_{\mathbf{u}=\tilde{\mathbf{u}}}.$$

The resulting marginal likelihood approximation in Equation (3.25) is the first order Laplace approximation, and it is equivalent to integrating the random effects out (see Lee and Nelder (2001) for further details).

### Joint estimation of fixed and random effect

For the estimation of the fixed effects, we assume that the dispersion parameter vector  $\boldsymbol{\theta}$  is fixed and try to maximise the approximated log-likelihood of the model defined in Equation (3.25). For fixed  $\boldsymbol{\theta}$ , we denote the approximated log-likelihood as

$$l(\boldsymbol{\beta}|\mathbf{y}, \tilde{\mathbf{u}}, \boldsymbol{\theta}) = A(\boldsymbol{\beta}) + h(\boldsymbol{\beta}|\mathbf{y}, \tilde{\mathbf{u}}, \boldsymbol{\theta}). \quad (3.27)$$

Hence, the score equations of the fixed parameters of the model correspond to

$$S(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\beta}|\mathbf{y}, \tilde{\mathbf{u}}, \boldsymbol{\theta}) = \frac{A(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} + \frac{\partial h(\boldsymbol{\beta}|\mathbf{y}, \tilde{\mathbf{u}}, \boldsymbol{\theta})}{\partial \boldsymbol{\beta}}. \quad (3.28)$$

In GLMMs, Breslow and Clayton (1993) showed that  $A(\boldsymbol{\beta})$  depends on  $\boldsymbol{\beta}$  through the variance of the working vector or weight matrices defined in Equation (3.20). Assuming that this variance or weight matrix varies slowly as a function of  $\boldsymbol{\beta}$ , they proposed to ignore the term  $\partial A(\boldsymbol{\beta})/\partial \boldsymbol{\beta}$  in obtaining the marginal MLEs.

Given that the beta-binomial distribution does not belong to the exponential family, the normalization of the density function through the working vector theory cannot be applied directly. Hence, the previous statement is not directly extensible in this case. However, in BBmm approach we have that the adjustment term is defined as

$$A(\boldsymbol{\beta}) = -\frac{1}{2} \log |\mathbf{M}| = -\frac{1}{2} \log |\mathbf{Z}' \mathbf{W} \mathbf{S} \mathbf{Z} - \mathbf{D}^{-1}| \quad (3.29)$$

where  $\mathbf{S} = \text{diag}(p_i(1-p_i))$ ,  $\mathbf{W} = \text{diag}(w_i)$ ,  $w_i = -v_i p_i(1-p_i) + \xi_i(1-2p_i)$ ,

$$\begin{cases} \xi_i = \sum_{k=0}^{y_i-1} \frac{1}{p_i + k\phi} - \sum_{k=0}^{m_i-y_i-1} \frac{1}{1-p_i + k\phi} \\ v_i = \sum_{k=0}^{y_i-1} \frac{1}{(p_i + k\phi)^2} + \sum_{k=0}^{m_i-y_i-1} \frac{1}{(1-p_i + k\phi)^2} \end{cases} \quad (3.30)$$

and  $p_i = 1/[1 + \exp(-\mathbf{x}'_i \boldsymbol{\beta} - \mathbf{z}'_i \mathbf{u})]$  for  $i = 1, \dots, n$ , being all the previous formulas

evaluated at  $\mathbf{u} = \tilde{\mathbf{u}}$  (see Appendix D.2, Equation (D.11) for further details). Therefore, similar to GLMMs, we have shown in Equation (3.29) that in BBmm approach the adjustment term,  $A(\cdot)$ , only depends on  $\beta$  through the weight matrices  $\mathbf{W}$  and  $\mathbf{S}$ .

In order to show that the weight matrices vary slowly as a function of the fixed effects, we randomly fix some values of a covariate  $X$ , dispersion parameter vector  $\theta$  and random effects  $\mathbf{u}$ , and compare the joint-likelihood in Equation (3.26), the adjustment term in Equation (3.29) and the sum, or equivalently, the approximation of the log-likelihood of the model in Equation (3.27), as functions of  $\beta$ . Figure 3.1 shows the distribution of each mentioned function and the value where they reach the maximum. It can be appreciated that the low scale of the adjustment term does not alter the summation and that we get identical distributions for the joint likelihood and the approximated likelihood. Therefore, as it is shown in the left-hand side figure, both functions get the maximum in the same point, meaning that the effect of the adjustment term is redundant, and that the effects of the fixed effects in the weight matrices is insignificant.

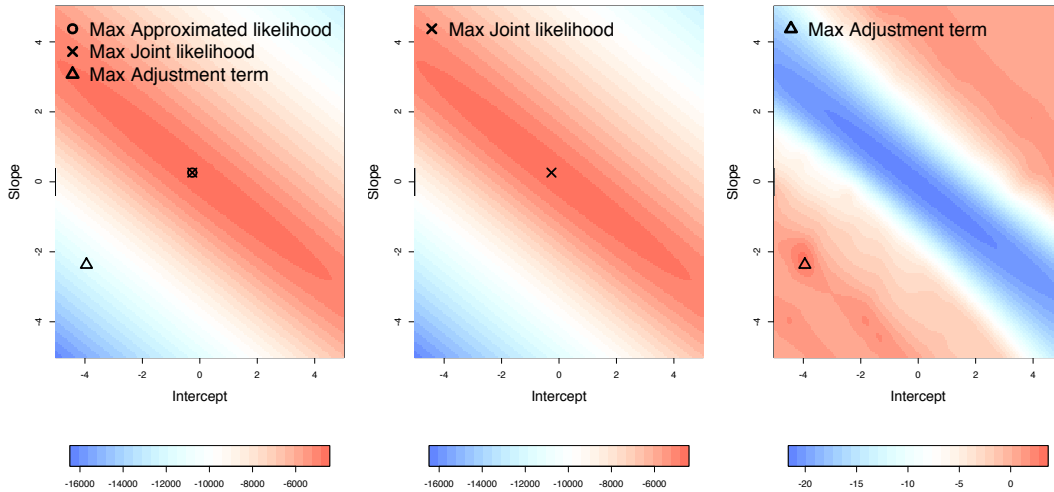


Figure 3.1: Distribution of the approximated marginal likelihood, joint log-likelihood and the penalisation term. The value where each function gets the maximum is also displayed.

Consequently, we can assume that the adjustment term in Equation (3.29) is flat with respect to  $\beta$  and, following Breslow and Clayton (1993) argument, ignore it in

the maximisation process of the fixed effects in the model.

We have shown that the adjusted term in the approximated log-likelihood in Equation (3.25) does not carry (almost) any information about the fixed effects, and hence, that all the information regarding  $\boldsymbol{\beta}$  is collected by the joint likelihood in Equation (3.26). The adjusted term in the approximated log-likelihood does not depend on  $\boldsymbol{\beta}$ , and hence, the random effects are canonical for the fixed effects. Lee et al. (2006) proved that if the random effects in the model are canonical for the fixed effects, then the MLE of the fixed effects through the marginal likelihood coincides with the MLE from the joint maximiser of the joint log-likelihood. Therefore, based on the previous statement, we can derive the MLE of the fixed effects by the joint maximisation of the log-likelihood presented in Equation (3.26).

Differentiation of the joint log-likelihood in Equation (3.26) with respect to  $\boldsymbol{\beta}$  and  $\mathbf{u}$  leads to the next score equations for the mean parameters,

$$\begin{cases} S(\boldsymbol{\beta}) = \frac{\partial h_1}{\partial \mathbf{p}} \frac{\partial \mathbf{p}}{\partial \boldsymbol{\beta}} = \boldsymbol{\xi}' \mathbf{S} \mathbf{X} \\ S(\mathbf{u}) = \frac{\partial h_1}{\partial \mathbf{p}} \frac{\partial \mathbf{p}}{\partial \mathbf{u}} + \frac{\partial h_2}{\partial \mathbf{u}} = \boldsymbol{\xi}' \mathbf{S} \mathbf{Z} - \mathbf{u}' \mathbf{D}^{-1} \end{cases}$$

where  $\boldsymbol{\xi} = (\xi_1, \dots, \xi_n)'$  and  $\xi_i$  and  $\mathbf{S}$  have been defined in Equation (3.30), and  $h_1$  and  $h_2$  correspond to the log-density function of the beta-binomial and normal distributions respectively (see Appendix D.2 for further details).

Different numerical algorithms may be used to solve the previous equations iteratively, such as the Newton-Raphson procedure. However, similarly to the cross-sectional model, we derive a estimation algorithm based on the delta method for several reasons. As it was explained in Section 2.2.1, due to the complexity of the beta-binomial density function and the fact that it does not belong to the exponential family, estimation process based on the Newton-Raphson algorithm could be hard to obtain, and even sometimes inappropriate (see Section 2.2.1). For instance, the second derivatives of the joint log-likelihood with respect to the fixed and random effects are quite complicated and the application of the expectation operator to conclude to the Fisher scoring algorithm intractable.

First, we assume that  $\mathbf{u}$  is fixed and try to get the estimation of  $\boldsymbol{\beta}$  for those fixed realizations of the random effects. Following the delta algorithm described in Equation (2.8), we have that the approximation of the matrix of the second

derivatives corresponds to

$$\frac{\partial^2 h}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} \approx \frac{\partial \mathbf{p}'}{\partial \boldsymbol{\beta}} \frac{\partial^2 h}{\partial \mathbf{p} \partial \mathbf{p}'} \frac{\partial \mathbf{p}}{\partial \boldsymbol{\beta}}.$$

Consequently, we have that

$$\frac{\partial^2 h}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} \approx -\mathbf{X}' \mathbf{S} \mathbf{V} \mathbf{S} \mathbf{X},$$

where  $\mathbf{V} = \text{diag}(v_1, \dots, v_n)$  and  $v_i = -\partial \xi_i / \partial p_i$  has been defined in Equation (3.30) for  $i = 1, \dots, n$  (see Appendix D.2, Equation (D.6) and Equation (D.8) for further details).

At this point, we can apply the Fisher scoring algorithm where the negative of the second derivatives is replaced by its expectation in a Newton-Raphson procedure. However, as it was explained in the cross-sectional beta-binomial regression model in Section 2.2.1, the expectation of  $\mathbf{V}$  is intractable, and therefore, the observed Fisher information must be used in the algorithm instead. This adjustment will usually increase the rate of convergence, though the algorithm may be less stable if the starting points are far from the maximum (Jørgensen, 1984). Therefore, we have that the estimation equations for the fixed effects are defined as

$$\hat{\boldsymbol{\beta}}^{(r+1)} = \hat{\boldsymbol{\beta}}^{(r)} - (-\mathbf{X}' \mathbf{S} \mathbf{V} \mathbf{S} \mathbf{X})^{-1} \mathbf{X}' \mathbf{S} \boldsymbol{\xi}.$$

where  $\mathbf{S}$ ,  $\mathbf{V}$  and  $\boldsymbol{\xi}$  are evaluated at the current  $r$  value of  $\boldsymbol{\beta}$ . Hence, after some easy operations we have that

$$\hat{\boldsymbol{\beta}}^{(r+1)} = (\mathbf{X}' \mathbf{S} \mathbf{V} \mathbf{S} \mathbf{X})^{-1} \mathbf{X}' \mathbf{S} \mathbf{V} \boldsymbol{\nu}_\beta, \quad (3.31)$$

where  $\boldsymbol{\nu}_\beta = \mathbf{X} \hat{\boldsymbol{\beta}}^{(r)} + (\mathbf{S} \mathbf{V})^{-1} \boldsymbol{\xi}$  being  $\mathbf{V}$  and  $\boldsymbol{\xi}$  evaluated at  $\hat{\boldsymbol{\beta}}^{(r)}$ .

Once fixed the estimated  $\hat{\boldsymbol{\beta}}$ , the same procedure can be developed to get the estimation equations for the random effects. Given that the development of the estimation procedure for the random effects is almost equal to the one performed for the fixed effects, we will only show the final estimating equation which corresponds to

$$\hat{\mathbf{u}}^{(r+1)} = (\mathbf{Z}' \mathbf{S} \mathbf{V} \mathbf{S} \mathbf{Z} + \mathbf{D}^{-1})^{-1} \mathbf{Z}' \mathbf{S} \mathbf{V} \boldsymbol{\nu}_u, \quad (3.32)$$

where  $\boldsymbol{\nu}_u = \mathbf{Z} \hat{\mathbf{u}}^{(r)} + (\mathbf{Z} \mathbf{V})^{-1} \boldsymbol{\xi}$  being  $\boldsymbol{\xi}$  and  $\mathbf{V}$  evaluated at  $\hat{\mathbf{u}}^{(r)}$ . More details about the derivatives of the approximated log-likelihood can be found in Appendix D.2.

### Estimation of variance components

For the estimation of the dispersion or variance component vector  $\boldsymbol{\theta}$ , the (marginal) maximum log-likelihood estimator might be substantially biased due to the estimates of  $\boldsymbol{\beta}$  (Lee and Nelder, 2001). Many authors have defined different likelihood adjustments to perform inferences in the variance components in several situations in the literature. For instance, Patterson and Thompson (1971) developed a restricted or residual maximum likelihood (REML) criteria in LMMs and Breslow and Clayton (1993) extended the approach to GLMMs. However, Lee and Nelder (1996) developed a more general approach based on the so called adjusted profile h-likelihood, which they proved that it is equivalent to the previously defined likelihood adjustments in each situation. The adjusted profile h-likelihood is defined as

$$h_p = h_A \Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}, \mathbf{u}=\hat{\mathbf{u}}},$$

where  $h_A$  is an adjusted h-likelihood defined as

$$h_A = h + \frac{1}{2} \log\{\det(2\pi \mathbf{H}^{-1})\}, \quad (3.33)$$

where  $h$  is the joint log-likelihood defined in Equation (3.26) and  $\mathbf{H}$  is the corresponding Hessian matrix of the model. The performed penalisation on the joint log-likelihood is equivalent to integrating the random effects out, as in the first order Laplace approximation in Equation (3.25), and then, eliminating nuisance fixed effects  $\boldsymbol{\beta}$  by conditioning on the maximum likelihood estimates  $\hat{\boldsymbol{\beta}}$  (Lee and Nelder, 2001). Therefore, maximum adjusted profile h-likelihood estimators can be derived for variance parameters by solving iteratively

$$\frac{\partial h_A}{\partial \theta_i} \Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}, \mathbf{u}=\hat{\mathbf{u}}} = 0,$$

where  $\hat{\boldsymbol{\beta}}$  and  $\hat{\mathbf{u}}$  are evaluated in each iteration.

As shown in Chapter 2, the delta algorithm offers an approximation procedure of the second derivatives of the log-likelihood, which defines the Hessian matrix of the model as

$$H \approx \begin{pmatrix} \mathbf{X}'\mathbf{S}\mathbf{V}\mathbf{S}\mathbf{X} & \mathbf{X}'\mathbf{S}\mathbf{V}\mathbf{S}\mathbf{Z} \\ \mathbf{Z}'\mathbf{S}\mathbf{V}\mathbf{S}\mathbf{X} & \mathbf{Z}'\mathbf{S}\mathbf{V}\mathbf{S}\mathbf{Z} + \mathbf{D}^{-1} \end{pmatrix}, \quad (3.34)$$

where all the terms involving the formula have been previously defined (see Appendix D.2, Equations (D.5), (D.6), (D.8) and (D.10) for further details). It is worth

noticing that  $\mathbf{D}$  is the only term in the Hessian matrix that depends on  $\boldsymbol{\lambda}$ , the vector of the variance components of the random effects. Besides, unlike the usual GLMM or even HGLM where the dispersion parameter of the conditional distribution can be explicitly taken out from the Hessian matrix, in the BBmm the matrix  $\mathbf{V}$  depends implicitly on the dispersion parameter  $\phi$ . Consequently, the computation of the score equation for the dispersion parameter of the conditional beta-binomial distribution is computationally more expensive than in models where the conditional distribution belongs to the exponential family.

Hence, the score equations for the variance parameters of the BBmm approach are defined as

$$\frac{\partial h_p}{\partial \theta_i} = \frac{\partial h(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{u})}{\partial \theta_i} \Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}, \mathbf{u}=\hat{\mathbf{u}}} + \frac{1}{2} \frac{\partial \log(\det(\mathbf{H}^{-1}))}{\partial \theta_i}, \quad (3.35)$$

for  $i = 1, \dots, k+1$ , where  $k$  is the number of parameters needed for specifying  $\mathbf{D}$ , i.e. the length of the vector  $\boldsymbol{\lambda}$ . Notice that  $\theta_i$  refers to either the beta-binomial dispersion parameter  $\phi$  or a parameter of the vector of variance components  $\boldsymbol{\lambda}$ .

Property D.4 (see Appendix D.1) offers an easy decomposition of the second term in Equation (3.35) as,

$$\frac{\partial \log(\det(\mathbf{H}^{-1}))}{\partial \theta_i} = -\frac{\partial \log(\det(\mathbf{H}))}{\partial \theta_i} = -\text{trace} \left[ \mathbf{H}^{-1} \frac{\partial \mathbf{H}}{\partial \theta_i} \right],$$

which simplifies the score equation of the variance components. In addition the score equations can be even more simplified depending on the variance parameter we are trying to estimate.

On the one hand, as it has been mentioned, due to the fact that the dispersion parameter  $\phi$  cannot be taken out from the Hessian matrix, the score equation for  $\phi$  is more complicated than in usual approaches. In BBmm approach, the score equation for the dispersion parameter  $\phi$  of the conditioned beta-binomial distribution is defined as

$$\begin{aligned} \frac{\partial h_p}{\partial \phi} = & \sum_{i=0}^n \left[ \sum_{k=0}^{y_i-1} \frac{k}{p_i + k\phi} + \sum_{k=0}^{m_i-y_i-1} \frac{k}{1-p_i + k\phi} - \sum_{k=0}^{m_i-1} \frac{k}{1+k\phi} \right] \\ & + \text{trace} \left[ \mathbf{H}^{-1} \begin{pmatrix} \mathbf{X}'\mathbf{S}\mathbf{J}\mathbf{S}\mathbf{X} & \mathbf{X}'\mathbf{S}\mathbf{J}\mathbf{S}\mathbf{Z} \\ \mathbf{Z}'\mathbf{S}\mathbf{J}\mathbf{S}\mathbf{X} & \mathbf{Z}'\mathbf{S}\mathbf{J}\mathbf{S}\mathbf{Z} \end{pmatrix} \right], \end{aligned} \quad (3.36)$$

where  $\mathbf{J} = \text{diag}(j_i)$ , and

$$j_i = \frac{1}{2} \frac{\partial v_i}{\partial \phi} = \sum_{k=0}^{y_i-1} \frac{k}{(p_i + k\phi)^3} + \sum_{k=0}^{m_i-y_i-1} \frac{k}{(1 - p_i + k\phi)^3}$$

for  $i = 1, \dots, n$ . Numerical algorithms, such as Newton-Raphson can be applied to get the maximum adjusted profile h-likelihood of the dispersion parameter of the conditioned beta-binomial distribution.

On the other hand, by Property D.6, the score equation for the variance parameters of the random effects is defined as

$$S(\theta_i) = -\frac{1}{2} \text{trace} \left[ \mathbf{D}^{-1} \frac{\partial \mathbf{D}}{\partial \theta_i} \right] + \frac{1}{2} \mathbf{u}' \mathbf{D}^{-1} \frac{\partial \mathbf{D}}{\partial \theta_i} \mathbf{D}^{-1} \mathbf{u} + \frac{1}{2} \text{trace} \left[ \mathbf{P} \mathbf{D}^{-1} \frac{\partial \mathbf{D}}{\partial \theta_i} \mathbf{D}^{-1} \right] \quad (3.37)$$

where  $\mathbf{P}$  is the right bottom block of  $\mathbf{H}^{-1}$  defined as

$$\mathbf{P} = \left[ \mathbf{Z}' \mathbf{S} \mathbf{V} \mathbf{S} \mathbf{Z} + \mathbf{D}^{-1} - \mathbf{Z}' \mathbf{S} \mathbf{V} \mathbf{S} \mathbf{X} (\mathbf{X}' \mathbf{S} \mathbf{V} \mathbf{S} \mathbf{X})^{-1} \mathbf{X}' \mathbf{S} \mathbf{V} \mathbf{S} \mathbf{Z} \right]^{-1}. \quad (3.38)$$

### Example 3.3: One independent random component

The easiest situation is the assumption of an unique independent random component. Suppose that the random effects  $\mathbf{u} = (u_1, \dots, u_q)'$  are iid with a normal distribution of mean 0 and variance  $\sigma_u^2$ . Namely, the variance-covariance matrix of the random effects is defined as

$$\mathbf{D} = \sigma_u^2 \mathbf{I}_q.$$

Thus, after some calculus, following Equation (3.37) we have that

$$S(\sigma_u) = -\sigma_u^{-1} \text{trace} [\mathbf{I}_q] + \sigma_u^{-3} \mathbf{u}' \mathbf{I}_q \mathbf{u} + \sigma_u^{-3} \text{trace} [\mathbf{P}],$$

and hence, the maximum adjusted profile estimator of  $\sigma_u$  is achieved by the iterative use of

$$\hat{\sigma}_u^{(r+1)2} = \frac{1}{q} \hat{\mathbf{u}}' \hat{\mathbf{u}} + \frac{1}{q} \text{trace} [\mathbf{P}],$$

where  $\mathbf{P}$  has been defined in Equation (3.38) and the matrices involving the formula are evaluated at  $\hat{\sigma}_u^{(r)}$ , the estimate of  $\sigma_u$  in the  $r$ th iteration.



□

**Example 3.4: Two independent random components**

Another possible scenario is the assumption of two independent random components. Assume that we have  $\mathbf{u}_1 = (u_{11}, \dots, u_{1q_1})'$  and  $\mathbf{u}_2 = (u_{21}, \dots, u_{2q_2})'$  where they are independent with variance-covariance matrices  $\mathbf{D}_1 = \sigma_1^2 \mathbf{I}_{q_1}$  and  $\mathbf{D}_2 = \sigma_2^2 \mathbf{I}_{q_2}$  respectively. Similarly, without loss of generalisation, we could assume that the model consists of a unique random component  $\mathbf{u} = (\mathbf{u}'_1, \mathbf{u}'_2)'$  where the variance-covariance matrix is defined as

$$\mathbf{D} = \begin{pmatrix} \mathbf{D}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_2 \end{pmatrix} = \begin{pmatrix} \sigma_1^2 \mathbf{I}_{q_1} & \mathbf{0} \\ \mathbf{0} & \sigma_2^2 \mathbf{I}_{q_2} \end{pmatrix}.$$

Following Equation (3.37), we have that the score equations for each dispersion parameter  $\sigma_1$  and  $\sigma_2$  are defined as

$$\begin{aligned} S(\sigma_1) &= -\sigma_1^{-1} \text{trace} \begin{pmatrix} \mathbf{I}_{q_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} + \sigma_1^{-3} \mathbf{u}' \begin{pmatrix} \mathbf{I}_{q_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \mathbf{u} + \sigma_1^{-3} \text{trace} \left[ \mathbf{P} \begin{pmatrix} \mathbf{I}_{q_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \right] \\ &= -\frac{q_1}{\sigma_1} + \frac{1}{\sigma_1^3} \mathbf{u}'_1 \mathbf{u}_1 + \frac{1}{\sigma_1^3} \text{trace}_{(1:q_1)} [\mathbf{P}] \end{aligned}$$

and

$$\begin{aligned} S(\sigma_2) &= -\sigma_2^{-1} \text{trace} \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{q_2} \end{pmatrix} + \sigma_2^{-3} \mathbf{u}' \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{q_2} \end{pmatrix} \mathbf{u} + \sigma_2^{-3} \text{trace} \left[ \mathbf{P} \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{q_2} \end{pmatrix} \right] \\ &= -\frac{q_2}{\sigma_2} + \frac{1}{\sigma_2^3} \mathbf{u}'_2 \mathbf{u}_2 + \frac{1}{\sigma_2^3} \text{trace}_{(q_1+1:q_2)} [\mathbf{P}], \end{aligned}$$

where  $\text{trace}_{(a:b)}$  corresponds to the trace of the matrix only considering the diagonal elements that are between the  $a$ th and  $b$ th position, i.e.

$$\text{trace}_{(a:b)}[\mathbf{P}] = \sum_{i=a}^b P_{ii}.$$

Notice that as the variance parameters are independent the score equations

correspond to the first example where only a random component was assumed. Estimation of  $\sigma_1$  and  $\sigma_2$  is performed by the iterative use of

$$\begin{aligned}\hat{\sigma}_1^{(r+1)^2} &= \frac{1}{q_1} \hat{\mathbf{u}}_1' \hat{\mathbf{u}}_1 + \frac{1}{q_1} \text{trace}_{(1:q_1)} [\mathbf{P}], \\ \hat{\sigma}_2^{(r+1)^2} &= \frac{1}{q_2} \hat{\mathbf{u}}_2' \hat{\mathbf{u}}_2 + \frac{1}{q_2} \text{trace}_{((q_1+1):(q_1+q_2))} [\mathbf{P}],\end{aligned}$$

where matrices involving the formulae are evaluated at  $\hat{\sigma}_i^{(r)}$ , the estimate of  $\sigma_i$  in the  $r$ th iteration,  $i = 1, 2$ .

Following Example 3.4, it is straightforward to generalise the estimating equations for any number of independent random components in the model.

□

### Example 3.5: Two correlated random components

It is common in some studies the assumption that two random effects are correlated with each other (Hedeker and Gibbons, 2006). For instance, assume that we have a study where  $n$  individuals are measured at  $t_i$  scenarios,  $i = 1, \dots, n$ , and that we construct the following linear predictor of the model,

$$\eta_{ij} = \mathbf{x}'_{ij} \boldsymbol{\beta} + \mathbf{z}'_{1ij} \mathbf{u}_{1i} + \mathbf{z}'_{2ij} \mathbf{u}_{2i},$$

for  $i = 1, \dots, n$  and  $j = 1, \dots, t_i$  or in matrix notation,

$$\boldsymbol{\eta} = \mathbf{X} \boldsymbol{\beta} + \mathbf{Z}_1 \mathbf{u}_1 + \mathbf{Z}_2 \mathbf{u}_2,$$

where  $\mathbf{u}_1 = (u_{11}, \dots, u_{1n})'$  and  $\mathbf{u}_2 = (u_{21}, \dots, u_{2n})'$  are two individual level random effects. Equivalently, we can rewrite the model for a unique random component as,

$$\boldsymbol{\eta} = \mathbf{X} \boldsymbol{\beta} + \mathbf{Z} \mathbf{u}$$

where  $\mathbf{Z} = [\mathbf{Z}_1 \ \mathbf{Z}_2]$  and  $\mathbf{u} = (\mathbf{u}'_1, \mathbf{u}'_2)'$ .

As well as in the previous examples, we assume that  $\mathbf{u}_1$  and  $\mathbf{u}_2$  are normally distributed random effects with mean  $\mathbf{0}$  and variance-covariance matrix  $\mathbf{D}_1 = \sigma_1^2 \mathbf{I}_n$  and  $\mathbf{D}_2 = \sigma_2^2 \mathbf{I}_n$  respectively. However, in this example we assume that the first random component is correlated with the second random component

for each individual. Namely, we assume that the variance-covariance matrix of the unique random component  $\mathbf{u}$  is defined as

$$\mathbf{D} = \begin{pmatrix} \sigma_1^2 \mathbf{I}_n & \rho \mathbf{I}_n \\ \rho \mathbf{I}_n & \sigma_2^2 \mathbf{I}_n \end{pmatrix},$$

where  $\rho$  is the covariance term between the two random effects for the same individual.

We know by Property D.6 in Appendix D.1 that the inverse of  $\mathbf{D}$  is defined as

$$\mathbf{D}^{-1} = \begin{pmatrix} (\sigma_1^2 - \rho^2 \sigma_2^{-2})^{-1} \mathbf{I}_n & -(\sigma_1^2 - \rho^2 \sigma_2^{-2})^{-1} \rho \sigma_2^{-2} \mathbf{I}_n \\ -(\sigma_2^2 - \rho^2 \sigma_1^{-2})^{-1} \rho \sigma_1^{-2} \mathbf{I}_n & (\sigma_2^2 - \rho^2 \sigma_1^{-2})^{-1} \mathbf{I}_n \end{pmatrix},$$

which is much more complicated than in the previous examples.

For instance, in order to evidence the complexity of the estimation of the dispersion parameters when correlation is included, we are going to show the score equation for  $\sigma_1$ . Before applying Equation (3.37) some calculus must be done, such as

$$\mathbf{D}^{-1} \frac{\partial \mathbf{D}}{\partial \sigma_1} = \mathbf{D}^{-1} \begin{pmatrix} 2\sigma_1 \mathbf{I}_n & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} = 2\sigma_1 \begin{pmatrix} (\sigma_1^2 - \rho^2 \sigma_2^{-2})^{-1} \mathbf{I}_n & \mathbf{0} \\ (\sigma_2^2 - \rho^2 \sigma_1^{-2})^{-1} \rho \sigma_1^{-2} \mathbf{I}_n & \mathbf{0} \end{pmatrix},$$

and,

$$\begin{aligned} \mathbf{D}^{-1} \frac{\partial \mathbf{D}}{\partial \sigma_1} \mathbf{D}^{-1} &= 2\sigma_1 \begin{pmatrix} (\sigma_1^2 - \rho^2 \sigma_2^{-2})^{-1} \mathbf{I}_n & \mathbf{0} \\ (\sigma_2^2 - \rho^2 \sigma_1^{-2})^{-1} \rho \sigma_1^{-2} \mathbf{I}_n & \mathbf{0} \end{pmatrix} \mathbf{D}^{-1} \\ &= \frac{2\sigma_1}{\sigma_1^2 - \rho^2 \sigma_2^{-2}} \\ &\quad \times \begin{pmatrix} (\sigma_1 - \rho^2 \sigma_2^{-2})^{-1} \mathbf{I}_n & -(\sigma_1^2 - \rho^2 \sigma_2^{-2})^{-1} \rho \sigma_2^{-2} \mathbf{I}_n \\ -(\sigma_2^{-2} - \rho^2 \sigma_1^{-2})^{-1} \rho \sigma_1^{-2} \mathbf{I}_n & (\sigma_2^2 - \rho^2 \sigma_1^{-2})^{-1} \rho^2 \sigma_1^{-2} \sigma_2^{-2} \mathbf{I}_n \end{pmatrix}. \end{aligned}$$

Therefore, we obtain that the score equation for the variance parameter of the first random component  $u_1$  is defined as

$$S(\sigma_1) = -\frac{n\sigma_1}{\sigma_1^2 - \rho^2 \sigma_2^{-2}} + \frac{\sigma_1}{\sigma_1^2 - \rho^2 \sigma_2^{-2}} \mathbf{u}' \check{\mathbf{D}}_1 \mathbf{u} + \frac{\sigma_1}{\sigma_1^2 - \rho^2 \sigma_2^{-2}} \text{trace} \left[ \mathbf{P} \check{\mathbf{D}}_1 \right],$$

where

$$\check{D}_1 = \begin{pmatrix} (\sigma_1 - \rho^2 \sigma_2^{-2})^{-1} \mathbf{I}_n & -(\sigma_1^2 - \rho^2 \sigma_2^{-2})^{-1} \rho \sigma_2^{-2} \mathbf{I}_n \\ -(\sigma_2^{-2} - \rho^2 \sigma_1^{-2})^{-1} \rho \sigma_1^{-2} \mathbf{I}_n & (\sigma_2^{-2} - \rho^2 \sigma_1^{-2})^{-1} \rho^2 \sigma_1^{-2} \sigma_2^{-2} \mathbf{I}_n \end{pmatrix}.$$

It can be appreciated the complexity of the solution of the equation due to all terms involving  $\sigma_1$ . The score equations for the rest of the variance parameters  $\sigma_2$  and  $\rho$  are quite similar to the presented equation. Therefore, numerical algorithms, such as Newton-Raphson, should be used to solve the score equations of the variance parameters of the random effects.

□

### Estimation algorithm

The estimation algorithm of the BBmm approach can be formulated in the following way:

- Step 1.** Give initial values for  $\hat{\beta}^{(0)}$ ,  $\hat{u}^{(0)}$  and  $\hat{\theta}^{(0)} = (\phi^{(0)}, \lambda^{(0)})$ .
- Step 2.** Set  $r_1, r_2 \leftarrow 0$ .
- Step 3.** Fix  $\hat{\theta}^{(r_2)}$ :
- Step 3.1.** fix  $\hat{u}^{(r_1)}$  and iterate Equation (3.31) until convergence of  $\hat{\beta}^{(r_1+1)}$ ,
- Step 3.2.** fix  $\hat{\beta}^{(r_1+1)}$  and iterate Equation (3.32) until convergence of  $\hat{u}^{(r_1+1)}$ ,
- Step 3.3.** set  $r_1 \leftarrow r_1 + 1$ ,
- Step 3.4.** iterate between **Step (3.1.)**, **Step (3.2.)** and **Step (3.3.)** until convergence.
- Step 4.** Fix  $\hat{\beta}^{(r_1)}$  and  $\hat{u}^{(r_1)}$ :
- Step 4.1.** fix  $\lambda^{(r_2)}$  and estimate  $\phi^{(r_2+1)}$  iterating Equation (3.36),
- Step 4.2.** fix  $\phi^{(r_2+1)}$  and estimate  $\lambda^{(r_2+1)}$  iterating Equation (3.37),
- Step 5.**  $r_2 \leftarrow r_2 + 1$ .
- Step 6.** Iterate between **Step 3.**, **Step 4.** and **Step 5.** until convergence.

### 3.3.3 Inference

In order to perform inference of the fixed effects, we must provide a procedure to estimate the variance of the estimation of  $\beta$ . The following property proved by Lee and Nelder (1996) will allow us to estimate  $\text{Var}[\hat{\beta}]$  in a simple and closed form.

**Property 3.1.** *Let*

$$\mathbf{V} = \begin{pmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22} \end{pmatrix} = n \begin{pmatrix} \text{Var}[\hat{\beta}] & \text{Cov}[\hat{\beta}, \hat{\mathbf{u}} - \mathbf{u}] \\ \text{Cov}[\hat{\mathbf{u}} - \mathbf{u}, \hat{\beta}] & \text{Var}[\hat{\mathbf{u}} - \mathbf{u}] \end{pmatrix} \quad \text{and} \quad \mathbf{M} = \frac{1}{n} \begin{pmatrix} \mathbf{B} & \mathbf{C} \\ \mathbf{C}' & \mathbf{D} \end{pmatrix},$$

where  $\mathbf{B}, \mathbf{C}$  and  $\mathbf{D}$  are matrices such that

$$\mathbf{B}_{ij} = -\frac{\partial^2 h}{\partial \beta_i \partial \beta_j} \Big|_{\beta=\hat{\beta}, \mathbf{u}=\hat{\mathbf{u}}}, \quad \mathbf{C}_{ij} = -\frac{\partial^2 h}{\partial \beta_i \partial u_j} \Big|_{\beta=\hat{\beta}, \mathbf{u}=\hat{\mathbf{u}}} \quad \text{and} \quad \mathbf{D}_{ij} = -\frac{\partial^2 h}{\partial u_i \partial u_j} \Big|_{\beta=\hat{\beta}, \mathbf{u}=\hat{\mathbf{u}}}.$$

Then, if  $\mathbb{E}[\mathbf{M}]$  is non-singular, under appropriate regularly conditions,  $\mathbf{M}^{-1}$  converges to  $\mathbf{V}$  as  $n \rightarrow \infty$ .

In addition, this holds when entries of  $\mathbf{M}$  are replaced by corresponding expectations since  $\mathbf{B}$ ,  $\mathbf{C}$  and  $\mathbf{D}$  are sums of matrices. Therefore, the variance-covariance matrix of the estimates in BBmm approach is given by the inverse of the Hessian matrix  $\mathbf{H}$  defined in Equation (3.34). Nevertheless, Lee and Nelder (1996) proved that when the realized values of the random effects are known (estimated) the estimation of  $\text{Var}[\hat{\beta}]$  corresponds to the inverse of the first block of  $\mathbf{H}$ . Therefore, in BBmm approach we have that

$$\text{Var}[\hat{\beta}] = (\mathbf{X}' \mathbf{S} \mathbf{V} \mathbf{S} \mathbf{X})^{-1}. \quad (3.39)$$

Hence, the estimation of the variance of the MLE of the fixed effects takes into account the information loss caused by estimating the random effects.

Danaher (1987) showed that the beta-binomial distribution satisfies the regularity conditions in Definition 1.1 in Chapter 1. Hence, we have that the MLE of the regression parameters in BBmm approach follows a normal distribution. Therefore, the relationship in Equation (3.39) allows us performing statistical test on the estimates of  $\beta = (\beta_1, \dots, \beta_{p+1})$ , such as the Wald's test (Pawitan, 2001). The Wald's test is used to test

$$\begin{cases} H_0 : & \beta_i = \beta_0, \\ H_1 : & \beta_i \neq \beta_0 \end{cases}, \quad i = 1, \dots, p+1$$

assuming that under  $H_0$  the Wald statistic

$$z = \frac{\hat{\beta}_i - \beta_0}{\sqrt{\text{Var}[\hat{\beta}_i]}}$$

follows a normal distribution with mean 0 and variance 1.

### Deviance

When dealing with random effects regression models, Lee and Nelder (1996) proposed the use of three different deviance functions depending on the model component we want to test. First, they proposed the deviance based on the joint likelihood in Equation (3.26), i.e.  $-2h$ , for testing random effects. Second, the deviance consisting of the marginal likelihood,  $-2\log L$ , for testing the fixed effects, and, finally, the deviance based on the marginal likelihood conditional on the fixed effects,  $-2\log f_\theta(\mathbf{y}|\hat{\boldsymbol{\beta}})$  for testing the variance components where  $f_\theta(\mathbf{y}|\hat{\boldsymbol{\beta}})$  is the marginal density function of the model. However, in models where the marginal likelihood is hard to obtain, for instance the BBmm, they proposed the use of approximation likelihoods instead. Consequently, for testing fixed effects Laplace approximation of the marginal likelihood defined Equation (3.25) can be used, whereas for testing the variance components the adjusted profile likelihood defined in Equation (3.33) should be used instead.

Nevertheless, the main utility of the deviance function lies in the goodness-of-fit criterion. The same authors defined a scaled deviance as

$$D(\mathbf{y}, \hat{\boldsymbol{\mu}}) = -2 [l(\hat{\boldsymbol{\mu}}, \phi|\mathbf{y}, \mathbf{u}) - l(\mathbf{y}, \phi|\mathbf{y}, \mathbf{u})] \quad (3.40)$$

where  $l(\hat{\boldsymbol{\mu}}, \phi|\mathbf{y}, \mathbf{u}) = \log f(\mathbf{y}|\mathbf{u}, \hat{\boldsymbol{\beta}})$  and  $\boldsymbol{\mu} = \mathbb{E}[\mathbf{y}|\mathbf{u}]$ , for testing the goodness-of-fit of the model. In addition, they proposed the following formula to estimate the degrees of freedom of the model

$$\text{d.f.} = n - \text{trace}(\mathbf{H}^{-1}\mathbf{H}^*),$$

where in BBmm approach  $\mathbf{H}^*$  is defined as

$$\mathbf{H}^* = \begin{pmatrix} \mathbf{X}'\mathbf{S}\mathbf{V}\mathbf{S}\mathbf{X} & \mathbf{X}'\mathbf{S}\mathbf{V}\mathbf{S}\mathbf{Z} \\ \mathbf{Z}'\mathbf{S}\mathbf{V}\mathbf{S}\mathbf{X} & \mathbf{Z}'\mathbf{S}\mathbf{V}\mathbf{S}\mathbf{Z} \end{pmatrix}.$$

They conclude that if the computed scaled deviance is much larger than the estimated degrees of freedom, we may suspect the absence of some necessary fixed or random effects in the linear predictor of the model. Notice that the scaled deviance is based on the conditioned density function of the model, so it cannot be used for testing dispersion parameters (Lee and Nelder, 1996).

### 3.3.4 Latent variable interpretation

In linear regression models, either fixed or mixed effects models, it is straightforward to determine with the variance-covariance structure of the residuals. However, notice that in non-linear regression models, the specification of the linear predictor does not allow the inclusion of any error term. In this context, the latent variable theory results of main interest as it allows estimating the variability that exists between the errors of underlying latent variables associated with each observation.

For instance, assume that there is an unobservable continuous random variable  $Y_i^*$  defined in the real line  $\mathbb{R}$ , such that the binary variable  $Y_i$  takes the value 1 if and only if  $Y_i^*$  exceeds a certain threshold  $\gamma$ . We will denote  $Y_i^*$  as the latent variable. Therefore, we have the following relationship,

$$\pi_i = \Pr(Y_i = 1) = \Pr(Y_i^* > \gamma).$$

Without loss of generality, as the location and scale of  $Y_i^*$  are arbitrary, we take the threshold to be zero and standardize the latent variable  $Y_i^*$  to have standard deviation equal to one.

Suppose now that the outcome depends on some given covariates  $X_1, \dots, X_p$ . In order to model the dependence of the correlated data, we define the following linear model for the latent variable,

$$Y_i^* = \mathbf{x}'_i \boldsymbol{\beta} + \epsilon_i,$$

where  $\epsilon_i$  is the error term, which it is assumed to have a distribution with cumulative density function  $F(\epsilon_i)$ .

Under this model, the probability of observing a positive outcome is given by

$$\pi_i = \Pr(Y_i^* > 0) = \Pr(\epsilon_i > -\eta_i) = 1 - F(-\eta_i),$$

where  $\eta_i = \mathbf{x}'_i \boldsymbol{\beta}$  is the linear predictor. If the distribution of the error term  $\epsilon_i$  is

symmetric about zero, we have that  $F(x) = 1 - F(-x)$ , and consequently

$$\pi_i = F(\eta_i),$$

which defines a GLM for binary responses with a link function equal to  $F^{-1}(\cdot)$ .

Depending on the features of the data to be analyzed there is a wide variety of link functions available in the literature. In particular, when dealing with binary or binomial data it is common to link the expectation of the dependent variable with the given covariates by means of a logit function as it corresponds to the canonical link function (see Section 2.1.1 for more details). The logit link function is defined as

$$\eta_i = \log \frac{\pi_i}{1 - \pi_i},$$

where the cumulative density function of the logistic distribution is defined as

$$\pi_i = F(\eta_i) = \frac{e^{\eta_i}}{1 + e^{\eta_i}},$$

being  $-\infty < \eta_i < \infty$ . Therefore, we can derive the standard density function as,

$$f(x) = \frac{\partial F(x)}{\partial x} = \frac{e^x}{(1 + e^x)^2},$$

which maintains the following symmetric property

$$\begin{aligned} f(-x) &= \frac{e^{-x}}{(1 + e^{-x})^2} = \frac{1}{e^x(1 + e^{-x})^2} = \frac{1}{(1 + e^x)(1 + e^{-x})} \\ &= \frac{1}{e^{-x}(1 + e^x)^2} = \frac{e^x}{(1 + e^x)^2} = f(x). \end{aligned}$$

Moreover, the first and second order moment of the distributions are defined as

$$\begin{aligned} \mathbb{E}[x] &= \int_{-\infty}^{\infty} x f(x) dx = \int_{-\infty}^{\infty} \frac{x e^x}{(1 + e^x)^2} = \left. \frac{x e^x}{1 + e^x} - \log(1 + e^x) \right]_{-\infty}^{\infty} = 0, \\ \text{Var}[x] &= \int_{-\infty}^{\infty} x^2 f(x) dx = \int_{-\infty}^{\infty} \frac{x^2 e^x}{(1 + e^x)^2} = \frac{\pi^2}{3}. \end{aligned}$$

Consequently, the standard logistic distribution is symmetric, has mean zero and variance  $\pi^2/3$ . The shape is very close to the normal distribution but it has heavier tails as it can be appreciated in Figure 3.2 where both distributions shapes are displayed.



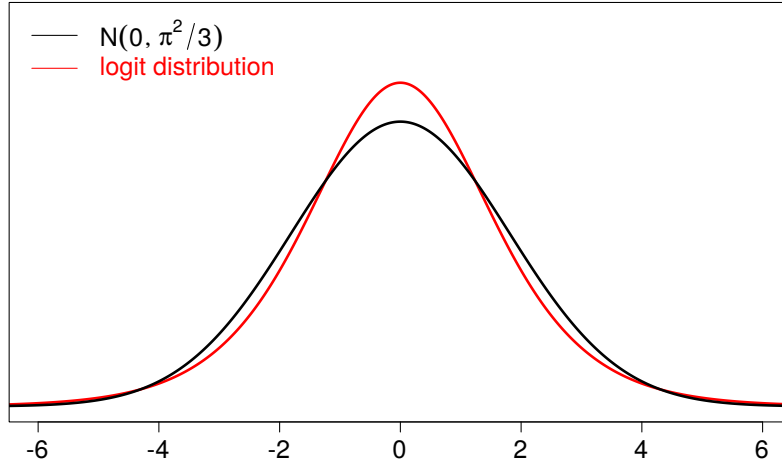


Figure 3.2: Distribution of the standard logistic distribution together with the normal distribution of mean 0 and variance  $\pi^2/3$ .

Therefore, the coefficients in a regression model where the logit link function has been applied can be interpreted not only in terms of log-odds, but also as effects of the covariates on a latent variable that follows a LM with logistic residuals. One of the main contributions of this new specification of a regression model based on the logit link function is that we can conclude with correlations between latent variables associated with observations.

For instance, assume that we have a longitudinal study where  $n$  individuals are measured in  $t_i$  time points,  $i = 1, \dots, n$ . Moreover, assume that the BBmm approach, which is based on a logit link function as shown in Equation (3.23), is used for the modelling the data. Assume that  $Y_{ij}^*$  is the latent variable associated with the observation of the  $i$ th individual on the  $j$ th observations,  $i = 1, \dots, n$  and  $j = 1, \dots, t_i$ . As explained before we can link the latent variable with the linear predictor concluding that

$$Y_{ij}^* = \mathbf{x}'_{ij}\boldsymbol{\beta} + u_i + \epsilon_{ij},$$

where  $u_i \sim \mathcal{N}(0, \sigma_u^2)$  are independent random effects and  $\epsilon_{ij}$  are independent error terms that follow a logistic distribution, being  $u_i$  and  $\epsilon_{ij}$  independent  $i = 1, \dots, n$  and  $j = 1, \dots, t_i$ . Therefore, the expectation, the variance and the covariance between

two latent variables of the same individual in BBmm approach are defined as,

$$\mathbb{E}[Y_{ij}^*] = \mathbb{E}[\mathbf{x}'_{ij}\boldsymbol{\beta} + u_i + \epsilon_{ij}] = \mathbf{x}'_{ij}\boldsymbol{\beta}$$

$$\text{Var}[Y_{ij}^*] = \text{Var}[\mathbf{x}'_{ij}\boldsymbol{\beta} + u_i + \epsilon_{ij}] = \text{Var}[u_i] + \text{Var}[\epsilon_{ij}] = \sigma_u^2 + \frac{\pi^2}{3}$$

$$\begin{aligned} \text{Cov}[Y_{ij}^*, Y_{ik}^*] &= \text{Cov}[\mathbf{x}'_{ij}\boldsymbol{\beta} + u_i + \epsilon_{ij}, \mathbf{x}'_{ik}\boldsymbol{\beta} + u_i + \epsilon_{ik}] = \text{Cov}[u_i + \epsilon_{ij}, u_i + \epsilon_{ik}] \\ &= \text{Var}[u_i] + \text{Cov}[\epsilon_{ij}, \epsilon_{ik}] = \begin{cases} \sigma_u^2 & \text{if } j \neq k \\ \sigma_u^2 + \pi^2/3 & \text{if } j = k \end{cases} \end{aligned}$$

where in a matrix notation corresponds to

$$\mathbb{E}[\mathbf{Y}_i^*] = \mathbb{E}[\mathbf{x}'_i\boldsymbol{\beta} + \mathbf{1}_{t_i}u_i + \boldsymbol{\epsilon}_i] = \mathbf{x}'_i\boldsymbol{\beta}$$

$$\text{Var}[\mathbf{Y}_i^*] = \text{Var}[\mathbf{x}'_i\boldsymbol{\beta} + \mathbf{1}_{t_i}u_i + \boldsymbol{\epsilon}_i] = \sigma_u^2\mathbf{J}_{t_i} + \frac{\pi^2}{3}\mathbf{I}_{t_i}$$

where  $\mathbf{Y}_i^* = (Y_{i1}^*, \dots, Y_{it_i}^*)'$ ,  $\mathbf{1}_{t_i}$  is a  $t_i$  length vector of 1s,  $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \dots, \epsilon_{it_i})'$ ,  $\mathbf{I}_{t_i}$  is the  $t_i \times t_i$  identity matrix and  $\mathbf{J}_{t_i}$  is a  $t_i \times t_i$  matrix of 1s. Consequently, the inclusion of a random intercept per individual defines a specific correlation structure between the latent variables associated with the observations of the same individual. Different linear predictors, i.e. different random effect structures, lead to different correlation structures. Specially, it is worth noticing that in longitudinal models where random slopes are included, correlation between different latent variables is not fixed, but instead depends on the slope variable, which it is usually defined as the time variable. Therefore, the introduction of random effects that interact with the time variable defines a correlation structure that changes over time, which measures the dynamic relationship between the covariates and the outcome variables.

### 3.3.5 Similar approaches in the literature

It has been mentioned before that PROs have some special features, such as correlation within patient responses and overdispersion, which make the fit through exponential family distributions inappropriate (Arostegui et al., 2007). We have defined in Section 3.2 the most widely used mixed-effects models when dealing with longitudinal data, however, these regression approaches strict the model assumption to exponential family distributions and Gaussian random effects. Due to the fact that in cross-sectional data the analysis of PROs through regression models based

on exponential family distributions is not appropriate enough, the extension to the inclusion of Gaussian random effects for longitudinal studies will fail in the same assumption, as Gaussian random effects are only included to model the hierarchical structure of the correlated data.

In the literature, although few, there exist some flexible models which allow the correct analysis of PROs in a longitudinal framework, at least in the model assumption. There are two different ways of solving the fitting problem caused by the features of PROs in mixed-effects framework: (i) include non-Gaussian random effects in a GLMM approach or, (ii) assume conditional distributions that do not belong to the exponential family. Following, we define the most widely used regression models based on the cited two approaches.

We have defined in Section 2.2.2 the so-called HGLMs, where additional non-Gaussian random effects can be included in the linear predictor of a classical GLM (Lee and Nelder, 1996). Therefore, in PROs longitudinal framework, these models assume that conditional on some beta and Gaussian random effects, the outcomes follow a binomial distribution, defining the so-called binomial-beta-normal regression model. That way, the beta random effects account for the PROs characteristics, while the Gaussian random effects accommodate the hierarchical structure of the longitudinal data. However, in Section 2 we have shown that in independent data framework the performance of the marginal approach is more appropriate than the conditional beta-binomial regression approach in terms of parameter estimation and statistical significance. Therefore, in a longitudinal framework, even if Gaussian random effects are introduced in the linear predictor of the `BBhglm` to accommodate the correlation given by the repeated measurements, the incapacity of the conditional model approach to assess the effect of some covariates may have on the PROs, will lead again to the same estimation problems (see Chapter 2, Section 2.4 for more details).

Based on a similar approach, Molenberghs et al. (2010) developed a general methodology which combines both mixed-effects models (GLMMs) and non-Gaussian effects to accommodate overdispersion. Similar to the HGLMs the election of the distribution of the non-normal random effects is based on the conjugacy assumption. Nevertheless, in contrast to HGLM where the conjugate random effects are included as additive terms in the linear predictor of a GLMM, in the combined model the effects multiply the parameter being modelled.

For instance, Molenberghs et al. (2012) defined the beta-binomial model with

logit link as

$$y_i \sim \text{Bin}(m_i, p_i = \theta_i \kappa_i),$$

$$\kappa_i = \frac{\exp(\mathbf{x}'_i \boldsymbol{\beta})}{1 + \exp(\mathbf{x}'_i \boldsymbol{\beta})} \quad (3.41)$$

where  $\theta_i \sim \text{Beta}(1/\alpha, 1/\alpha)$ . When normal random effects are included in Equation (3.41), closed forms cannot be reached for the mean or the variance of the model. In addition, only when beta dispersion random effects are assumed in the model conjugacy applies. Therefore, the authors defined the concept of strong conjugacy as a way of expressing in which cases conjugacy remains when normal random effects are included in the linear predictor. The strong conjugacy allows the construction of combined models, i.e. the inclusion of Gaussian random effects and conjugate multiplicative effects in the linear predictor of a GLM. This approach leads to the following definition of the combined beta-binomial model,

$$\kappa_i = \frac{\exp(\mathbf{x}'_i \boldsymbol{\beta} + \mathbf{z}'_i \mathbf{u})}{1 + \exp(\mathbf{x}'_i \boldsymbol{\beta} + \mathbf{z}'_i \mathbf{u})}, \quad (3.42)$$

where  $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{D})$  are the usual Gaussian random effects. However, it ends up that the combined beta-binomial regression does not satisfy the strong conjugacy criteria if the logistic function is used. The authors avoided the strong conjugacy problem by defining the logit model from the probit link function which does satisfy the strong conjugacy. The authors used that

$$\frac{\exp(y)}{1 + \exp(y)} \approx \Phi(cy) \quad (3.43)$$

where  $\Phi(\cdot)$  is the normal cumulative density function and  $c = (16\sqrt{3})/(15\pi)$ .

Summarizing, the definition in Equation (3.42) performs the model assuming a binomial distribution of the outcome and including two different random effects, different in the distribution and in the form they are included in the linear predictor. Therefore, it is straightforward to assess the differences that exists between the model in Equation (3.42) and our proposed BBmm approach. From our point of view it seems more reasonable to assume the marginal beta-binomial distribution as given rather than assuming a conditional binomial distribution. In Chapter 2, we showed that in independent data the marginal beta-binomial regression approach turned to be more appropriate than the conditional approach in terms of the estimation and

interpretation of the effects of the covariates. Therefore, although we do not directly compare our proposal with the model in Equation (3.41), similar results may be obtained due to the conditional feature of the approach. In addition, the extension to the combined models in Equation (3.42) include two random effects that interact in the linear predictor, which may mask the interpretation of the subject-specific realizations and the variance of each random component.

Until now, we have defined mixed-effects model approaches based on the first criteria, i.e. the addition of non-Gaussian random effects in the linear predictor of a GLMM to account for the PROs characteristics. However, there exists in the literature a widely used methodology based on the second criteria, i.e. the assumption of non-exponential family distributions. The so-called generalised additive model for location, scale and shape (GAMLSS) was developed by Rigby and Stasinopoulos (2005) and it is a very flexible methodology which not only deals with a very wide range of distributions, but also allows the inclusion of random effects to accommodate correlation of longitudinal data in all the parameters of the given distribution. The beta-binomial GAMLSS specification corresponds to the proposed BBmm approach, however differences in the estimation procedure lead to different parameter estimates as it will be shown.

Due to the similarity that exists in the beta-binomial mixed-effects model definition between the proposed BBmm and GAMLSS approaches, we will define GAMLSS in more detail in the following lines. The GAMLSS approach models the vector of parameters  $\boldsymbol{\zeta} = (\zeta_1, \dots, \zeta_s)'$  of a general population probability density function  $f(y|\boldsymbol{\zeta})$ . It assumes that, for  $j = 1, \dots, s$ , each distribution parameter  $\zeta_j$  is connected to some given covariates and random effects, and hence, that, for  $i = 1, \dots, n$ , the observations  $y_i$  are independent given those random effects.

Let  $\mathbf{y} = (y_1, \dots, y_n)'$  be the observations of the response variables and, for  $j = 1, \dots, s$ ,  $g_j(\cdot)$  be a known monotonic link function connecting  $\zeta_j$  to the given covariates and random effects through

$$g_j(\boldsymbol{\zeta}_j) = \boldsymbol{\eta}_j = \mathbf{X}_j \boldsymbol{\beta}_j + \sum_{k=1}^{K_j} \mathbf{Z}_{kj} \mathbf{u}_{kj}, \quad (3.44)$$

where  $\boldsymbol{\zeta}_j$  and  $\boldsymbol{\eta}_j$  are vectors of length  $n$ ,  $\boldsymbol{\beta}_j = (\beta_{1j}, \dots, \beta_{p_jj})'$  is the fixed effects parameter vector of length  $p_j$ ,  $\mathbf{X}_j$  is a known design matrix of order  $n \times p_j$  composed by the given covariates,  $\mathbf{Z}_{kj}$  is a fixed known  $n \times q_{kj}$  design matrix composed by the random structure of the model and  $\mathbf{u}_{kj}$  is a random vector of length  $q_{kj}$ . The model

assumes that, for each  $j = 1, \dots, s$  and  $k = 1, \dots, K_j$ , the random effects vector  $\mathbf{u}_{kj}$  has a normal distribution with mean  $\mathbf{0}$  and variance-covariance matrix  $\mathbf{D}_{kj}$ , where  $\mathbf{D}_{kj}$  depends on a vector of hyperparameters  $\boldsymbol{\lambda}_{kj}$ . Notice that for each distribution parameter  $\zeta_j$ ,  $j = 1, \dots, s$ , the random effects vectors  $\mathbf{u}_{kj}$ ,  $k = 1, \dots, K_j$ , can be combined into a single vector  $\mathbf{u}_j = (\mathbf{u}'_{1j}, \dots, \mathbf{u}'_{K_j j})'$  with a single design matrix  $\mathbf{Z}_j = [\mathbf{Z}_1 \cdots \mathbf{Z}_{K_j}]$ . Therefore, the formulation of the GAMLSS approach is now defined as

$$g_j(\zeta_j) = \boldsymbol{\eta}_j = \mathbf{X}_j \boldsymbol{\beta}_j + \mathbf{Z}_j \mathbf{u}_j,$$

where  $g_j(\cdot)$ ,  $\boldsymbol{\eta}_j$ ,  $\boldsymbol{\beta}_j$  and  $\mathbf{X}_j$  have been previously defined in Equation (3.44), but now  $\mathbf{u}_j$  is a random effects vector of length  $q_j = \sum_{k=1}^{K_j} q_{kj}$  and  $\mathbf{Z}_j$  is a known  $n \times q_j$  design matrix of the random effects.

In BBmm approach we have considered that the only parameter that depends on the given covariates is the mean or probability parameter and hence, we have assumed that the dispersion parameter  $\phi$  is equal for all the individuals. Therefore, in terms of the BBmm approach specification, GAMLSS beta-binomial regression model is defined as

$$\begin{cases} g_1(\zeta_1) = \boldsymbol{\eta}_1 = \mathbf{X} \boldsymbol{\beta}_1 + \mathbf{Z} \mathbf{u} \\ g_2(\zeta_2) = \boldsymbol{\eta}_2 = \beta_2 \end{cases} \quad (3.45)$$

where, on the one hand,  $\boldsymbol{\zeta}_1 = (\zeta_{1i}, \dots, \zeta_{1n})'$  is the location parameter assumed different for each individual, which corresponds to the probability parameter of the conditional beta-binomial distribution  $\mathbf{p}$  and  $g_1(\cdot)$  is the logistic link function. On the other hand,  $\zeta_2$  is the scale parameter assumed equal for all the individuals, which corresponds to the dispersion parameter of the conditional beta-binomial distribution  $\phi$  and  $g_2(\cdot)$  is the logarithm link function. As it can be appreciated the model assumptions match with the BBmm approach, however, we will show that there exist remarkable differences in the estimation procedure that lead to misleading conclusions.

Finally, it is worth noticing that Wu et al. (2017) proposed a longitudinal beta-binomial model for overdispersed binomial data where they estimated regression parameters under a probit model using the GEE approach (Zeger and Liang, 1986). It has been stated in Section 3.1.2 that mixed-effects model approaches are more appropriate for analyzing longitudinal data as they are more flexible regarding drop out (missing data), but specially, because they do allow for subject-specific effects which could be quite useful to understand individual variability in the longitudinal

response process and predicting responses for a given subject or a set of individuals from a particular grouping hierarchy. Therefore, as we have mentioned before, we will only focus on mixed-effects regression models for the analysis of longitudinal PROs.

### 3.3.6 Comparison to other approaches

It is straightforward to find out differences between BBmm and previously cited regression approaches where the conditional distribution belongs to the exponential family and non-Gaussian random effects are included in the linear predictor. However, it is not straightforward to appreciate the differences among BBmm and approaches considering non-exponential family distributions, such as GAMLSS. Indeed, as we have mentioned before, the main difference between BBmm and GAMLSS methodologies remains in the estimation process. In fact, while the estimation procedure of BBmm approach is based on a classical full likelihood framework, GAMLSS is based on an empirical Bayesian argument to make the inference.

As it was shown in Section 3.3.5, the GAMLSS approach is able to model all the parameters of the conditional distribution by the inclusion of fixed and random effects in the linear predictors. In particular, the beta-binomial distribution consists of two parameters, the probability parameter  $p$  and the dispersion parameter  $\phi$ . Therefore, GAMLSS approach allows for the modelling of both parameters through fixed and random effects. However, in order to compare like with like the GAMLSS and the BBmm approaches, the linear predictor corresponding to the dispersion parameter is restricted to the inclusion of a unique fixed effect. Let us denote  $\beta^*$  the vector that contains the fixed effects  $\beta_1$ , for modelling the probability parameter, and  $\beta_2$ , for the dispersion parameter (see Equation (3.45) for more details).

The estimation procedure of the fixed and random effects (in all the linear predictors) in GAMLSS methodology is done by means of a posterior mode estimation (Berger, 1985). Indeed, the model assumes that the joint distribution of all the parameters involving the model is given by

$$f(\mathbf{y}, \beta^*, \mathbf{u}, \boldsymbol{\lambda}) = f(\mathbf{y}|\beta^*, \mathbf{u})f(\mathbf{u}|\boldsymbol{\lambda})f(\boldsymbol{\lambda})f(\beta^*),$$

where  $f(\mathbf{y}|\beta^*, \mathbf{u})$  corresponds to the beta-binomial distribution and  $f(\mathbf{u}|\boldsymbol{\lambda})$  is the normal distribution of the random effects, and  $f(\boldsymbol{\lambda})$  and  $f(\beta^*)$  are appropriate prior distributions of  $\boldsymbol{\lambda}$  and  $\beta^*$  parameters respectively. Assuming that the hyperparameters  $\boldsymbol{\lambda}$ , i.e. the variance components of the random effects, are fixed and, assuming

a constant improper prior for  $\beta^*$ , then the posterior distribution for the fixed and random effects is given by

$$f(\beta^*, \mathbf{u} | \mathbf{y}, \boldsymbol{\lambda}) \propto f(\mathbf{y} | \beta^*, \mathbf{u}) f(\mathbf{u} | \boldsymbol{\lambda}). \quad (3.46)$$

Therefore, the estimation of  $\beta^*$  and  $\mathbf{u}$  in GAMLSS approach is done by the joint maximisation of the posterior distribution in Equation (3.46). Notice that the logarithm of the defined posterior distribution for  $\beta^*$  and  $\mathbf{u}$  coincides with the joint log-likelihood in BBmm presented in Equation (3.26), i.e.

$$\log f(\beta^*, \mathbf{u} | \mathbf{y}, \boldsymbol{\lambda}) \propto \log f(\mathbf{y} | \beta^*, \mathbf{u}) + \log f(\mathbf{u} | \boldsymbol{\lambda}) = h(\boldsymbol{\beta}, \boldsymbol{\theta} | \mathbf{y}, \mathbf{u}),$$

where  $\boldsymbol{\beta}$  is the fixed effect in BBmm approach which corresponds to  $\beta_1$  in GAMLSS approach and  $\phi$  (or  $\beta_2$  in GAMLSS) is included in the variance components vector  $\boldsymbol{\theta}$ . Consequently, the estimation of the fixed and random effects in GAMLSS is done by maximising the joint likelihood exactly as in BBmm approach. However, notice that there is a crucial difference. In GAMLSS approach the dispersion parameter  $\phi$  is modelled by a unique fixed effect ( $\exp(\beta_2)$ ), and hence, as the rest of the fixed effects, it is estimated by maximising the joint log-likelihood (posterior likelihood). On the contrary, in BBmm approach the dispersion parameter  $\phi$  is considered as a variance parameter and included in the variance parameter vector  $\boldsymbol{\theta}$ . Lee and Nelder (1996) showed that the estimates of dispersion parameters must be done by a penalisation of the joint likelihood in order to avoid the bias owing to the estimates of the fixed and random effects. The idea is based on the assumption that only parameters that are canonical can be estimated jointly. Therefore, in contrast to GAMLSS, in BBmm approach the joint log-likelihood is penalised before maximisation with respect to  $\boldsymbol{\theta}$ , and hence to  $\phi$ , is performed. Indeed, based on the idea of avoiding the bias of previous estimations of fixed and random effects, even in GAMLSS, the joint log-likelihood (posterior likelihood) is penalised for the estimation of the hyperparameters  $\boldsymbol{\lambda}$ .

### 3.4 Simulation study

In this section, we carry out a simulation study in order to evaluate the performance of BBmm and GAMLSS approaches when analyzing beta-binomial mixed-effects models. In order to perform the simulation study, we use our R-package `PROreg` presented more in detail in Chapter 5, while the GAMLSS approach is implemented



with the `gamlss` R-package version 5.0-2 (Stasinopoulos and Rigby, 2007).

We have generated 100 random realizations of 200 observations of a dependent variable  $Y$ , which conditional on some simulated random effects follows a beta-binomial distribution with fixed maximum number of scores  $m$ , probability parameter  $p$  and dispersion parameter  $\phi$ . For the sake of clarity, we only consider a single covariate in the linear predictor, which follows a normal distribution of mean 1 and standard deviation 2. Consequently, only two fixed effects (in BBmm terminology) have been considered,  $\beta_0 = 1$  and  $\beta_1 = -1.5$ . Furthermore, we have considered 50 realizations of the random effects assuming a normal distribution of mean 0 and standard deviation  $\sigma$ , where each component is randomly connected from 1 to 9 observations. Note that, according to the general notation in Equation (3.22), in this case we consider the vector of the variance components of the random effects  $\boldsymbol{\lambda}$  equal to  $\sigma$  and  $\mathbf{D} = \sigma^2 \mathbf{I}_{50}$ . Therefore, the model is defined as

$$\text{logit}(\mathbf{p}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}$$

where  $\boldsymbol{\beta} = (\beta_0, \beta_1)$ ,  $\mathbf{u}$  is the random effects vector of length 50,

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_{50} \end{pmatrix}_{200 \times 2} \quad \text{where } \mathbf{X}_i = \begin{pmatrix} 1 & x_{i,1} \\ \vdots & \vdots \\ 1 & x_{i,n_i} \end{pmatrix}_{n_i \times 2} \quad \text{and } \mathbf{Z} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{n_2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1}_{n_{50}} \end{pmatrix}_{200 \times 50}$$

where  $\mathbf{1}_{n_i}$  is a column vector of 1s of length  $n_i$ , being  $n_i \in \{1, \dots, 9\}$  and  $\sum_{i=1}^{50} n_i = 200$ . Indeed, this case study can be considered as a longitudinal study where each of the 50 individuals has from 1 to 9 repeated realizations of an event.

The simulation study has been divided in several scenarios depending on the variability of both the random effects and the conditional beta-binomial distribution. We consider three possible values,  $\{0.5, 1, 1.5\}$ , for the dispersion parameters  $\phi$  and  $\sigma$ , and hence, a total of 9 scenarios with all the possible combinations are defined.

Estimates of all the parameters have been obtained with both methodologies. However, only results for the slope,  $\beta_1$ , and the beta-binomial dispersion parameter  $\phi$  will be shown due to the following reasons. First, for simplicity, clarity and brevity, given that including the same analysis for all the parameters can make the reading quite dense and mask relevant conclusions. Second, to study the relationship between the outcome variable with individual or patient characteristics has

become one of the primary aims of many longitudinal studies in medical or biological framework. Consequently, we consider that the slope,  $\beta_1$ , is the focus of many studies, and hence, it needs an in-depth analysis in order to check the efficiency of the algorithms. Finally, as we have mentioned in Section 3.3.6, the main difference between the estimation approaches lies in the estimation of the dispersion parameter of the conditional beta-binomial distribution. Hence, in order to look for differences between the algorithms, we also show a detailed analysis of the estimation of  $\log(\phi)$ . Nevertheless, adequacy of the estimates for the other parameters has been also checked.

Table 3.1 shows the results of the simulations for the mentioned parameters. On the one hand, the mean, ASD, ESD, EMS and coverage probability of the 95% Wald confidence intervals for the estimates of the slope  $\beta_1$  are shown. On the other hand, the mean, ASD, ESD and EMS of the estimates of the logarithm of the conditional beta-binomial dispersion parameter  $\phi$  are also shown for the defined scenarios.

Results show that the ASD and ESD in BBmm approach remain quite similar in all the scenarios. However, as regards to the GAMLSS approach, results are more contradictory. Even in the lowest dispersion case, i.e.  $\sigma = 0.5, \phi = 0.5$ , the ASD doubles the value of the ESD. Hence, GAMLSS approach inflates the standard deviation of the estimates and consequently, confidence intervals for the slope parameter are larger than they should, which makes the coverage probability pointless. Furthermore, although in GAMLSS confidence intervals are over-estimated, it can be appreciated in Table 3.1 that as variance parameter  $\sigma$  increases BBmm approach gets much better results in terms of coverage percentage of the slope. For instance, in the largest variability scenario ( $\phi = 1.5$  and  $\sigma = 1.5$ ) the coverage probability of the GAMLSS approach is equal to 33%, while in the BBmm approach it is 83%. Therefore, we can conclude that as the dispersion parameter of the random effects increases, results provided by the GAMLSS approach worsen in terms of the significance of the parameters.

Regarding the EMS of the parameters, Table 3.1 shows that when the dispersion of the random effects is low ( $\sigma = 0.5$ ) both methodologies perform similarly. However, when the variance parameter  $\sigma$  increases there exist differences between the approaches. EMS in the BBmm approach remains quite constant in all the scenarios, nevertheless in the GAMLSS approach they increase as the dispersion parameter of the random effects increases.

Figure 3.3 shows the estimates of parameters  $\beta_1$  and  $\log(\phi)$  for the different scenarios. In the right panel, we can observe the resulting estimations for the slope

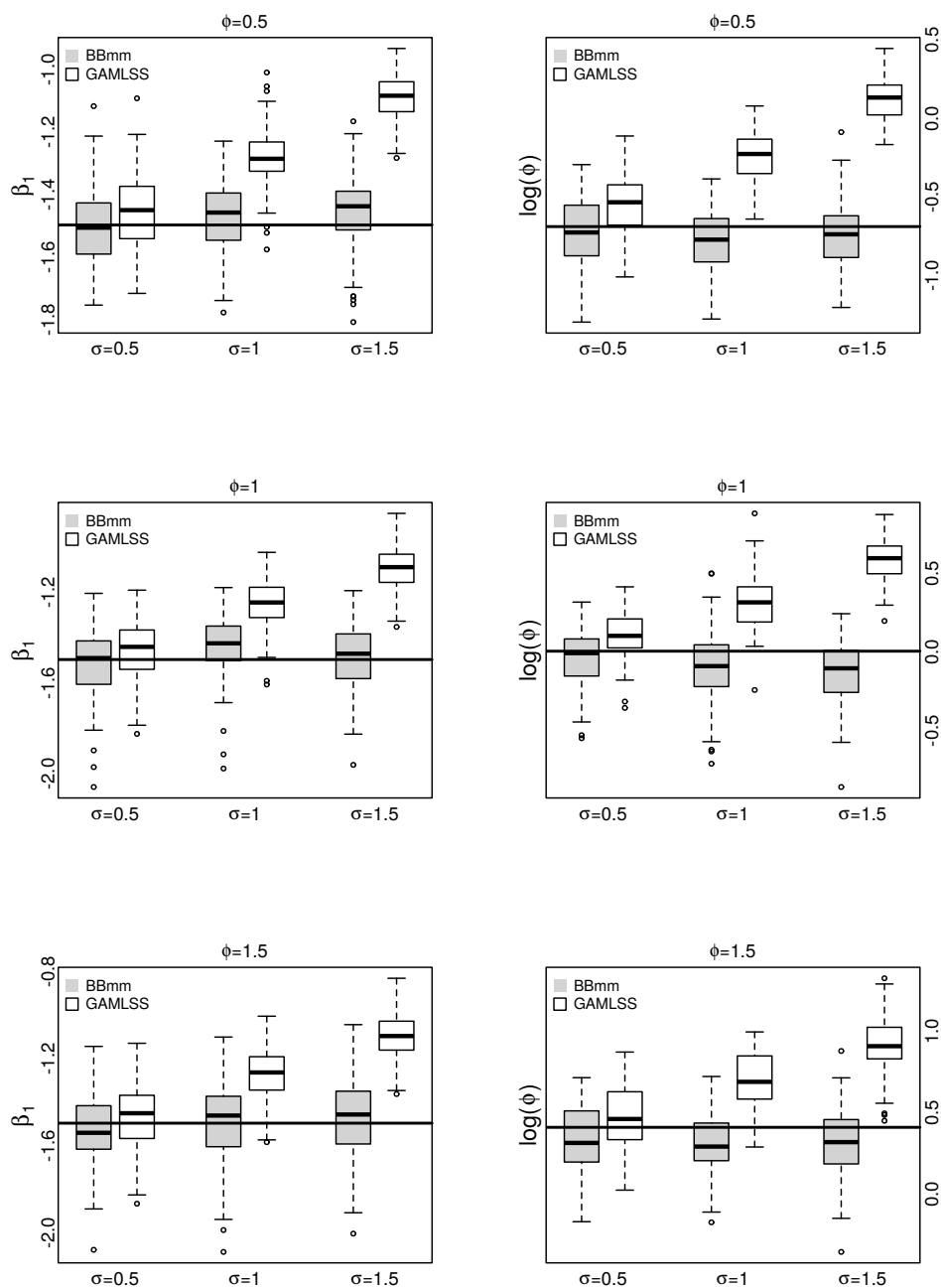


Figure 3.3: Boxplots of the estimates of the parameters  $\beta_1$  and  $\log(\phi)$ . In the right panel the plots show the estimates for  $\beta_1$ , while in the left panel estimates for  $\log(\phi)$  appear. Each group of boxplots describes three scenarios corresponding to the all 9 scenarios.

Table 3.1: Results for 100 converged simulations with 200 number of observations.

Scenarios			$\beta_1$					$\log(\phi)$			
$\sigma$	$\phi$	Method	Mean	ESD	ASD	EMS	CP	Mean	ESD	ASD	EMS
0.5	0.5	BBmm	-1.500	0.127	0.110	0.016	92	-0.727	0.232	0.162	0.055
		GAMLSS	-1.449	0.120	0.244	0.017	99	-0.549	0.195	0.395	0.059
	1	BBmm	-1.520	0.151	0.131	0.023	91	-0.051	0.182	0.163	0.036
		GAMLSS	-1.456	0.128	0.256	0.019	100	0.106	0.143	0.400	0.032
	1.5	BBmm	-1.530	0.166	0.143	0.028	88	0.323	0.208	0.171	0.050
		GAMLSS	-1.469	0.147	0.255	0.023	100	0.466	0.186	0.410	0.038
1	0.5	BBmm	-1.472	0.113	0.105	0.014	91	-0.774	0.193	0.165	0.044
		GAMLSS	-1.280	0.096	0.215	0.058	95	-0.254	0.151	0.387	0.215
	1	BBmm	-1.441	0.145	0.121	0.025	85	-0.091	0.227	0.164	0.060
		GAMLSS	-1.255	0.112	0.214	0.073	90	0.311	0.171	0.394	0.126
	1.5	BBmm	-1.494	0.180	0.137	0.032	88	0.296	0.180	0.174	0.044
		GAMLSS	-1.279	0.126	0.217	0.065	89	0.686	0.169	0.408	0.107
1.5	0.5	BBmm	-1.462	0.117	0.100	0.015	90	-0.752	0.217	0.171	0.051
		GAMLSS	-1.092	0.074	0.182	0.172	30	0.121	0.137	0.382	0.681
	1	BBmm	-1.483	0.142	0.119	0.021	93	-0.314	0.198	0.171	0.057
		GAMLSS	-1.097	0.093	0.186	0.171	35	0.585	0.134	0.394	0.361
	1.5	BBmm	-1.469	0.180	0.132	0.033	83	0.309	0.208	0.178	0.052
		GAMLSS	-1.098	0.114	0.187	0.174	33	0.904	0.165	0.407	0.275

ESD: Empirical Standard Deviation; ASD: Average Standard Deviation; EMS: Expected Mean  
Square errors; CP: Coverage Probability of 95%.

$\beta_1$ , while in the left panel estimates for the logarithm of the beta-binomial dispersion parameter  $\phi$  are shown. It is worth mentioning that BBmm approach gives similar bias of the slope parameter in all the scenarios, and hence, we can conclude that it is a stable methodology as the variance of the data do not alter its performance. However, it can be noticed that as the variance parameter  $\sigma$  increases, the bias of the estimate of the slope parameter in GAMLSS increases as well. Besides, similar results can be obtained from the estimation of the dispersion parameter of the beta-binomial distribution  $\phi$ . Apparently, the increase of  $\sigma$  do not affect the estimations of  $\phi$  through BBmm approach, but, on the contrary, it worsens the performance of GAMLSS approach considerably.

In summary, we have shown that, not only estimates of the slope, but also estimates of the dispersion parameter  $\phi$  are more accurate in terms of bias and

expected mean square error in BBmm approach than in GAMLSS. In addition, we also have evidenced the over-estimation of the standard deviation of the estimates in GAMLSS approach, which could conclude on inappropriate statistical inference. Consequently, based on the results of the simulation study we can state that, although both methodologies perform similarly with low variance parameter  $\sigma$ , estimates with BBmm approach are better in general as  $\sigma$  increases. Hence, we propose the use of BBmm methodology as an unified way to analyse real data by a beta-binomial mixed-effect model.

### 3.5 Application to real data

In this section, we apply the developed beta-binomial mixed-effects model approach to the data introduced in Chapter 1.

First of all, as an illustration of the applicability of the proposed regression approach, we analyse the Paquid research programme data (see Section 1.3.2). As it was mentioned, the Paquid research programme was developed to measure the cognitive status and incidence of dementia and Alzheimer's disease in elderly people in South-Western France. It contains MMSE measurements which have already been fully described in Section 1.2.3. For the real application, we use a subsample of the Paquid dataset which is freely available at `1cmm` R-package (Proust-Lima et al., 2017). In order to show the adequacy of the proposed methodology, we validate the obtained results in terms of the effects of the covariates in the cognitive status comparing them with the results in the literature when analysing the incidence of dementia and Alzheimer.

Finally, once the applicability of the proposed methodology has been illustrated by the analysis of the Paquid data, we focus our attention on the COPD Study introduced in Section 1.3.1. However, compared with the cross-sectional analysis of the COPD Study in Chapter 2, in this case, we will analyse longitudinal measurements provided by the SGRQ which was introduced in Section 1.2.2. Therefore, we will not only restrict the analysis to the first measurement of each individual and instead, all the longitudinal features of the study will be considered.

#### 3.5.1 Mini Mental Score Examination: Paquid Research Programme

It is well known that MMSE scores have a skewed distribution, accumulating values at one or both edges (0 and 30) of the scale (Folstein et al., 1988). In fact, we

have shown in Figure 1.7 in Section 1.4.3, that the distribution of the MMSE scores of the Paquid program subsample tend to accumulate in the right side of the scale, displaying a J-shaped curve, which exponential family distributions are not able to fit appropriately. Instead, the beta-binomial distribution has been shown to be a good candidate due to its distributional flexibility allowing for a reasonable fit compared to the exponential family distributions. Therefore, we consider the analysis of the Paquid research programme using a mixed-effects regression model based on the beta-binomial distribution. The data were collected in a longitudinal context where the measurements of each patient were observed along time. This process led to the so-called repeated measurements and consequently, to non-independent structures in the data. Hence, in order to accommodate the correlation that may exist among longitudinal measurements provided by each patient, we include individual level random effects in the linear predictor of the model. That way, we assume that the  $n_i$  observations of the  $i$ th patient are iid drawn from a beta-binomial distribution. Following BBmm approach, we link the probability parameter of the beta-binomial distribution with the given covariates and random effects by means of a logit link function. In other words, we assume that,

$$\begin{aligned} y_{ij}|u_i &\sim \text{BB}(m_i, p_{ij}, \phi) \quad \text{indep. } j = 1, \dots, n_i \\ \eta_{ij} &= \text{logit}(p_{ij}) = \mathbf{x}'_{ij}\boldsymbol{\beta} + u_i, \end{aligned} \tag{3.47}$$

where  $\boldsymbol{\beta}$  are the fixed effects,  $\mathbf{x}_{ij}$  is a row of the full rank design matrix composed by the covariates and  $u_i$  is the random component attributed to the  $n_i$  observations of the  $i$ th individual,  $i = 1, \dots, n$ ,  $n = 498$  and  $j = 1, \dots, n_i$ . Hence, we assume that observations from different individuals are independent, while the observations from the same individual are connected through a random component. Therefore, we propose a random intercept model where the random effects vector  $\mathbf{u} = (u_1, \dots, u_n)'$  is drawn from a multivariate normal distribution with mean  $\mathbf{0}$  and variance-covariance matrix  $\mathbf{D} = \sigma_u^2 \mathbf{I}_n$ .

If we consider the latent variable interpretation described in Section 3.3.4, we can assume that a continuous latent variable  $Y_{ij}^*$  underlines the observed  $y_{ij}$  for the  $i$ th individual in time  $j$ ,  $i = 1, \dots, n$  and  $j = 1, \dots, n_i$ . That way, the variance of the random effects  $u$ ,  $\sigma_u^2$ , represents the between-subject variation of the model, and the variance of the logistic distribution of the errors in the model for  $y_{ij}^*$ ,  $\pi^2/3$ , represents the variation within subjects (McCullagh and Nelder, 1989). Consequently, correlation between two latent observations of the same individual,  $y_{ik}^*$  and  $y_{ij}^*$ , is

defined as  $\rho = \sigma_u^2 / (\sigma_u^2 + \pi^2/3)$ .

Table 3.2: Results for the application of the random intercepts beta-binomial mixed-effects model in the Paquid research programme subsample.

Covariates	Levels	Estimate	SD	p-value
(Intercept)	-	3.580	0.216	<0.001
Dependency level	<i>no</i>	-	-	-
	<i>mild</i>	-0.094	0.043	0.027
	<i>moderate</i>	-0.358	0.046	<0.001
	<i>severe</i>	-1.145	0.061	<0.001
Age	-	-0.050	0.003	<0.001
Age at dementia diagnosis	-	0.043	0.003	<0.001
Dementia diagnosis	<i>no</i>	-	-	-
	<i>yes</i>	-0.570	0.028	<0.001
Age at the entry of the cohort	-	-0.015	0.003	<0.001
Educational level	<i>no</i>	-	-	-
	<i>yes</i>	0.604	0.028	<0.001
$\log(\phi)$	-	-6.567	0.806	-
$\sigma$	-	0.615	0.043	-

SD: Standard deviation.

Table 3.2 illustrates the results after the application of the BBmm approach in the Paquid data subsample. It can be appreciated that all the covariates in the model are statistically significant, concluding that they alter the cognitive status of patients. The standard deviation of the heterogeneity between the individuals at baseline is 0.615. The total deviance of the model is 1898.1, and the null deviance or deviance of the null model is equal to 6791.1, where the p-value associated with the goodness-of-fit deviance test is lower than 0.001. Several conclusions can be obtained from the clinical interpretation of data in Table 3.2. First, as it was expected, a more severe dependency status worsens the cognitive status of the subject. Second, the increase of the age also worsens the cognitive mental status of the patients. Indeed, an increase of 5-year increases by  $1/\exp(-5 \times 0.050) = 1.284$  the odds of having a smaller MMSE score or worse cognitive status. Moreover, both the diagnosis of

dementia ( $\beta_{\text{dem}} = -0.570$ ) and a later entry in the study ( $\beta_{\text{age.init}} = -0.015$ ) also increase the risk of having worse cognitive status (lower MMSE score). However, a later dementia diagnosis ( $\beta_{\text{age.dem}} = 0.043$ ) and having at least primary school educational level ( $\beta_{\text{CEP}} = 0.604$ ) increases the odds of having a larger MMSE score, i.e. a better cognitive status. The estimated standard deviation of the random effects is 0.615, which means that the correlation between latent observations of the same individual is  $0.615^2 / (0.615^2 + \pi^2/3) = 0.103$ . Additionally, Table 3.2 shows that the estimation of the dispersion parameter of the beta-binomial distribution is  $\phi = 0.001$  ( $\log(\phi) = -6.567$ ), indicating that dispersion of the outcome data decreases when the covariates are included in the model (the beta-binomial distributional fit performed in Chapter 1 led to  $\phi = 0.113$ ).

The obtained results are similar to the results available in the literature regarding the effect of the analysed covariates in the MMSE scores. For instance, in other study, Matallana et al. (2011) stated that there exists a correlation between the educational level and the cognitive status of the patients and, in fact, they showed that a higher educational level is related with a higher score in the MMSE questionnaire.

On the whole, we have applied the proposed methodology to the Paquid research programme and we have reached clinically valid results which are in line with the available literature for the analysis of the MMSE questionnaire. In addition, regarding the interpretation of the covariate effect, we have shown that the BBmm approach offers an easy quantification of the effects in terms of odds-ratio which is very familiar to clinical researchers.

### 3.5.2 St. George's Respiratory Questionnaire: COPD Study

One of the main objectives of the COPD Study was to assess the evolution of the health-status of the patients during the cohort. Although in the COPD Study longitudinal data was picked up by means of two different PRO questionnaires, we will only show the analysis of the lung disease specific, the SGRQ (see Section (1.2.2)). We believe that results for one instrument are enough to illustrate practical validity of the proposed methodology. Moreover, pneumologists were interested in measurements provided by the SGRQ due to the fact that it is an specific respiratory questionnaire, and therefore, it could show the clinical evolution of patients more precisely than the SF-36. They were interested to know if there is a substantial (statistically and clinically significant) worsening, or at least a change, in the health-status of COPD patients as time goes by, and in that case, quantify the evolution



and detect variables related to it. More detailed results over time will be discussed jointly with clinical researchers in a future work.

In Section 1.2.2, we have described the SGRQ, which decomposes the health-status of the patients in three dimensions: *activity*, *symptoms* and *impacts*. We have mentioned that originally, each dimension is bounded in a 0 – 100 scale. However, as it was displayed in Figure 1.6 the normal distribution does not offer an appropriate fit of the scores, and additionally, it does not match with the boundary condition, leaving tails out of the 0 – 100 scale. Therefore, we have proposed a recoding procedure of the scores in order to apply the beta-binomial distribution. Based on the result provided by Jones (2005), there exists a clinically significant change in the score only if a four points threshold is exceeded. Consequently, we have divided the SGRQ scores in four length subintervals and then recode them as it was shown in Table 1.5 in Chapter 1.

As regards to the statistical methodology, we will use the BBmm approach to assess the health evolution of the COPD patients. For that purpose, for each dimension a longitudinal study is carried out where subject-specific random intercepts and slopes are included in the model. In fact, the model is defined as follows: assume that we have  $n$  individuals where each individual provides  $k_i$  measurements over time,  $i = 1, \dots, n$ . We denote as  $y_{ij}$  the observation of the  $i$ th individual in the  $j$ th time point which, conditioned by the random effects  $\mathbf{u}$  and  $\mathbf{v}$ , is iid drawn from a beta-binomial distribution with parameters  $m_i$ ,  $p_{ij}$  and  $\phi$ ,  $i = 1, \dots, n$  and  $j = 1, \dots, k_i$ . In addition, we assume that we measure the day each response was provided and include them in a year-scale in the time covariate  $t$ . Finally, we apply the following model to the data

$$\eta_{ij} = \log \left( \frac{p_{ij}}{1 - p_{ij}} \right) = (\beta_0 + u_i) + (\beta_1 + v_i)t_{ij}, \quad (3.48)$$

where  $\eta_{ij}$  is the linear predictor of the model,  $\beta_0$  and  $\beta_1$  are the regression parameters,  $u_i \sim \mathcal{N}(0, \sigma_u)$  and  $v_i \sim \mathcal{N}(0, \sigma_v)$  are the random effects of the model and  $p_{ij}$  is the probability parameter of the beta-binomial distribution  $i = 1, \dots, n$ ,  $j = 1, \dots, k_i$ . For the sake of clarity, we are going to describe the meaning of each parameter in Equation (3.48):

- holding the contribution of each individual  $u_i$  and  $v_i$ ,  $\beta_0$  stands for the intercept of the regression, i.e. the expected health-status of the population at baseline.
- holding the contribution of each individual  $u_i$  and  $v_i$ ,  $\beta_1$  stands for the slope

of the regression, i.e. the expected evolution of the health-status of the population.

- $u_i$  represents the difference of the population's and  $i$ th individual's health-status at baseline.
- $v_i$  represents the difference of the health-status evolution between the  $i$ th individual and the population.
- $\sigma_u$  determines the standard deviation of the health-status of the individuals at baseline.
- $\sigma_v$  determines the standard deviation that exists in the health-status evolution between the patients.

Table 3.3 shows the results of the model in Equation (3.48) applied to the three SRGQ dimensions of the COPD Study. It displays the estimation, standard deviation and p-values associated with the estimation of each regression parameter. Additionally, it shows the odds-ratio for the estimation of the slope. Table 3.3 also displays the estimation and standard deviation of the dispersion parameters of the model  $\sigma_u$ ,  $\sigma_v$  and  $\log(\phi)$ .

It must be mentioned that the algorithm failed to converge showing that the estimation of  $\sigma_v$  tended to zero for the *impacts* dimension. Therefore, we adjusted the model in Equation (3.48) removing the random slopes effects from the linear predictor and evaluated it again in the mentioned dimension. Consequently, the estimation of  $\sigma_v$  is not displayed in Table 3.3 for *impacts*.

Regarding the interpretation of the results, Table 3.3 shows that the evolution of the patients differs from one dimension to the others. For the *activity* component of the SGRQ, the estimated coefficient of time in years is 0.027. Therefore, the odds-ratio of having a worsening in *activity* dimension is 1.027 ( $\exp(0.0270) = 1.027$ ) for each year of evolution. In terms of the interpretation, each year of evolution is associated with an odds-ratio of 1.028 for a worse *activity*, in other words, each year of evolution is associated with a 2.8% worsening in *activity*, which is statistically significant ( $p = 0.001$ ). Moreover, the standard deviation of the random effects are  $\sigma_u = 1.180$  and  $\sigma_v = 0.123$  for the random intercepts and slopes respectively. On the one hand, it means that centred in  $-0.186$  the standard deviation of the scores of the individuals in *activity* is 1.180 at baseline. On the other hand, as time goes by, and centred in 0.027, the slopes of the regression (evolutions) for different individuals have a standard deviation equal to 0.123.

Table 3.3: Results for the beta-binomial longitudinal model applied to the three SGRQ dimensions provided in the COPD Study.

Dimension	Covariate	Estimate	SD	OR	p-value
<i>Activity</i>					
	(Intercept)	-0.186	0.026	–	<0.001
	Year	0.027	0.008	1.027	0.001
	$\sigma_u$	1.180	0.037	–	–
	$\sigma_v$	0.123	0.006	–	–
	$\log(\phi)$	-3.295	0.086	–	–
<i>Symptoms</i>					
	(Intercept)	-0.306	0.029	–	<0.001
	Year	0.007	0.009	1.007	0.493
	$\sigma_u$	0.760	0.025	–	–
	$\sigma_v$	0.093	0.006	–	–
	$\log(\phi)$	-2.610	0.057	–	–
<i>Impacts</i>					
	(Intercept)	-1.320	0.023	–	<0.001
	Year	0.003	0.007	1.003	0.673
	$\sigma_u$	0.972	0.030	–	–
	$\log(\phi)$	-4.050	0.115	–	–

SD: Standard deviation; OR: Odds-ratio.

For the *symptoms* component of the SGRQ, the estimated coefficient of time in years is 0.007. Therefore, the odds-ratio of having a worsening in the *symptoms* dimension is 1.007 ( $\exp(0.007) = 1.007$ ) for each year of evolution. In terms of the interpretation, each year of evolution is associated with an odds-ratio of 1.007 for worse *symptoms*, in other words, each year of evolution is associated with a 0.7% worsening in *symptoms*, which is not statistically significant ( $p = 0.493$ ). Regarding the variability of the random effects, on the one hand, the standard deviation of the random intercepts or status at baseline is 0.760, which shows less variability than *activity*. On the other hand, centred in 0.007, the standard deviation of the random slopes (evolutions) is 0.093.

Finally, for the *impacts* component, the estimated coefficient of time in years is 0.003. Therefore, the odds-ratio of having a worsening in the *impacts* dimension is 1.003 ( $\exp(0.003) = 1.003$ ) for each year of evolution. In terms of the interpretation, each year of evolution is associated with an odds-ratio of 1.003 for worse impacts, in

other words, each year of evolution is associated with a 0.3% worsening in *impacts*, which is not statistically significant ( $p = 0.673$ ). As it has been mentioned, the standard deviation of the random slopes was equal to zero, meaning that there is no variability in the evolution among patients. Therefore, the model is implemented with a single random component (random intercepts) whose standard deviation is equal to 0.972. In conclusion, in the *impacts* dimension the expected population evolution is not statistically significant and, moreover, there are no differences between the evolution trend of different patients.

It must be mentioned that the standard deviations of the random slopes in both *activity* ( $\sigma_v = 0.123$ ) and *symptoms* ( $\sigma_v = 0.093$ ) dimensions are much larger than the expected population evolution ( $\hat{\beta}_1 = 0.027$  and  $\hat{\beta}_1 = 0.007$  respectively). Consequently, although in *activity* the population evolution is expected to worsen, there are even patients that improve their health-status. In *symptoms* we may not expect any change in the status of patients however, due to the previous argument, there are patients that improve their quality of life while others deteriorate. In fact, this shows the variability that exists in the PRO analysis due to the implicit subjective nature of the outcomes. Figure 3.4 displays the expected population evolution and the evolution of each patient in *activity* dimension. It is straightforward to notice in the figure the variability that exists between the baseline health-status and evolution in the data.

Once the unadjusted evolution of the patients has been assessed, we will fit the same longitudinal model described in Equation (3.48), but adjusted by covariates. In Chapter 1, we have mentioned that Esteban et al. (2016) divided the patients participating in the COPD Study in four subtypes, where each subtype was associated with different health-status and characteristics. Therefore, in order to measure the evolution over time of the patients taking into account the special characteristics they may have, we are going to fit the following model,

$$\eta_{ij} = \log \left( \frac{p_{ij}}{1 - p_{ij}} \right) = (\beta_0 + u_i + \beta_1 \mathbf{S}_i) + (\beta_2 + v_i + \beta_3 \mathbf{S}_i) t_{ij}, \quad (3.49)$$

where  $\mathbf{S}_i$  is a vector consisting of three binary variables which indicate the subtype the  $i$ th patient belongs and  $\beta_1$  and  $\beta_3$  consist of 3 components for the B, C and D subtypes (the A subtype has been taken as reference). As it can be appreciated in Equation (3.49), we do not consider the option that individuals transit over subtypes as the objective is to assess the evolution of patients that belonged to a specific subtype at baseline.

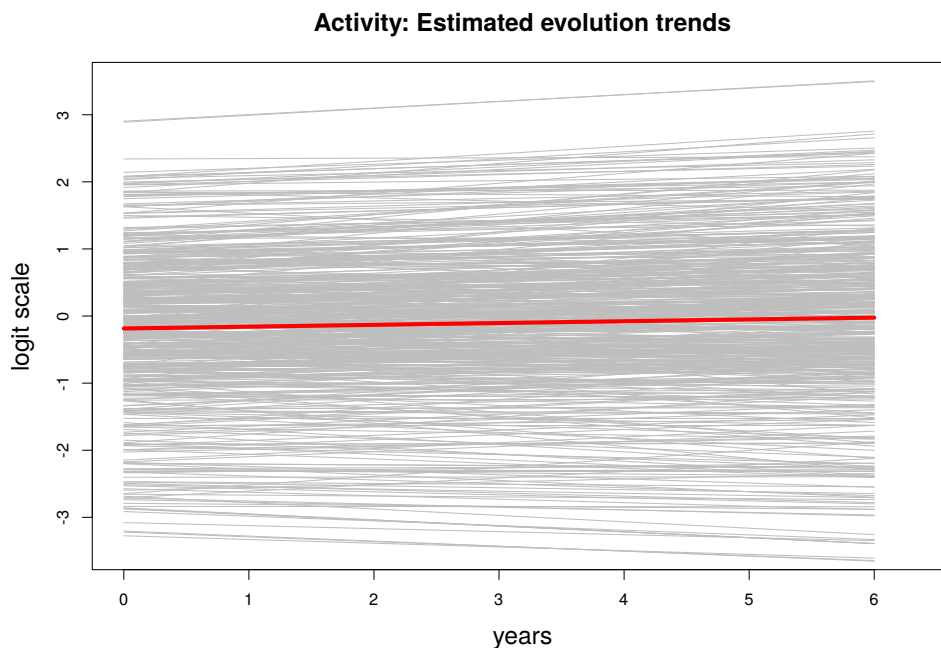


Figure 3.4: Estimated evolution trends in *activity* dimension for the individuals in the cohort. The red line stands for the expected evolution, whereas the grey lines refer to each individual.

Table 3.4 displays the estimation, standard deviation and p-values associated with the estimation of each regression parameter. Additionally, it also shows the exponential of the fixed effects coefficients that could be directly interpreted as odds-ratios. Table 3.4 also displays the estimation and standard deviation of the dispersion parameters of the model  $\sigma_u$ ,  $\sigma_v$  and  $\log(\phi)$ . Regarding the interpretation of the results, Table 3.4 shows that the evolution of the patients differs from one dimension to the others.

Several conclusions can be obtained from Table 3.4. For instance, it is worth mentioning that the only dimension where there were differences between the evolution of patients within subtypes was *symptoms*. In the other two, patients in each subtype evolved equally over time hence, we did not include random slopes as  $\sigma_v$  tended to zero in the **BBmm** algorithm. Table 3.4 shows that although there are statistically significant differences between the subtypes at baseline in all the dimensions, all of them evolve equally. Moreover, the only dimension where the evolution of patients is statistically significant is *activity* (p-value= 0.038 for subtype

Table 3.4: Results for the longitudinal beta-binomial model including cluster classification applied to the three SGRQ dimensions provided in the COPD Study.

Dimension	Covariate	Estimate	SD	OR	p-value
<i>Activity</i>					
	(Intercept)	-0.900	0.050	–	< 0.001
	Year	0.031	0.015	1.031	0.038
	Subtype B	0.787	0.066	2.197	< 0.001
	Subtype C	1.801	0.089	6.056	< 0.001
	Subtype D	0.780	0.085	2.18	< 0.001
	Year $\times$ Subtype B	0.013	0.021	–	0.530
	Year $\times$ Subtype C	-0.023	0.029	–	0.422
	Year $\times$ Subtype D	-0.012	0.029	–	0.691
	$\sigma_u$	1.032	0.032	–	–
	$\log(\phi)$	-3.013	0.073	–	–
<i>Symptoms</i>					
	(Intercept)	-0.561	0.053	–	< 0.001
	Year	0.004	0.016	1.004	0.825
	Subtype B	0.282	0.071	1.326	< 0.001
	Subtype C	0.740	0.090	2.096	< 0.001
	Subtype D	0.189	0.091	1.208	0.037
	Year $\times$ Subtype B	0.026	0.022	–	0.239
	Year $\times$ Subtype C	-0.033	0.030	–	0.263
	Year $\times$ Subtype D	-0.001	0.031	–	0.985
	$\sigma_u$	0.726	0.024	–	–
	$\sigma_v$	0.096	0.005	–	–
	$\log(\phi)$	-2.619	0.058	–	–
<i>Impacts</i>					
	(Intercept)	-1.485	0.050	–	< 0.001
	Year	0.010	0.015	1.010	0.503
	Subtype B	0.497	0.065	1.644	< 0.001
	Subtype C	1.304	0.078	3.684	< 0.001
	Subtype D	0.563	0.082	1.756	< 0.001
	Year $\times$ Subtype B	0.013	0.020	–	0.510
	Year $\times$ Subtype C	-0.047	0.025	–	0.063
	Year $\times$ Subtype D	-0.013	0.028	–	0.643
	$\sigma_u$	0.983	0.031	–	–
	$\log(\phi)$	-3.547	0.095	–	–

SD: Standard deviation; OR: Odds-ratio. Subtype A was stated as reference.

A and differences between subtypes were not statistically significant). For instance, considering the *activity* dimension the odds-ratios for worse initial activity were 2.2, 6.2 and 2.2 for subtypes B, C and D versus subtype A being all of them statistically significant. As regards to evolution, the exponential of the fixed effects coefficients cannot be interpreted directly as odds-ratios, due to the presence of interaction

terms in the model. As an example, each year of evolution is associated with a 3.1% of worsening activity for subjects on subtype A ( $OR = \exp(0.031) = 1.031$ ), whereas it is associated with a 4.5% of worsening activity for subjects on subtype B ( $OR = \exp(0.031 + 0.013) = 1.045$ ), although this difference was not statistically significant ( $p\text{-value} = 0.530$ ). The rest of the effects would be interpreted in a similar way.

Therefore, in general terms, we can state that the evolution of COPD patients does not depend on the initial health-status defined by the cluster classification, as all of them evolve similarly over time.

### 3.6 Conclusions and discussion

Mixed-effects regression models are a useful technique to deal with many types of data with hierarchical or non-independent structures. LMMs (see Section 3.2.2), but specially GLMMs (see Section 3.2.3) are in practice the most used methodologies for analysing longitudinal or correlated data. GLMMs restrict their conditional assumptions to exponential family distributions, which cover most of the real scenarios, and include Gaussian random effects to accommodate the hierarchical structure of the data. In Chapter 1, we have shown that PROs consist of some specific characteristics which make the fit by exponential family distributions inadequate. In addition, regression based on a mixture between two exponential family distributions, the beta-binomial distribution, has been proposed as a good alternative to analyse PROs in an independent data framework. However, the beta-binomial distribution does not belong to the exponential family and, consequently, GLMM estimation and inference theory cannot be directly applied.

In Chapter 2, we have proposed the use of BBreg approach for the analysis of independent PROs in a regression context. In fact, we have shown through a simulation study that BBreg obtains appropriate results in terms of parameter estimation and variance when applied to highly dispersed binomial data. However, the model does not allow the analysis of dependent outcomes and hence, it cannot be applied in correlated data, longitudinal studies for instance. Therefore, in order to provide a solution to non-independence, in this chapter we have developed the BBmm approach. Basically, the BBmm is the extension of the BBreg to the inclusion of Gaussian random effects in the linear predictor of the model. The random effects accommodate the correlation that could exist between different outcomes which allows BBmm approach to provide a new insight to perform hierarchical regression

analysis based on the beta-binomial distribution. We have developed an estimation algorithm based on the first order Laplace approximation for the random and fixed effects and a penalised profile likelihood for the dispersion parameters of the model.

Based on a similar approach, the assumption of a conditional beta-binomial distribution, GAMLSS performs the model defined in Equation (3.23). GAMLSS is a general methodology that does not restrict the distributional assumption to the exponential family and instead, any type of density function can be analysed in a mixed-effects approach. GAMLSS can model all the parameters involving a given distribution and, as it was explained in Section 3.3.6, it is based on an empirical Bayesian argument to make the inference. For instance, GAMLSS links the probability parameter  $p$  and the dispersion parameter  $\phi$  of the beta-binomial distribution with some given covariates and random effects by means of logit and log link functions, respectively. For the estimation of the fixed and random effects in the model, the joint distribution of all the parameters in the model is maximised. However, the estimation procedure does not make any difference between fixed effects belonging to the probability parameter or the dispersion parameter. Consequently, compared to BBmm approach, GAMLSS does not penalise the likelihood function when estimating  $\phi$ , avoiding the property that only canonical parameters should be estimated jointly. In order to compare the performance of both regression approaches, a simulation study has been carried out in some controlled scenarios. Results showed that the penalisation of the likelihood when estimating the dispersion parameter  $\phi$  improves the results in terms of bias and coverage percentage. In this chapter, we have focused on the development of a new estimating procedure for a mixed-effects regression model based on the beta-binomial distribution. However, as it was mentioned, GAMLSS does not restrict the conditional assumption to a unique distribution or family of distributions. Therefore, although simulation only for the beta-binomial distribution have been performed, the penalisation of the likelihood to estimate non-canonical parameters can be useful in other situations.

On the contrary, based on a conditional approach, there exist in the literature two methodologies that can deal with highly dispersed hierarchical binomial data. On the one hand, in Chapter 2 we have introduced the HGLM, which can be extended from a cross-sectional to a hierarchical framework. Avoiding the comparison between the BBreg and HGLM approaches carried out in Chapter 2, we are now going to focus our discussion on the interpretation of the random effects when the binomial-beta HGLM is extended to the inclusion of Gaussian random effects. For instance, assume that we perform a random intercepts model such as in the Paquid application



in Section 3.5.1. The beta-binomial-normal model with random intercepts for the  $j$ th observation of the  $i$ th individual will be defined as

$$\begin{aligned} Y_{ij}|u_i, \omega_i &\sim \text{BB}(m_i, p_{ij}, \phi) \\ \eta_{ij} = \text{logit}(p_{ij}) &= \mathbf{x}'_{ij}\boldsymbol{\beta} + v_i + \omega_i, \end{aligned} \quad (3.50)$$

where  $v_i = \text{logit}(u_i/(1 - u_i))$  and  $u_i \sim \text{Beta}(1/\alpha, 1/\alpha)$  are the same as in the cross-sectional model, but we include  $\omega_i \sim \mathcal{N}(0, \sigma_\omega)$  random effects. Under this model,  $v_i$  is the random effect that accounts for the overdispersion in the binomial data, whereas  $\omega_i$  accommodates the hierarchical structure. However, it seems that the model could have identifiability problems as both random effects affect the linear predictor in the same way. Additionally, the interpretation of the standard deviation of the intercept random effect, as the heterogeneity at baseline, can be masked due to the fact that results are obtained conditioned on the transformation of the beta random effects.

On the other hand, Molenberghs et al. (2010) defined the so-called combined models which include conjugate and normal random effects. However, unlike the HGLM approach, the conjugate random effects do not enter additively in the linear predictor and instead, they multiply its transformation through the antilogit function (see Section 3.3.5). Compared to BBmm model definition in Equation (3.23), Molenberghs et al. (2012) stated that the parameters in the combined model have a different meaning, as they are interpreted conditional on the assumed random-effects structure. In addition, differences may be very noticeable when binomial measurements are collected repeatedly over time or in any other hierarchical fashion. Indeed, from our point of view, not only the interpretation of the fixed effects, but specially the interpretation of the random effects is much more intuitive in the BBmm approach as it follows classical GLMM perspective. If we fit a random intercept combined beta-binomial-normal model based on Equation (3.41) and Equation (3.42), similar to the HGLM approach, we would not be able to interpret the normal random realizations as individual baseline differences because another beta random realization multiplies the transformation of the linear predictor. Moreover, the standard deviation of the random intercepts would not represent the heterogeneity of the individuals at baseline any more.

Finally, we have applied the BBmm approach in two different real datasets. The analysis of the Paquid data has been performed as an illustration of the applicability of our proposal. However, much more effort has been dedicated in the longitudinal

---

analysis of the COPD Study. In fact, this was the motivation of this thesis as there was no available methodology which could analyse the evolution of patients with COPD in an appropriate way. Section 3.5.2 shows that the BBmm approach offers clinically valid and relevant results in terms of the evolution of patients with COPD. Moreover, all the results obtained from the longitudinal analysis of both questionnaires included in the COPD Study, the SF-36 and the SGRQ, are being discussed with clinical researchers and they will be included in a clinical publication that is in preparation.



---

---

## CHAPTER 4

---


# MULTIVARIATE ANALYSIS

*“Evolution has ensured that our brains just aren’t equipped to visualise 11 dimensions directly. However, from a purely mathematical point of view it’s just as easy to think in 11 dimensions, as it is to think in three or four.”*

---

Stephen Hawking , 1942-

*The article based on the work developed in this chapter is under preparation for submitting it to a journal.*

 *Najera-Zuloaga, J., Lee, D.-J. and Arostegui, I. (2017). Multivariate analysis of patient reported-outcomes based on beta-binomial mixed-effects model approach. (under preparation)*

### 4.1 Introduction

In the previous chapters we have developed regression models based on the beta-binomial distribution for dealing with different statistical challenges that usually occur in PRO analysis. In Chapter 2, we have proposed a marginal beta-binomial regression model (denoted as BBreg) for analysing overdispersed binomial data in an independent data framework. However, in many scientific applications, we often need to analyse data resulting from experiments in which outcomes have been measured repeatedly on a set of units, leading to the so-called repeated measurements. A correct statistical analysis of such data should account for the hierarchical nature of the data, allowing those measurements within subjects to be correlated, while

observations from different individuals are independent. Marginal beta-binomial regression approach is not applicable in this case, as it assumes the independence of the outcomes being modelled. Therefore, in Chapter 3 we have developed a beta-binomial mixed-effect model (denoted as BBmm) for the correct analysis of overdispersed hierarchical binomial data, hierarchical meaning that there exists a correlation structure. Application of both presented methodologies has been addressed to the analysis of PROs, which usually show an integer and bounded dispersed feature.

In Chapter 1 we have introduced different questionnaires for assessing different types of PROs, such as cognitive status, HRQoL and so on. We have explained that there are generic or disease-specific questionnaires and consequently, that each questionnaire is commonly applied in different populations. We have also detailed that questionnaires usually decompose the health aspect they are assessing in different dimensions, providing a multi-dimensional insight of the health-status of the patients. Until now, either by a cross-sectional (Chapter 2) or longitudinal (Chapter 3) approach, we have analysed each of the dimensions separately, and results have been obtained dimension by dimension.

In some cases, independent analysis of each outcome separately is enough to respond all research questions, subject-matter questions for example. However, when the interest lies in assessing the relation between some covariates and all the dimensions simultaneously, in comparing longitudinal trends between dimensions, or in the association between the dimensions and how that association evolves over time, joint analysis of all the outcomes is preferable. In fact, dimensions are constructed by responses from the same individuals, then, it is reasonable to think that there may exist some correlation among different dimension scores for the same patient. Consequently, the joint analysis of all the dimensions could accommodate some extra variability that the independent model cannot account, and therefore, improve the modelling technique. Several regression approaches have been proposed in the literature for the joint analysis of different responses, some of them focusing on longitudinal data framework (Verbeke et al., 2014). In this chapter we propose a regression approach based on the beta-binomial distribution for the joint analysis of more than one dimension, both in cross-sectional and longitudinal framework.

In terms of notation, from now on and the rest of the chapter, we will refer to the joint analysis of more than one response variable as *multivariate* analysis, whereas the separately analysis of each outcome variable will be considered as *univariate* analysis.

This chapter is organised as follows. In Section 4.2, we present a short review

of some multivariate analysis approaches proposed in the literature, especially for the joint analysis of different dimensions containing repeated measurements over time. In Section 4.3, we introduce a model proposal for the joint analysis of all the PRO dimensions provided by a questionnaire. Following, considering the proposed model approach we carry out a multivariate analysis of the questionnaires applied in the COPD data which have been analysed in Chapter 2 and Chapter 3 in a univariate approach. Finally, in Section 4.5, we present a discussion about the proposed methodology and introduce some limitations that allow us to think about some future work.

## 4.2 Background

A number of approaches have been proposed in the literature for the joint modelling of multivariate longitudinal outcomes (Verbeke et al., 2014). These model approaches belong to different modelling traditions and their construction is based on different arguments. In fact, they differ in a number of formal characteristics, such as the structure (balanced or unbalanced), the scale of the observed outcomes (continuous, ordinal or binary), or the way the association between and across outcomes is modelled (with or without latent variables). In fact, the latter aspect is important given that the use of latent variables allows for more flexible data structures but usually also has some limitations with respect to the interpretation of the parameters.

For the remainder of the chapter, assume that  $\mathbf{Y}^1, \dots, \mathbf{Y}^k$  are the vectors of random variables associated with the longitudinal measurements for the  $k$  dimensions and let  $\mathbf{Y}$  be the vector of all random variables.

The first modelling approach attempts to specify directly the joint density  $f(\mathbf{y})$  of  $\mathbf{Y}$ . Specification of a marginal model for  $\mathbf{Y}$  requires assumptions about the marginal association among the repeated measurements within each of the dimensions, but also must include assumptions on the association between elements of any two dimensions  $\mathbf{Y}^l$  and  $\mathbf{Y}^s$ , where  $l \neq s$ . However, when the dimensions  $\mathbf{Y}^l$ ,  $l = 1, \dots, k$ , are of different types (e.g. continuous or discrete) and/or in the case of unbalanced data this approach becomes hard to deal with. When the data are discrete, likelihood-based marginal models can be formulated (Molenberghs and Verbeke, 2005), but unlike the Gaussian case, they are difficult to implement unless  $tk$  is sufficiently small (Verbeke et al., 2014), where  $t$  is the number of time points at which measurements have been taken. In COPD Study (see Chapter 1) for exam-

ple, measurements of different individuals are provided in different time points. In fact, the number of different days where an outcome was measured, is equal to 656 and taking into account that a maximum of 4 measurements is provided by each patient, the data can be considered highly unbalanced. Therefore, due to continuity assumption of the time variable, the number of instances where observations were measured is very high, making the specification of the joint density extremely hard.

One way to avoid the direct specification of a joint distribution for  $\mathbf{Y}$  is to model a subject's measurement on a given outcome at a particular time point, conditioned on all other  $tk - 1$  observations. In a longitudinal context, it is often considered natural only to condition on the past, which can be done through the so-called transition models. Transition models for discrete longitudinal data (Diggle et al., 2002) consider the time course as a sequence of states and that transition probabilities to be in a specific state at a particular time point depends on the previous time point(s), but extensions to multivariate longitudinal data are possible (Zeng and Cook, 2007). However, these extensions differ in the way the associations among observations are modelled for the different outcomes. In fact, with more than two outcomes, many possible factorizations are possible, all potentially leading to different results (Diggle et al., 2002). Hence, transition models are often not considered as the preferred choice to analyse high-dimensional multivariate longitudinal data. In addition, similar to the specification of the marginal density function approach, the assumption that the time, measured as continuous, is divided in instances leads to highly unbalanced data which complicates the model development.

It is known that when either the number of dimensions or the number of instances is high, the models defined above lead to several estimation problems (Verbeke et al., 2014). When modelling high-dimensional longitudinal data, one option is to use one or more latent variables for the outcome dimension, reducing the dimensionality of the multivariate vector of outcomes. For example in PRO context, the model assumes that the observed dimensions are measuring one or more underlying concepts characterising the health-status the questionnaire is trying to assess. The general idea is to use a factor-analytic, or principal-component type, analysis to first reduce the dimensionality of the response vector and to use standard longitudinal models for the analysis of the principal factors (Oort, 2001). Although dimensionality problems are solved by applying the reduction, the disadvantage is that we cannot always interpret the resulting outcome and therefore, the model cannot be properly interpreted. In addition, this methodology does not correct the unbalancedness problem of the presented multivariate analysis approaches.

However, there exists a very flexible class of models often used for the analysis of multivariate longitudinal data, the family of mixed-effects models. In fact, we have made use of this approach when analysing, in a univariate approach, longitudinal data in Chapter 3. This approach assumes that the observations represent realizations of a latent subject-specific trajectory which can be modelled parsimoniously using relatively a small number of subject-specific parameters. Such models have the main advantage that they do not assume balancedness, allowing for different number of observations per individual and/or measurements of different individuals taken at different time points.

In the next section we present two different mixed-effects models for analysing longitudinal data in a multivariate approach. We detail the models proposal, advantages and disadvantages of each approach and offer an ‘easy’ way to compare univariate and multivariate regression parameters in some circumstances.

### 4.3 Beta-binomial mixed-effects model approach

In this section we present a multivariate regression approach based on mixed-effects models for the joint analysis of correlated dimensions drawn from a beta-binomial distribution. Therefore, based on the mixed-effects theory, we consider the BBmm general approach developed in Chapter 3 for performing the estimation and inference of the multivariate model proposal. The idea consist of using random effects to accommodate not only the correlation of the repeated measurements within the dimensions, but also the correlation that could exist among the different dimensions. Although many authors have proposed the use of random-effects models for multivariate repeated measurements (Reinsel, 1984; MacCallum et al., 1997), it must be pointed out that examples in the context of multivariate nonlinear or GLMM are less common (Verbeke et al., 2014).

It is worth noticing that the estimation and inference theory for BBmm approach has already been developed in Chapter 3. Hence, the objective of this section will be the proposal of a model definition adapted to the characteristics of multivariate longitudinal beta-binomial data.

#### 4.3.1 Shared random effects approach

McCulloch (2008) proposed the shared random effects approach for jointly modelling multiple outcomes of mixed types. He defined the model in a cross-sectional



context however, its extension to longitudinal data is straightforward. The basic idea is to use a random effect to build in the correlation between the dimensions, assuming that conditional on the random effect  $\mathbf{u}$  the dimensions are independent. The conditionally independent assumption reflects the belief that a common set  $\mathbf{u}$  of underlying characteristics of the individual governs the outcomes process.

For instance, consider that we have two vectors of beta-binomial random variables  $\mathbf{Y}^1$  and  $\mathbf{Y}^2$  associated with the measurements of the same  $n$  patients for different dimensions, and assume that we are interested in modelling them as a function of a covariate  $X$ . For the sake of clarity, we only consider two dimensions although extension to more dimensions is immediate. If we analyse the outcomes separately, as done in Chapter 2, we get that

$$\begin{aligned} Y_i^1 &\sim \text{BB}(m_i^1, p_i^1, \phi^1) \text{ indep. } i = 1, \dots, n \\ \text{logit}(p_i^1) &= \beta_0^1 + \beta_1^1 x_i, \\ Y_i^2 &\sim \text{BB}(m_i^2, p_i^2, \phi^2) \text{ indep. } i = 1, \dots, n \\ \text{logit}(p_i^2) &= \beta_0^2 + \beta_1^2 x_i. \end{aligned} \tag{4.1}$$

However, while Equation (4.1) would be sufficient for separate analysis, it does not accommodate the correlation between  $Y_i^1$  and  $Y_i^2$ ,  $i = 1, \dots, n$ . An option to skip the problem is to introduce a random effect per individual that will be shared by both dimensions. Equation (4.1) is modified accordingly by modelling the distributions conditional on the random effect  $\mathbf{u} = (u_1, \dots, u_n)'$ ,

$$\begin{aligned} Y_i^1 | u_i &\sim \text{BB}(m_i^1, p_i^1, \phi^1) \text{ indep. } i = 1, \dots, n \\ \text{logit}(p_i^1) &= \gamma_0^1 + \gamma_1^1 x_i + u_i, \\ Y_i^2 | u_i &\sim \text{BB}(m_i^2, p_i^2, \phi^2) \text{ indep. } i = 1, \dots, n \\ \text{logit}(p_i^2) &= \gamma_0^2 + \gamma_1^2 x_i + \lambda u_i, \\ u_i &\sim \mathcal{N}(0, \sigma_u^2). \end{aligned} \tag{4.2}$$

In general model formulation, the  $\lambda$  multiplying  $u_i$  in the equation for  $Y_i^2$  is included to account for the fact that the linear predictors for  $\mathbf{Y}^1$  and  $\mathbf{Y}^2$  may be measured on different scales. This formulation is useful when, for instance, we are modelling outcomes following different distributions and hence, the random effects are included in different transformations of the linear predictor (e.g. logit, probit, logarithm, exponential). However, in this case, we are modelling outcomes drawn by the same distribution using the same transformation of the conditioned expectation for the

linear predictor construction. Therefore, as both linear predictor are defined in the same scale, we can assume that  $\lambda = 1$ .

Although we have defined the model assuming that all the dimensions follow a beta-binomial distribution, one of the main characteristics of the shared random effects approach is that dimensions can be drawn from different distributions and hence, different discrete and continuous variables can be analysed jointly. In addition, the number of dimensions involving the analysis does not alter the dimensionality of the joint density or likelihood integration as increasing the number of dimensions does not alter the number of random effects in the model,

$$f(\mathbf{y}) = \int f(y^1, \dots, y^k | \mathbf{u}) f(\mathbf{u}) d\mathbf{u} = \int \prod_{l=1}^k f(y^l | \mathbf{u}) f(\mathbf{u}) d\mathbf{u},$$

where  $k$  is the number of dimensions. In fact, this is the main advantage of the shared random effects models compared to other commonly used approaches.

In order to compare fixed effects in the separate or joint model, care must be taken. As it was discussed in Chapter 2, we cannot compare like with like fixed effects of marginal and conditional models. In fact, if we try to calculate the marginal mean of each dimensions for the BBmm approach we end up that there is not a closed form. However, we could apply the approximated relationship between the logit and the probit link functions defined in Equation (3.43) to conclude with an approximated comparison of the fixed effects in both approaches. Let us assume that  $\Phi(X) = \Pr\{Z < X|X\}$ , where  $Z \sim \mathcal{N}(0, 1)$  and  $\Phi()$  is the standard normal cumulative density function. On the one hand, we have that

$$\begin{aligned} \mathbb{E}[Y_i^l] &= \mathbb{E} \left[ \frac{\exp(\gamma_0^l + \gamma_1^l x_i + u_i)}{1 + \exp(\gamma_0^l + \gamma_1^l x_i + u_i)} \right] \\ &\approx \mathbb{E} \left[ \Phi(c(\gamma_0^l + \gamma_1^l x_i + u_i)) \right] \\ &= \mathbb{E} \left[ \Pr \left( Z < c(\gamma_0^l + \gamma_1^l x_i + u_i) | u_i \right) \right] \\ &= \Pr \left[ Z < c(\gamma_0^l + \gamma_1^l x_i + u_i) \right] \\ &= \Pr \left[ \frac{Z - cu_i}{\sqrt{1 + c\sigma_u}} < \frac{c}{\sqrt{1 + c\sigma_u}} (\gamma_0^l + \gamma_1^l x_i) \right] \\ &= \Phi \left( \frac{c}{\sqrt{1 + c\sigma_u}} (\gamma_0^l + \gamma_1^l x_i) \right), \end{aligned} \tag{4.3}$$

where the fourth identity holds because the expected value of the conditional prob-

ability is the unconditional probability (McCulloch, 2008) and  $c = (16\sqrt{3})/(15\pi)$ ,  $l = 1, 2$ . On the other hand, based on the marginal model, it is straightforward to conclude that

$$\mathbb{E}[Y_i^l] = \mathbb{E} \left[ \frac{\exp(\beta_0^l + \beta_1^l x_i)}{1 + \exp(\beta_0^l + \beta_1^l x_i)} \right] \approx \Phi \left( c(\beta_0^l + \beta_1^l x_i) \right). \quad (4.4)$$

Therefore, based on Equation (4.3) and Equation (4.4), we can compare the fixed effects in the univariate and multivariate approaches by

$$\beta_r^l = \frac{\gamma_r^l}{\sqrt{1 + c\sigma_u}}, \quad r = 0, 1. \quad (4.5)$$

In terms of the BBmm approach definition in Equation (3.23), we can rewrite the multivariate longitudinal (or hierarchical) model in the following way. Assume that we have  $n$  different individuals which report  $t_i$  outcomes repeatedly in  $k$  dimensions, so

$$\begin{aligned} Y_{ij}^l | u_i &\sim \text{BB}(m_{ij}^l, p_{ij}^l, \phi^l) \text{ indep.} \\ \eta_{ij}^l &= \text{logit}(p_{ij}^l) = \beta_0^l + \beta_1^l x_{ij} + z_{ij} u_i \\ \mathbf{u} &= (u_1, \dots, u_n)' \sim \mathcal{N}(0, \mathbf{D}) \end{aligned}$$

$i = 1, \dots, n, j = 1, \dots, t_i, l = 1, \dots, k$ . Or equivalently, in a matrix way

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}$$

where,

$$\begin{aligned} \mathbf{X} &= \mathbf{I}_k \otimes \mathbf{X}^*, \quad \mathbf{X}_i^* = \begin{pmatrix} 1 & x_{ij1} & \cdots & x_{ijp} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{it_i1} & \cdots & x_{it_i p} \end{pmatrix}, \\ \mathbf{Z} &= \mathbf{1}_k \otimes \mathbf{Z}^*, \quad \mathbf{Z}^* = \begin{pmatrix} \mathbf{1}_{t_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{t_2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1}_{t_n} \end{pmatrix}, \\ \boldsymbol{\beta} &= (\boldsymbol{\beta}^1, \dots, \boldsymbol{\beta}^k)', \quad \boldsymbol{\beta}^l = (\beta_0^l, \dots, \beta_p^l)', \\ \mathbf{u} &= (u_1, \dots, u_n)', \end{aligned}$$

being  $p$  the number of covariates in the model and  $\mathbf{1}_s$  a 1s vector of length  $s$ ,  $i = 1, \dots, n$ ,  $j = 1, \dots, t_i$  and  $l = 1, \dots, k$ . Notice that compared to the BBmm approach defined in Chapter 3, in the multivariate model we allow  $\phi^l$  the dispersion parameter of the beta-binomial distribution vary from one dimension to the other.

One of the disadvantages of the shared random effects approach when analysing repeated measurements jointly, is that the correlation between different dimensions is determined by the dimension's intrinsic correlation given by the repeated measurements. For example, assume that  $\mathbf{Y}^1$  and  $\mathbf{Y}^2$  are vectors of random variables representing different dimensions which contain repeated measurements for a set of  $n$  individuals and that we apply a longitudinal model with shared random intercepts,

$$\begin{aligned}\boldsymbol{\eta}^1 &= \text{logit}(\mathbf{p}^1) = \beta_0 \mathbf{1}_n + \mathbf{u} + \beta_1 \mathbf{t}, \\ \boldsymbol{\eta}^2 &= \text{logit}(\mathbf{p}^2) = \beta_2 \mathbf{1}_n + \mathbf{u} + \beta_3 \mathbf{t},\end{aligned}$$

where  $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \sigma_u \mathbf{I}_n)$  and  $\mathbf{p}^1$  and  $\mathbf{p}^2$  are the probability parameters of the conditional  $\mathbf{Y}^1$  and  $\mathbf{Y}^2$  beta-binomial vectors of variables respectively. Based on the latent approach developed in Chapter 3 Section 3.3.4, we can define the latent variables associated with each dimension as

$$\begin{aligned}\mathbf{y}^{1*} &= \beta_0 \mathbf{1}_n + \mathbf{u} + \beta_1 \mathbf{t} + \boldsymbol{\epsilon}^1 \\ \mathbf{y}^{2*} &= \beta_2 \mathbf{1}_n + \mathbf{u} + \beta_3 \mathbf{t} + \boldsymbol{\epsilon}^2\end{aligned}$$

where  $\boldsymbol{\epsilon}^l$  are independent error terms that follow a logistic distribution,  $l = 1, 2$ . We can calculate the correlation between the latent variables associated with each dimension for the  $i$ th individual as

$$\begin{aligned}\text{Corr}[Y_{ij}^{1*}, Y_{is}^{2*}] &= \frac{\text{Cov}[Y_{ij}^{1*}, Y_{is}^{2*}]}{\sqrt{\text{Var}[Y_{ij}^{1*}]} \sqrt{\text{Var}[Y_{is}^{2*}]}} \\ &= \frac{\text{Cov}[\beta_0 + u_i + \beta_1 t_{ij} + \epsilon_{ij}^1, \beta_2 + u_i + \beta_3 t_{is} + \epsilon_{is}^2]}{\sqrt{\text{Var}[\beta_0 + u_i + \beta_1 t_{ij} + \epsilon_{ij}^1]} \sqrt{\text{Var}[\beta_2 + u_i + \beta_3 t_{is} + \epsilon_{is}^2]}} \quad (4.6) \\ &= \frac{\text{Var}[u_i]}{\sqrt{\sigma_u^2 + \pi^2/3} \sqrt{\sigma_u^2 + \pi^2/3}} \\ &= \frac{\sigma_u^2}{\sigma_u^2 + \pi^2/3},\end{aligned}$$

where  $1 \leq j \leq s \leq t_i$  and  $j \neq s$ .

Additionally, we have that the correlation between observations of the same individual in each dimension is defined as

$$\text{Corr} [Y_{ij}^{1*}, Y_{is}^{1*}] = \frac{\sigma_u^2}{\sigma_u^2 + \pi^2/3}, \quad (4.7)$$

where  $1 \leq j \neq s \leq t_i$ ,  $l = 1, 2$ .

Therefore, the correlation structure within each dimension dictates the association between the different dimensions. For example, the model would not allow that  $\mathbf{Y}^1$  and  $\mathbf{Y}^2$  to be independent if repeated measurements of  $\mathbf{Y}^1$  are strongly correlated. In fact, the assumption that the error components follow a logistic distribution does not impose any restriction and, if any distribution has been chosen, we would have reached the same conclusion (Verbeke et al., 2014). The restriction of the correlation can be relaxed by allowing different but correlated random effects for the various dimensions.

### 4.3.2 Correlated random effects approach

In the previous section we have described that the shared random effects approach contains many desirable properties such as (i) allowing for different types of outcomes; (ii) allowing for (highly) unbalanced data; and (iii) ‘easy’ transition of the comparison of the fixed effects from the multivariate to the univariate model. However, we have mentioned that in some occasions the rigidity of the imposed correlation structures makes them unrealistic to be useful in real practise. This can be solved by allowing for different random effects for each dimension as,

$$\begin{aligned} Y_{ij}^l | u_i^l &\sim \text{BB}(m_i^l, p_{ij}^l, \phi^l) \text{ indep.} \\ \eta_{ij}^l &= \text{logit}(p_{ij}^l) = \beta_0^l + \beta_1^l x_{ij} + z_{ij} u_i^l \\ \mathbf{u} &= (\mathbf{u}^1, \dots, \mathbf{u}^k)' \sim \mathcal{N}(\mathbf{0}, \mathbf{D}) \end{aligned} \quad (4.8)$$

where  $\mathbf{u}^l = (u_1^l, \dots, u_n^l)'$ ,  $i = 1, \dots, n$ ,  $j = 1, \dots, t_i$ ,  $l = 1, \dots, k$ .

This new model relaxes the shared random effects assumption in the sense that the correlation structure of the individual outcomes no longer dictates the association between pairs of measurements from different dimensions. Under the conditional

independence the joint marginal density is given by

$$\begin{aligned} f(\mathbf{y}) &= \int f(\mathbf{y}^1, \dots, \mathbf{y}^k | \mathbf{u}) f(\mathbf{u}) d\mathbf{u} \\ &= \int \cdots \int \prod_{l=1}^k f(\mathbf{y}^l | \mathbf{u}^l) f(\mathbf{u}^1, \dots, \mathbf{u}^k) d\mathbf{u}^1 \cdots d\mathbf{u}^k, \end{aligned} \quad (4.9)$$

which increases the dimensionality as more dimensions and, hence, more random effects, are included in the model. In fact, the Laplace approximation developed in Chapter 3, Section 3.3.2, for the calculation of the marginal likelihood may involve in inappropriate approximations if the dimensionality of the integral is high (Shun and McCullagh, 1995). Moreover, due to the complexity of the beta-binomial distribution, when more than two dimensions are analysed the estimation procedure developed in Chapter 3 is not tractable any more. In fact, many PRO questionnaires decompose the health-status that are assessing in more than two dimensions. Consequently, in thesis, the real data will be analysed by the proposed shared random effects approach.

## 4.4 Application to COPD Study

In this section, we perform the COPD Study analysis carried out in Chapter 2 and Chapter 3 in a multivariate approach. On the one hand, we are going to adjust a model for the joint analysis of the eight dimensions provided by SF-36 in a cross-sectional fashion and compare the results with the univariate analysis provided in Chapter 2. On the other hand, we carry out a longitudinal joint analysis of the three dimensions provided by the SGRQ. We will compare the obtained results with the results provided in Chapter 3, where each of the dimension was analysed separately in a univariate approach.

### 4.4.1 Cross-sectional analysis

In this section, we analyse the eight dimensions provided by the SF-36 Health Survey in the COPD Study simultaneously, using a multidimensional approach based on the shared random effects model. In fact, the shared random effects approach is applied for accommodating the correlation that may exist between measurements of the same patients in different dimensions. In Chapter 2 we have analysed each of the dimension separately in an univariate approach. Consequently, in order to

maintain the model assumption for assessing the differences that may exist between the univariate and multivariate models, following Chapter 2, exclusively data from the first visit to the outpatient clinic is going to be considered. Consequently, the joint analysis of the SF-36 dimensions is going to be performed in a cross-sectional framework.

In Section 4.3.1 we have shown that the shared random effects approach is based on a general mixed-effects model where the model is specified in a particular way. Therefore, estimation and inference of the parameters involving the model do not require any extension of the BBmm approach developed in Chapter 3, with the exception of the dispersion parameter of the beta-binomial distribution, which can vary for different dimensions. In fact, the model we are going to apply for the joint analysis of the eight dimensions of the SF-36 in COPD Study is defined as

$$Y_i^l | u_i \sim \text{BB}(m_i^l, p_i^l, \phi^l) \quad \text{indep.} \quad (4.10)$$

$$\eta_i^l = \log \frac{p_i^l}{1 - p_i^l} = \mathbf{x}_i' \boldsymbol{\beta}^l + u_i$$

where  $\mathbf{x}_i$  is the  $i$ th row of a full rank matrix composed by the given covariates,  $\boldsymbol{\beta}^l$  are the regression parameters for each dimension and  $u_i$  are independent random effects being  $u_i \sim \mathcal{N}(0, \sigma_u^2)$ ,  $i = 1, \dots, 543$ ,  $l = 1, \dots, 8$ . Notice that in the COPD Study the number of individuals is equal to 543. The model defined in Equation (4.10) can be redefined following the generic notation of BBmm approach as

$$\mathbf{Y} | \mathbf{u} \sim \text{BB}(\mathbf{m}, \mathbf{p}, \boldsymbol{\phi}) \quad \text{indep.}$$

$$\boldsymbol{\eta} = \log \frac{\mathbf{p}}{1 - \mathbf{p}} = \mathbf{X}^* \boldsymbol{\beta} + \mathbf{u}$$

where  $\mathbf{X} = \mathbf{I}_8 \otimes \mathbf{X}$  being  $\mathbf{X}$  the model matrix composed by the given covariates defined in Equation (4.10),  $\mathbf{Y} = (\mathbf{Y}^{1'}, \dots, \mathbf{Y}^{8'})'$  being  $\mathbf{Y}^l = (Y_1^l, \dots, Y_{543}^l)$  the outcomes variable vector of the  $l$ th dimension,  $\boldsymbol{\beta} = (\boldsymbol{\beta}^{1'}, \dots, \boldsymbol{\beta}^{8'})'$  a vector containing the regression coefficients involving the linear predictor of each dimension,  $\boldsymbol{\phi} = (\phi^1, \dots, \phi^8)'$  the vector of the dispersion parameters of the beta-binomial distribution and  $\mathbf{u} = (u_1, \dots, u_{543})'$  independent normally distributed random effects with variance-covariance matrix equal to  $\sigma_u^2 \mathbf{I}_{543}$ . Hence, the BBmm approach developed in Chapter 3 has been used to fit the model.

Table 4.1 and Table 4.2 show the results for the univariate and multivariate analysis of the SF-36 dimensions in COPD data. Although in the multivariate

approach all the dimensions have been analysed jointly, due to the lack of space, results are displayed dividing physical dimensions (Table 4.1) and mental dimensions (Table 4.2). For the variable selection, we retain in the model those covariates whose influence in the outcome variable was statistically significant ( $p$ -value  $< 0.05$ ) in at least one of the approaches. The tables show the estimated value of the regression parameters  $\beta$  together with the standard deviation and the  $p$ -value associated with the significance of the estimation in both approaches. Additionally, they display the estimation and standard deviation of the vector of dispersion parameters of the beta-binomial distributions,  $\phi$ .

The first conclusion we address from Table 4.1 and Table 4.2 is that the estimated values of the regression parameters are quite similar in the two approaches, both in magnitude and sign. Therefore, they lead to similar interpretations regarding the effect of the covariates in the analysed dimensions. However, there are some differences that could lead to contradictory results in terms of the significance of the effect of the covariates. As it can be appreciated, the inclusion of a shared random effect that accounts for the correlation between different dimensions reduces the dispersion parameter of the beta-binomial distribution, as the variance structure is split in two components. Consequently, the standard deviations associated with the estimates of the regression parameters differ, being lower for the multivariate approach in most of the dimensions (6 out of 8). Therefore, different conclusions will be obtained for instance in *mental health* dimension if a univariate or a multivariate approach would have been implemented. In fact, age and FEV1% covariates do not have a statistically significant effect ( $p$ -values equal to 0.051 and 0.069 respectively) in the univariate approach, while in the joint analysis of the dimensions both covariates offer statistically significant effects ( $p$ -values equal to 0.045 and 0.023 respectively).

Summarizing, we can state that, although the joint analysis of cross-sectional dimensions do not alter the estimates of the regression parameters in the model, there could be differences in the variance of the estimates leading to contradictory results in terms of the significance of the effect. Therefore, in cases where it is possible, we recommend the use of the multivariate approach as it offers a better measurement of the variability of the data splitting it in two components, the correlation and the overdispersion.



Table 4.1: Multivariate and univariate analysis of the eight dimensions provided by the SF-36 in COPD patients. Results for the physical dimensions are displayed.

	Multivariate ( $\sigma = 0.535$ )			Univariate (Chapter 2)		
	Estimate	SD	p-value	Estimate	SD	p-value
<b>Physical functioning</b>						
Intercept	-0.478	0.386	0.215	-1.091	0.455	0.017
<i>Dyspnea2</i>	-0.652	0.096	< 0.001	-0.580	0.112	< 0.001
<i>Dyspnea3</i>	-1.417	0.102	< 0.001	-1.281	0.120	< 0.001
<i>Dyspnea4</i>	-2.432	0.151	< 0.001	-2.207	0.176	< 0.001
<i>Depression</i>	-0.521	0.109	< 0.001	-0.544	0.130	< 0.001
<i>Anxiety</i>	-0.415	0.076	< 0.001	-0.404	0.090	< 0.001
<i>Age</i>	0.009	0.003	0.005	0.012	0.004	0.002
<i>BMI</i>	-0.021	0.006	< 0.001	-0.018	0.007	0.009
<i>FEV</i>	0.007	0.002	0.001	0.006	0.003	0.012
<i>Sex</i>	0.467	0.130	< 0.001	0.461	0.155	0.003
<i>WalkingTest</i>	0.003	< 0.001	< 0.001	0.004	< 0.001	< 0.001
log( $\phi$ )	-3.845	0.214	—	-2.826	0.116	—
<b>Role physical</b>						
Intercept	1.310	0.707	< 0.001	1.100	0.690	0.111
<i>Dyspnea2</i>	-1.366	0.354	< 0.001	-1.296	0.340	< 0.001
<i>Dyspnea3</i>	-2.478	0.356	< 0.001	-2.349	0.341	< 0.001
<i>Dyspnea4</i>	-3.429	0.417	< 0.001	-3.291	0.405	< 0.001
<i>Anxiety</i>	-0.812	0.204	< 0.001	-0.732	0.199	< 0.001
<i>Age</i>	0.020	0.009	0.030	0.021	0.009	0.020
log( $\phi$ )	-0.306	0.117	—	-0.103	0.109	—
<b>Bodily pain</b>						
Intercept	2.575	0.393	< 0.001	2.454	0.396	< 0.001
<i>Dyspnea2</i>	-0.495	0.227	0.030	-0.504	0.227	0.027
<i>Dyspnea3</i>	-0.837	0.239	< 0.001	-0.811	0.239	0.001
<i>Dyspnea4</i>	-1.160	0.312	< 0.001	-1.091	0.316	0.001
<i>Anxiety</i>	-0.652	0.162	< 0.001	-0.612	0.163	< 0.001
<i>FEV</i>	-0.013	0.005	0.010	-0.012	0.005	0.020
log( $\phi$ )	-0.742	0.089	—	-0.559	0.085	—
<b>General health</b>						
Intercept	-0.950	0.307	0.002	-1.005	0.343	0.004
<i>Dyspnea2</i>	-0.412	0.099	< 0.001	-0.391	0.110	< 0.001
<i>Dyspnea3</i>	-1.012	0.108	< 0.001	-0.969	0.120	< 0.001
<i>Dyspnea4</i>	-1.235	0.155	< 0.001	-1.175	0.173	< 0.001
<i>Depression</i>	-0.539	0.139	< 0.001	-0.560	0.156	< 0.001
<i>Anxiety</i>	-0.390	0.094	< 0.001	-0.362	0.105	0.001
<i>Age</i>	0.018	0.004	< 0.001	0.018	0.004	< 0.001
<i>FEV</i>	0.006	0.003	0.041	0.005	0.003	0.056
log( $\phi$ )	-2.636	0.108	—	-2.297	0.117	—

SD: Standard Deviation; BMI: Body Mass Index; FEV1%: Forced Expiratory Volume in one second in percentile.

Table 4.2: Multivariate and univariate analysis of the eight dimensions provided by the SF-36 in COPD patients. Results for the mental dimensions are displayed.

	Multivariate ( $\sigma = 0.535$ )			Univariate (Chapter 2)		
	Estimate	SD	p-value	Estimate	SD	p-value
<b>Vitality</b>						
<i>Intercept</i>	0.417	0.286	0.145	0.342	0.335	0.308
<i>Dyspnea2</i>	-0.792	0.119	< 0.001	-0.765	0.137	< 0.001
<i>Dyspnea3</i>	-1.592	0.123	< 0.001	-1.544	0.142	< 0.001
<i>Dyspnea4</i>	-1.947	0.157	< 0.001	-1.893	0.183	< 0.001
<i>Depression</i>	-0.928	0.142	< 0.001	-0.927	0.166	< 0.001
<i>Anxiety</i>	-0.644	0.094	< 0.001	-0.598	0.112	< 0.001
<i>Age</i>	0.018	0.004	< 0.001	0.018	0.005	< 0.001
$\log(\phi)$	-2.557	0.107	–	-1.891	0.084	–
<b>Social functioning</b>						
<i>Intercept</i>	2.859	0.249	< 0.001	2.674	0.250	< 0.001
<i>Dyspnea2</i>	-0.657	0.266	0.014	-0.565	0.268	0.035
<i>Dyspnea3</i>	-1.402	0.265	< 0.001	-1.254	0.267	< 0.001
<i>Dyspnea4</i>	-1.780	0.310	< 0.001	-1.598	0.314	< 0.001
<i>Depression</i>	-0.459	0.225	0.041	-0.471	0.232	0.043
<i>Anxiety</i>	-1.232	0.158	< 0.001	-1.179	0.163	< 0.001
$\log(\phi)$	-1.278	0.120	–	-1.036	0.109	–
<b>Role emotional</b>						
<i>Intercept</i>	2.924	0.404	< 0.001	2.769	0.390	< 0.001
<i>Dyspnea2</i>	-0.664	0.434	0.126	-0.612	0.418	0.143
<i>Dyspnea3</i>	-1.495	0.429	< 0.001	-1.376	0.414	0.001
<i>Dyspnea4</i>	-2.164	0.485	< 0.001	-2.042	0.467	< 0.001
<i>Anxiety</i>	-1.705	0.233	< 0.001	-1.642	0.227	< 0.001
$\log(\phi)$	0.579	0.154	–	0.669	0.150	–
<b>Mental health</b>						
<i>Intercept</i>	1.229	0.409	0.003	0.994	0.460	0.031
<i>Dyspnea2</i>	-0.380	0.128	0.003	-0.333	0.142	0.019
<i>Dyspnea3</i>	-0.872	0.134	< 0.001	-0.793	0.149	< 0.001
<i>Dyspnea4</i>	-1.225	0.174	< 0.001	-1.135	0.196	< 0.001
<i>Depression</i>	-0.852	0.137	< 0.001	-0.866	0.157	< 0.001
<i>Anxiety</i>	-1.302	0.095	< 0.001	-1.244	0.108	< 0.001
<i>Age</i>	0.011	0.004	0.045	0.010	0.005	0.051
<i>BMI</i>	0.018	0.008	0.024	0.018	0.009	0.044
<i>FEV</i>	-0.007	0.003	0.023	-0.006	0.003	0.069
$\log(\phi)$	-3.046	0.177	–	-2.297	0.117	–

SD: Standard Deviation; FEV1%: Forced Expiratory Volume in one second in percentile; BMI: Body Mass Index.

#### 4.4.2 Longitudinal analysis

The second real data application of the beta-binomial shared random effects approach is performed in a longitudinal or repeated measurements context. Similar to the joint analysis of the SF-36 Health Survey, one of the objectives of this section is the comparison of the longitudinal analysis of some dimensions in univariate and multivariate approaches. For this aim, have considered the analysis of the SGRQ in COPD Study which was carried out in Chapter 3, where each of the three dimensions was analysed separately. As a result of the univariate analysis, several conclusions were addressed in terms of the evolution of the patients with COPD. Therefore, in this section we will perform the same analysis of the dimensions but analysing them all together in a multivariate approach. Therefore, differences in the results in terms of the evolution of the patients provided by both approaches will be compared.

In Chapter 3, we have analysed each dimension separately with a random intercept and slope model. In order to maintain the model assumption, we propose a shared random intercept and slope effects model for the joint analysis of the dimensions provided by the SGRQ. In fact, the model could be defined for any longitudinal multivariate analysis where random intercepts and slopes would be shared. For instance, assume that we have  $n$  individuals measured in  $k$  dimensions where the  $l$ th dimension contains  $t_i^l$  repeated measurements of the  $i$ th individual,  $i = 1, \dots, n$ ,  $l = 1, \dots, k$ . Then, we can propose the following multivariate random intercept and slope model based on BBmm approach,

$$Y_{ij}^l | u_i, v_i \sim \text{BB}(m_i^l, p_{ij}^l, \phi^l) \quad \text{indep.} \quad (4.11)$$

$$\eta_{ij}^l = \log \frac{p_{ij}^l}{1 - p_{ij}^l} = (\beta_0^l + u_i) + (\beta_1^l + v_i)T_{ij}$$

where  $Y_{ij}^l$  is the random variable associated with the  $j$ th repeated measurement of the  $i$ th individual in the  $l$ th dimension,  $T_{ij}$  is the time in years from the beginning of the study where the  $j$ th observation of the  $i$ th individual was performed,  $\beta_0^l$  is the overall intercept,  $\beta_1^l$  is the overall slope and  $u_i$  and  $v_i$  are the random intercept and slope of the  $i$ th individual respectively assuming that  $\boldsymbol{\omega}_i = (u_i, v_i) \sim \mathcal{N}(\mathbf{0}, \mathbf{D})$ , where

$$\mathbf{D} = \begin{pmatrix} \sigma_u^2 & 0 \\ 0 & \sigma_v^2 \end{pmatrix},$$

$l = 1, \dots, k$ ,  $i = 1, \dots, n$ ,  $j = 1, \dots, t_{ij}^l$ .

Table 4.3 shows the results of the multivariate longitudinal model proposed in Equation (4.11) applied to *activity*, *impacts* and *symptoms* dimensions of the SGRQ in the COPD Study. It also displays the longitudinal univariate results of each dimension. The main aim of this analysis is to measure the evolution of the patients with COPD and hence, contrary to the previous COPD multivariate analysis in Section 4.4.1, all the repeated measurements of each individual have been introduced in the model. In addition, in order to assess the evolution of the patients, such as performed in Chapter 3 Section 3.5.2 and as it was showed in Equation (4.11), only time has been introduced as a covariate in the model.

Table 4.3 displays the estimates and standard deviations of all the parameters in the model together with the p-values associated with the effect of the covariates. It must be noticed that in the *impacts* dimension there is no estimation of the random slope in the univariate model because, as it was pointed out in Chapter 3, the estimation of  $\sigma_v$  tended to zero.

Table 4.3: Multivariate and univariate longitudinal analysis of the SGRQ in COPD Study.

	Multivariate			Univariate (Chapter 3)		
	Estimate	SD	p-value	Estimate	SD	p-value
<b>Activity</b>						
<i>Intercept</i>	-0.151	0.030	< 0.001	-0.185	0.026	< 0.001
<i>Time</i>	0.012	0.010	0.212	0.027	0.008	0.001
$\log(\phi)$	-2.627	0.058	–	-3.305	0.087	–
$\sigma_u$	0.941	0.029	–	1.180	0.037	–
$\sigma_v$	0.130	0.005	–	0.126	0.006	–
<b>Impacts</b>						
<i>Intercept</i>	-0.962	0.025	< 0.001	-0.993	0.026	< 0.001
<i>Time</i>	-0.003	0.008	0.715	0.004	0.008	0.670
$\log(\phi)$	-3.627	0.095	–	-3.541	0.095	–
$\sigma_u$	0.941	0.029	–	1.060	0.033	–
$\sigma_v$	0.130	0.005	–	–	–	–
<b>Symptoms</b>						
<i>Intercept</i>	-0.356	0.033	0.000	-0.305	0.029	0.000
<i>Time</i>	0.025	0.011	0.019	0.006	0.009	0.499
$\log(\phi)$	-2.228	0.045	–	-2.617	0.058	–
$\sigma_u$	0.941	0.029	–	0.759	0.025	–
$\sigma_v$	0.130	0.005	–	0.097	0.006	–

SD: Standard Deviation;  $\sigma_u$ : standard deviation of the random intercepts;  $\sigma_v$ : standard deviation of the random slopes.

The first conclusion we address from Table 4.3 is that, although estimates regarding the intercept are quite similar in both approaches, there are considerable differences in both the estimation and significance of the time covariate. In fact, it can be appreciated in Table 4.3 that in the univariate analysis the unique dimension where the evolution of the patients is statistically significant is the *activity*, whereas in the multidimensional approach it is the *symptoms*. Therefore, it seems that both regression approaches lead to different conclusions about the evolution of the patients with COPD. However, in our opinion, some details must be taken into account before performing that comparison, and we will discuss them in the next section.

## 4.5 Discussion and future work

In this chapter we have proposed a model based on a beta-binomial mixed-effect model approach for the joint analysis of different but correlated dimensions provided by different PRO questionnaires. The model is based on the shared random effects approach proposed by McCulloch (2008) and as the name indicates, several random effects are shared by different dimensions in order to accommodate the correlation. The model can be applied to cross-sectional as well as longitudinal (hierarchical) multivariate data. The advantage of the model is that, compared to other commonly used multivariate analysis techniques, such as transition or reduction approaches, dimensions drawn from different distributions can be jointly analysed and especially, that the unbalanceness of the data, or continuity of the time variable, is not a problem. In addition, compared to the correlated random effects model introduced in Section 4.3.2, we do not have to concern about the dimensionality of the model when a large amount of correlated dimensions are analysed, because the random effects in the model are the same regardless of the number of dimensions.

In Section 4.4.1, we have shown in Table 4.1 and Table 4.2 that in cross-sectional application the shared random effects approach lead to similar estimates compared to the univariate approach, however the measurement of the correlation between dimensions accommodates an extra variability. In fact, the estimates of the dispersion parameter of the beta-binomial distribution reduces in most of the times and consequently, standard deviations of the estimates became smaller. Hence, variables that in the univariate analysis were treated as non-informative, turned out to have a statistically significant effect in the multivariate approach.

On the contrary, in Section 4.4.2, Table 4.3 displays a completely different situa-

tion for the longitudinal application where variables that were statistically significant in the univariate analysis are not significant in the multivariate analysis and vice versa. These results lead to completely misleading conclusions about the evolution of patients with COPD. In Section 4.3.1, we have developed a procedure to compare regression coefficients in univariate and multivariate approaches in cross-sectional framework. Equation (4.5) shows that the relationship between regression parameters lies in a multiplicative term that depends on the variance of the random effects. However, as we will show in the longitudinal context the relationship is not so direct.

In shared random effects context, if we try to compare the fixed effects between both univariate and multivariate longitudinal approaches, the marginal expectation of an outcome in the multivariate approach is obtained by

$$\begin{aligned}
\mathbb{E}[Y_{ij}^l] &= \mathbb{E} \left[ \frac{\exp((\gamma_0^l + u_i) + (\gamma_1^l + v_i)T_{ij})}{1 + \exp((\gamma_0^l + u_i) + (\gamma_1^l + v_i)T_{ij})} \right] \\
&\approx \mathbb{E} \left[ \Phi \left( c((\gamma_0^l + u_i) + (\gamma_1^l + v_i)T_{ij}) \right) \right] \\
&= \mathbb{E} \left[ \Pr \left( Z < c \left( (\gamma_0^l + u_i) + (\gamma_1^l + v_i)T_{ij} \right) \mid u_i, v_i \right) \right] \\
&= \Pr \left[ Z < c \left( (\gamma_0^l + u_i) + (\gamma_1^l + v_i)T_{ij} \right) \right] \\
&= \Pr \left[ \frac{Z - c(u_i + v_i T_{ij})}{\sqrt{1 + c(\sigma_u + \sigma_v T_{ij})}} < \frac{c}{\sqrt{1 + c(\sigma_u + \sigma_v T_{ij})}} (\gamma_0^l + \gamma_1^l T_{ij}) \right] \\
&= \Phi \left( \frac{c}{\sqrt{1 + c(\sigma_u + \sigma_v T_{ij})}} (\gamma_0^l + \gamma_1^l T_{ij}) \right),
\end{aligned}$$

and in the univariate approach is given by

$$\begin{aligned}
\mathbb{E}[Y_{ij}^l] &= \mathbb{E} \left[ \frac{\exp((\beta_0^l + u_i^l) + (\beta_1^l + v_i^l)T_{ij})}{1 + \exp((\beta_0^l + u_i^l) + (\beta_1^l + v_i^l)T_{ij})} \right] \\
&\approx \Phi \left( \frac{c}{\sqrt{1 + c(\sigma_u^l + \sigma_v^l T_{ij})}} (\beta_0^l + \beta_1^l T_{ij}) \right),
\end{aligned}$$

where  $c = 16\sqrt{3}/(15\pi)$ ,  $l = 1, \dots, k$ ,  $i = 1, \dots, n$ ,  $j = 1, \dots, t_{ij}^l$ . Notice that, while in the first marginal expectation the random effects  $u_i$  and  $v_i$  correspond to the shared random effect model, in the second equation they correspond to the univariate analysis.

Therefore, we have the following relationship

$$\frac{1}{\sqrt{1 + c(\sigma_u + \sigma_v T_{ij})}}(\gamma_0^l + \gamma_1^l T_{ij}) = \frac{1}{\sqrt{1 + c(\sigma_u^l + \sigma_v^l T_{ij})}}(\beta_0^l + \beta_1^l T_{ij}), \quad (4.12)$$

which as it can be appreciated, does not allow a straightforward comparison of the parameters  $\beta$ . Consequently, unlike in the cross-sectional comparison in Equation (4.5), the estimates of the fixed effects can differ considerably in univariate and multivariate approaches. In fact, the shared random effects approach considers that the subject-specific intercept and slope is equal in all the dimensions, i.e. that the initial status and evolution of each patient differs in the same magnitude from the population's overall initial status and evolution in all the dimensions. Therefore, based on this assumptions, compared to the univariate approach, modifications must be done in the estimation of the fixed effects in order to the model be correct. In addition, due to the shared random effects assumption, each of the approach leads to different estimations of the dispersion parameters  $\phi$ ,  $\sigma_u$  and  $\sigma_v$ . Indeed, compared to the cross-sectional framework, as it has been shown in Equation (4.6) and Equation (4.7), each random effect must account for two correlations, the within dimension and between dimensions. Consequently, as it can be appreciated in Table 4.3, the longitudinal multivariate approach enlarges the estimated value of  $\phi$  in most of the dimensions, increasing the standard deviation of the estimation of the fixed effects. Hence, different estimations and standard deviations can lead to different significance test, leading to different interpretation of the effects of the covariates.

In fact, we have mentioned that the shared random effects modelling approach could have some inconveniences when dealing with longitudinal data as the same random effects must accommodate the correlation between dimensions and within dimensions and in some cases, the considered assumption could be too rigid. In Section 4.3 we have detailed that the correlation between two latent variables associated with two different dimensions is determined by the correlation within dimensions, which may be inappropriate in cases where there is a high correlation between the repeated measurements, but dimensions are slightly correlated. The assumption that the initial value and the evolution of each individual behaves equally with respect to the population overall trend in all the dimensions may be quite unrealistic indeed.

---

---

## CHAPTER 5

---


# SOFTWARE DEVELOPMENT

*“We have the duty of formulating, of summarizing, and of communicating our conclusions, in intelligible form, in recognition of the right of other free minds to utilize them in making their own decisions.”*

---

Ronald Fisher, 1890-1962

*The article based on the work developed in this chapter is under preparation for submitting it to a journal.*

 **Najera-Zuloaga, J., Lee, D.-J. and Arostegui, I. (2017) PROreg: An R Package for Analysis of Patient Reported Outcomes. (under preparation)**

This chapter is focused on the implementation of the different analysis models we have described and developed throughout the thesis in R language. We have repeatedly mentioned that the beta-binomial distribution does not belong to the exponential family and consequently, classical estimation procedures cannot be applied. That is why there is not much literature regarding modelling approaches dealing with the beta-binomial distribution. Hence, due to the fact that few methods have been developed, implementation of algorithms based on the beta-binomial distribution are not common.

The objective of this chapter is divided in two parts. In the first part, we introduce and describe the R-packages available in the literature for the beta-binomial distribution. As far as we know, it does not exist an R-package containing all the methodology developed in this thesis. In fact, most of the packages we will mention



restrict their performance to distributional fit and consequently, they do not offer any regression approach based on the beta-binomial distribution. Therefore, the second objective is to implement, among others, the statistical models developed in this thesis in a R-package called `PROreg`, as there is no software implementation for some of regression approaches. Additionally, the package contains many other functions focus in the analysis of PROs, such as binomial with dispersion parameter distribution based regression approaches or questionnaire specific functions. Indeed, the objective of `PROreg` is to unify in a single R-package all the required statistical methodology to analyse PROs, or any overdispersed binomial form, discrete and bounded data.

## 5.1 R-packages for the beta-binomial distribution

There do not exist many functions that offer statistical analysis based on the beta-binomial distribution in public domain software, especially in R language. In fact, most of the packages containing functions to deal with the beta-binomial distribution restrict their applicability to distributional fits. Therefore, we think that the existing R implementation has limitations when trying to analyse in both, independent or hierarchical data framework, beta-binomial data through a regression approach, favouring the use of GLMMs or GLMs, or even LMMs.

Table 5.1 shows the different R-packages which contain functions for dealing with the beta-binomial distribution in a variety of analysis approaches. In the table, it can be appreciated that although all the packages contain functions to calculate the density and generate random realizations from a beta-binomial distribution for a given set of parameters, most of them do not extend the analysis to an estimation framework. Indeed, the `VGAM` and `gamlss` are the only packages which can perform regression models based on the beta-binomial distribution, and furthermore, the `VGAM` limits its applicability to independent data. Therefore, we conclude that the `gamlss` is the most complete package regarding the beta-binomial distribution as it contains functions that cover all the analysis situations displayed in Table 5.1. However, it has been mentioned and discussed in Chapter 2 and Chapter 3 that although the `gamlss()` cross-sectional regression analysis is based on the marginal beta-binomial regression approach, there are substantial differences between the estimation procedure of the proposed `BBmm` and `gamlss()` implementations. In fact, in Chapter 3 it was shown that the penalisation of the profile likelihood of the dispersion parameter of the beta-binomial in `BBmm` approach, improves the results

of the parameter estimation in terms of reducing the bias. However, there is no R-package that performs the BBmm approach developed in Chapter 3.

Table 5.1: R-packages for beta-binomial analysis.

R-package	Version	Distribution			Regression	
		Dens.	Rand.	Est.	Indep.	Corr.
<code>rmutil</code>	1.1.0	<code>dbetabinom()</code>	<code>rbetabinom()</code>	–	–	–
<code>TailRank</code>	3.1.3	<code>dbb()</code>	<code>rbb()</code>	–	–	–
<code>emdbook</code>	1.3.9	<code>dbetabinom()</code>	<code>rbetabinom()</code>	–	–	–
<code>VGAM</code>	1.0-4	<code>dbetabinom()</code>	<code>rbetabinom()</code>	<code>vglm()</code>	<code>vglm()</code>	–
<code>gamlss</code>	5.0-2	<code>dBB</code>	<code>rBB</code>	<code>gamlss()</code>	<code>gamlss()</code>	<code>gamlss()</code>

Dens.: Density function; Rand: Random realizations; Est.: Estimation of the distribution; Inped.: Independent (cross-sectional) framework; Corr.: Correlated (hierarchical) framework; Mult.: Multivariate (shared random effects approach).

## 5.2 PROreg R-package

PROreg R-package is available at <https://cran.r-project.org/web/packages/PROreg/>. The name of the package stands for ‘Patient Reported Outcomes Regression’ and it offers a variety of tools, such as specific plots and regression model approaches, for analysing different patient reported questionnaires. However, although it is focused on the analysis of PROs, any binomial form outcome can be analysed using PROreg. In fact, it contains regression models based on the binomial, binomial with dispersion parameter and beta-binomial distribution for the correct analysis of any discrete and bounded data. In addition, regression models have been developed in either cross-sectional or hierarchical data, which increases its applicability in many real situations.

There are many packages such as `stats`, that offer the analysis of regression models in cross-sectional or hierarchical data based on the binomial or binomial with dispersion parameter distributions (`family=quasibinomial`). In fact, the novelty of the package lies in the variety of regression models based on the beta-binomial distribution which are not so typical, especially in hierarchical framework. Apart from the mentioned regression models, PROreg package contains additional plotting or questionnaire-specific recoding algorithms. Indeed, the objective of the package is to unify all the requires techniques when dealing with PROs in a regression context in an R-package. Therefore, the PROreg R-apackage consists of the following main

functions:

Table 5.2: Different functions available in `PR0reg` R-package

Beta-binomial	Binomial	Additional functions
<code>BB()</code>	<code>BI()</code>	<code>SF36rec()</code>
<code>·dBB()</code>	<code>·dBI()</code>	<code>HRQoLplot</code>
<code>·rBB()</code>	<code>·rBI()</code>	
<code>BBest()</code> <sup>†</sup>	<code>BIest()</code> <sup>†</sup>	
<code>BBreg()</code> <sup>†‡</sup>	<code>BIreg()</code> <sup>†‡</sup>	
<code>BBmm()</code> <sup>†‡</sup>	<code>·BIiwlS</code>	
<code>·EffectsEst_BBDelta()</code>	<code>BIImm()</code> <sup>†‡</sup>	
<code>·EffectsEst_NR()</code>		
<code>·BBmm_VarEst()</code>		

<sup>†</sup> stands for functions that contain `print()`; <sup>‡</sup> stands for functions that contain `summary()` and `print.summary()`.

`PR0reg` package contains dependencies from other packages: `fmsb`, `RColorBrewer`, `car`, `matrixcalc`, `rootSolve`, `numDeriv`, `Matrix`.

As mentioned before, the novelty of the package lies in the different regression approaches that can be performed based on the beta-binomial distribution. Therefore, following we are going to fully describe, both theoretically and practically, the functions for analysing beta-binomial distributed data.

### 5.2.1 BB: The beta-binomial distribution

#### Description

Density and random generation for the beta-binomial distribution.

#### Usage

```
dBB(m,p,phi)
rBB(k,m,p,phi)
```

### Arguments

<code>k</code>	number of simulations.
<code>m</code>	maximum score number in each beta-binomial observation.
<code>p</code>	probability parameter of the beta-binomial distribution.
<code>phi</code>	dispersion parameter of the beta-binomial distribution.

### Details

The beta-binomial distribution consists of a finite sum of Bernoulli dependent variables whose probability parameter is random and follows a beta distribution. Assume that we have  $Y_j$  a set of variables,  $j = 1, \dots, m$ , with  $m$  integer, that conditioned on a random variable  $u$ , are independent and follow a Bernoulli distribution with probability parameter  $u$ . On the other hand, the random variable  $u$  follows a beta distribution with parameter  $p/\phi$  and  $(1-p)/\phi$ . Namely,

$$Y_j \sim \text{Ber}(u), \quad u \sim \text{Beta}(p/\phi, (1-p)/\phi),$$

where  $0 < p < 1$  and  $\phi > 0$ . The first and second order marginal moments of this distribution are defined as

$$\mathbb{E}[Y_j] = p, \quad \text{Var}[Y_j] = p(1-p),$$

and correlation between observations is defined as

$$\text{Corr}[Y_j, Y_k] = \phi/(1+\phi),$$

where  $j, k = 1, \dots, m$  are different. Consequently,  $\phi$  can be considered as a dispersion parameter.

If we sum up all the variables we will define a new variable which follows a new distribution that is called beta-binomial, and it is defined as follows. The variable  $Y$  follows a beta-binomial distribution with parameters  $m$ ,  $p$  and  $\phi$  if

$$Y|u \sim \text{Bin}(m, u), \quad u \sim \text{Beta}(p/\phi, (1-p)/\phi).$$

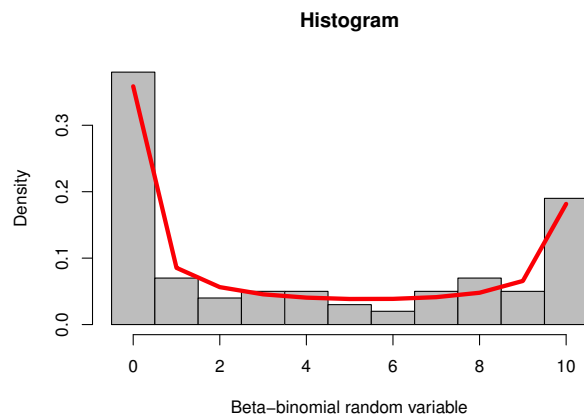
## Value

`dBB` gives the density of a beta-binomial distribution with the defined `m`, `p` and `phi` parameters.

`rBB` generates `k` random observations based on a beta-binomial distribution with the defined `m`, `p` and `phi` parameters.

## Examples

```
> set.seed(12)
> # We define the parameters for the simulation.
> m <- 10
> p <- 0.4
> phi <- 1.8
>
> # We perform k beta-binomial simulations for those parameters.
> k <- 100
> bb <- rBB(k,m,p,phi)
> bb
 [1] 0 10 10 9 10 10 8 0 10 4 7 10 1 0 0 10 8 5 3 0 9 2 4
[24] 0 7 0 1 0 8 6 5 0 0 0 8 0 0 0 1 10 9 7 10 1 10 3
[47] 8 4 0 8 0 10 0 2 0 1 10 10 0 4 9 2 0 8 0 0 3 9 7
[70] 4 0 5 0 0 6 7 0 1 10 0 0 0 0 0 0 3 0 0 2 0 0 3
[93] 10 0 10 10 10 0 10 1
>
> # We estimate the probability of each point for the fixed parameters.
> dd <- dBB(m,p,phi)
> # We are going to plot the histogram of the created variable,
> # and using dBB() function we are going to fit the distribution:
> hist(bb,col="grey",breaks=seq(-0.5,m+0.5,1),probability=TRUE,
      main="Histogram",xlab="Beta-binomial random variable")
```



## 5.2.2 BBest: Estimation of the parameters of a beta-binomial distribution

### Description

This function performs the estimation of the parameters of a beta-binomial distribution for the given data and maximum score number in each observation.

There are two different approaches available for performing the estimation of the parameters: (i) method of moments, and, (ii) maximum likelihood approach.

### Usage

```
BBest(y,m,method="MM")
```

### Arguments

<b>y</b>	response variable which follows a beta-binomial distribution.
<b>m</b>	maximum score number in each beta-binomial observation.
<b>method</b>	the method used for performing the estimation of the probability and dispersion parameters of a beta-binomial distribution, "MM" for the method of moments and "MLE" for maximum likelihood estimation. Default "MM".

### Details

BBest function performs the estimation and inference of the parameters of a beta-binomial distribution for the given data. The estimations can be performed using two different approaches, the method of moments (MM) and the maximum likelihood estimation (MLE) approach.

The density function of a given observation  $y$  that follows a beta-binomial distribution with parameters  $m$ ,  $p$  and  $\phi$  is defined as

$$f(y|p, \phi) = \binom{m}{y} \frac{\Gamma(p/\phi + y)}{\Gamma(p/\phi)} \frac{\Gamma((1-p)/\phi + m - y)}{\Gamma((1-p)/\phi)} \frac{\Gamma(1/\phi)}{\Gamma(1/\phi + m)}.$$

The first and second order moments are defined as

$$\begin{aligned}\mathbb{E}[Y] &= mp, \\ \mathbb{V}\text{ar}[Y] &= mp(1-p) \left[ 1 + (m-1) \frac{\phi}{1+\phi} \right].\end{aligned}$$

Hence, if  $\mathbf{y} = (y_1, \dots, y_n)'$  is the given data, we can conclude with the MM from the previous as

$$\begin{aligned}\hat{p}_{MM} &= \mathbb{E}[\mathbf{y}]/m, \\ \hat{\phi}_{MM} &= \frac{\mathbb{V}\text{ar}[\mathbf{y}] - m\hat{p}_{MM}(1 - \hat{p}_{MM})}{m^2\hat{p}_{MM}(1 - \hat{p}_{MM}) - \mathbb{V}\text{ar}[\mathbf{y}]}\end{aligned}$$

where  $\mathbb{E}[\mathbf{y}]$  is the sample mean and  $\mathbb{V}\text{ar}[\mathbf{y}]$  is the sample variance.

On the other hand, the MLE of both parameters is obtained from the maximisation of the log-likelihood of the model defined as

$$\begin{aligned}\log L(p, \phi | \mathbf{y}) &= \sum_{i=1}^n \left[ \log \binom{m}{y_i} + \sum_{k=0}^{y_i-1} \log(p + k\phi) + \sum_{k=0}^{m-y_i-1} \log(1 - p + k\phi) \right. \\ &\quad \left. - \sum_{k=0}^{m_i} \log(1 + k\phi) \right],\end{aligned}$$

where if  $y_i = 0$  then  $\sum_{k=0}^{y_i-1} \log(p + k\phi) = 0$  and if  $y_i = m_i$  then  $\sum_{k=0}^{m-y_i-1} \log(1 - p + k\phi) = 0$ . Therefore, the MLE of  $p$  and  $\phi$  are obtained solving the following score equations,

$$\begin{aligned}S(p) &= \frac{\partial}{\partial p} \log L(p, \phi | \mathbf{y}) = \sum_{i=1}^n \left[ \sum_{k=0}^{y_i-1} \frac{1}{p + k\phi} + \sum_{k=0}^{m-y_i-1} \frac{1}{1 - p + k\phi} \right] = 0 \\ S(\phi) &= \frac{\partial}{\partial \phi} \log L(p, \phi | \mathbf{y}) = \sum_{i=1}^n \left[ \sum_{k=0}^{y_i-1} \frac{k}{p + k\phi} + \sum_{k=0}^{m-y_i-1} \frac{k}{1 - p + k\phi} \right. \\ &\quad \left. - \sum_{k=0}^m \frac{k}{1 + k\phi} \right] = 0\end{aligned}$$

Numerical algorithms are required to solve both equations. In the `PROreg` package we have implemented a numerical approximation algorithm based on the Newton-Raphson procedure.

## Value

`BBest` returns an object of class 'BBest'.

The function `summary` (i.e., `summary.BBest`) can be used to obtain or print a summary of the results.

<code>p</code>	estimated probability parameter of the beta-binomial distribution.
<code>phi</code>	estimated dispersion parameter of the beta-binomial distribution.
<code>pVar</code>	variance of the estimation of the probability parameter <code>p</code> .
<code>psi</code>	estimation of the logarithm of the dispersion parameter <code>phi</code> .
<code>psiVar</code>	variance of the estimation of the logarithm of the dispersion parameter <code>psi</code> .
<code>m</code>	maximum score number in each beta-binomial observation.
<code>balanced</code>	if the response variable is balanced it returns 'yes', otherwise 'no'.
<code>method</code>	the used approach for performing the estimations.

## Examples

```
> y <- rBB(k,m,p,phi)
>
> # Performing the estimation of the parameters
>
> # Method of moments:
> MM <- BBest(y,m)
>
> MM
The probability parameter of the beta-binomial distribution: 0.7124
The dispersion parameter of the beta-binomial distribution: 1.629548

Balanced data, maximum score number: 10
>
> # Maximum likelihood approach
> MLE <- BBest(y,m,method="MLE")
>
> MLE
The probability parameter of the beta-binomial distribution: 0.7066453
The dispersion parameter of the beta-binomial distribution: 1.621879

Balanced data, maximum score number: 10
```



### 5.2.3 BBreg: Fit a marginal beta-binomial regression model

#### Description

BBreg function fits a marginal beta-binomial regression model, i.e., it links the probability parameter of a beta-binomial distribution with the given covariates by means of a logistic link function. The estimation of the parameters in the model is done via maximum likelihood estimation.

#### Usage

```
BBreg(formula,m,data,maxiter=100)
```

#### Arguments

<b>formula</b>	an object of class 'formula' (or one that can be coerced to that class): a symbolic description of the model to be fitted.
<b>m</b>	maximum score number in each beta-binomial observation.
<b>data</b>	an optional data frame, list or environment (or object coercible by as.data.frame to a data frame) containing the variables in the model. If not found in data, the variables are taken from environment (formula).
<b>maxiter</b>	the maximum number of iterations in the estimation process. Default 100.

#### Details

The BBreg function performs a marginal beta-binomial regression model for a given outcome and covariates. Assume that we have a set of variables  $\mathbf{Y} = (Y_1, \dots, Y_n)'$  which follow a beta-binomial distribution which we want to model as a function of some covariates  $X_1, \dots, X_k$ . Hence, similar to the logistic regression, we can connect the probability parameter of the beta-binomial distribution with the given covariates by means of a logistic link function as

$$\eta_i = \text{logit}(p_i) = \log \frac{p_i}{1 - p_i} = \mathbf{x}'_i \boldsymbol{\beta},$$

where  $\mathbf{x}_i$  is the  $i$ th row of a full rank matrix  $\mathbf{X}$  composed by the given covariates and  $\boldsymbol{\beta}$  are the regression coefficients.

The estimation of the parameters in the model is done via maximum likelihood estimation. The log-likelihood of the model for a set of observations  $\mathbf{y} = (y_1, \dots, y_n)'$  is defined as

$$\log L(\boldsymbol{\beta}, \phi | \mathbf{y}) = \sum_{i=1}^n \left[ \log \binom{m_i}{y_i} + \sum_{k=0}^{y_i-1} \log(p_i + k\phi) + \sum_{k=0}^{m_i-y_i-1} \log(1 - p_i + k\phi) - \sum_{k=0}^{m_i} \log(1 + k\phi) \right],$$

where  $\boldsymbol{\beta}$  enters in the equation through  $p_i$ .

Regarding the estimation procedures, the regression coefficients  $\boldsymbol{\beta}$  are calculated solving the following score function,

$$S(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} \log L(\boldsymbol{\beta}, \phi | \mathbf{y}) = \boldsymbol{\xi}' \mathbf{S} \mathbf{X}$$

where  $\mathbf{S} = \text{diag}(p_1(1-p_1), \dots, p_n(1-p_n))$  and  $\boldsymbol{\xi} = (\xi_1, \dots, \xi_n)'$  being

$$\xi_i = \sum_{k=0}^{y_i-1} \frac{1}{p_i + k\phi} - \sum_{k=0}^{m_i-y_i-1} \frac{1}{1 - p_i + k\phi}.$$

In this package we have developed an estimation procedure based on the delta algorithm which leads to the following iterative estimation equation,

$$\hat{\boldsymbol{\beta}}^{(r+1)} = (\mathbf{X}' \mathbf{S} \mathbf{V} \mathbf{S} \mathbf{X})^{-1} \mathbf{X}' \mathbf{S} \mathbf{V} \mathbf{S} \boldsymbol{\nu},$$

where  $\boldsymbol{\nu} = \mathbf{X} \boldsymbol{\beta}^{(r)} + (\mathbf{S} \mathbf{V})^{-1} \boldsymbol{\xi}$ , and the previous matrices are evaluated at  $\hat{\boldsymbol{\beta}}^{(r)}$ .

The estimates of  $\boldsymbol{\beta}$  are functions of  $\phi$ . Hence, if we replace  $\boldsymbol{\beta}$  with  $\hat{\boldsymbol{\beta}}$  or, equivalently,  $\mathbf{p}$  with  $\hat{\mathbf{p}}$ , in the log-likelihood function, we obtain the profile log-likelihood with respect to  $\phi$ . In **BBreg** function we have developed a Newton-Raphson based algorithm for the estimation of  $\phi$ .

## Value

**BBreg** returns an object of class 'BBreg'.

The function `summary` (i.e., `summary.BBreg`) can be used to obtain or print a summary of the results.

<code>coefficients</code>	the estimated value of the regression coefficients.
<code>vcov</code>	the variance-covariance matrix of the estimated coefficients of the regression.
<code>phi</code>	the estimation of the dispersion parameter of the beta-binomial distribution.
<code>psi</code>	the estimation of the logarithm of the dispersion parameter of the beta-binomial distribution.
<code>psi.var</code>	the variance of the estimated logarithm of the dispersion parameter of the beta-binomial distribution.
<code>conv</code>	convergence of the methodology. If the method has converged it returns 'yes', otherwise 'no'.
<code>fitted.values</code>	the fitted mean values of the model.
<code>deviance</code>	the deviance of the model.
<code>df</code>	degrees of freedom of the model.
<code>null.deviance</code>	null-deviance, the deviance for the null model. The null model will only include an intercept as the estimation of the probability parameter.
<code>null.df</code>	the degrees of freedom for the null model.
<code>iter</code>	number of iterations in the estimation process.
<code>X</code>	the model matrix.
<code>y</code>	the dependent response variable in the model.
<code>m</code>	maximum score number in each beta-binomial observation.
<code>balanced</code>	if the response beta-binomial variable is balanced it returns 'yes', otherwise 'no'.
<code>nObs</code>	number of observations.
<code>call</code>	the matched call.
<code>formula</code>	the formula supplied.

### Examples:

```

> # We simulate a covariate, fix the parameters of the beta-binomial
> # distribution and simulate a response variable. Then we apply the
> # model, and try to get the same values.
>
> set.seed(18)
> k <- 1000
> m <- 10
> x <- rnorm(k,5,3)

```

```

> beta <- c(-10,2)
> p <- 1/(1+exp(-(beta[1]+beta[2]*x)))
> phi <- 1.2
> y <- rBB(k,m,p,phi)
>
> model <- BBreg(y~x,m)
> model
Call: BBreg(formula = y ~ x, m = m)

Beta coefficients:
      Intercept          x
[1,] -10.13476  2.032702

Dispersion parameter: 1.264069

Deviance: 1236.094 on 997 degrees of freedom
Null deviance: 3437.819 on 999 degrees of freedom

Balanced data, maximum score in the trials: 10

```

### 5.2.4 BBmm: Fit a beta-binomial mixed-effects regression model

#### Description

BBmm function performs beta-binomial mixed-effects models, i.e., it allows the inclusion of Gaussian random effects in the linear predictor of a marginal beta-binomial logistic regression model in order to accommodate the correlation among the outcomes. It allows the joint estimation of more than one outcome vector in a multivariate framework.

Each component of the model can be specified by means of two different ways:

*Outcomes:* (i) determining the `fixed.formula` argument, or (ii) including the vector of the outcomes `y`.

*Fixed part:* (i) determining the `fixed.formula` argument, or (ii) specifying the model matrix of the covariates `X`.

*Random part:* (i) determining the `random.formula` argument, or (ii) specifying the model matrix of the random effects, `Z`, and determining the number of random effects in each random component, `nRandComp`.

The estimation of the fixed and random effects in the model can be done by means of two approaches: (i) BB-Delta, the delta algorithm developed for the beta-binomial mixed-effects model, and (ii) using the NR R-package. The selected method must be specified in the arguments of the function.

## Usage

```
BBmm(fixed.formula=NULL,X=NULL,y=NULL,random.formula=NULL,Z=NULL,
nRandComp=NULL,m,data=list(),method="BBNR",maxiter=50,show=FALSE,
nDim=1)
```

## Arguments

<code>fixed.formula</code>	an object of class 'formula' (or one that can be coerced to that class): a symbolic description of the fixed part of the model to be fitted.
<code>X</code>	design matrix composed by the given covariates in the model. It must be only specified in cases where the <code>fixed.formula</code> argument is not determined.
<code>y</code>	the vector of the outcomes that are going to be modelled as a function of the covariates. It must be only specified in cases where the <code>fixed.formula</code> argument is not determined.
<code>random.formula</code>	an object of class "formula" (or one that can be coerced to that class): a symbolic description of the random part of the model to be fitted.
<code>Z</code>	design matrix composed by the correlation, or random effects structure, of the model. It must be only specified in cases where the <code>random.formula</code> argument is not determined.
<code>nRandComp</code>	the number of random effects in each random component of the model. It must be specified as a vector where the $i$ th value corresponds with the number of random effects of the $i$ th random component. It must be only included when the random structure of the model is described through the matrix of the random effects <code>Z</code> .
<code>m</code>	maximum score number in each beta-binomial observation.
<code>data</code>	an optional data frame, list or environment (or object coercible by <code>as.data.frame</code> to a data frame) containing the variables in the model. If not found in data, the variables are taken from environment (formula).
<code>method</code>	the method for the estimation of the fixed and random effects in the model. Two options are available: (i) 'BB-Delta', the delta algorithm developed for the estimation procedure of the beta-binomial mixed-effects regression approach; (ii) 'NR', general Newton-Raphson algorithm for estimating the root of a set of $n$ (nonlinear) equations.
<code>maxiter</code>	the maximum number of iterations in the estimation process. Default 50.
<code>show</code>	logical, if TRUE, then the tolerance of the stop criterion together with the maximum difference of the fixed effects, beta-binomial log-dispersion parameter and random effects standard deviation with respect to the previous estimation is shown in each iteration.
<code>nDim</code>	number of dimensions that are going to be jointly analysed. Default 1.

### Details

`BBmm` function performs a beta-binomial mixed effects models. It extends the marginal beta-binomial logistic regression to the inclusion of random effects in the linear predictor of the model. It is assumed that, conditional on some Gaussian random effects  $\mathbf{u}$ , each variable of the the response variable vector  $\mathbf{Y}$  follows a beta-binomial distribution of parameters  $m_i$ ,  $p_i$  and  $\phi_i$ ,

$$Y_i|\mathbf{u} \sim \text{BB}(m_i, p_i, \phi_i), \quad \mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{D})$$

where

$$\boldsymbol{\eta} = \log \left( \frac{\mathbf{p}}{1 - \mathbf{p}} \right) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}$$

being  $\mathbf{X}$  and  $\mathbf{Z}$  model matrices composed by the given covariates and random structure respectively and  $\mathbf{D}(\boldsymbol{\lambda})$  is determined by some dispersion parameters  $\boldsymbol{\lambda}$ , which are included in the parameter vector  $\boldsymbol{\theta} = (\phi, \boldsymbol{\lambda}')'$ .

The estimation of the fixed regression parameters  $\boldsymbol{\beta}$  and the prediction of the random effects  $\mathbf{u}$  is done via the maximum likelihood, where the marginal likelihood of the model is approximated though the joint-likelihood by a first order Laplace approximation,

$$l(\boldsymbol{\beta}, \mathbf{u}, \boldsymbol{\theta}|\mathbf{y}) \approx \log f(\mathbf{y}|\boldsymbol{\beta}, \mathbf{u}, \boldsymbol{\theta}) + \log f(\mathbf{u}|\boldsymbol{\theta}). \quad (5.1)$$

The previous formula does not have a closed form and numerical methods are needed for developing a estimation procedure. Two approaches are available in the `BBmm` function in order to perform the fixed and random effects estimation: (i) A special case of the delta algorithm developed for the beta-binomial mixed-effects model estimation, and (ii) a general Newton-Raphson algorithm.

The estimation of the dispersion parameters  $\boldsymbol{\theta}$  by the joint-likelihood may be substantially biased due to the previous estimation of the fixed and random effects. Consequently, a penalisation of the joint-likelihood must be performed in order to get an unbiased estimation of the dispersion parameters. Lee and Nelder (1996) proposed the adjusted profile h-likelihood for the correct estimation of the dispersion parameters in mixed-effects model framework,

$$h(\boldsymbol{\theta}|\hat{\boldsymbol{\beta}}, \hat{\mathbf{u}}, \mathbf{y}) = \log f(\mathbf{y}|\boldsymbol{\beta}, \mathbf{u}, \boldsymbol{\theta}) + \log f(\mathbf{u}|\boldsymbol{\theta}) + \frac{1}{2} \log [\det (2\pi\mathbf{H}^{-1})],$$

where  $\mathbf{H}$  is the Hessian matrix of the model, i.e. the second derivatives of the

log-likelihood with respect to  $\beta$  and  $u$ .

The BBmm methodology iterates between the estimations of the regression and dispersion parameters until convergence is reached. The convergence is reached when the tolerance of the model is lower than  $10^{-6}$ , where the tolerance for the  $(r + 1)$ th iteration is defined as

$$\text{tolerance}^{(r+1)} = \frac{\sum_{i=1}^n \left[ \eta_i^{(r)} - \eta_i^{(r+1)} \right]^2}{\sum_{i=1}^n \left[ \eta_i^{(r+1)} \right]^2}.$$

## Value

BBmm returns an object of class 'BBmm'.

The function `summary` (i.e., `summary.BBmm`) can be used to obtain or print a summary of the results.

<code>fixed.coef</code>	estimated value of the fixed effects of the regression.
<code>fixed.vcov</code>	the variance-covariance matrix of the estimated fixed effects of the regression.
<code>random.coef</code>	predicted random effects of the regression.
<code>sigma.coef</code>	estimated value of the standard deviations of the random effects.
<code>sigma.var</code>	variance of the estimation of the standard deviation of the random effects.
<code>phi.coef</code>	estimated value of the dispersion parameter of the beta-binomial distribution.
<code>psi.coef</code>	estimated value of the logarithm of the dispersion parameter of the beta-binomial distribution.
<code>psi.var</code>	variance of the estimation of the logarithm of the dispersion parameter of the beta-binomial distribution.
<code>fitted.values</code>	the fitted mean values of the probability parameter of the beta-binomial distribution.
<code>conv</code>	convergence of the methodology. If the method has converged it returns 'yes', otherwise 'no'.
<code>deviance</code>	deviance of the model.
<code>df</code>	degrees of freedom of the model.
<code>null.deviance</code>	null-deviance, deviance of the null model. The null model will only include an intercept as the estimation of the probability parameter.

<b>null.df</b>	degrees of freedom of the null model.
<b>nRand</b>	number of random effects.
<b>nComp</b>	number of random components.
<b>nRandComp</b>	number of random effects in each random component of the model.
<b>namesRand</b>	names of the random components.
<b>iter</b>	number of iterations in the estimation method.
<b>nObs</b>	number of observations in the data.
<b>y</b>	the vector of the outcomes that are going to be modelled as a function of the covariates.
<b>X</b>	design matrix composed by the given covariates in the model.
<b>Z</b>	design matrix composed by the correlation, or random effects structure, of the model.
<b>D</b>	variance-covariance matrix of the random effects.
<b>balanced</b>	if the response beta-binomial variable is balanced it returns 'yes', otherwise 'no'.
<b>m</b>	maximum score number in each beta-binomial observation.
<b>nDim</b>	number of dimensions that are going to be jointly analysed.
<b>call</b>	the matched call.
<b>formula</b>	the fixed and random supplied formulas. It only provides the formula if it has been previously specified in the arguments of the function. The first formula corresponds to the fixed part of the model, while the second formula corresponds to the random structure.

## Examples

```

> set.seed(15)
> # Defining the parameters
> nObs <- 500 # 500 observations
> m <- 10# balanced data, maximum score number equal to 10.
> nRandComp <- c(70,50) # number of random effects in each random component
> phi <- 1.1 # dispersion parameter of the beta-binomial distribution
> sigma1 <- 1.2 # standard deviation of the first random effect
> sigma2 <- 0.5 # standard deviation of the second random effect
> beta <- c(-1,3.25) # the fixed effects
>
> # Simulate
> x <- rnorm(nObs,0.5,1.5) # the covariate
> u1 <- rnorm(nRandComp[1],0,sigma1) # first random effects
> u2 <- rnorm(nRandComp[2],0,sigma2) # second random effects
> u <- as.vector(c(u1,u2))

```



```

> # Desing matrices
> X <- model.matrix(~x)
> z1 <- as.factor(rBB(nObs,nRandComp[1]-1,0.5,2)) # correlation structure for
the first random effect
> z2 <- as.factor(rBB(nObs,nRandComp[2]-1,0.5,2)) # correlation structure for
the second random effect
> Z1 <- model.matrix(~z1-1) # model matrix of the correlation structure for
the first random effect
> Z2 <- model.matrix(~z2-1) # model matrix of the correlation structure for
the second random effect
> Z <- cbind(Z1,Z2) # correlation structure of both random effects
>
> # Calculate the linear predictor and simulate the outcome variable
> eta <- X%*%beta+Z%*%u # the linear predictor
> p <- exp(eta)/(1+exp(eta)) # apply the antilogit to the linear predictor
> y <- rBB(nObs,m,p,phi) # simulate the outcome variable
>
> # Apply the model
> model <- BBmm1(fixed.formula=y~x,random.formula=~z1+z2,m=m)
Iteration number: 1
Iteration number: 2
Iteration number: 3
Iteration number: 4
Iteration number: 5
Iteration number: 6
Iteration number: 7
Iteration number: 8
Iteration number: 9
Iteration number: 10
Iteration number: 11
Iteration number: 12
Iteration number: 13
Iteration number: 14
Iteration number: 15
> summary(model)
Call: BBmm1(fixed.formula = y ~ x, random.formula = ~z1 + z2, m = m)

Fixed effects coefficients:

              Estimate   StdErr t.value  p.value
(Intercept) -1.02274    0.13551 -7.5474  4.44e-14 ***
x              3.26362    0.20147 16.1990 < 2.2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

-----
Random effects dispersion parameter(s):

      Estimate   StdErr
z1 1.2337309 0.1435245
z2 0.5323015 0.1105570

```

```

-----
Logarithm of beta-binomial dispersion parameter log(phi):

      Estimate   StdErr
1 0.08360297 0.1357834
-----

Deviance of the model: 215.3224 ; with 495 degrees of freedom.
Deviance of the null model 1334.249 ; with 496 degrees of freedom.
Deviance goodness-of-fit test p-value: 0

Number of observations: 500
Number of iterations: 15
Balanced data, maximum score number: 10
Number of random effects in each random component: 70 50
Number of analysed dimensions: 1

```

### 5.2.5 Additional functions

Apart from the described functions for dealing with beta-binomial data in different scenarios, the PROreg package also contains some other functions. As it was shown in Table 5.2, functions based on the binomial distribution can also be performed, whose implementation is very similar to the presented functions. For instance we can make use of:

#### **BI: The binomial distribution**

Similar to `dBB()` and `rBB()` in Section 5.2.1 we can use `dBI()` and `rBI()` to compute the density and random generation of the binomial with optional dispersion parameter distribution respectively.

#### **BIest: Estimation of the parameters of a binomial distribution with optional dispersion parameter**

`BIest()` estimates the parameters of a binomial distribution for a given outcome. The estimation of a dispersion parameter, which accommodates overdispersion, can be selected. Implementation is similar to `BBest()` in Section 5.2.2.

#### **BIreg: Fit a logistic regression model with optional dispersion parameter**

`BIreg()` fits a logistic regression with optional dispersion parameter which in case, it is calculated through quasi-likelihood theory introduced in Chapter 1. The implementation is quite similar to `BBreg()` function in Section 5.2.3.

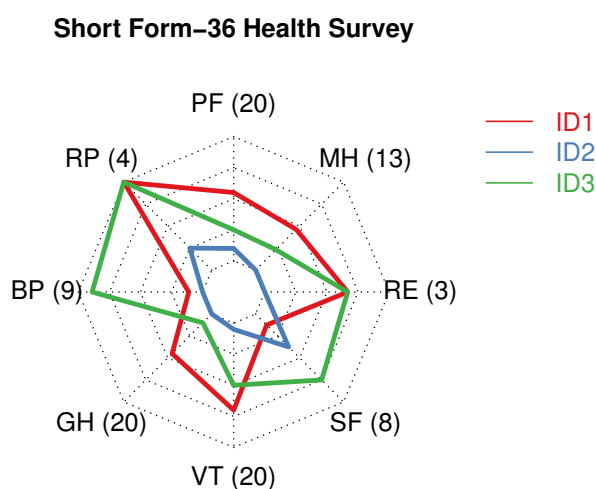
### BImm: Fit a logistic mixed-effects regression model

`BImm()` fits a logistic mixed-effects regression model where random effects are included in the linear predictor of a logistic regression. Implementation of the algorithm is quite similar to `BBmm()` introduced in Section 5.2.1.

Additionally, `PROreg` R-package includes some questionnaire-specific functions. For instance, two functions are available for the SF-36 Health Survey. First, the `SF36rec()` function performs the recoding of the SF-36 Health Survey based on Arostegui et al. (2013) (see Appendix A). The inputs of the function are (i) the dimension to be recoded and (ii) a scalar number from 1 to 8 which identifies the input vector with a SF-36 dimension. Second, the `HRQoLplot()` offers a nice visual descriptive analysis of all the dimensions provided by the SF-36. The HRQoL status of the individuals can be drawn by groups, clusters or covariates. Following we are going to display some code in order to show the easy application of the function and the nice output it provides.

```
> set.seed(9)
> # We insert the columns in the order that has been determined:
> m <- c(20,4,9,20,20,8,3,13)
> k <- 3
> p <- runif(3,0.2,0.8)
> phi <- runif(3,0.5,1)
> dat <- data.frame(rbind(ID1=rBB(8,m,p[1],phi[1]),
                          ID2=rBB(8,m,p[2],phi[2]),
                          ID3=rBB(8,m,p[3],phi[3])))

> colnames(dat) <- c("RF", "RP", "BP", "GH", "VT", "SF", "RE", "MH")
> HRQoLplot(dat,TRUE)
```



---

---

## CHAPTER 6

---

### CONCLUSIONS AND FURTHER WORK

*“There is nothing permanent except change”*

---

Heraclitus of Ephesus, c. 535– c. 475

PROs are important supplements to traditional medical outcomes which play an important role in health care and understanding health outcomes as they offer a new insight of the health-status of patients. PROs are increasingly used as primary outcomes and have gradually become an important element and a crucial source for monitoring disease condition or assessing the effectiveness of treatment, especially in chronic diseases or some health problems such as subjective discomfort and psychological distress (Chang, 2007). In fact, the U.S. Food and Drug Administration (FDA) has recommended that objective indicators combined with PROs are considered a more comprehensive form of outcome evaluation since 2006 (Speight and Barendse, 2010). Measuring PROs can help determine the burden of preventable diseases, injuries, and disabilities, and it can provide valuable new insights into the relationships between PROs and risk factors.

The measurement of the effect that some risk factors may have on the health-status of patients provided by PROs, requires the development of regression models. Most of the regression techniques developed in the literature, such as GLMs (McCullagh and Nelder, 1989) or GLMMs (McCulloch and Searle, 2001), are based on the assumption that the outcome variable follows a distribution from the exponential family. In fact, exponential family distributions offer good properties which simplify the estimation and inference procedure of regression models in many situa-

tions. However, it has been shown in the literature that PROs are usually displayed following a discrete and bounded U, J or inverse J-shaped distribution, accumulating values in one or both edges of the scale (Arostegui et al., 2007; Najera-Zuloaga et al., 2017). Hence, exponential family distributions, and especially, due to the nature of PROs, the binomial distribution, offers an inappropriate distributional fit to PROs. In order to overcome the inappropriate distributional fit provided by exponential family distributions, the beta-binomial distribution has been proposed in the literature as a good candidate for modelling the shape of PROs (Arostegui et al., 2007). The beta-binomial distribution is defined as a binomial distribution whose probability parameter is random and drawn from a beta distribution. Hence, it does not belong to the exponential family. Therefore, regression models based on the beta-binomial distribution are not so common in the literature.

The first objective of the thesis was to establish methods to provide inference in a cross-sectional beta-binomial regression model for the analysis of PROs. In Chapter 2, we have shown that the beta-binomial distribution can be defined by means of two different approaches, marginally or conditionally, where each approach leads to a different regression model, BBreg and BBhglm. In that chapter, we have compared the performance of both approaches when analysing PROs. It is well-known that marginal and the conditional approaches provide different regression results as they are modelling different expectations (Lee and Nelder, 2004). The marginal approach refers to the whole population and it models the marginal expectation. On the contrary, conditional approach refers to individuals and it models the conditional expectation. However, we have shown through a real data application and a simulation study that covariates that have a marginally statistical significant effect in PROs turn out not to be statistically significant in the conditional approach. The objective of PRO studies is usually the measurement of the effect of risk factors on the health-status of some patients, and therefore, conclusions are drawn from the whole population suffering from the disease. Therefore, the use of the conditional approach can sometimes mask the effect of some covariates, making them not relevant when they are statistically significant. Moreover, the assumption of a covariate that does not affect the individuals, but affects the population can be misleading. Following the argument by Senn in the comments to the paper by Lee and Nelder (2004), “After all, if the treatment cannot affect individuals, it has no effect on populations...”. Therefore, we suggest the use of the marginal BBreg approach for the analysis of PROs when there is not correlation in the data.

The second objective of the thesis was the extension to the of the cross-sectional beta-binomial model to a hierarchical regression framework. In Chapter 3, we have developed the BBmm approach which consists of the inclusion of Gaussian random effects in the linear predictor of BBreg. That way, the Gaussian random effects account for the hierarchical, or correlation, structure of the data while the beta-binomial distribution accommodates the properties of PROs. We propose a procedure based on the Laplace approximation and the Delta method for the estimation of all the parameters involving the model. The proposal consists of the estimation of fixed and random effects in the model through the maximisation of the joint likelihood, while the dispersion parameters are estimated by maximising a penalised profile likelihood proposed by Lee and Nelder (2001). We compare the performance of BBmm with other similar approaches in the literature. For instance, BBmm has been compared to hierarchical generalised linear models (HGLMs)(Lee and Nelder, 1996) and combined models (Molenberghs et al., 2010). However, special attention has been focused on the comparison of the BBmm with the generalised additive models for location, scale and shape (GAMLSS) by Rigby and Stasinopoulos (2005), due to the similarities between both modelling approaches. The main difference between GAMLSS and BBmm approaches is the penalisation of the profile likelihood in order to estimate the dispersion parameter of the beta-binomial distribution. We have shown through a simulation study that the proposed penalisation not only improves the estimation of the dispersion parameter of the beta binomial distribution in terms of bias, but also the estimation of the rest of parameters in the model. Therefore, we conclude that BBmm approach is more convenient than GAMLSS to make inference based on the beta-binomial mixed-effects model. Although, BBmm estimation procedure has been restricted to the beta-binomial distribution case, the penalisation of the profile likelihood in order to estimate dispersion (or non-canonical) parameters of the model, it can be useful in any other situation. However, the previous hypothesis goes beyond the objectives of this thesis and therefore, it will be considered as future work.

PROs are usually measured using questionnaires that decompose the health aspect they are assessing in different dimensions. Until now, either in a cross-sectional or hierarchical scenario, the models we have proposed analysed each health dimension separately. However, it seems logical to think that dimensions that contain measurements provided by the same individuals may have a correlation. Therefore, the third objective was the proposal of a multivariate regression model based on the beta-binomial distribution, which has been carried out in Chapter 4. Due to

the advantages that mixed-effects model framework provide, we have considered the shared random effects approach (McCulloch, 2008) and we have applied it to the beta-binomial mixed-effects models developed in Chapter 3. Therefore, the shared random effects model consist of including the same random effects in the linear predictor of each dimension and, in such a way that they allow for a correlation structure among them. We have obtained quite consistent results for cross-sectional studies and some preliminary results when analysing longitudinal data.

The development of statistical models may be pointless unless a tool would be provided for researchers. Therefore, the fourth objective of the thesis consisted of the development of an R-package which compiles, among others, all the methodologies proposed in this thesis. The name of the implemented R-package is **PROreg** (*Patient-Reported Outcomes regression analysis*) and it is already available at CRAN <https://cran.r-project.org/web/packages/PROreg/>. The description of the main functions as well as any other information has been provided in Chapter 5. Figure 6.1 shows the number of downloads of the R-package since it was uploaded to CRAN untill october 2017.

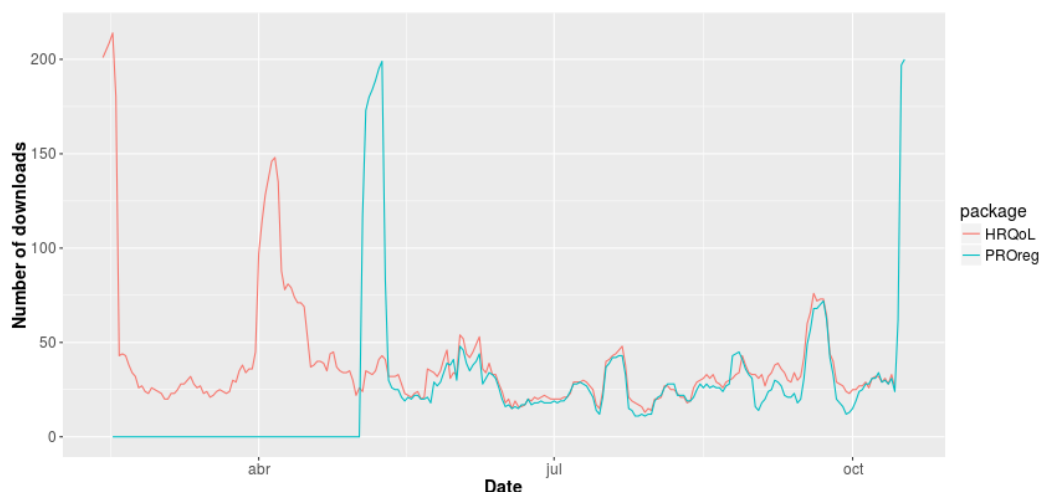


Figure 6.1: Number of weekly downloads of **PROreg** and **HRQoL** R-packages. Originally in Najera-Zuloaga et al. (2017) the package was called **HRQoL**. This plot was obtained from the shiny application available at <https://dgrtwo.shinyapps.io/cranview/>

Finally, researchers at the Respiratory Service at Galdakao Hospital in Spain designed a longitudinal study where the health-status of patients with COPD was repeatedly measured. The objective was to measure the health-status and evolu-

tion of COPD patients who were followed for up to five years. Therefore, the last objective of the thesis was to achieve clinically relevant and useful results regarding the evolution and COPD patients, and to detect significant relationships between their health-status and risk factors. This last objective was developed in Chapter 2 and Chapter 4, where relationship between the health-status of COPD patients and covariates is assessed in a univariate and a multivariate framework. On the contrary, the evolution of the patients was developed in Chapter 3 and Chapter 4 analysing the dimensions separately and jointly. For instance, we conclude that patients with COPD does not evolve in their symptomatology and disease impact as time goes by and moreover, that the initial status at baseline does not influence that evolution. On the contrary, the activity of COPD patients is expected to worsen over time, although the initial severity of the disease does not determine the evolution, meaning that all the patients evolve similarly.

As some final remarks, we would like to mention that part of the modelling approaches presented in this thesis have some limitations that we will explore in further research. On the one hand, we shown in Chapter 4 that although the multivariate shared random effects approach offers valid results in cross-sectional data, in longitudinal data they can be quite restrictive and hence, and they can lead to misleading conclusions. In fact, the assumption that all the patients distance in the same quantity in baseline and evolution from the overall population trend in all the dimensions can be quite unrealistic. In order to avoid the this assumption, one may consider a model with correlated random effects, where each dimension depends on different random effects that are correlated among dimensions. The correlated random effects model is much more flexible than the shared random effects approach, however, when more than two dimensions are analysed, the approximation of the marginal likelihood through the Laplace approximation becomes cumbersome. Most of the PRO questionnaires provides more than two dimensions, and consequently, the correlated random effects approach would not be useful due to dimensionality problems. Fieuws and Verbeke (2006) proposed an inference strategy in order to avoid the high dimensionality of the marginal likelihood of the model. Their proposal was based on the idea that all parameters in the joint model can be estimated by fitting all bivariate models implied in the multivariate model. More specifically, for all  $\binom{k}{2}$   $(\mathbf{Y}^l, \mathbf{Y}^s)$  dimension pairs,  $1 \leq l \leq s \leq k$ , they proposed to fit using MLE the model defined by

$$f(\mathbf{y}^l, \mathbf{y}^s) = \int \int f(\mathbf{y}^l | \mathbf{u}^l) f(\mathbf{y}^s | \mathbf{u}^s) d\mathbf{u}^l d\mathbf{u}^s,$$



where  $\mathbf{u}^l$  and  $\mathbf{u}^s$  are random effects associated with  $\mathbf{Y}^l$  and  $\mathbf{Y}^s$  dimensions. Then, they proposed to obtain the estimates of the correlated random effects model by averaging over the results for the  $\binom{k}{2}$  pairwise model fits. Obviously, the resulting estimates do not maximise the likelihood, hence inference does not follow from classical maximum likelihood theory. Instead, Fieuws and Verbeke (2006) showed that pseudo-likelihood theory could be used to derive the asymptotic distribution of the obtained estimates. Alternatively, the proposed estimation procedure based on the replacement of the log-likelihood of the original model by a sum of implied marginal or conditional log-likelihoods is also referred to as composite likelihood (Lindsay, 1988). However, the implementation of the pairwise estimation procedure for the beta-binomial distribution goes beyond the scope of this thesis, and hence, it is left as future work. On the other hand, in Chapter 3, we have shown the evolution of patients with COPD adjusted by patients' subtypes. Nowadays, we are working with clinicians in order to incorporate covariates that can have both clinically and statistically significant effect in the evolution of COPD patients.

## Further work

The research that has been undertaken for this thesis has highlighted a number of topics on which further research would be beneficial in the statistical analysis of PROs. For instance:

1. *The jointly modelling of the conditional mean and variance components in hierarchical beta-binomial data.*

An important issue that needs to be addressed when modelling hierarchical data, is how to correctly account for both individual and group level variation that might be present in the data. The main focus in (generalised) linear mixed effects models has been on modelling the mean structure of the data while treating the variances/covariances as nuisance parameters. Taking into account the structure but misspecifying or oversimplifying the covariance structure can still have a serious impact on the efficiency of the mean parameters. In some circumstances, like the mean, the variance may depend on a set of explanatory variables, McCullagh and Nelder (1989) proposed the joint modelling of mean and dispersion in the GLM framework. However, the literature on joint modelling for generalised linear mixed models (GLMMs) is rather limited. Pan and McKenzie (2007) proposed modelling the conditional vari-

ance using the Cholesky Decomposition of the covariance matrix, for the linear mixed model. Lee et al. (2006) proposed a joint mean–covariance modelling for hierarchical data via HGLMs with structured dispersions and the so-called double HGLMs Lee and Nelder (2006). Modelling dispersion and conditional variance in the beta-binomial context would be of interest for further research.

2. *Joint modelling of survival and longitudinal PROs.*

In many longitudinal studies measurements also may include the time at which an event of particular interest occurs (e.g. death, development of a disease or drop out from the study). These outcomes are often separately analysed; however, in many instances, a joint modelling approach is either required or may produce a better insight into the mechanisms that underlie the phenomenon under study (see Tsiatis and Davidian, 2004, for a detailed overview). Usually, the joint distribution of the event times and the longitudinal measurements is modelled via a set of random effects that are assumed to account for the associations between these two outcomes (see Hsieh et al., 2006; Rizopoulos et al., 2009, for a review). Related to this topic, a joint modelling of PROs and survival analysis of chronic disease patients will be of interest for further research where the longitudinal modelling is based on the beta-binomial mixed effects model developed in this thesis.

3. *Testing for variance components in the beta-binomial mixed-effects model.*

Testing zero variance components is one of the most challenging problems in the context of linear mixed-effects models. The usual asymptotic Chi-square distribution of the likelihood ratio and score statistics under this null hypothesis is incorrect because the null is on the boundary of the parameter space. Previous work by Greven et al. (2008) addressed this issue and provide Restricted Likelihood Ratio Tests. Other authors introduce a simple test statistic based on the variance least square estimator of variance components (Drikvandi et al., 2013). In this thesis, we did not explore this issue when testing for the variance components in BBmm, although it would be an interesting topic for further research.

4. *Inclusion of non-linear covariate effects.*

The real data applications presented in this thesis did not show evidence of non-linear covariate effects. However, when the relationship between the response variable and covariate effects are non-linear, the inclusion of non-linear

effects in the BBmm under the modelling framework presented in this thesis is straightforward by means of smooth effects with penalised splines reformulated as a mixed-effects models (Ruppert et al., 2003; Wood, 2006). The estimation of the variance components with the estimation procedure developed in this thesis would be of great interest for researchers in PROs.

5. *Incorporation of new features and efficient algorithms in PROreg R-package.*

As part of the further research in the analysis of PROs, the incorporation of new features and more efficient implementations of the BBreg and BBm functions will be considered for further research.

# Appendices



---

---

## APPENDIX A

---

### RECODING PROCESS OF THE SHORT FORM-36 HEALTH SURVEY

The SF-36 Health Survey recoding methodology is fully explained in Arostegui et al. (2013) however, for the sake of clarity as it was mentioned in Section 1.2.1, we are going to define more in detail the recoding process in this appendix. For simplicity and brevity of exposition, we only show results for three of the eight health dimensions of the SF-36. The selected three dimensions (*physical functioning*, *mental health* and *role emotional*) illustrate different distribution shapes as shown in Figure 1.1 and a wide range of number of possible values (see Table 1.1).

Table A.1 shows the possible values of the raw and standardized original scores, as well as the sub-interval division of the 0-100 scale and the final recoded scores for *physical functioning*, *mental health* and *role emotional* dimensions. Arostegui et al. (2013) evaluated and validated the proposed method of recoding the scores provided by the SF-36 Health Survey. They showed that the recoding has a natural interpretation, not only for ordinal scores but also for questionnaires with many dimensions and different profiles, where a common method of analysis is desired, such as the SF-36. Briefly, let  $Y$  denote the original standardized score observed in  $[0, 100]$  and  $Z$  the recoded ordinal score, from 0 to  $m$ , where  $Z \sim \text{Bin}(m, p)$ . Thus,  $Z$  could be interpreted as grouped data for a dichotomous outcome that represents the number of successes in  $m$  binomial trials and  $p$  represents the probability of success in each trial. In the HRQoL context,  $Z$  is interpreted as the number of ‘points’ that an individual has,  $p$  as the probability of obtaining one point more and  $m$  represents

the maximum number of points that can be obtained.

Table A.1: Recoding methodology for *role emotional*, *mental health* and *physical functioning* dimensions of the SF-36.

<i>Role emotional</i>				<i>Mental health</i>			
Raw	Stdr.	Inter.	Recoded	Raw	Stdr.	Inter.	Recoded
3	0	[0, 16.67]	0	5	0	[0,2]	0
4	33.3	(16.67, 50]	1	6	4	(2,10]	1
5	66.7	(50, 83.33]	2	7	8		
6	100	(83.33, 100]	3	8	12	(10,18]	2
				9	16		
				10	20	(18, 26]	3
				11	24		
				12	28	(26,34]	4
				13	32		
				14	36	(34,42]	5
				15	40		
				16	44	(42,50]	6
				17	48		
				18	52	(50,58]	7
				19	56		
				20	60	(58,66]	8
				21	64		
				22	68	(66,74]	9
				23	72		
				24	76	(74,82]	10
				25	80		
				26	84	(82,90]	11
				27	88		
				28	92	(90,98]	12
				29	96		
				30	100	(98,100]	13

Raw: Raw scores; Stdr.: Standardized original scores; Inter.: The subinterval division of the 0 – 100 scale; Recoded: Recoding of the values. The decomposition in raw and original standardized scores of the HRQoL dimensions is developed in Ware et al. (1993), while the subinterval division and recoding process are explicitly explained in Arostegui et al. (2013).

---

---

## APPENDIX B

---

### ITERATIVE ESTIMATION METHODOLOGIES

Numerical algorithms to find the maximum likelihood estimates and standard errors are crucial for regression analysis including non-normal outcomes. In these cases, the computation of the marginal likelihood has not a closed form, and hence, likelihood approximations are needed. Moreover, even with approximated likelihoods, there is no explicit formula to obtain both maximum likelihood estimations and standard errors. It turns out that there is one general algorithms, called the iterative weighted least squares (IWLS), that works reliably for exponential family distributions. Although there are many ways to derive an IWLS procedure, we will focus on the development of the methodology by the Newton-Raphson technique. Therefore, we are going to, first, briefly introduce the well known Newton-Raphson procedure and then, we are going to derive the IWLS algorithm

#### B.1 Newton-Raphson procedure

The Newton-Raphson procedure is a powerful technique for solving equations of the form  $g(x)$  numerically. Although, it is widely used for finding out the root of a function, it is simply based on the idea of a linear approximation.

Let define  $r$  as the root of the equation  $g(x) = 0$ . Assume that  $x_0$ , the initial value, is a good estimate for  $r$ , good in terms of distance. Hence, we have that  $r = x_0 + h$ , where  $h$  is not a large number. Since the true root is  $r$  and  $h = r - x_0$ ,



the number  $h$  measures how far the estimate  $x_0$  is from the truth.

Considering that  $h$  is small we can use the linear or tangent approximation to conclude that

$$0 = g(r) = g(x_0 + h) \approx g(x_0) + hg'(x_0)$$

and, therefore, unless  $g'(x_0)$  is close to 0,

$$h \approx -\frac{g(x_0)}{g'(x_0)}$$

and hence, we have that

$$r = x_0 + h \approx x_0 - \frac{g(x_0)}{g'(x_0)}$$

Consequently, we get a new improved estimate  $x_1$  in the following way

$$x_1 = x_0 - \frac{g(x_0)}{g'(x_0)}$$

The repetition of the previous formula, getting new improved estimates for  $r$ , results in the following iterative procedure: if  $x_n$  is the current estimate, then the next estimate  $x_{n+1}$  is define as

$$x_{n+1} = x_n - \frac{g(x_n)}{g'(x_n)}$$

## B.2 Iterative weighted least squares algorithm

Now we are going to extend the Newton-Raphson procedure to the GLM framework (see Chapter 2 Section 2.1.1). First, we start with a log-likelihood contribution of an observation  $y_i$  of the exponential family form,

$$\log L(\theta_i, \phi | y_i) = \frac{y_i \theta_i - A(\theta_i)}{\phi} + c(y_i, \phi)$$

where  $\theta_i$  is the so-called canonical parameter,  $\phi$  is the dispersion parameter and  $c(\cdot)$  and  $A(\cdot)$  are known functions. The exponential family distributions have very useful characteristics, which make them very flexible for estimation procedures. The most relevant feature of the exponential family distributions, compared with other more rigid models, is that the variance is not closely defined by the mean,

$$\mu_i = \mathbb{E}[Y_i] = A'(\theta_i)$$

$$\text{Var}[Y_i] = \phi A''(\theta_i) = \phi \frac{\partial}{\partial \theta_i} \mathbb{E}[Y_i] = \phi v(\mu_i)$$

Let us assume that we have  $n$  independent observations  $y_i$  and consider that the connection between the given covariates and the observed responses is defined as

$$h(\mu_i) = \mathbf{x}'_i \boldsymbol{\beta}$$

where  $h(\cdot)$  is called the link function. If we try to get the estimate of  $\boldsymbol{\beta}$ , for a fixed value of  $\phi$ , we get the following score equation

$$S(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} \log L(\theta_i | \phi, \mathbf{y}) = \phi^{-1} \sum_{i=1}^n \frac{\partial \theta_i}{\partial \boldsymbol{\beta}} (y_i - A'(\theta_i))$$

while the Fisher information matrix is defined as

$$I(\boldsymbol{\beta}) = \phi^{-1} \sum_{i=1}^n \left[ -\frac{\partial^2 \theta_i}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} (y_i - A'(\theta_i)) + \frac{\partial \theta_i}{\partial \boldsymbol{\beta}} \frac{\partial \theta_i}{\partial \boldsymbol{\beta}'} A''(\theta_i) \right]$$

which in general a very complex expression.

If we use the exponential family property,  $A'(\theta_i) = \mu_i$ , we have that

$$A''(\theta_i) = \partial \mu_i / \partial \theta_i = v_i$$

$$\frac{\partial \theta_i}{\partial \boldsymbol{\beta}} = \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial h} \frac{\partial h}{\partial \boldsymbol{\beta}} = v_i^{-1} \frac{\partial \mu_i}{\partial h} \mathbf{x}_i$$

Hence, the second term of  $I(\boldsymbol{\beta})$  reduces to

$$\mathbf{U} \equiv \sum_{i=1}^n \left[ \left( \frac{\partial h}{\partial \mu_i} \right)^2 \phi v_i \right]^{-1} \mathbf{x}_i \mathbf{x}'_i$$

In the same way we have that,

$$S(\boldsymbol{\beta}) = \sum_{i=1}^n \left[ \left( \frac{\partial h}{\partial \mu_i} \right)^2 \phi v_i \right]^{-1} \mathbf{x}_i \frac{\partial h}{\partial \mu_i} (y_i - \mu_i)$$

Furthermore, if we use a canonical link function in the model, i.e., if we use a link function such as  $\theta_i = h(\mu_i) = \mathbf{x}'_i \boldsymbol{\beta}$ , we have that

$$\frac{\partial \theta_i}{\partial \boldsymbol{\beta}} = \mathbf{x}_i$$

$$\frac{\partial^2 \theta_i}{\partial \beta \partial \beta'} = 0$$

and consequently,  $I(\beta) = \mathbf{U}$ , from where we obtain a Newton-Raphson update equal to

$$\beta^1 = \beta^0 + \mathbf{U}^{-1} S(\beta)$$

If we define  $\mathbf{X}$  as the design matrix of the covariates and  $\Sigma$  as a diagonal matrix with elements

$$\Sigma_{ii} = \left( \frac{\partial h}{\partial \mu_i} \right)^2 \phi v_i \quad (\text{B.1})$$

we have that  $\mathbf{U} = \mathbf{X} \Sigma^{-1} \mathbf{X}'$  and

$$S(\beta) = \mathbf{X}' \Sigma^{-1} \frac{\partial h}{\partial \mu} (\mathbf{y} - \mu)$$

where  $\partial h / \partial \mu (\mathbf{y} - \mu)$  is a vector that contains the elements  $\partial h / \partial \mu_i (y_i - \mu_i)$ .

Finally, we can rewrite the  $\beta$  update formula as

$$\begin{aligned} \beta^1 &= \beta^0 + (\mathbf{X}' \Sigma^{-1} \mathbf{X})^{-1} \mathbf{X}' \Sigma^{-1} \frac{\partial h}{\partial \mu} (\mathbf{y} - \mu) \\ &= (\mathbf{X}' \Sigma^{-1} \mathbf{X})^{-1} \mathbf{X}' \Sigma^{-1} \left[ \mathbf{X} \beta^0 + \frac{\partial h}{\partial \mu} (\mathbf{y} - \mu) \right] \\ &= (\mathbf{X}' \Sigma^{-1} \mathbf{X})^{-1} \mathbf{X}' \Sigma^{-1} \mathbf{y}^w \end{aligned}$$

where  $\mathbf{y}^w$  is called the working vector. Indeed, it is a vector with elements

$$y_i^w = \mathbf{x}'_i \beta^0 + \frac{\partial h}{\partial \mu_i} (y_i - \mu_i)$$

where all the unknown parameters are evaluated in their current values.

The updating formula can be connected to a quadratic approximation of the log-likelihood. Indeed,

$$\begin{aligned} \mathbb{E}[Y_i] &= \mathbf{x}'_i \beta^0 \\ \text{Var}[Y_i] &= \text{Var} \left[ \frac{\partial h}{\partial \mu_i} y_i \right] = \left( \frac{\partial h}{\partial \mu_i} \right)^2 \text{Var}[y_i] = \left( \frac{\partial h}{\partial \mu_i} \right)^2 \phi v_i = \Sigma_{ii} \end{aligned}$$

Indeed, starting with  $\beta^0$ , the log-likelihood of any distribution of the exponential family is approximated by

$$-\frac{1}{2} \log |\Sigma| - \frac{1}{2} (\mathbf{y}^w - \mathbf{X} \beta)' \Sigma^{-1} (\mathbf{Y} - \mathbf{X} \beta)$$

with  $\mathbf{y}^w$  and  $\mathbf{\Sigma}$  defined above.

At convergence, we can evaluate the standard errors for the estimates from the inverse of

$$I(\boldsymbol{\beta}) = \mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X}$$

where the variance matrix is evaluated using the estimates  $\hat{\boldsymbol{\beta}}$  and  $\hat{\phi}$ .



---

---

## APPENDIX C

---

### LAPLACE APPROXIMATION

Laplace method is a technique used to approximate integrals of the form,

$$\int_a^b e^{Mf(x)} dx$$

where  $f(x)$  is a twice-differentiable function,  $M$  is a large number and the integral endpoints  $a$  and  $b$  could be possibly infinite.

Assume that the function  $f(x)$  has a unique global maximum at  $x_0$ . Then, the value  $f(x_0)$  will be larger than other values  $f(x)$ . If we multiply this function by a large number  $M$ , the ratio between  $Mf(x_0)$  and  $Mf(x)$  will stay the same, but it will grow exponentially in the function  $\exp(Mf(x))$ .

If we expand  $f(x)$  around  $x_0$  by Taylor's theorem,

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2}f''(x_0)(x - x_0)^2 + R \quad (\text{C.1})$$

where  $R = \mathcal{O}(x - x_0)^3$ .

since  $f(\cdot)$  has a global maximum at  $x_0$ , and since  $x_0$  is not an endpoint, it is a stationary point hence, the derivative of  $f(\cdot)$  vanishes at  $x_0$ . Consequently, the second term in Equation (C.1) equals to zero and we get that

$$f(x) \approx f(x_0) - \frac{1}{2}|f''(x_0)|(x - x_0)^2$$

for  $x$  close to  $x_0$ . Notice that the second derivative is negative at the global maxi-

mum. The assumptions made ensure the accuracy of the approximation

$$\int_a^b e^{Mf(x)} dx \approx e^{Mf(x_0)} \int_a^b e^{-M|f''(x_0)|(x-x_0)^2/2} dx \quad (\text{C.2})$$

The integral in Equation (C.2) is a Gaussian integral if the limits of the integration go from  $-\infty$  to  $\infty$  (which can be assumed because the exponential decays very fast away from  $x_0$ ), and thus it can be calculated. In fact, we find that

$$\int_a^b e^{Mf(x)} dx \approx \sqrt{\frac{2\pi}{M|f''(x_0)|}} e^{Mf(x_0)} \quad \text{as} \quad M \rightarrow \infty. \quad (\text{C.3})$$

---

---

## APPENDIX D

---

### MATRIX DIFFERENTIATION

In this Appendix we collect some useful formulas of matrix calculus that appear through the thesis.

#### D.1 Properties and definitions

**Definition D.1.** *Through this work we are going to use the numerator notation for the matrix derivatives. Hence, if  $y$  and  $x$  are scalars,  $\mathbf{y}_{n \times 1}$  and  $\mathbf{x}_{t \times 1}$  are vectors and  $\mathbf{Y}_{n \times m}$  and  $\mathbf{X}_{t \times k}$  are matrices, we define the derivatives in the following way:*

*By scalar:*

$$\frac{\partial y}{\partial x} \quad \frac{\partial \mathbf{y}}{\partial x} = \begin{pmatrix} \frac{\partial y_1}{\partial x} \\ \vdots \\ \frac{\partial y_n}{\partial x} \end{pmatrix} \quad \frac{\partial \mathbf{Y}}{\partial x} = \begin{pmatrix} \frac{\partial y_{11}}{\partial x} & \cdots & \frac{\partial y_{1m}}{\partial x} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_{n1}}{\partial x} & \cdots & \frac{\partial y_{nm}}{\partial x} \end{pmatrix}$$

*By vector:*

$$\frac{\partial y}{\partial \mathbf{x}} = \left( \frac{\partial y}{\partial x_1} \quad \cdots \quad \frac{\partial y}{\partial x_t} \right) \quad \frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_t} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_n}{\partial x_1} & \cdots & \frac{\partial y_n}{\partial x_t} \end{pmatrix}$$



By matrix:

$$\frac{y}{\partial \mathbf{X}} = \begin{pmatrix} \frac{\partial y}{\partial x_{11}} & \cdots & \frac{\partial y}{\partial x_{t1}} \\ \vdots & \ddots & \vdots \\ \frac{\partial y}{\partial x_{1k}} & \cdots & \frac{\partial y}{\partial x_{tk}} \end{pmatrix}$$

**Property D.1.** We present commonly used derivatives properties, which are not difficult to prove. Assume that the scalar  $a$ , vector  $\mathbf{a}$  and matrices  $\mathbf{A}$  and  $\mathbf{B}$  are not functions of  $x$  or  $\mathbf{x}$ .

- Derivatives of scalar, vector and matrix by scalar.

$$(SS1) \quad \frac{\partial(u+v)}{\partial x} = \frac{\partial u}{\partial x} + \frac{\partial v}{\partial x}$$

$$(SS2) \quad \frac{\partial uv}{\partial x} = u \frac{\partial v}{\partial x} + v \frac{\partial u}{\partial x} \quad (\text{product rule})$$

$$(SS3) \quad \frac{\partial g(u)}{\partial x} = \frac{\partial g(u)}{\partial u} \frac{\partial u}{\partial x} \quad (\text{chain rule})$$

$$(VS1) \quad \frac{\partial a\mathbf{u}}{\partial x} = a \frac{\partial \mathbf{u}}{\partial x}$$

$$(VS2) \quad \frac{\partial \mathbf{A}\mathbf{u}}{\partial x} = \mathbf{A} \frac{\partial \mathbf{u}}{\partial x}$$

$$(VS3) \quad \frac{\partial \mathbf{u}'}{\partial x} = \left( \frac{\partial \mathbf{u}}{\partial x} \right)'$$

$$(VS4) \quad \frac{\partial(\mathbf{u} + \mathbf{v})}{\partial x} = \frac{\partial \mathbf{u}}{\partial x} + \frac{\partial \mathbf{v}}{\partial x}$$

$$(VS5) \quad \frac{\partial g(\mathbf{u})}{\partial x} = \frac{\partial g(\mathbf{u})}{\partial \mathbf{u}} \frac{\partial \mathbf{u}}{\partial x} \quad (\text{chain rule})$$

$$(MS1) \quad \frac{\partial a\mathbf{U}}{\partial x} = a \frac{\partial \mathbf{U}}{\partial x}$$

$$(MS2) \quad \frac{\partial \mathbf{A}\mathbf{U}\mathbf{B}}{\partial x} = \mathbf{A} \frac{\partial \mathbf{U}}{\partial x} \mathbf{B}$$

$$(MS3) \quad \frac{\partial(\mathbf{U} + \mathbf{V})}{\partial x} = \frac{\partial \mathbf{U}}{\partial x} + \frac{\partial \mathbf{V}}{\partial x}$$

$$(MS4) \quad \frac{\partial \mathbf{U}\mathbf{V}}{\partial x} = \mathbf{U} \frac{\partial \mathbf{V}}{\partial x} + \frac{\partial \mathbf{U}}{\partial x} \mathbf{V}$$

- Derivatives of scalar and vector by vector.

$$(SV1) \quad \frac{\partial a u}{\partial \mathbf{x}} = a \frac{\partial u}{\partial \mathbf{x}}$$

$$(SV2) \quad \frac{\partial(u+v)}{\partial \mathbf{x}} = \frac{\partial u}{\partial \mathbf{x}} + \frac{\partial v}{\partial \mathbf{x}}$$

$$(SV3) \frac{\partial uv}{\partial \mathbf{x}} = u \frac{\partial v}{\partial \mathbf{x}} + v \frac{\partial u}{\partial \mathbf{x}} \text{ (product rule)}$$

$$(SV4) \frac{\partial g(u)}{\partial \mathbf{x}} = \frac{\partial g(u)}{\partial u} \frac{\partial u}{\partial \mathbf{x}} \text{ (chain rule)}$$

$$(SV5) \frac{\partial \mathbf{u}'\mathbf{v}}{\partial \mathbf{x}} = \mathbf{u}' \frac{\partial \mathbf{v}}{\partial \mathbf{x}} + \mathbf{v}' \frac{\partial \mathbf{u}}{\partial \mathbf{x}} \text{ (product rule)}$$

$$(SV6) \frac{\partial \mathbf{a}'\mathbf{x}}{\partial \mathbf{x}} = \frac{\partial \mathbf{x}'\mathbf{a}}{\partial \mathbf{x}} = \mathbf{a}'$$

$$(SV7) \frac{\partial \mathbf{x}'\mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{x}'$$

$$(SV8) \frac{\partial \mathbf{x}'\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = \mathbf{x}'(\mathbf{A} + \mathbf{A}')$$

$$(VV1) \frac{\partial a\mathbf{u}}{\partial \mathbf{x}} = a \frac{\partial \mathbf{u}}{\partial \mathbf{x}} + \mathbf{u} \frac{\partial a}{\partial \mathbf{x}} \text{ (product rule)}$$

$$(VV2) \frac{\partial \mathbf{A}\mathbf{u}}{\partial \mathbf{x}} = \mathbf{A} \frac{\partial \mathbf{u}}{\partial \mathbf{x}}$$

$$(VV3) \frac{\partial (\mathbf{u} + \mathbf{v})}{\partial \mathbf{x}} = \frac{\partial \mathbf{u}}{\partial \mathbf{x}} + \frac{\partial \mathbf{v}}{\partial \mathbf{x}}$$

$$(VV4) \frac{\partial g(\mathbf{u})}{\partial \mathbf{x}} = \frac{\partial g(\mathbf{u})}{\partial \mathbf{u}} \frac{\partial \mathbf{u}}{\partial \mathbf{x}} \text{ (chain rule)}$$

$$(VV5) \frac{\partial \mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = \mathbf{A}$$

$$(VV6) \frac{\mathbf{x}'\mathbf{A}}{\partial \mathbf{x}} = \mathbf{A}'$$

- *Derivatives of scalar by matrix.*

$$(SM1) \frac{\partial a\mathbf{u}}{\partial \mathbf{X}} = a \frac{\partial \mathbf{u}}{\partial \mathbf{X}}$$

$$(SM2) \frac{\partial (u + v)}{\partial \mathbf{X}} = \frac{\partial u}{\partial \mathbf{X}} + \frac{\partial v}{\partial \mathbf{X}}$$

$$(SM3) \frac{\partial uv}{\partial \mathbf{X}} = v \frac{\partial u}{\partial \mathbf{X}} + u \frac{\partial v}{\partial \mathbf{X}} \text{ (product rule)}$$

$$(SM4) \frac{\partial g(u)}{\partial \mathbf{X}} = \frac{g(u)}{\partial u} \frac{\partial u}{\partial \mathbf{X}} \text{ (chain rule)}$$

**Property D.2 (Woodbury matrix identity).**

$$(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} = \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{B}(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{C}\mathbf{A}^{-1}$$

for matrices  $\mathbf{A}$ ,  $\mathbf{B}$ ,  $\mathbf{C}$  and  $\mathbf{D}$  with appropriate dimensions and assuming  $\mathbf{A}$ ,  $\mathbf{C}$ ,  $\mathbf{D}$  and  $\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C}$  are invertible.

Alternative names for this formula are the matrix inversion lemma, Sherman–Morrison–Woodbury formula or just Woodbury formula. The proof of the formula can be found at Harville (1997) Theorem 19.2.8 in Equation (2.22).

**Property D.3 (Jacobi’s formula).**

$$\frac{\partial \det(\mathbf{H})}{\partial x} = \text{trace} \left[ \text{adj}(\mathbf{H}) \frac{\partial \mathbf{H}}{\partial x} \right]$$

and, moreover, if  $\mathbf{H}$  is nonsingular we have that

$$\frac{\partial \det(\mathbf{H})}{\partial x} = \det(\mathbf{H}) \text{trace} \left( \mathbf{H}^{-1} \frac{\partial \mathbf{H}}{\partial x} \right).$$

The proof can be found at Harville (1997) Equation (8.5).

**Property D.4.** Let be  $\mathbf{H}$  a nonsingular matrix, then

$$\frac{\partial \log \det(\mathbf{H})}{\partial x} = \text{trace} \left( \mathbf{H}^{-1} \frac{\partial \mathbf{H}}{\partial x} \right)$$

The proof can be found at Harville (1997) Equation (8.6).

**Property D.5.**

$$\frac{\partial \mathbf{H}^{-1}}{\partial x} = -\mathbf{H}^{-1} \frac{\partial \mathbf{H}}{\partial x} \mathbf{H}^{-1}$$

The proof can be found at Harville (1997) Equation (8.15).

**Property D.6 (Inverse of a block matrix).** If

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix},$$

then

$$\mathbf{\Sigma}^{-1} = \begin{pmatrix} (\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} & -(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}\mathbf{B}\mathbf{D}^{-1} \\ (\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{C}\mathbf{A}^{-1} & (\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1} \end{pmatrix}.$$

The proof of the formula can be found at Harville (1997) Theorem 8.5.11 in Equation (5.17a).

## D.2 Matrix differentiation of the proposed models

In this section we present all the matrix analysis of the BBreg and BBmm approaches developed in Chapter 2 and Chapter 3 respectively.

### BBreg approach

The log-likelihood contribution of the BBreg approach is given by (see Chapter 2, Equation (2.5))

$$\log L(\mathbf{p}, \phi | \mathbf{y}) = \sum_{i=1}^n \left[ \log \binom{m_i}{y_i} + \sum_{k=0}^{y_i-1} \log(p_i + k\phi) + \sum_{k=0}^{m_i-y_i-1} \log(1-p_i+k\phi) - \sum_{k=0}^{m_i} \log(1+k\phi) \right].$$

Therefore, we present all the derivatives that are used in the estimation procedure in Chapter 2, i.e.:

$$\begin{aligned} \frac{\partial p_i}{\partial \beta_j} &= \frac{1}{\partial \beta_j} \left( \frac{\exp(\mathbf{x}'_i \boldsymbol{\beta})}{1 + \exp(\mathbf{x}'_i \boldsymbol{\beta})} \right) = \frac{x_{ij} \exp(-\mathbf{x}'_i \boldsymbol{\beta})}{(1 + \exp(-\mathbf{x}'_i \boldsymbol{\beta}))^2} = x_{ij} p_i (1 - p_i), \\ \frac{\partial \mathbf{p}}{\partial \boldsymbol{\beta}} &= \begin{pmatrix} \frac{\partial p_1}{\partial \beta_1} & \cdots & \frac{\partial p_1}{\partial \beta_p} \\ \vdots & \ddots & \vdots \\ \frac{\partial p_n}{\partial \beta_1} & \cdots & \frac{\partial p_n}{\partial \beta_p} \end{pmatrix} = \mathbf{S} \mathbf{X}, \end{aligned} \quad (\text{D.1})$$

where  $\mathbf{S} = \text{diag}(p_1(1-p_1), \dots, p_n(1-p_n))$ .

$$\begin{aligned} \frac{\partial \log L}{\partial p_i} &= \sum_{k=0}^{y_i-1} \frac{1}{p_i + k\phi} + \sum_{k=0}^{m_i-y_i-1} \frac{1}{1-p_i+k\phi} = \xi_i, \\ \frac{\partial \log L}{\partial \mathbf{p}} &= \left( \frac{\partial \log L}{\partial p_1} \quad \cdots \quad \frac{\partial \log L}{\partial p_n} \right) = \boldsymbol{\xi}', \end{aligned} \quad (\text{D.2})$$

where  $\boldsymbol{\xi} = (\xi_1, \dots, \xi_n)'$ .

$$\frac{\partial^2 \log L}{\partial \mathbf{p} \partial \mathbf{p}'} = \frac{\partial}{\partial \mathbf{p}'} \boldsymbol{\xi}' = \begin{pmatrix} \frac{\partial \xi_1}{\partial \mathbf{p}'} & \cdots & \frac{\partial \xi_n}{\partial \mathbf{p}'} \\ \vdots & \ddots & \vdots \\ \frac{\partial \xi_1}{\partial \mathbf{p}_n} & \cdots & \frac{\partial \xi_n}{\partial \mathbf{p}_n} \end{pmatrix} = -\mathbf{V}, \quad (\text{D.3})$$

$$v_{ij} = -\frac{\partial \xi_i}{\partial p_j} = \begin{cases} 0, & \text{if } i \neq j \\ \sum_{k=0}^{y_i-1} \frac{1}{(p_i + k\phi)^2} + \sum_{k=0}^{m_i-y_i-1} \frac{1}{(1-p_i + k\phi)^2}, & \text{if } i = j. \end{cases}$$

$$\frac{\partial \log L}{\partial \phi} = \sum_{i=1}^n \left[ \sum_{k=0}^{y_i-1} \frac{k}{p_i + k\phi} + \sum_{k=0}^{m_i-y_i-1} \frac{k}{1-p_i + k\phi} - \sum_{k=0}^{m_i-1} \frac{k}{1+k\phi} \right]. \quad (\text{D.4})$$

### BBmm approach

The approximated marginal likelihood of the model in Equation (3.25) is defined as

$$\log L(\boldsymbol{\beta}, \boldsymbol{\theta} | \mathbf{y}) \approx h(\boldsymbol{\beta}, \boldsymbol{\theta} | \mathbf{y}, \tilde{\mathbf{u}}) - \frac{1}{2} \log |\mathbf{M}|$$

where

$$h(\boldsymbol{\beta}, \boldsymbol{\theta} | \mathbf{y}, \mathbf{u}) = \underbrace{\sum_{i=1}^n \log f_{y|u}(y_i | \boldsymbol{\beta}, \phi, \mathbf{u})}_{h_1} + \underbrace{\log f_{\mathbf{u}}(\mathbf{u} | \boldsymbol{\lambda})}_{h_2}$$

$$h_1 = \sum_{i=1}^n \left[ \sum_{k=0}^{y_i-1} \log(p_i + k\phi) + \sum_{k=0}^{m_i-y_i-1} \log(1-p_i + k\phi) - \sum_{k=0}^{m_i-1} \log(1+k\phi) \right]$$

$$h_2 = -\frac{1}{2} \log |\mathbf{D}| - \frac{1}{2} \mathbf{u}' \mathbf{D}^{-1} \mathbf{u},$$

and

$$\mathbf{M} = \frac{\partial^2 h}{\partial \mathbf{u} \partial \mathbf{u}'} \Big|_{\mathbf{u}=\tilde{\mathbf{u}}}.$$

where  $\tilde{\mathbf{u}}$  is the solution of  $\partial h / \partial \mathbf{u} = 0$ .

Therefore, we present all the derivatives that are used in the estimation procedure in Chapter 3, i.e.:

$$\begin{aligned}\frac{\partial p_i}{\partial u_j} &= \frac{1}{\partial u_j} \left( \frac{\exp(\mathbf{x}'_i \boldsymbol{\beta} + \mathbf{z}'_i \mathbf{u})}{1 + \exp(\mathbf{x}'_i \boldsymbol{\beta} + \mathbf{z}'_i \mathbf{u})} \right) = z_{ij} p_i (1 - p_i) \\ \frac{\partial \mathbf{p}}{\partial \mathbf{u}} &= \begin{pmatrix} \frac{\partial p_1}{\partial u_1} & \cdots & \frac{\partial p_1}{\partial u_q} \\ \vdots & \ddots & \vdots \\ \frac{\partial p_n}{\partial u_1} & \cdots & \frac{\partial p_n}{\partial u_q} \end{pmatrix} = \mathbf{S} \mathbf{Z}\end{aligned}\quad (\text{D.5})$$

where  $\mathbf{S} = \text{diag}(p_1(1 - p_1), \dots, p_n(1 - p_n))$ .

$$\begin{aligned}\frac{\partial p_i}{\partial \beta_j} &= \frac{1}{\partial \beta_j} \left( \frac{\exp(\mathbf{x}'_i \boldsymbol{\beta} + \mathbf{z}'_i \mathbf{u})}{1 + \exp(\mathbf{x}'_i \boldsymbol{\beta} + \mathbf{z}'_i \mathbf{u})} \right) = x_{ij} p_i (1 - p_i) \\ \frac{\partial \mathbf{p}}{\partial \boldsymbol{\beta}} &= \begin{pmatrix} \frac{\partial p_1}{\partial \beta_1} & \cdots & \frac{\partial p_1}{\partial \beta_p} \\ \vdots & \ddots & \vdots \\ \frac{\partial p_n}{\partial \beta_1} & \cdots & \frac{\partial p_n}{\partial \beta_p} \end{pmatrix} = \mathbf{S} \mathbf{X}\end{aligned}\quad (\text{D.6})$$

$$\begin{aligned}\frac{\partial h_1}{\partial p_i} &= \sum_{k=0}^{y_i-1} \frac{1}{p_i + k\phi} + \sum_{k=0}^{m_i-y_i-1} \frac{1}{1 - p_i + k\phi} = \xi_i \\ \frac{\partial h_1}{\partial \mathbf{p}} &= \begin{pmatrix} \frac{\partial h_1}{\partial p_1} & \cdots & \frac{\partial h_1}{\partial p_n} \end{pmatrix} = \boldsymbol{\xi}'\end{aligned}\quad (\text{D.7})$$

$$\begin{aligned}v_{ij} &= -\frac{\partial \xi_i}{\partial p_j} = \begin{cases} 0, & \text{if } i \neq j \\ \sum_{k=0}^{y_i-1} \frac{1}{(p_i + k\phi)^2} + \sum_{k=0}^{m_i-y_i-1} \frac{1}{(1 - p_i + k\phi)^2}, & \text{if } i = j. \end{cases} \\ \frac{\partial^2 h_1}{\partial \mathbf{p} \partial \mathbf{p}'} &= \frac{\partial}{\partial \mathbf{p}'} \boldsymbol{\xi}' = \begin{pmatrix} \frac{\xi_1}{\partial p_1} & \cdots & \frac{\xi_n}{\partial p_1} \\ \vdots & \ddots & \vdots \\ \frac{\xi_1}{\partial p_n} & \cdots & \frac{\xi_n}{\partial p_n} \end{pmatrix} = \begin{pmatrix} -v_{11} & & \\ & \ddots & \\ & & -v_{nn} \end{pmatrix} = -\mathbf{V}\end{aligned}\quad (\text{D.8})$$

By Property D.1 (SV8),

$$\frac{\partial h_2}{\partial \mathbf{u}} = \frac{1}{2} \mathbf{u}' (\mathbf{D}^{-1} + \mathbf{D} - \mathbf{1}) = \mathbf{u}' \mathbf{D}^{-1}\quad (\text{D.9})$$

$$\frac{\partial^2 h_2}{\partial \mathbf{u} \partial \mathbf{u}'} = \frac{\partial}{\partial \mathbf{u}'} \mathbf{u}' \mathbf{D}^{-1} = \mathbf{D}^{-1}\quad (\text{D.10})$$

By Equations (D.7), (D.5), (D.7) and (D.10),

$$\begin{aligned}
\frac{\partial^2 h}{\partial \mathbf{u} \partial \mathbf{u}'} &= \frac{\partial^2 h_1}{\partial \mathbf{u} \partial \mathbf{u}'} - \frac{\partial^2 h_2}{\partial \mathbf{u} \partial \mathbf{u}'} = \frac{\partial}{\partial \mathbf{u}'} \left[ \frac{\partial h_1}{\partial \mathbf{p}} \frac{\partial \mathbf{p}}{\partial \mathbf{u}} \right] - \mathbf{D}^{-1} \\
&= \frac{\partial}{\partial \mathbf{u}'} [\boldsymbol{\xi}' \mathbf{S} \mathbf{Z}] - \mathbf{D}^{-1} = \frac{\partial \mathbf{p}}{\partial \mathbf{u}'} \frac{\partial}{\partial \mathbf{p}} [\boldsymbol{\xi}' \mathbf{S}] \mathbf{Z} - \mathbf{D}^{-1} \\
&= \mathbf{Z}' \mathbf{S} \begin{pmatrix} \frac{\partial}{\partial p_1} \xi_1 p_1 (1 - p_1) & \cdots & \frac{\partial}{\partial p_n} \xi_1 p_1 (1 - p_1) \\ \vdots & \ddots & \vdots \\ \frac{\partial}{\partial p_1} \xi_n p_n (1 - p_n) & \cdots & \frac{\partial}{\partial p_n} \xi_n p_n (1 - p_n) \end{pmatrix} \mathbf{Z} - \mathbf{D}^{-1} \\
&= \mathbf{Z}' \mathbf{S} \mathbf{W} \mathbf{Z} - \mathbf{D}^{-1}
\end{aligned} \tag{D.11}$$

where  $\mathbf{W} = \text{diag}(w_{11}, \dots, w_{nn})$  being

$$w_{ij} = \frac{\partial}{\partial p_j} \xi_i p_i (1 - p_i) = \begin{cases} 0, & \text{if } i \neq j \\ -v_i p_i (1 - p_i) + \xi_i (1 - 2p_i), & \text{if } i = j. \end{cases}$$

---

## BIBLIOGRAPHY

- American Thoracic Society (2002). ATS statement: guidelines for the six-minute walk test. *American Journal of Respiratory and Critical Care Medicine*, 166:111–7.
- Anderson, G. F., Hurst, J., Hussey, P. S., and Jee-Hughes, M. (2000). Health spending and outcomes: trends in OECD countries. *Health Affairs*, 19(3):150–157.
- Arostegui, I. and Núñez-Antón, V. (2008). Statistical aspects of the health related quality of life questionnaire Short Form-36 (SF-36). *Estadística Española*, 50(167):147–192.
- Arostegui, I., Núñez-Antón, V., and Quintana, J. M. (2007). Analysis of the Short Form-36 (SF-36): The beta-binomial distribution approach. *Statistics in Medicine*, 26:1318–1342.
- Arostegui, I., Núñez-Antón, V., and Quintana, J. M. (2012). Statistical approaches to analyse patient-reported outcomes as response variables: An application to health-related quality of life. *Statistical Methods in Medical Research*, 21:189–214.
- Arostegui, I., Núñez-Antón, V., and Quintana, J. M. (2013). On the recoding of continuous and bounded indexes to a binomial form: an applications to quality-of-life scores. *Journal of Applied Statistics*, 40:563–582.
- Au, H. J., Ringash, J., Brundage, M., Palmer, M., Richardson, H., Meyer, R. M., and NCIC CTG Quality of Life Committee (2010). Added value of health-related



- quality of life measurement in cancer clinical trials: the experience of the NCIC CTG. *Expert Review of Pharmacoeconomics & Outcomes Research*, 10(2):119–28.
- Basu, A. and Manca, A. (2012). Regression estimators for generic health-related quality of life and quality-adjusted life years. *Medical Decision Making*, 32(1):56–69.
- Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer, New York.
- Bock, R. D. (1985). *Multivariate Statistical Methods in Behavioural Research*. Scientific Software International, New York, USA.
- Bock, R. D. (1989). *Multilevel Analysis of Educational Data*. Academic Press, San Diego, CA, USA.
- Breiman, L. and Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlations (with discussion). *Journal of the American Statistical Association*, 80(391):580–619.
- Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88:9–25.
- Buckner, T. W., Wang, J., DeWalt, D. A., Jacobs, S., Reeve, B. B., and Hinds, P. S. (2014). Patterns of symptoms and functional impairments in children with cancer. *Pediatric Blood and Cancer*, 61(7):1282–1288.
- Buist, A. S., Vollmer, W. M., and McBurnie, M. A. (2008). Worldwide burden of COPD in high-and low-income countries. Part I. The burden of obstructive lung disease (BOLD) initiative. *International Journal of Tuberculosis and Lung Disease*, 12:703–708.
- Chan, A. (2001). Singapores changing structure and the policy implications for financial security, employment, living arrangements and health care. *Asian Meta-centre Research Paper*, 2:1–32.
- Chang, C. H. (2007). Patient-reported outcomes measurement and management with innovative methodologies and technologies. *Quality of Life Research*, 16:157–66.
- Cosio, B. G. and Agustí, A. (2010). Update in Chronic Obstructive Pulmonary Disease 2009. *American Journal of Respiratory and Critical Care Medicine*, 181:655–60.

- Cox, D. R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society, Series B*, 49(1):1–39.
- Croog, S. H., Levine, S., Testa, M. A., Brown, B., Bulpitt, C. J., Jenkins, C. D., Klerman, G. L., and Williams, G. H. (1986). The effects of antihypertensive therapy on the quality of life. *New England Journal of Medicine*, 314:1657–64.
- Danaher, P. J. (1987). *Estimating multidimensional tables from survey data: Predicting magazine audiences*. PhD thesis, Florida State University.
- De Leeuw, J. and Kreft, I. (1986). Random coefficient models for multilevel analysis. *Journal of Educational and Behavioral Statistics*, 11(1):57–85.
- Dempster, A. P., Rubin, D. B., and Tsutakawa, R. K. (1981). Estimation in covariance component models. *Journal of the American Statistical Association*, 76:341–353.
- Diggle, P., Heagerty, P., Liang, K.-Y., and Zeger, S. (2002). *Analysis of longitudinal data*. Clarendon Press.
- Dobrozsi, S. and Panepinto, J. (2015). Patient-reported outcomes in clinical practice. *Hematology. American Society of Hematology. Education Program.*, 2015:501–6.
- Doll, H., Duprat-Lomo, I., Ammerman, E., and Sagnier, P. P. (2003). Validity of the St Georges Respiratory Questionnaire at acute exacerbation of chronic bronchitis: Comparison with the Nottingham Health Profile. *Quality of Life Research*, 12:117–32.
- Drikvandi, R., Verbeke, G., Khodadadi, A., and Partovi Nia, V. (2013). Testing multiple variance components in linear mixed-effects models. *Biostatistics*, 14(1):144–159.
- Esteban, C., Arostegui, A., Aburto, M., Moraza, J., Quintana, J. M., Garca-Loizaga, A., Basualdo, L. V., Aramburu, A., Aizpiri, S., Uranga, A., and Capelastegui, A. (2016). Chronic obstructive pulmonary disease subtypes. Transitions over time. *PLoS One*, 11(9):e0161710.
- Ferri, C. P., Prince, M., Brayne, C., Brodaty, H., Fratiglioni, L., Ganguli, M., Hall, K., Hasegawa, K., Hendrie, H., Huang, Y., Jorm, A., Mathers, C., Menezes, P. R., Rimmer, E., and Sczufca, M. (2005). Global prevalence of dementia: A Delphi consensus study. *Lancet*, 366(9503):2112–2117.

- Fieuws, S. and Verbeke, G. (2006). Pairwise fitting of mixed models for the joint modelling of multivariate longitudinal profiles. *Biometrics*, 62(2):424–431.
- Fisch, M. J., Lee, J. W., Weiss, M., Wagner, L. I., Chang, V. T., Cella, D., Manola, J. B., Minasian, L. M., McCaskill-Stevens, W., Mendoza, T. R., and Cleeland, C. S. (2012). Prospective, observational study of pain and analgesic prescribing in medical oncology outpatients with breast, colorectal, lung, or prostate cancer. *Journal of Clinical Oncology*, 30(16):1980–8.
- Fitzmaurice, G., Davidian, M., Verbeke, G., and Molenberghs, G. (2008). *Longitudinal Data Analysis: A handbook of modern statistical methods*. Chapman & Hall/CRC.
- Folstein, M. F., Folstein, S. F., and McHugh, P. R. (1975). Mini-Mental State: A practical method for grading the cognitive state of patients for clinicians. *Journal of Psychiatric Research*, 12:189–198.
- Folstein, M. F., Folstein, S. F., and McHugh, P. R. (1988). Key papers in geriatric psychiatry. Mini-Mental State: A practical method for grading the cognitive state of patients for clinicians. *International Journal of Geriatric Psychiatry*, 13:285–294.
- Forcina, A. and Franconi, L. (1988). Regression analysis with the beta-binomial distribution. *Rivista di Statistica Applicata*, 21(1).
- Garratt, A. M., Schmidt, L., Mackintosh, A., and Fitzpatrick, R. (2002). Quality of life measurement: Bibliographic study of patient assessed health outcome measures. *British Medical Journal*, 324:1417–1421.
- Gibbons, R. D., Hedeker, D., Waternaux, C., and Davis, J. M. (1988). Random regression models: a comprehensive approach to the analysis of longitudinal psychiatric data. *Psychopharmacology Bulletin*, 24(3):438–43.
- Goldsmith, S. B. (1972). The status of health status indicators. *Health Services Reports*, 87:212–220.
- Goldstein, H. (1995). *Multilevel Statistical Models, Second Edition*. Halstead Press, New York, USA.

- Greven, S., Crainiceanu, C., Küchenhoff, H., and Peters, A. (2008). Restricted likelihood ratio testing for zero variance components in linear mixed models. *Journal of Computational and Graphical Statistics*, 17(4):870–891.
- Halbert, R. J., Isonaka, S., George, D., and Iqbal, A. (2003). Interpreting COPD prevalence estimates: what is the true burden of disease? *Chest*, 123:1684–92.
- Harville, D. A. (1997). *Matrix Algebra From a Statistical's Perspective*. Springer, Yorktown Heights, N Y, 1 edition.
- Hedeker, D. and Gibbons, R. D. (2006). *Longitudinal Data Analysis*. John Wiley & Sons, INC.
- Hsieh, F., Tseng, Y. K., and Wang, J. L. (2006). Joint modeling of survival and longitudinal data: Likelihood approach revisited. *Biometrics*, 62(4):1037–1043.
- Hui, S. L. and Berger, J. O. (1983). Empirical bayes estimation of rates in longitudinal studies. *Journal of the American Statistical Association*, 78(384):753–760.
- Hunger, M., Baumert, J., and Holle, R. (2011). Analysis of SF-6D index data: Is beta regression appropriate? *Value in Health*, 14(5):759–767.
- Hunt, S. M., McEwen, J., and McKenna, S. P. (1985). Measuring health status: A new tool for clinicians and epidemiologists. *Journal of the Royal College of General Practitioners*, 35:185–188.
- Izem, R., Kammerman, L. A., and Komo, S. (2014). Statistical challenges in drug approval trials that use patient-reported outcomes. *Statistical Methods in Medical Research*, 23:398–408.
- Jones, P. W. (2002). Interpreting thresholds for a clinically significant change in health status in asthma and COPD. *European Respiratory Journal*, 19:398–404.
- Jones, P. W. (2005). St. George's Respiratory Questionnaire: MCID. *Journal of Chronic Obstructive Pulmonary Disease*, 2:75–79.
- Jones, P. W. and Higenbottam, T. (2007). Quantifying of severity of exacerbations in chronic obstructive pulmonary disease: Adaptations to the definition to allow quantification. *Proceedings of the American Thoracic Society*, 4:597–601.
- Jones, P. W., Quirk, F. H., and Baveystock, C. M. (1991). The St George's Respiratory Questionnaire. *Respiratory Medicine*, 85:25–31.

- Jones, P. W., Quirk, F. H., Baveystock, C. M., and Littlejohns, P. (1992). A self-complete measure of health status for chronic airflow limitation. The St. George's Respiratory Questionnaire. *American Review of Respiratory Disease*, 145(6):1321–7.
- Jørgensen, B. (1984). The Delta Algorithm and GLIM. *International Statistical Review*, 52:283–300.
- Jørgensen, B. (1987). Exponential dispersion models (with discussion). *Journal of the Royal Statistical Society, Series B*, 49:127–162.
- Jørgensen, B. (1997). *The Theory of Dispersion Models*. Chapman & Hall.
- Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, 38(4):963–74.
- Lee, Y. and Nelder, J. A. (1996). Hierarchical generalized linear models. *Journal of the Royal Statistical Society, Series B*, 58:619–678.
- Lee, Y. and Nelder, J. A. (2001). Hierarchical generalized linear models: A synthesis of generalized linear models, random-effects and structured dispersions. *Biometrika*, 88:987–1006.
- Lee, Y. and Nelder, J. A. (2004). Conditional and marginal models: Another view. *Statistical Science*, 19(2):219–238.
- Lee, Y. and Nelder, J. A. (2006). Double hierarchical generalized linear models (with discussion). *Applied Statistics*, 55(2):139–185.
- Lee, Y., Nelder, J. A., and Pawitan, Y. (2006). *Generalized Linear Models with Random Effects: Unified Analysis via H-likelihood*. Chapman & Hall/CRC.
- Letenneur, L., Commenges, D., Dartigues, J. F., and Barberger-Gateau, P. (1994). Incidence of dementia and Alzheimer's disease in elderly community residents of south-western France. *International Journal of Epidemiology*, 23:1256–1261.
- Lindsay, B. (1988). Composite likelihood methods. *Contemporary Mathematics*, 80:221–239.
- Longford, N. T. (1987). A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. *Biometrika*, 74(4):817–827.

- MacCallum, R., Kim, C., Malarkey, W., and Kiecolt-Glaser, J. (1997). Studying multivariate change using multilevel models and latent curve models. *Multivariate Behavioral Research*, 32:215–253.
- Mahler, D. A., Ward, J., Waterman, L. A., McCusker, C., Zuwallack, R., and Baird, J. C. (2009). Patient-reported dyspnea in COPD reliability and association with stage of disease. *Chest*, 136(6):1473–9.
- Martin, G. M. (2009). Defeating dementia. *Nature*, 431:247–248.
- Matallana, D., de Santacruz, C., Cano, C., Reyes, P., Samper-Ternent, R., Markides, K. S., Ottenbacher, K. J., and Reyes-Ortiz, C. A. (2011). The relationship between education level and Mini Mental State Examination domains among older Mexican Americans. *Journal of Geriatric Psychiatry and Neurology*, 24(1):9–18.
- McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models (Second edition)*. Chapman & Hall, London.
- McCulloch, C. E. (2008). Joint modelling of mixed outcome types using latent variables. *Statistical Methods in Medical Research*, 17(1):53–73.
- McCulloch, C. E. and Searle, S. R. (2001). *Generalized, Linear and Mixed Models*. John Wiley & Sons, INC., Hoboken, NJ, USA.
- McDowell, I. and Newell, C. (1987). *Measuring Health: A Guide to Rating Scales and Questionnaires*. Oxford University Press.
- Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. Springer, New York.
- Molenberghs, G., Verbeke, G., Demétrio, C., and Vieira, A. (2010). A family of generalized linear models for repeated measures with normal and conjugate random effects. *Statistical Science*, 25(3):325–347.
- Molenberghs, G., Verbeke, G., Iddi, S., and Demétrio, C. (2012). A combined beta and normal random-effects model for repeated, overdispersed binary and binomial data. *Journal of Multivariate Analysis*, 111:94–109.
- Montazeri, A. (2008). Health-related quality of life in breast cancer patients: a bibliographic review of the literature from 1974 to 2007. *Journal of Experimental and Clinical Cancer Research*, 29:27–32.

- Montazeri, A. (2009). Quality of life data as prognostic indicators of survival in cancer patients: an overview of the literature from 1982 to 2008. *Health and Quality of Life Outcomes*, 7(102):1–21.
- Murray, C. J. and López, A. D. (1997). Alternative projections of mortality and disability by cause 1990-2020: Global burden of disease study. *Lancet*, 349:1498–504.
- Najera-Zuloaga, J., Lee, D.-J., and Arostegui, I. (2017). Comparison of beta-binomial regression model approaches to analyze health-related quality of life data. *Statistical Methods in Medical Research (in press)*.
- Oort, F. (2001). Three-mode models for multivariate longitudinal data. *British Journal of Mathematical and Statistical Psychology*, 54:49–78.
- Pal, B., Murti, K., Siddiqui, N. A., Das, P., Lal, C. S., Babu, R., Rastogi, M. K., and Pandey, K. (2017). Assessment of quality of life in patients with post kalaazar dermal leishmaniasis. *Health and Quality of Life Outcomes*, 15(1):148.
- Pan, J. and McKenzie, G. (2007). Modelling conditional covariance in the linear mixed model. *Statistical Modelling*, 7(1):49–71.
- Patterson, H. D. and Thompson, R. (1971). Recovery of interblock information when block sizes are unequal. *Biometrika*, 58:545–554.
- Pauwels, R. A. and Rabe, K. F. (2004). Burden and clinical features of chronic obstructive pulmonary disease (COPD). *Lancet*, 364(9434):613–20.
- Pawitan, Y. (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford University Press, USA.
- Peruzza, S., Sergi, G., Vianello, A., Pisent, C., Tiozzo, F., Manzan, A., Coin, A., Inelmen, E. M., and Enzi, G. (2003). Chronic obstructive pulmonary disease (COPD) in elderly subjects: impact on functional status and quality of life. *Respiratory Medicine*, 97(6):612–7.
- Proust-Lima, C., Phillipps, V., Diakite, A., and Liquet, B. (2017). Estimation of extended mixed models using latent class and latent process: the R package lcmdm. *Journal of Statistical Software*, 78(2):1–56.

- Quirk, F. H. and Jones, P. W. (1990). Patients' perception of distress due to symptoms and effects of asthma on daily living and an investigation of possible influential factors. *Clinical Science*, 79:17–21.
- Raubenbush, S. W. and Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods, Second Edition*. Sage, Newbury Park, CA, USA.
- Reinsel, G. (1984). Estimation and prediction in a multivariate random effects generalized linear model. *Journal of the American Statistical Association*, 79:406–414.
- Rigby, R. A. and Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape. *Applied Statistics*, 54:507–554.
- Rizopoulos, D., Verbeke, G., and Lesaffre, E. (2009). Fully exponential laplace approximations for the joint modelling of survival and longitudinal data. *Journal of the Royal Statistical Society, Series B*, 71:637–654.
- Robinson, G. K. (1991). That BLUP is a Good Thing: The Estimation of Random Effects. *Statistical Science*, 6(1):15–32.
- Ronnegard, L., Shen, X., and Alam, M. (2010). hglm: A package for fitting hierarchical generalized linear models. *The R Journal*, 2(2):20–28.
- Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.
- Sánchez-García, S., Gallegos-Carrillo, K., Espinel-Bermudez, M. C., Doubova, S. V., Sánchez-Arenas, R., García-Peña, C., Salá, A., and Briseño-Fabian, S. C. (2017). Comparison of quality of life among community-dwelling older adults with the frailty phenotype. *Quality of Life Research*. Epub ahead of print.
- Shun, Z. and McCullagh, P. (1995). Laplace approximation of high dimensional integrals. *Journal of the Royal Statistical Society, Series B*, 57(4):749–760.
- Singh, S. J., Sodergren, S. C., Hyland, M. E., Williams, J., and Morgan, M. D. (2001). A comparison of three disease-specific and two generic health-status measures to evaluate the outcome of pulmonary rehabilitation in COPD. *Respiratory Medicine*, 95(1):71–1.



- Smithson, M. and Verkuilen, J. (2006). A better lemon squeezer? maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods*, 11(1):54–71.
- Speight, J. and Barendse, S. M. (2010). FDA guidance on patient reported outcomes. *British Medical Journal*, 340:c2921.
- Stansfeld, S., Roberts, R., and Foot, S. (1997). Assessing the validity of the SF-36 general health survey. *Quality of Life Research*, 6:217–224.
- Stasinopoulos, D. M. and Rigby, R. A. (2007). Generalized additive models for location scale and shape (GAMLSS) in R. *Journal of Statistical Software*, 23:1–56.
- Tsiatis, A. and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: An overview. *Statistica Sinica*, 14:809–834.
- US Department of Health and Human Services (2012). *Guidance for industry: patient-reported outcome measures: use in medical product development to support labeling claims*.
- Verbeke, G., Fieuws, S., Molenberghs, G., and Davidian, M. (2014). The analysis of multivariate longitudinal data: A review. *Statistical Methods in Medical Research*, 23(1):42–59.
- Verkuilen, J. and Smithson, M. (2012). Mixed and mixture regression models for continuous bounded responses using the beta distribution. *Journal of Educational and Behavioral Statistics*, 37(1):82–113.
- Ware, J. E., Kosinski, M. A., and Keller, S. D. (1994). *SF-36 Physical and Mental Health Summary Scales: A Users Manual*. The Health Institute, New England Medical Center, Boston.
- Ware, J. E., Snow, K. K., Kosinski, M. A., and Gandek, B. (1993). *SF36 Health Survey, Manual and Interpretation Guides*. The Health Institute, New England Medical Center.
- Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*, 61(3):439–447.

- WHO, World Health Statistics report (2017). World Health Statistics 2017: Monitoring health for the SDGs, sustainable development goals. Technical report, World Health Organization.
- Wiklund, I. (2004). Assessment of patient-reported outcomes in clinical trials: The example of health-related quality of life. *Fundamental and Clinical Pharmacology*, 18:351–63.
- Wilson, I. B. and Cleary, P. D. (1995). Linking clinical variables with health-related quality of life. A conceptual model of patient outcomes. *Journal of the American Medical Association*, 273:59–65.
- Winer, B. J. (1971). *Statistical Principles in Experimental Design, 2nd edition*. McGraw-Hill, New York, USA.
- Wittchen, H. U., Jacobi, F., Rehm, J., Gustavsson, A., Svensson, M., Jönsson, B., Olesen, J., Allgulander, C., Alonso, J., Faravelli, C., Fratiglioni, L., Jennum, P., Lieb, R., Maercker, A., van Os, J., Preisig, M., Salvador-Carulla, L., Simon, R., and Steinhausen, H. C. (2011). The size and burden of mental disorders and other disorders of the brain in Europe 2010. *European Neuropsychopharmacology*, 21(9):655–679.
- Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R*. Chapman & Hall/CRC Texts in Statistical Science.
- Wu, H., Zhang, Y., and Long, J. D. (2017). Longitudinal beta-binomial modeling using GEE for overdispersed binomial data. *Statistics in Medicine*, 36:1029–1040.
- Zeger, S. L. and Liang, K.-Y. (1982). Random effects models for longitudinal data. *Biometrics*, 38:784–92.
- Zeger, S. L. and Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, 42:121–130.
- Zeng, L. and Cook, R. J. (2007). Transition models for multivariate longitudinal binary data. *Journal of the American Statistical Association*, 102(477):211–223.
- Zigmond, A. S. and Snaith, R. P. (1983). The hospital anxiety and depression scale. *Acta Psychiatrica Scandinavica*, 67:361–370.