eman ta zabal zazu
UPV EHU
ZIENTZIA ETA TEKNOLOGIA FAKULTATEA
FACULTAD DE CIENCIA Y TECNOLOGÍA
50 URTE AÑOS 1968 - 2018
Biba Zientzia! Ciencia Viva

Hosmer-Lemeshow testean erabilitako talde-kopuruaren azterketa simulazioen bidez

Gradu Amaierako Lana
Matematikako Gradua

Ane Moreno Oya

Irantzu Barrio Beraza
Irakasleak zuzendutako lana

Leioa, 2022ko ekainaren 22a

Gaien Aurkibidea

Sarrera

Gaur egun, erregresio logistikoa datu bitarrak modelizatzeko gehien erabiltzen den tresnetako bat da. Izan ere, eredu mota honek doitzen dituen balioak emaitza interesekoa dugun kategoriakoa izateko probabilitatea da, interpretatzeko erraza dena. Behin eredua doituta, ezinbestekoa da estimatutako probabilitateek behatutako datuak adierazten dituztela egiaztatzea. Hosmer-Lemeshow testa ereduaren doikuntza-egokitasuna neurtzeko erabiltzen den metodo bat da eta ereduak estimatutako probabilitateen zehaztasuna neurtzen du.

Hosmer-Lemeshow testak lagin tamainarekin lotutako hainbat muga ditu eta honen isla testak denboran zehar jasandako aldaketak dira. Izan ere, azken urteetan, hainbat eraldaketa proposatu dira eragozpenak gainditzeko helburuarekin (ikusi [1], [2], [3], [4]). Test honetan, estimatutako probabilitateak sailkatzeko g talde kopuru jakin bat erabiltzen da.

Lan honetan, g talde kopuruaren arabera, testaren egonkortasuna nola aldatzen den aztertu dugu eta, horretarako, `R` programa erabili dugu.

Erabakien aldakuntza aztertzeko simulazioak erabili ditugu, eszenario ezberdinak planteatuz. Eszenario bakoitzean, erregresio logistikoko ereduak doitu ditugu eta g talde kopuruak erabakietan eragindako aldaketak ikertu ditugu.

1. kapituluan, erregresio logistikoa lantzen da. Erregresio logistikoko ereduen adierazpen orokorra eta hauek zehazteko erabiltzen diren koefizienteak lortzeko prozedura azaltzen dira. Ereduen doikuntza-egokitasuna zer den azaltzen da, Hosmer-Lemeshow testari sarrera emanez.

2. kapituluan, Hosmer-Lemeshow testa azaltzen da. Testa aplikatzeko erabiltzen diren parametroak azaltzen dira testaren eragozpenekin batera. Gainera, Hosmer-Lemeshow test eraldatua azaltzen da.

3. kapituluan, egindako simulazioak azaltzen dira lortutako emaitzekin batera.

4. kapituluan, elkartutako informazioa eta simulazioen emaitzak erabiliz, ondorioak azaltzen dira.

Amaitzeko, bibliografia eta, eranskin gisa, erabilitako `R` kodea daude. Bertan simulazioak egiteko sortu diren funtzioak azaltzen dira.

1. Kapitulua

Erregresio logistikoa

1.1 Sarrera

Erregresio ereduen helburua erantzun aldagaiaren eta aldagai askeen arteko erlazioa adierazteko interpretagarria den eta doikuntza egokiena duen eredua lortzea da. Erregresio motarik ezagunena erregresio lineala da, non menpeko aldagaia jarraitua eta normala dela suposatzen den. Erregresio logistikoa linealetik bereizten duen ezaugarri nagusia erregresio logistikoan erantzun aldagaia bitarra edo dikotomikoa dela da.

Erregresio logistikoaren abantaila handietako bat gure erantzun aldagaiarentzat probabilitate eredu bat sortzen duela da. Beste era batera esanda, eredu logistiko batean doitutako balioak ez dira bitarrak, baizik eta emaitza intereseko kategoriakoa izateko probabilitatea. Hori dela eta, datu bitarrak estatistikoki modelizatzeko gehien erabiltzen den metodoetako bat da. Emaitza bitarrak ia ikasketa eremu guztietan aurkitzen dira, esate baterako, medikuntzan, ekonomian eta psikologian.

1.2 Erregresio logistiko bakuna

Izan bitez Y eta X bi zorizko aldagai, Y erantzun aldagai bitarra eta X aldagai askea izanik. Interesekoa dugun ezaugarriaren presentzia adierazteko Y aldagaiak 1 balioa hartzen duela suposatuko dugu. Beraz, $Y = 1$ arrakasta izendatuko dugu eta $Y = 0$, porrota.

Izan bedi $p(X) = P(Y = 1|X)$ arrakastaren probabilitate baldintzatua. Orduan, $Y \sim Bernoulli(p(X))$ betetzen da. Eredua eraikitzeko $p(X)$-ren *logit* transformazioa erabiliko dugu,

$$g(X) = logit[p(X)] = \ln\left[\frac{p(X)}{1-p(X)}\right] = \beta_0 + \beta_1 X \tag{1.1}$$

lortuz. Orduan, erregresio logistiko bakunaren eredua honakoa da:

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}. \tag{1.2}$$

$g(X)$ funtzioak erregresio linealeko hainbat propietate betetzen ditu: lineala da, X-ren arabera, $(-\infty, \infty)$ tarteko balioak har ditzake eta jarraitua izan daiteke.

1.2.1 Ereduaren doikuntza

Demagun (X, Y) aldagaien n tamainako lagin askea dugula, $\{(x_i, y_i)\}_{i=1}^n$. Eredua zehazteko, behatutako datuetan oinarrituta, β_0 eta β_1 balioak estimatu beharra dago. Horretarako, egiantz handieneko metodoa erabiltzen da. Metodo honi esker, β_0 eta β_1-ren estimazioak lortzen ditugu non behatutako datuak lortzeko probabilitatea maximizatzen den [5].

Orduan, $\boldsymbol{\beta} = (\beta_0, \beta_1)^t$ eta $p(x_i) = p(X = x_i)$ izendatuz, egiantz handieneko funtzioa hurrengoa da:

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} p(x_i)^{y_i}(1-p(x_i))^{1-y_i}. \tag{1.3}$$

Izan ere,

$$P(Y = y_1, \ldots, Y = y_n) = \prod_{i=1}^{n} P(Y = y_i)$$

betetzen da lagina askea delako eta, $Y \sim Bernoulli(p(X))$ denez,

$$\prod_{i=1}^{n} P(Y = y_i) = \prod_{i=1}^{n} p(x_i)^{y_i}(1-p(x_i))^{1-y_i}.$$

Lehenago aipatu denez, metodoaren helburua egiantz handieneko funtzioa maximizatzea da. Hala ere, matematikoki errazagoa da (1.3) ekuazioaren logaritmoa lantzea. Defini dezagun egiantz handieneko logaritmoa:

$$L(\boldsymbol{\beta}) = \ln[l(\boldsymbol{\beta})] = \sum_{i=1}^{n} y_i \ln[p(x_i)] + (1-y_i)\ln[1-p(x_i)]. \tag{1.4}$$

$\boldsymbol{\beta}$ estimatzeko, (1.4) diferentziatuko dugu β_0 eta β_1 koefizienteekiko. Ondorioz, bi egiantz-ekuazio lortuko ditugu:

$$\sum_{i=1}^{n} y_i - p(x_i) = 0 \tag{1.5}$$

eta

$$\sum_{i=1}^{n} x_i\,(y_i - p(x_i)) = 0. \tag{1.6}$$

(1.5) eta (1.6) ekuazioak β_0 eta β_1 aldagaietan ez-linealak direnez, hauek ebazteko zenbakizko metodoak erabiltzen dira. Newton-Raphson metodoa erabiltzea gomendatzen da [6].

$\boldsymbol{\beta}$-ren estimatzailea $\hat{\boldsymbol{\beta}}$ izendatuko dugu. Beraz, (x_i, y_i) indibiduoaren arrakastarako probabilitatearen estimazioa $\hat{\boldsymbol{\beta}}$ (1.2) ekuazioan ordezkatuz lortzen da:

$$\hat{p}(x_i) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x_i}}{1 + e^{\hat{\beta}_0+\hat{\beta}_1 x_i}}. \tag{1.7}$$

1.3 Erregresio logistiko anizkoitza

Aurreko atalean erregresio logistikoa azaldu dugu aldagai azaltzaile bakarra dagoen kasurako. Atal honetan, aldagai independente gehiago erabiliko ditugu.

Izan bitez $X_1, \ldots, X_p$ p zorizko aldagai aske eta $\mathbf{X} = (X_1, \ldots, X_p)^t$, Y erantzun aldagai bitarra izanik. Izan bedi $p(\mathbf{X}) = P(Y = 1|\mathbf{X})$ arrakastaren probabilitate baldintzatua. Beraz, $Y \sim Bernoulli(p(\mathbf{X}))$ betetzen da. Ereduaren *logit* transformazioa

$$g(\mathbf{X}) = logit[p(\mathbf{X})] = \ln\left[\frac{p(\mathbf{X})}{1 - p(\mathbf{X})}\right] = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p \tag{1.8}$$

da, non

$$p(\mathbf{X}) = \frac{e^{g(\mathbf{X})}}{1 + e^{g(\mathbf{X})}} \tag{1.9}$$

den.

Aldagai askeren bat diskretua bada, desegokia da jarraitua izango balitz bezala erabiltzea. Egoera honetan, *dummy* aldagaiak erabiliko ditugu.

Orokorrean, aldagai diskretu batek m balio hartzen baditu, $m-1$ *dummy* aldagai beharko ditugu. *Dummy* aldagaien kopurua $m - 1$ da ereduan β_0 gai konstantea dagoelako.

Suposa dezagun ereduko j. aldagaiak m balio dituela. *Dummy* aldagaiak $D_{j_1}, \ldots, D_{j_{m-1}}$ izendatuko ditugu eta dagozkien koefizienteak $\beta_{j_1}, \ldots, \beta_{j_{m-1}}$ izango dira, hurrenez hurren. Orduan, X_j aldagaiak i. balioa hartzen badu, D_{j_i} aldagaiak 1 balioa hartuko du, gainerakoek 0 balioa duten bitartean.

Beraz, *logit* transformazioa, j. aldagaia diskretua izanik,

$$g(\mathbf{X}) = \ln\left[\frac{p(\mathbf{X})}{1 - p(\mathbf{X})}\right] = \beta_0+\beta_1 X_1+\ldots+\ldots+\sum_{l=1}^{m-1} \beta_{j_l} D_{j_l}+\ldots+\beta_p X_p \tag{1.10}$$

da.

Hemendik aurrera, notazioa errazteko, aldagai askeak jarraituak edo dikotomikoak direla suposatuko dugu.

1.3.1 Ereduaren doikuntza

Eredua zehazteko $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^t$ bektorearen balioak estimatzea beharrezkoa da eta, horretarako, erregresio logistiko bakunean bezala, egiantz handieneko metodoa erabiltzen da.

Demagun $(\mathbf{X}, Y)$ aldagaien n tamainako lagin askea dugula, $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$. Orduan, egiantz handieneko funtzioa hurrengoa da:

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} p(\mathbf{x}_i)^{y_i}(1 - p(\mathbf{x}_i))^{1-y_i}, \tag{1.11}$$

non $p(\mathbf{x}_i) = p(\mathbf{X} = \mathbf{x}_i)$ den. Aldagai azaltzaile bakarreko ereduan bezala, (1.11) ekuazioaren logaritmoa landuko dugu. Egiantz handieneko logaritmoa ondoko eran definituko dugu:

$$L(\boldsymbol{\beta}) = \ln[l(\boldsymbol{\beta})] = \sum_{i=1}^{n} y_i \ln[p(\mathbf{x}_i)] + (1 - y_i) \ln[1 - p(\mathbf{x}_i)]. \tag{1.12}$$

$\boldsymbol{\beta}$-ren estimazioa lortzeko, (1.12) ekuazioa diferentziatuko dugu β_j bakoitzarekiko, $j = 0, \ldots, p$ izanik. Beraz, guztira $p + 1$ egiantz-ekuazio lortuko ditugu. Hurrengo eran adieraz daitezke:

$$\sum_{i=1}^{n} y_i - p(\mathbf{x}_i) = 0 \tag{1.13}$$

eta

$$\sum_{i=1}^{n} x_{ij} \left(y_i - p(\mathbf{x}_i)\right) = 0, \tag{1.14}$$

non $j = 1, \ldots, p$ eta $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^t$ diren.

(1.13) eta (1.14) ekuazioak $\beta_0, \ldots, \beta_p$ aldagaietan ez-linealak direnez, hauek ebazteko zenbakizko metodoak erabiltzen dira. Newton-Raphson metodoa erabiltzea gomendatzen da [6].

$\boldsymbol{\beta}$-ren estimatzailea $\hat{\boldsymbol{\beta}}$ denotatuko dugu eta $(\mathbf{x}_i, y_i)$ indibiduoaren arrakastaren probabilitatea $\hat{\boldsymbol{\beta}}$ (1.8) ekuazioan ordezkatuz lortzen da:

$$p(\mathbf{x}_i) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_p x_{ip}}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_p x_{ip}}}. \tag{1.15}$$

1.4 Ereduaren doikuntza-egokitasuna

Behin eredua eraikita, aldagai guztien eragina adierazgarria dela suposatuz, jakin nahi dugu ea estimatzen dituen probabilitateek behatutako erantzun aldagaiaren balioak islatzen dituzten. Honi doikuntza-egokitasuna deritzo.

Doikuntza-egokitasuna neurtzeko estatistikoak absolutuak dira, hau da, ezin dira erabili eredu ezberdinak konparatzeko.

Eredu baten doikuntza-egokitasuna aztertzeko bi irizpide nagusi daude: kalibrazioa eta diskriminazioa. Kalibrazio ona duen eredu batek zehaztasunez estimatzen ditu probabilitateak. Bestalde, diskriminazio ona duen eredu batek doitasunez bereizten du noiz gertatzen den interesekoa dugun gertaera.

Hosmer-Lemeshow (HL) testa doitutako ereduaren kalibrazioa neurtzeko erabiltzen da.

2. Kapitulua

Hosmer-Lemeshow testa

Suposa dezagun gure ereduan p aldagai aske ditugula, $X_1, \ldots, X_p$, eta Y erantzun aldagaia dela. $\mathbf{X} = (X_1, \ldots, X_p)^t$ izanik, demagun $(\mathbf{X}, Y)$ aldagaien lagin aske bat dugula, $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$. Orduan, $\mathbf{X}$ aldagai-bektorearen behatutako balio ezberdinen kopurua J denotatuko dugu. Balioren bat errepikatzen bada, orduan $J < n$ izango da. Izan bedi $\mathbf{x}^j$ $\mathbf{X}$-ren behatutako j. balioa adierazten duen bektorea, $j = 1, \ldots, J$ izanik. $\mathbf{x}_i = \mathbf{x}^j$ betetzen duten indibiduoen kopurua m_j bitartez adieraziko dugu, non $j = 1, \ldots, J$. Ohartu $m_j = \sum_{i=1}^n I(\mathbf{x}_i = \mathbf{x}^j)$ dela. Beraz, $\sum_{j=1}^J m_j = n$ betetzen da. Izan bedi y^j $y_i = 1$ balioa duten banako kopurua $\mathbf{x}_i = \mathbf{x}^j$ betetzen dutenen artean. Beraz, $y^j = \sum_{i=1}^n I(y_i = 1 | \mathbf{x}_i = \mathbf{x}^j)$ da. Orduan, $\sum_{j=1}^J y^j = n_1$ non n_1 lagineko arrakasta kopuru totala den.

2.1 Pearsonen Khi Karratua

Pearsonen Khi Karratua behatutako eta itxarotako probabilitateen arteko diferentzia neurtzeko erabiltzen da eta HL testaren oinarria da.

Pearsonen hondarra $\mathbf{x}_i = \mathbf{x}^j$ betetzen duten banakoentzat

$$r_j = \frac{y^j - m_j \hat{p}_j}{\sqrt{m_j \hat{p}_j (1 - \hat{p}_j)}} \tag{2.1}$$

da, non $\hat{p}_j = \hat{p}(\mathbf{x}^j) = \dfrac{e^{\hat{g}(\mathbf{x}^j)}}{1 + e^{\hat{g}(\mathbf{x}^j)}}$ den. Orduan, hondar hauetan oinarrituta, Pearsonen Khi Karratu estatistikoa

$$\chi^2 = \sum_{j=1}^J (r_j)^2 = \sum_{j=1}^J \left(\frac{y^j - m_j \hat{p}_j}{\sqrt{m_j \hat{p}_j (1 - \hat{p}_j)}} \right)^2 = \sum_{j=1}^J \frac{(y^j - m_j \hat{p}_j)^2}{m_j \hat{p}_j (1 - \hat{p}_j)} \tag{2.2}$$

eran definitzen da.

2.2 Hosmer-Lemeshow pertzentilen testa

Demagun $J = n$ dela. Orduan, estimatutako probabilitateak txikienetik handienera ordenatu ostean, hauek sailkatzeko bi estrategia proposatzen dira:

(i) Estimatutako probabilitateak pertzentilak erabilita taldekatzea. Mozketa puntu hauei arriskuko pertzentilak deritze. Termino hau osasun zientzien ikerketatik dator, non askotan arrakastak gaixotasun baten presentzia adierazten duen.

(ii) Estimatutako probabilitateak aldez aurretik finkatutako balioen arabera taldekatzea.

Lehenengo metodoarekin probabilitateak sailkatzeko $g = 10$ talde erabilita, lehenengo taldean $n'_1 = n/10$ estimatutako probabilitate baxuenak dituzten banakoak edukiko genituzke eta azken taldean, $n'_{10} = n/10$ estimatutako probabilitate altuenak dituzten banakoak. Bigarren metodoarekin $g = 10$ talde erabilita, $t/10$ mozketa puntuak lortzen ditugu $t = 1, \ldots, 9$ izanik. Orduan, 1. taldean $1/10$ eta $2/10$ arteko estimatutako probabilitatea duten indibiduoak egongo lirateke eta 10. taldean, $9/10$ eta 1 arteko estimatutako probabilitatea dutenak.

Aipatutako taldekatzeko bi metodoetatik edozein aukeratuta ere, HL testean erabilitako estatistikoa jarraian definituko dugun $\hat{C}$ estatistikoa da. $g = 10$ kasuan, $\hat{C}$ kalkulatzeko erabiltzen diren talde bakoitzeko behatutako eta itxarotako maiztasunak 2.1. taulan ageri dira, d_l arriskuko dezilak izanik $l = 1, \ldots, 10$.

Izan bedi c_k k. taldeko behaketa kopurua. Orduan, k. taldean, arrakastarako itxarotako eta behatutako maiztasunak

$$e_{1k} = \sum_{j=1}^{c_k} m_j \hat{p}_j \tag{2.3}$$

eta

$$o_{1k} = \sum_{j=1}^{c_k} y^j \tag{2.4}$$

dira, hurrenez hurren. Era berean,

$$e_{0k} = \sum_{j=1}^{c_k} (1 - m_j)\hat{p}_j \tag{2.5}$$

eta

$$o_{0k} = \sum_{j=1}^{c_k} (m_j - y^j) \tag{2.6}$$

k. taldeko porroterako itxarotako eta behatutako maiztasunak dira, hurrenez hurren.

Orduan, $\hat{C}$ estatistikoa χ^2 estatistikoa kalkulatuz lortzen da (2.2) ekuazioa jarraituz:

$$\hat{C} = \sum_{k=1}^{g} \left[\frac{(o_{1k} - e_{1k})^2}{e_{1k}} + \frac{(o_{0k} - e_{0k})^2}{e_{0k}} \right]. \tag{2.7}$$

$J = n$ eta aukeratutako ereduaren doikuntza egokia denean, $\hat{C}$ estatistikoak χ^2_{g-2} banaketa asintotikoa du. $J \approx n$ deneko kasuan, ez da zertan bete behar [5]. Gainera, simulazioen bidez frogatu da $\hat{C}$-ren banaketa χ^2_{g-2} banaketara hobeto doitzen dela pertzentilen taldekatze metodoa erabilita mozketa puntuen taldekatze teknikarekin baino, bereziki estimatutako probabilitateak txikiak direnean [7]. Hortaz, hemendik aurrera, $\hat{C}$ kalkulatzeko pertzentilen metodoa erabili dela suposatuko dugu.

Orduan,

$$\begin{cases} H_0 : \text{Ereduaren doikuntza egokia da.} \\ H_1 : \text{Ereduaren doikuntza ezegokia da.} \end{cases}$$

hipotesi-kontrastea egiteko erabiliko dugun estatistikoa $\hat{C}$ da. Ondorioz, *p-balioa*

$$p = \int_{\hat{C}}^{\infty} \chi^2_{g-2}(z)dz \tag{2.8}$$

izango da, non $\chi^2_{g-2}(z)$ χ^2 banaketaren dentsitate funtzioa den z-n ebaluatuta, askatasun-graduak $g - 2$ izanik.

Hortaz, α adierazgarritasun-maila finkaturik, $p > \alpha$ bada, gure ereduaren doikuntza egokia dela ez dugu errefusatuko. Bestelakoan, doikuntza ezegokia dela esango dugu.

Taldeak sortzerako orduan, behaketa kopurua txikia bada, gerta liteke ez ohartzea ereduaren doikuntza ezegokia dela. Gainera, arriskuko talde be-

	Y: menpeko aldagaia			
	$Y = 1$		$Y = 0$	
Taldeak	Behatutakoa	Itxarotakoa	Behatutakoa	Itxarotakoa
$\hat{p} < d_1$	o_{11}	e_{11}	o_{01}	e_{01}
$d_1 \leq \hat{p} < d_2$	o_{12}	e_{12}	o_{02}	e_{02}
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
$d_9 \leq \hat{p} < d_{10}$	$o_{1_{10}}$	$e_{1_{10}}$	$o_{0_{10}}$	$e_{0_{10}}$
Guztira	o_1	e_1	o_0	e_0

2.1. Taula: $\hat{C}$ kalkulatzeko erabili diren behatutako eta itxarotako maiztasunak $g = 10$ kasuan.

rean dauden banakoek aldagai azaltzaileetan balio ezberdinak eduki ditzakete. Amankomunean duten ezaugarri bakarra estimatutako probabilitateak antzekoak edukitzea izan daiteke.

Hurrengo eragozpena azaldu baino lehen, test baten ahalmena zer den gogoratuko dugu. Ahalmena II motako errorea ez egiteko probabilitatea da. Hau da, hipotesi alternatiboa (ereduaren doikuntza ez dela ona) onartzeko probabilitatea, eredua ondo doituta ez dagoenean. 2.2. taulan hipotesi-kontrasteetan har daitezkeen erabaki posibleak ikus daitezke.

Behaketa kopurua handia bada, milaka adibidez, gerta daiteke testak doikuntza oneko hipotesia baztertzea eredua zentzuzkoa eta klinikoki onargarria izan arren. Honen arrazoia gehiegizko ahalmen estatistikoa da, testak estimatutako eta itxarotako balioen arteko desberdintasun txikiak estatistikoki esanguratsu gisa sailka ditzake. Izan ere, testaren ahalmena behaketa kopuruarekin handitzen da.

Bestalde, lortutako *p-balioa* erabilitako g talde kopuruaren araberakoa da. Honek ikertzaileak komenigarriagoa duen talde kopurua aukeratzea eragin dezake, ondorioak alboratuz.

Beraz, subjektibitatea eta gehiegizko ahalmen estatistikoa saihesteko, behaketa kopurua 1000 eta 25000 artean badago, g talde kopurua

$$g = \max\left(10, \min\left\{\frac{n_1}{2}, \frac{n-n_1}{2}, 2+8\left(\frac{n}{1000}\right)^2\right\}\right) \tag{2.9}$$

ekuazioaren arabera aukeratzea proposatu dute Paul et al. (2013) ikertzaileek [1]. Behaketa kopurua 25000 baino altuagoa bada, test honen erabilpena ez da gomendagarria eta kopurua 1000 baino baxuagoa bada, $g = 10$ erabiltzea gomendatzen da. Gainera, talde bakoitzean behintzat 5 banako egotea aholkatzen da.

2.2.1 Hosmer-Lemeshow testaren eraldaketa

HL pertzentilen testaren erabilera ez da gomendatzen 25000 behaketa baino gehiago ditugunean. Arazo hori gainditzeko, HL test eraldatua proposatu dute Nattino et al. (2020) ikertzaileek [2].

Lehenago aipatu denez, ereduaren doikuntza egokia denean, (2.7) ekuazioan definitu dugun $\hat{C}$ estatistikoaren banaketa asintotikoa χ^2_{g-2} da. Aldiz,

	Doikuntza egokia	Doikuntza ezegokia
H_0 ez errefusatu	Erabaki zuzena	I motako errorea
H_0 errefusatu	II motako errorea	Erabaki zuzena

2.2. Taula: Hipotesi-kontrasteetan har daitezkeen erabaki ezberdinak.

ereduaren doikuntza ezegokia denean, $\hat{C}$ estatistikoaren banaketa asintotikoa $\chi^2_{g-2,\lambda}$ da: $g-2$ askatasun-graduko eta $\lambda \geq 0$ ez-zentralizazio parametrodun χ^2 ez-zentratua [8]. 2.1. irudian, λ-ren balio ezberdinetarako eta $g = 10$ ezarrita, χ^2_{g-2} eta $\chi^2_{g-2,\lambda}$ banaketen dentsitate funtzioak ikus daitezke.

Zenbat eta handiagoa izan behatutako eta itxarotako datuen arteko desberdintasuna, orduan eta handiagoa da λ. Halaber, ereduaren doikuntza egokia denean, $\lambda = 0$ eta $\chi^2_{g-2,\lambda} = \chi^2_{g-2}$. Hau da, hain zuzen ere, test eraldatuaren oinarria.

λ ez-zentralizazio parametroa lagin tamainaren menpekoa da; eredu bat populazio batean doitzen bada, laginaren tamaina handitzean, λ linealki hazten da lagin tamainaren arabera. Hori dela eta, λ ezin da zuzenean erabili ereduaren doikuntza-egokitasuna neurtzeko.

Defini dezagun

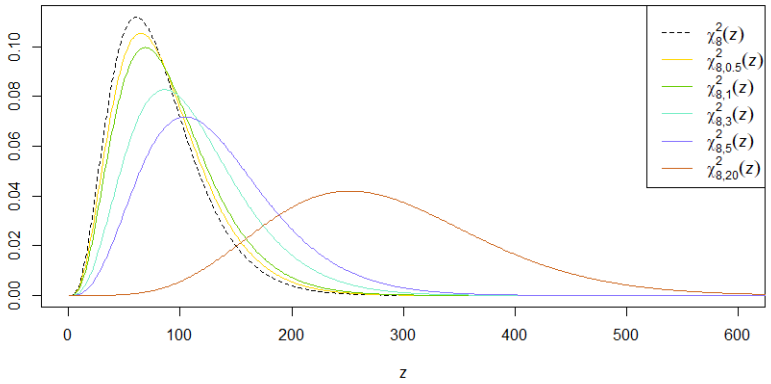
$$\epsilon = \sqrt{\frac{\lambda}{n}} \tag{2.10}$$

ez-zentralizazio parametro estandarizatua. Eredu jakin baterako ϵ asintotikoki konstantea da, ez da aldatzen lagin tamainarekin. Ondorioz, ϵ ereduaren doikuntza-egokitasuna neurtzeko erabil daiteke.

λ estimatzeko $\hat{\lambda} = \max\{0, \hat{C} - (g-2)\}$ [9] erabilita, ϵ-en estimatzailea

$$\hat{\epsilon} = \sqrt{\frac{\max\{0, \hat{C} - (g-2)\}}{n}} \tag{2.11}$$

dela lortzen dugu.



2.1. Irudia: χ^2_{g-2} eta $\chi^2_{g-2,\lambda}$ banaketen dentsitate funtzioak $g = 10$ eta $\lambda = 0.5,\ 1,\ 3,\ 5,\ 20$ direnean.

Orduan,

$$\begin{cases} H_0 : \epsilon \leq \epsilon_0 \\ H_1 : \epsilon > \epsilon_0 \end{cases}$$

hipotesi-kontrastea proposatzen da non ϵ_0 tolerantzia den. ϵ_0 aldez aurretik finkatu behar da eta txikia izan behar da, eredu onargarri batekin bat etor dadin.

$$\epsilon_0 = \sqrt{\frac{\chi^2_{g-2,\alpha} - (g-2)}{n_0}} \tag{2.12}$$

erabiltzea gomendatzen da, non $\chi^2_{g-2,\alpha}$ χ^2_{g-2} banaketaren $\%(1-\alpha)100$-kuantila eta n_0 erreferentziako lagin tamaina diren, $n_0 = 10^6$ erabiltzea gomendatzen da [2]. Beraz, ϵ_0 ϵ-en itxarotako balioa da HL testean $p = \alpha$ *p-balioa* lortu duen eredu batentzat, lagin tamaina milioi bat izanik. Ondorioz, test eraldatuak gure ereduaren doikuntza "mugako" ereduaren doikuntzarekin alderatzen du.

Bestalde, testaren *p-balioa*

$$p = \int_{\hat{C}}^{\infty} \chi^2_{g-2,\epsilon_0^2 n}(z)dz \tag{2.13}$$

da, non $\chi^2_{g-2,\epsilon_0^2 n}(z)$ $g-2$ askatasun-graduko eta $\lambda = \epsilon_0^2 n$ ez-zentralizazio parametroko χ^2 ez-zentratuaren dentsitate funtzioa den z-n ebaluatuta.

Beraz, $p > \alpha$ bada, gure ereduaren doikuntza onargarria dela ez dugu errefusatuko. Bestelakoan, doikuntza onargarria ez dela esango dugu.

Ereduaren doikuntza perfektua den ala ez egiaztatzea test familia honen muturreko kasua da ($\epsilon_0 = 0$) eta HL pertzentilen testaren baliokidea da. Izan ere,

$$\begin{cases} H_0 : \epsilon \leq 0 \\ H_1 : \epsilon > 0 \end{cases} \quad \text{eta} \quad \begin{cases} H_0 : \epsilon = 0 \\ H_1 : \epsilon > 0 \end{cases}$$

hipotesi-kontrasteak baliokideak dira $\epsilon \geq 0$ delako.

Test eraldatuaren *p-balioa*k HL pertzentilen testaren *p-balioa*k baino handiagoak dira eta hau emaitza desiratua da. Izan ere, test berri honen helburua doikuntza onargarria baina ez perfektua duten ereduen errefusatzea murriztea da.

Bi testak egoera berdinetan erabil daitezke eta, gainera, eraldatutako testa pertzentilen testaren ordez erabiltzea proposatzen da. Lagin tamaina txiki edo ertainetan, emaitza berdinak lortzen dira bi testak erabilita, lagin tamaina handietan emaitza nabarmenki ezberdinak lortzen diren bitartean [2].

HL testaren aldaketa gehiago proposatu dira (ikusi [3], [4]), baina lan honetan ez ditugu azalduko. Izan ere, azaldutako test eraldatua hipotesi-kontraste gisa planteatzen da, HL pertzentilen testa bezala.

3. Kapitulua

Simulazioak

3.1 Sarrera

Lan honen helburua, egoera jakin batzuetan, HL testean erabilitako g talde kopuruaren arabera testaren egonkortasuna aztertzea da. Horretarako, egoera ezberdinak planteatuta, hainbat simulazio egin ditugu.

Simulazioetan bi egoera nagusi aztertu ditugu:

- Ereduan aldagai azaltzaile bakarra edo bi aldagai azaltzaile egotea.
- Eredu teorikoan aldagai jarraituen eta $logit(p)$-ren arteko erlazio teorikoa lineala edo ez-lineala izatea.

3.2 Eredu bakun teorikoa

Hasteko, aldagai jarraitu bat eraiki dugu: Z. Z-ren gaixo (arrakasta) eta osasuntsuen (porrota) aldagaiak sortu ditugu, Z_G eta Z_O izendatuko ditugunak, hurrenez hurren. Bi aldagaiek banaketa normala jarraituko dute: $Z_G \sim N(1.5, \sigma_G)$ eta $Z_O \sim N(0, \sigma_O)$. $\sigma_G = 1$ definitu dugu eta σ_O^2 bariantzarentzat bi kasu bereiztu ditugu: $\sigma_O^2 = 1$ ($\sigma_O = 1$) edo $\sigma_O^2 = 0.5$ ($\sigma_O = \sqrt{0.5}$) izatea.

Orduan,

$$logit[p(Z)] = \beta_0 + \beta_1 Z \tag{3.1}$$

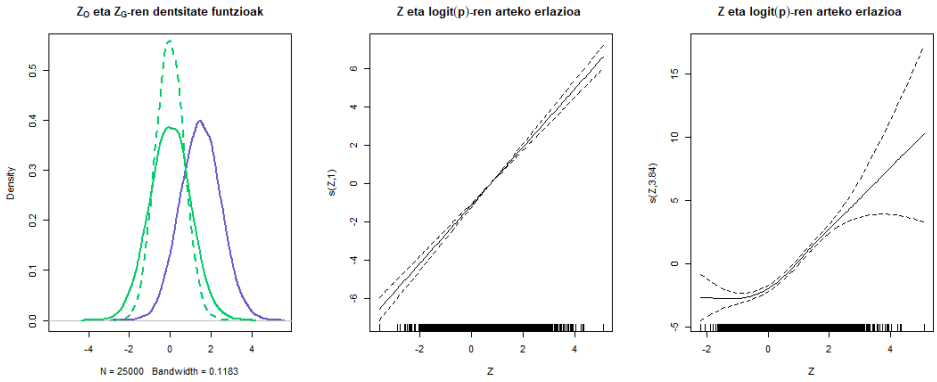
itxurako ereduak doitu ditugu.

Alde batetik, $\sigma_G = \sigma_O = 1$ direnean, Z eta $logit(p)$-ren arteko linealtasuna frogatuta dago [10].

Bestetik, $\sigma_G = 1$ eta $\sigma_O = \sqrt{0.5}$ direnean, Z eta $logit(p)$-ren arteko erlazioa ez-lineala dela frogatuta dago. Izan ere, Z aldagaiaren bariantza populazio osasuntsu eta gaixoan ezberdina da. Zehazki,

$$logit[p(Z)] \propto Z^2,\ Z$$

erlazioa betetzen da [10]. 3.1. irudian Z_O eta Z_G-ren dentsitate funtzioak ikus daitezke Z eta $logit(p)$-ren erlazioarekin batera.



(a) (b) (c)

3.1. Irudia: Z_G eta Z_O aldagaien dentsitate funtzioak Z eta $logit(p)$-ren arteko erlazioarekin batera $n = 2000$ eta prebalentzia 0.5 direnean. (a) irudian, Z_O-ren dentsitate funtzioa lerro berdea da: lerro jarraituak $\sigma_O = 1$ adierazten du eta etenak, $\sigma_O = \sqrt{0.5}$. Z_G-ren dentsitate funtzioa urdina da. (b) irudian, $\sigma_O = \sigma_G = 1$ dira eta (c) irudian, $\sigma_G = 1$ eta $\sigma_O = \sqrt{0.5}$.

3.3 Bi aldagai azaltzaile dituen eredu teorikoa

X_1 eta X_2 aldagai aske jarraituak kontsideratu ditugu bi aldagai azaltzaileko ereduak eraikitzeko. $\mathbf{X} = (X_1, X_2)^t$ definituz, $\mathbf{X}$-ren gaixo (arrakasta) eta osasuntsuen (porrota) bektoreak sortu ditugu: $\mathbf{X}_G$ eta $\mathbf{X}_O$. Bi bektoreak normalak izango dira: $\mathbf{X}_G \sim N(\boldsymbol{\mu}_G, \boldsymbol{\Sigma}_G)$ eta $\mathbf{X}_O \sim N(\boldsymbol{\mu}_O, \boldsymbol{\Sigma}_O)$, $\boldsymbol{\mu}_O = (0, 1)$ eta $\boldsymbol{\mu}_G = (1.5, 2)$ izanik. $\boldsymbol{\Sigma}_G = I_2$ eran definitu dugu eta $\boldsymbol{\Sigma}_O$ kobariantza matrizearentzat bi kasu bereiztu ditugu: $\boldsymbol{\Sigma}_O = I_2$ edo $\boldsymbol{\Sigma}_O = \begin{pmatrix} 0.5 & 0 \\ 0 & 1 \end{pmatrix}$ izatea.

Orduan, bi aldagaiak erabiliz

$$logit[p(\mathbf{X})] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \tag{3.2}$$

motako ereduak doitu ditugu.

Lehenengo kasuan, $\boldsymbol{\Sigma}_G = \boldsymbol{\Sigma}_O = I_2$ direnean, frogatuta dago X_1 eta X_2 aldagaien eta $logit(p)$-ren arteko erlazioa lineala dela [10]. 3.2. irudian X_{1G} eta X_{1O} aldagaien dentsitate funtzioak ikus daitezke X_1 eta X_2 aldagaien eta $logit(p)$-ren arteko erlazioarekin batera.

Bigarren kasuan, $\mathbf{\Sigma}_O = \begin{pmatrix} 0.5 & 0 \\ 0 & 1 \end{pmatrix}$ eta $\mathbf{\Sigma}_G = I_2$ direnean, X_2-ren eta $logit(p)$-ren arteko erlazioa lineala dela frogatuta dago, X_1-en eta $logit(p)$-ren arteko erlazioa ez-lineala den bitartean. Izan ere, X_1 aldagaiaren bariantza populazio osasuntsu eta gaixoan ezberdina da eta X_2-ren bariantza, ordea, berdina da bi populazioetan. Zehazki,

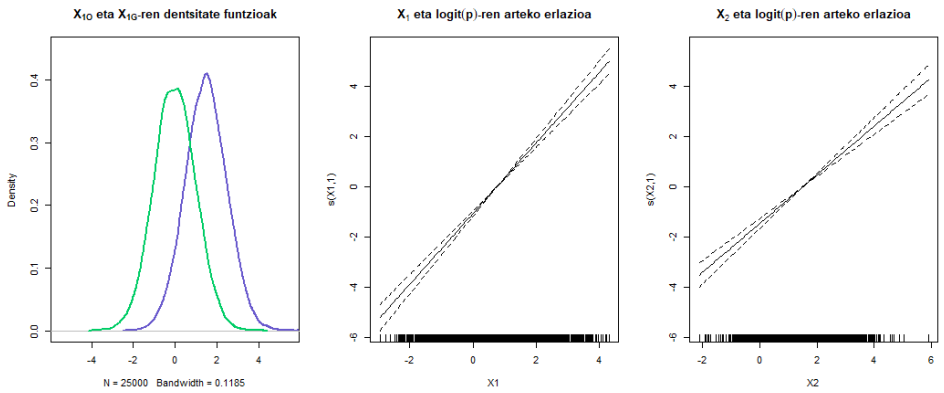
$$logit[p(\mathbf{X})] \propto X_1^2,\, X_1$$

betetzen da [10]. 3.3. irudian X_{1G} eta X_{1O} aldagaien dentsitate funtzioak ikus daitezke X_1 eta X_2 aldagaien eta $logit(p)$-ren arteko erlazioarekin batera.

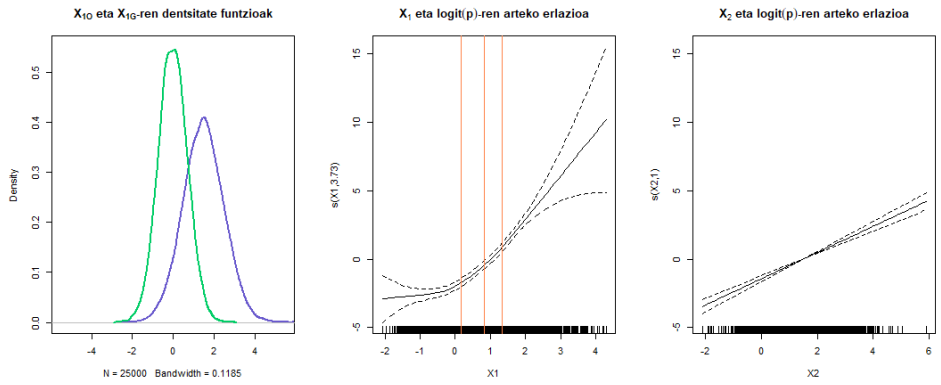
Hortaz, ez-linealtasunaren arazoa konpontzeko, X_1 2 eta 3 mozketa puntu erabilita kategorizatu dugu eta

$$logit[p(\mathbf{X})] = \beta_0 + \sum_{l=1}^{m-1} \beta_{1_l} D_{1_l} + \beta_2 X_2 \tag{3.3}$$

itxurako ereduak doitu ditugu non $m = 2, 3$ den.



3.2. Irudia: $\mathbf{\Sigma}_O = \mathbf{\Sigma}_G = I_2$ direnean, X_{1G} eta X_{1O} aldagaien dentsitate funtzioak eta X_1 eta X_2 aldagaien erlazioa $logit(p)$-rekiko $n = 2000$ eta prebalentzia 0.5 direnean. X_{1G}-ren dentsitate funtzioa lerro urdina da eta X_{1O}-rena, berdea.



(a) (b) (c)

3.3. Irudia: $\mathbf{\Sigma}_O = \begin{pmatrix} 0.5 & 0 \\ 0 & 1 \end{pmatrix}$ eta $\mathbf{\Sigma}_G = I_2$ direnean, X_{1G} eta X_{1O} aldagaien dentsitate funtzioak eta X_1 eta X_2 aldagaien erlazioa $logit(p)$-rekiko $n = 2000$ eta prebalentzia 0.5 direnean. (a) irudian, X_{1G}-ren dentsitate funtzioa urdina da eta X_{1O}-rena, berdea. (b) irudian, lerro bertikal gorriek X_1 kategorizatzeko erabili diren mozketa-puntuak adierazten dituzte.

Bestalde, HL testa $J = n$ kasurako garatuta dago eta aldagai bat kategorizatzean, J txikiagoa izango da. Simulazioetan horren eragina aztertu nahi izan da.

Laburbilduz, 4 eszenario ezberdin kontsideratu ditugu, 3.4 diagraman adierazita daudenak.

$$\begin{cases} \text{Eredu bakuna} \begin{cases} \sigma_G = \sigma_O \text{ Eszenario 1 (E1)} \\ \sigma_G \neq \sigma_O \text{ Eszenario 2 (E2)} \end{cases} \\ \text{Eredu anizkoitza} \begin{cases} \mathbf{\Sigma}_G = \mathbf{\Sigma}_O \text{ Eszenario 3 (E3)} \begin{cases} X_1 \text{ jarraitua (E3.1)} \\ X_1 \text{ kategorikoa (E3.2)} \end{cases} \\ \mathbf{\Sigma}_G \neq \mathbf{\Sigma}_O \text{ Eszenario 4 (E4)} \begin{cases} X_1 \text{ jarraitua (E4.1)} \\ X_1 \text{ kategorikoa (E4.2)} \end{cases} \end{cases} \end{cases} \tag{3.4}$$

Lau egoeretan 0.5 eta 0.9 prebalentzia duten 50000 behaketako laginak sortu ditugu. Lagin hauetatik 200, 500, 1200 eta 2000 tamainako azpilaginak sortu ditugu, kasu bakoitzean laginaren eta azpilaginaren prebalentzia

berdina mantenduz.

Sortutako azpilaginetan azaldutako ereduak doitu ditugu eta HL testa aplikatu dugu g talde kopurua aldatuz. $g = 5$tik $g = 10$ra aldatu dugu eta, talde kopuru gomendatua 10 baino altuagoa izanez gero, testa talde kopuru gomendatuarekin aplikatu dugu.

Kasu bakoitzean, prozesua 100 aldiz errepikatu dugu eta, adierazgarritasun-mailari $\alpha = 0.01, 0.05, 0.1$ balioak emanez, doikuntza egokiko hipotesia errefusatu ez deneko proportzioak kalkulatu ditugu. Ohartu E1, E3 eta E4.2 eszenarioetan H_0 egia dela eta, beraz, H_0 ez errefusatzeko proportzioa $1-\alpha$ (konfiantza-maila) inguru egotea espero dugula. Bestalde, E2 eta E4.1 eszenarioetan, H_1 egia da eta, ondorioz, H_0 ez errefusatzeko proportzioa II motako errorea egiteko probabilitatea da.

Aldagai jarraituak kategorizatzeko, `R` programako CatPredi paketea erabili dugu, ereduaren diskriminazio ahalmena maximizatuz.

Simulazio hauek *"Eredu aurresaleen balidazio tekniken konparaketa eta inplementazioa"* lanean proposatutakoak jarraituz egin dira [11].

3.4 Emaitzak

Hasteko, simulazioetan erabilitako parametro kopurua handia denez, emaitza nagusiak laburtuko ditugu. Aldagai azaltzaileen eta $logit(p)$-ren arteko erlazioa lineala denean, α finkaturik, $1-\alpha$ inguruko proportzioak lortu ditugu eta hau da, hain zuzen ere, espero genuena. Antzeko emaitzak lortu ditugu n lagin tamaina ezberdinetarako. Berriz, aldagai askeen erlazioa $logit(p)$-rekiko ez-lineala denean, batez ere tamaina txikiko laginetan, II motako errorearen probabilitate altuak (ahalmen txikia) lortu ditugu. Gainera, prebalentziaren arabera oso aldakorrak diren emaitzak lortu ditugu.

Orain, emaitzak taula eta grafikoen bitartez adieraziko ditugu. Emaitzak lehenago azaldutako eszenarioen arabera sailkatu ditugu. Ohartu ordenatu ardatzeko balioak grafiko multzoaren arabera aldatzen direla.

3.4.1 E1: eredu bakuna, $\sigma_G = \sigma_O$.

E1 eszenarioan, aldagai azaltzaile bakarra dugu ereduan eta $\sigma_G = \sigma_O$ betetzen da. Doitutako ereduaren itxura 3.1 ekuazioan ikus daiteke. Kasu honetan, frogatuta dago aldagai azaltzailearen eta $logit(p)$-ren arteko erlazioa lineala dela.

3.1 eredua doitzerakoan, HL testean doikuntza egokiko hipotesia errefusatu ez deneko proportzioak 3.1. taulan adierazi ditugu, α adierazgarritasun-mailari $\alpha = 0.01, 0.05, 0.1$ balioak emanez. Gainera, 3.4. irudian lortutako

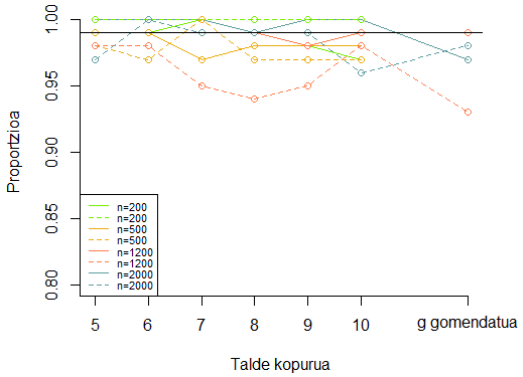
proportzioak grafikoki adierazi ditugu.

Aldagai azaltzailearen eta $logit(p)$-ren arteko erlazioa lineala denez, ereduaren doikuntza ona da eta proportzioek testaren konfiantza-maila adierazten dute. Gainera, α adierazgarritasun-maila finko baterako, espero duguna kalkulatutako proportzioak $1-\alpha$ izatea da.

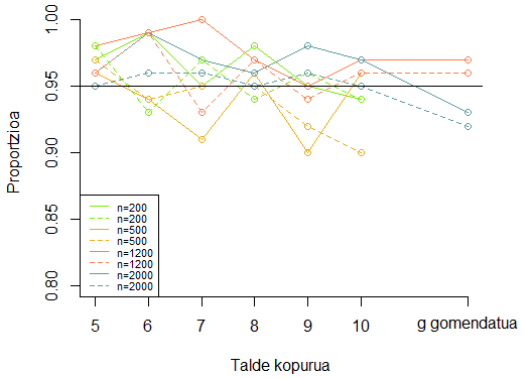
Orokorrean, espero duguna betetzen da; $\alpha = 0.01, 0.05, 0.1$ denean, $1-\alpha$ inguruko proportzioak lortu ditugu. Honetaz gain, ez dago proportzioen arteko desberdintasun nabarmenik lagin tamainari eta talde kopuruari erreparatzen badiogu. Ordea, aipatu beharra dago egoera batzuetan talde kopuru gomendatua ez den talde kopuru batekin $1-\alpha$ baliora gehiago hurbiltzen diren proportzioak lortu ditugula. Adibidez, gertaera hau $\alpha = 0.01$, $n = 1200$ eta prebalentzia 0.9 direneko kasuan ikus daiteke: $g = 10$ erabilita 0.98 lortu dugu eta $g = 14$ (talde kopuru gomendatua) erabilita, 0.93. Bestalde, kasu batzuetan, proportzioetan prebalentziaren araberako $1-\alpha$ balioarekiko aldaketak eman dira. Adibidez, $\alpha = 0.05$, $n = 1200$ eta prebalentzia 0.5 direnean, $1-\alpha$ baino proportzio altuagoak edo berdinak lortu ditugu. Aldiz, prebalentzia 0.9 denean, $1-\alpha$ baino proportzio altuagoak zein baxuagoak lortu ditugu. Adierazgarritasun-maila $\alpha = 0.1$ eta $n = 1200, 2000$ denean berdina lortu dugu. Gainera, bi kasu hauetan, osasuntsuen proportzioa 0.9 denean, oso proportzio baxuak lortu ditugu $1-\alpha$-rekin alderatuta. Bestalde, α handitzerakoan, proportzioak txikitzen dira eta hau espero daitekeena da proportzioek konfiantza-maila islatzen dutelako. Honetaz gain, α-ren balioa handitzean, proportzioen sakabanapena handitzen da.

α	Lagin tamaina	Prebalentzia	Talde kopurua (g)							
			5	**6**	**7**	**8**	**9**	**10**	**14**	**34**
0.01	200	0.5	0.99	0.99	1	0.99	0.98	0.97		
	500		0.99	0.99	0.97	0.98	0.98	0.98		
	1200		1	1	1	0.99	0.98	0.99	0.99	
	2000		1	1	1	0.99	1	1		0.97
	200	0.9	1	1	1	1	1	1		
	500		0.98	0.97	1	0.97	0.97	0.97		
	1200		0.98	0.98	0.95	0.94	0.95	0.98	0.93	
	2000		0.97	1	0.99	0.99	0.99	0.96		0.98
0.05	200	0.5	0.97	0.99	0.95	0.98	0.95	0.94		
	500		0.96	0.94	0.91	0.96	0.90	0.96		
	1200		0.98	0.99	1	0.97	0.95	0.97	0.97	
	2000		0.96	0.99	0.97	0.96	0.98	0.97		0.93
	200	0.9	0.98	0.93	0.97	0.94	0.96	0.94		
	500		0.97	0.94	0.95	0.95	0.92	0.90		
	1200		0.96	0.99	0.93	0.97	0.94	0.96	0.96	
	2000		0.95	0.96	0.96	0.95	0.96	0.95		0.92
0.1	200	0.5	0.88	0.95	0.90	0.94	0.92	0.89		
	500		0.93	0.91	0.87	0.91	0.85	0.91		
	1200		0.92	0.97	0.95	0.95	0.91	0.94	0.94	
	2000		0.92	0.91	0.96	0.93	0.96	0.92		0.90
	200	0.9	0.94	0.93	0.91	0.93	0.92	0.92		
	500		0.88	0.92	0.90	0.89	0.93	0.90		
	1200		0.88	0.97	0.93	0.93	0.80	0.89	0.90	
	2000		0.92	0.94	0.89	0.86	0.86	0.86		0.87

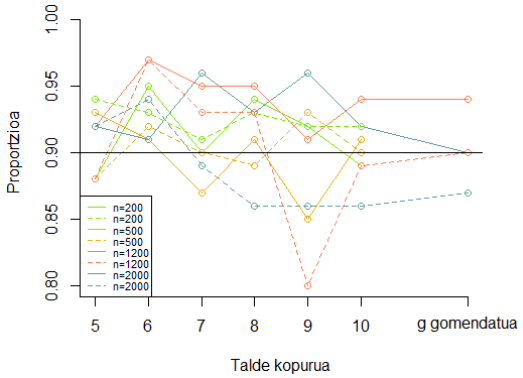
3.1. Taula: E1 eszenarioan, HL testak doikuntza egokiko hipotesia errefusatu ez dueneko proportzioa (konfiantza-maila). Urdinez irudikatu ditugu talde kopuru gomendatuari dagozkion proportzioak.



(a) $\alpha = 0.01$.



(b) $\alpha = 0.05$.



(c) $\alpha = 0.1$.

3.4. Irudia: E1 eszenarioan, HL testak doikuntza egokiko hipotesia errefusatu ez dueneko proportzioa (konfiantza-maila). Lerro jarraituak prebalentzia 0.5 adierazten du eta etenak, 0.9.

3.4.2 E2: eredu bakuna, $\sigma_G \neq \sigma_O$.

E2 eszenarioan, aldagai azaltzaile bakarra dugu ereduan eta $\sigma_G \neq \sigma_O$ dira. Doitutako ereduaren itxura 3.1 ekuazioan ikus daiteke. Egoera honetan, aldagai askearen eta $logit(p)$-ren arteko erlazioa ez-lineala dela frogatuta dago.

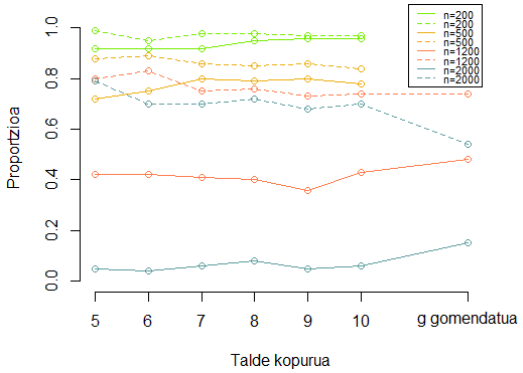
3.1 eredua doitu ostean, HL testak doikuntza egokiko hipotesia errefusatu ez dueneko proportzioak 3.2. taulan idatzi ditugu, $\alpha = 0.01, 0.05, 0.1$ izanda. 3.5. irudian emaitzak grafikoki adierazi ditugu.

Aldagai askearen eta $logit(p)$-ren arteko erlazioa ez-lineala denez, proportzioek II motako errorearen probabilitatea adierazten dute. Hortaz, espero duguna proportzio baxuak lortzea da.

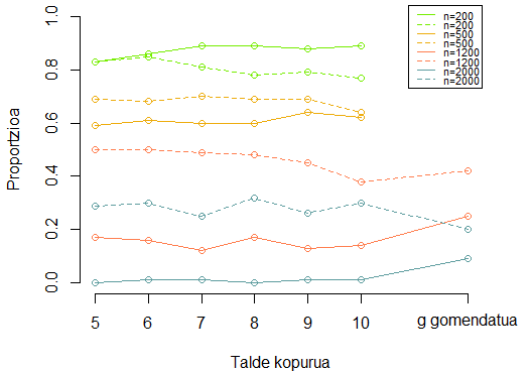
Espero genuena ez da betetzen: HL testak, orokorrean, doikuntza egokia ez du errefusatzen. Hau da, testaren ahalmena oso txikia da. Berezi-ki, $n = 200, 500$ denean, oso proportzio altuak lortzen ditugu. Adibidez, $\alpha = 0.01$, $n = 200$, prebalentzia 0.5 eta $g = 10$ (talde kopuru gomendatua) direnean, lortutako proportzioa 0.96 da. Egoera honetan, HL testak erabaki zuzena 4 aldiz hartu du 100tik. Gainera, proportzioak n lagin tamainarekin txikitzen dira. Honek zentzua du testaren ahalmena laginaren tamainarekin handitzen delako. Bestalde, kasu batzuetan talde kopuru gomendatua erabilita, proportzioak handitzen dira; $\alpha = 0.05$, $n = 1200$ eta prebalentzia 0.5 direnean, $g = 10$ aukeratuta 0.14 proportzioa lortu dugu $g = 14$-rekin (talde kopuru gomendatuarekin) 0.25 lortu dugun bitartean. Honetaz gain, $\alpha = 0.01$ izanda, prebalentzia 0.9 denean proportzio altuagoak lortzen ditugu prebalentzia 0.5 denean baino eta, lagin tamaina handitzean, hauen arteko diferentzia nabarmenki hazten da. $\alpha = 0.05$ eta $n = 200$ direnean, prebalentzia 0.5 denean prebalentzia handiagoak dira prebalentzia 0.9 direnean baino. Aldiz, $\alpha = 0.05$ eta $n = 500, 1200, 2000$ direnean, prebalentzia 0.9 denean proportzioak altuagoak dira prebalentzia 0.5 denean lortutako proportzioekin alderatuta. Gainera, n handitzean, haien arteko aldea asko handitzen da. Bestalde, α-ren balioak handitzean, proportzio baxuagoak lortu ditugu.

α	Lagin tamaina	Prebalentzia	Talde kopurua (g)							
			5	**6**	**7**	**8**	**9**	**10**	**14**	**34**
0.01	200	0.5	0.92	0.92	0.92	0.95	0.96	0.96		
	500		0.72	0.75	0.80	0.79	0.80	0.78		
	1200		0.42	0.42	0.41	0.40	0.36	0.43	0.48	
	2000		0.05	0.04	0.06	0.08	0.05	0.06		0.15
	200	0.9	0.99	0.95	0.98	0.98	0.97	0.97		
	500		0.88	0.89	0.86	0.85	0.86	0.84		
	1200		0.80	0.83	0.75	0.76	0.73	0.74	0.74	
	2000		0.79	0.70	0.70	0.72	0.68	0.70		0.54
0.05	200	0.5	0.83	0.86	0.89	0.89	0.88	0.89		
	500		0.59	0.61	0.60	0.60	0.64	0.62		
	1200		0.17	0.16	0.12	0.17	0.13	0.14	0.25	
	2000		0	0.01	0.01	0	0.01	0.01		0.09
	200	0.9	0.83	0.85	0.81	0.78	0.79	0.77		
	500		0.69	0.68	0.70	0.69	0.69	0.64		
	1200		0.50	0.50	0.49	0.48	0.45	0.38	0.42	
	2000		0.29	0.30	0.25	0.32	0.26	0.30		0.20
0.1	200	0.5	0.74	0.80	0.85	0.88	0.83	0.86		
	500		0.44	0.47	0.48	0.50	0.57	0.54		
	1200		0.06	0.13	0.07	0.07	0.09	0.10	0.12	
	2000		0	0	0	0	0.01	0		0.03
	200	0.9	0.69	0.70	0.62	0.66	0.63	0.63		
	500		0.60	0.56	0.51	0.49	0.51	0.47		
	1200		0.35	0.33	0.29	0.34	0.32	0.30	0.24	
	2000		0.18	0.18	0.16	0.12	0.14	0.09		0.09

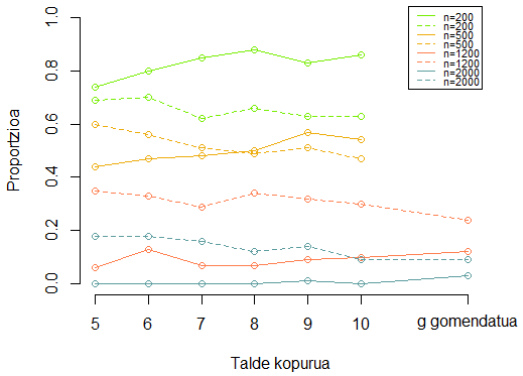
3.2. Taula: E2 eszenarioan, HL testak doikuntza egokiko hipotesia errefusatu ez dueneko proportzioa (II motako errorearen probabilitatea). Urdinez irudikatu ditugu talde kopuru gomendatuari dagozkion proportzioak.



(a) $\alpha = 0.01$.



(b) $\alpha = 0.05$.



(c) $\alpha = 0.1$.

3.5. Irudia: E2 eszenarioan, HL testak doikuntza egokiko hipotesia errefusatu ez dueneko proportzioa (II motako errorearen probabilitatea). Lerro jarraituak prebalentzia 0.5 adierazten du eta etenak, 0.9.

3.4.3 E3: eredu anizkoitza, $\mathbf{\Sigma}_O = \mathbf{\Sigma}_G$.

E3 eszenarioan, bi aldagai azaltzaile ditugu ereduan eta doikuntza egokia betetzen da. Doitutako ereduen itxura (3.2) eta (3.3) ekuazioetan ikus daitezke. Egoera honetan, bi aldagaiak jarraituak izanik, frogatuta dago aldagai azaltzaileen eta $logit(p)$-ren arteko erlazioa lineala dela.

X_1 jarraitua

3.2 eredua doitzerakoan, HL testak doikuntza egokiko hipotesia errefusatu ez dueneko proportzioak 3.3. taulan adierazi ditugu, α adierazgarritasun-maila $\alpha = 0.01, 0.05, 0.1$ ezarriz. 3.6. irudian lortutako emaitzak grafikoki adierazi ditugu.

Aldagai azaltzaileen eta $logit(p)$-ren arteko erlazioa lineala denez, proportzioek testaren konfiantza-maila islatuko dute eta, hortaz, espero duguna kalkulatutako proportzioak $1 - \alpha$ izatea da, α adierazgarritasun-maila izanda.

Gehienetan, espero duguna betetzen da. Kasu batzuetan, prebalentziaren arabera $1 - \alpha$ balioarekiko aldaketak ikusi ditugu. Adibidez, $\alpha = 0.05$ eta $n = 500$ direnean, $1 - \alpha$ baino proportzio baxuagoak lortu ditugu prebalentzia 0.9 denean eta, prebalentzia 0.5 denean, $1 - \alpha$ baino proportzio altuago zein baxuagoak lortu ditugu. $\alpha = 0.01$ eta $n = 2000$ direnean, prebalentzia 0.9 denean, $1-\alpha$ baino proportzio altuagoak lortu ditugu eta prebalentzia 0.5 denean, altuagoak zein baxuagoak. Honetaz gain, α-ren balioak handitzean, proportzioak txikitzen dira eta espero genuena da proportzioek konfiantza-maila adierazten dutelako. Gainera, α-ren balioa handitzean, proportzioen sakabanapena handitzen da.

X_1 kategorikoa

Bi aldagai azaltzaileak jarraituak direnean, aldagaien eta $logit(p)$-ren arteko erlazioa lineala da. Ondorioz, aldagai bat kategorizatzerakoan, ereduaren doikuntza ona izaten jarraitzea espero dugu eta, beraz, proportzioek testaren konfiantza-maila adierazten dute. Espero duguna proportzioak $1 - \alpha$ balioaren ingurukoak izatea da.

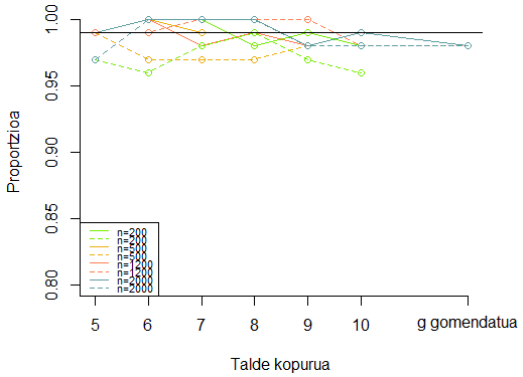
3.3 eredua doitzerakoan, HL testean doikuntza egokiko hipotesia errefusatu ez deneko proportzioak 3.4. taulan adierazi ditugu, $\alpha = 0.01, 0.05, 0.1$ izanik. 3.7. irudian lortutako emaitzak grafikoki adierazi ditugu.

Orokorrean, espero duguna betetzen da: $1 - \alpha$ inguruko balioak lortu ditugu $\alpha = 0.01, 0.05, 0.1$ denean. Gainera, $n = 200, 500, 1200, 2000$ denean, α-ren balioa finkatuta, 3 mozketa puntu erabilita proportzio altuagoak

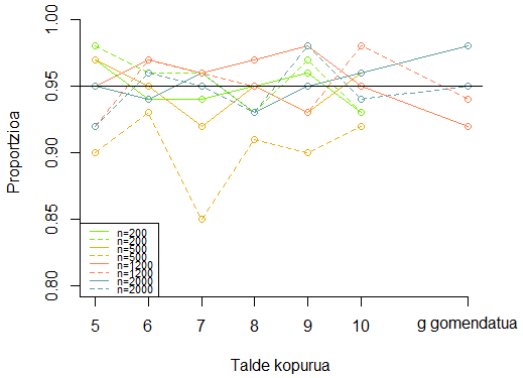
lortu ditugu 2 mozketa punturekin baino, prebalentzia 0.5 zein 0.9 izanik. Bestalde, α handitzean, proportzio sakabanatuagoak lortu ditugu.

α	Lagin tamaina	Prebalentzia	Talde kopurua (g)							
			5	**6**	**7**	**8**	**9**	**10**	**14**	**34**
0.01	200	0.5	0.99	1	1	0.98	0.99	0.98		
	500		0.99	1	0.99	0.99	0.98	0.99		
	1200		0.99	1	0.98	0.99	0.98	0.99	0.98	
	2000		0.99	1	1	1	0.98	0.99		0.98
	200	0.9	0.97	0.96	0.98	0.99	0.97	0.96		
	500		0.99	0.97	0.97	0.97	0.98	0.98		
	1200		0.99	0.99	1	1	1	0.98	0.98	
	2000		0.97	1	1	1	0.98	0.98		0.98
0.05	200	0.5	0.97	0.94	0.94	0.95	0.96	0.93		
	500		0.97	0.95	0.92	0.95	0.93	0.96		
	1200		0.95	0.97	0.96	0.97	0.98	0.95	0.92	
	2000		0.95	0.94	0.96	0.93	0.95	0.96		0.98
	200	0.9	0.98	0.96	0.96	0.93	0.97	0.93		
	500		0.90	0.93	0.85	0.91	0.90	0.92		
	1200		0.92	0.97	0.96	0.94	0.95	0.91	0.92	
	2000		0.92	0.96	0.95	0.93	0.98	0.94		0.95
0.1	200	0.5	0.87	0.90	0.87	0.92	0.90	0.87		
	500		0.89	0.91	0.87	0.90	0.92	0.94		
	1200		0.88	0.93	0.93	0.89	0.94	0.94	0.90	
	2000		0.87	0.89	0.90	0.91	0.91	0.93		0.93
	200	0.9	0.88	0.88	0.90	0.90	0.90	0.89		
	500		0.92	0.88	0.90	0.95	0.91	0.90		
	1200		0.93	0.96	0.87	0.94	0.92	0.93	0.92	
	2000		0.90	0.93	0.92	0.92	0.91	0.91		0.91

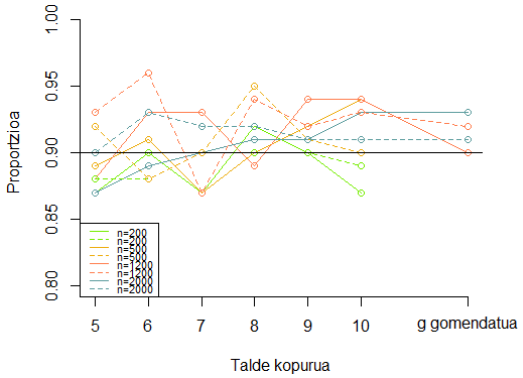
3.3. Taula: E3.1 eszenarioan, HL testak doikuntza egokiko hipotesia errefusatu ez dueneko proportzioa (konfiantza-maila). Urdinez irudikatu ditugu talde kopuru gomendatuari dagozkion proportzioak.



(a) $\alpha = 0.01$.



(b) $\alpha = 0.05$.

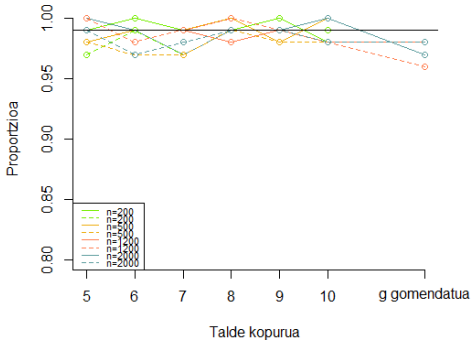


(c) $\alpha = 0.1$.

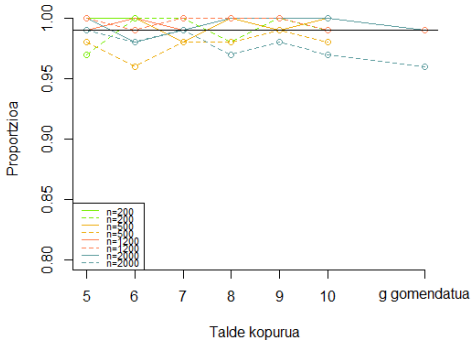
3.6. Irudia: E3.1 eszenarioan, HL testak doikuntza egokiko hipotesia errefusatu ez dueneko proportzioa (konfiantza-maila). Lerro jarraituak prebalentzia 0.5 adierazten du eta etenak, 0.9.

α	Lagin tamaina	Prebalentzia	Mozketa puntuak	Talde kopurua (g)							
				5	**6**	**7**	**8**	**9**	**10**	**14**	**34**
0.01	200	0.5	2	0.99	1	0.99	0.99	1	0.98		
	500			0.98	0.99	0.99	1	0.98	1		
	1200			1	0.99	0.99	0.98	0.99	0.98	0.98	
	2000			1	0.99	0.97	0.99	0.99	1		0.97
	200		3	1	1	0.99	1	1	1		
	500			0.99	1	0.98	1	0.99	1		
	1200			0.99	1	0.99	1	1	1	0.99	
	2000			1	0.98	0.99	1	1	1		0.99
	200	0.9	2	0.97	0.99	0.97	0.99	0.99	0.99		
	500			0.98	0.97	0.97	0.99	0.98	0.98		
	1200			1	0.98	0.99	1	0.99	0.98	0.96	
	2000			0.99	0.97	0.98	0.99	0.99	0.98		0.98
	200		3	0.97	1	1	0.98	1	0.99		
	500			0.98	0.96	0.98	0.98	0.99	0.98		
	1200			1	0.99	1	1	1	0.99	0.99	
	2000			0.99	0.98	0.99	0.97	0.98	0.97		0.96
0.05	200	0.5	2	0.95	0.95	0.96	0.92	0.95	0.92		
	500			0.96	0.95	0.97	0.98	0.94	0.94		
	1200			0.95	0.95	0.97	0.96	0.95	0.95	0.95	
	2000			0.97	0.97	0.96	0.92	0.97	0.97		0.90
	200		3	0.94	0.97	0.97	0.95	0.98	0.98		
	500			0.95	0.99	0.95	0.98	0.96	0.97		
	1200			0.95	0.96	0.99	0.96	0.99	0.99	0.97	
	2000			0.94	0.95	0.96	0.97	0.98	0.95		0.97
	200	0.9	2	0.94	0.98	0.97	0.97	0.97	0.94		
	500			0.93	0.92	0.97	0.98	0.95	0.93		
	1200			0.94	0.97	0.96	0.98	0.95	0.94	0.94	
	2000			0.94	0.95	0.95	0.96	0.97	0.95		0.94
	200		3	0.94	0.99	0.98	0.97	0.99	0.95		
	500			0.95	0.94	0.94	0.95	0.94	0.95		
	1200			0.97	0.98	0.99	0.98	0.99	0.97	0.97	
	2000			0.97	0.94	0.96	0.95	0.96	0.96		0.91
0.1	200	0.5	2	0.89	0.91	0.90	0.86	0.92	0.90		
	500			0.92	0.90	0.91	0.90	0.92	0.93		
	1200			0.89	0.87	0.90	0.93	0.87	0.89	0.92	
	2000			0.90	0.89	0.90	0.90	0.91	0.92		0.87
	200		3	0.91	0.92	0.92	0.92	0.95	0.95		
	500			0.93	0.94	0.91	0.95	0.94	0.92		
	1200			0.90	0.89	0.95	0.90	0.95	0.95	0.91	
	2000			0.92	0.92	0.96	0.91	0.95	0.93		0.91
	200	0.9	2	0.93	0.95	0.92	0.97	0.94	0.94		
	500			0.89	0.90	0.93	0.91	0.92	0.91		
	1200			0.90	0.94	0.90	0.91	0.92	0.89	0.92	
	2000			0.92	0.87	0.93	0.91	0.95	0.91		0.90
	200		3	0.93	0.97	0.98	0.96	0.98	0.95		
	500			0.90	0.90	0.92	0.92	0.92	0.88		
	1200			0.94	0.97	0.95	0.96	0.93	0.93	0.95	
	2000			0.94	0.90	0.93	0.89	0.93	0.93		0.87

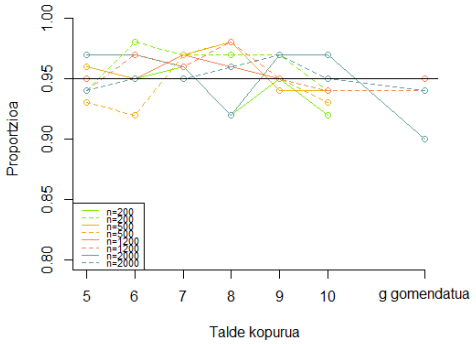
3.4. Taula: E3.2 eszenarioan, HL testak egokiko hipotesia errefusatu ez dueneko proportzioa (konfiantza-maila). Urdinez irudikatu ditugu talde kopuru gomendatuari dagozkion proportzioak.



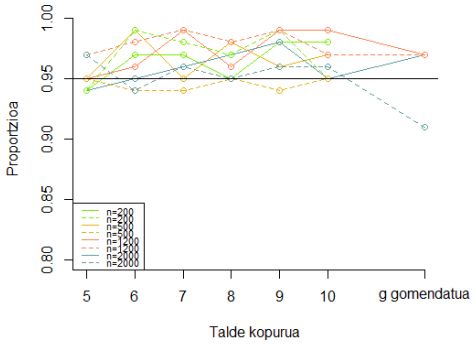
(a) $\alpha = 0.01$, 2 mozketa puntu.



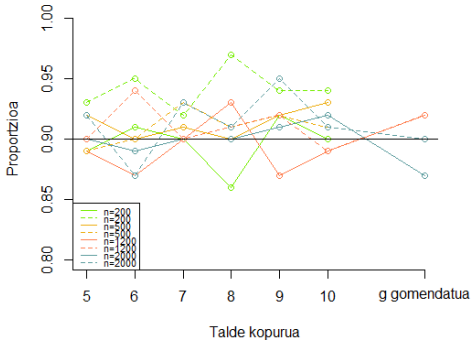
(b) $\alpha = 0.01$, 3 mozketa puntu.



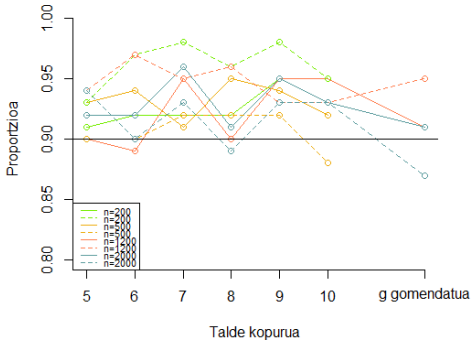
(c) $\alpha = 0.05$, 2 mozketa puntu.



(d) $\alpha = 0.05$, 3 mozketa puntu.



(e) $\alpha = 0.1$, 2 mozketa puntu.



(f) $\alpha = 0.1$, 3 mozketa puntu.

3.7. Irudia: E3.2 eszenarioan, HL testak doikuntza egokiko hipotesia errefusatu ez dueneko proportzioa (konfiantza-maila). Lerro jarraituak prebalentzia 0.5 adierazten du eta etenak, 0.9.

3.4.4 E4: eredu anizkoitza, $\mathbf{\Sigma}_O \neq \mathbf{\Sigma}_G$.

E4 eszenarioan, bi aldagai azaltzaile ditugu ereduan eta $\mathbf{\Sigma}_O \neq \mathbf{\Sigma}_G$ betetzen da. Doitutako ereduen itxura (3.2) eta (3.3) ekuazioetan ikus daitezke. 3.2 ereduan, X_1 aldagai azaltzailearen erlazioa $logit(p)$-rekiko ez-lineala da eta X_2-rena, lineala.

X_1 jarraitua

3.2 eredua doitu ostean, HL testean doikuntza egokiko hipotesia errefusatu ez deneko proportzioak 3.5. taulan adierazi ditugu. 3.8. irudian proportzioak grafikoki azaldu ditugu.

X_1 aldagai azaltzailearen erlazioa $logit(p)$-rekiko ez-lineala denez, doitutako eredua ez da zuzena eta, horrenbestez, proportzioek II motako errorearen probabilitatea islatzen dute. Beraz, proportzio baxuak lortzea espero dugu.

Bereziki $n = 200, 500$ denean, ordea, orokorrean lortutako probabilitateak altuak dira. Adibidez, $\alpha = 0.01$, $n = 200$, prebalentzia 0.5 eta $g = 10$ (talde kopuru gomendatua) denean, proportzioa 0.95 da. Beste era batera esanda, HL testak erabaki zuzena 5 aldiz hartu du 100tik. Honetaz gain, orokorrean, proportzioak n lagin tamaina handitzean, txikitu egiten dira. Gainera, $\alpha = 0.01, 0.05$ eta $n = 1200$ direnean, prebalentzia 0.5 baliotik 0.9 baliora handitzean, proportzio txikitu egin dira. $\alpha = 0.1$ eta $n = 1200$ denean, proportzioak handitu egiten dira prebalentzia txikitzean. $n = 2000$ denean, $\alpha = 0.01, 0.05$ izanik, 0.5 prebalentziari dagozkion proportzioak 0.9 prebalentziakoak baino baxuagoak dira, talde kopuru gomendatua erabiltzean izan ezik. Gainerako kasuetan, talde kopuru gomendatuaren erabilerak ez du eragin garrantzitsurik izan. $\alpha = 0.1$ denean, prebalentzia 0.9 denean, proportzio altuagoak lortzen ditugu prebalentzia 0.5 denean baino. Bestalde, $n = 200, 500$ denean, orokorrean, prebalentzia txikitzean, proportzio altuagoak lortu ditugu. Honetaz gain, α adierazgarritasun-maila handitzean, orokorrean, proportzio baxuagoak lortu ditugu.

X_1 kategorikoa

Lehenago ikusi denez, X_1 aldagai askearen erlazioa $logit(p)$-rekiko ez da lineala. Arazo hau gainditzeko, X_1 kategorizatu dugu eta, ondorioz, ereduaren doikuntza ona izango da. Hortaz, proportzioek testaren konfiantza-maila adierazten dute. Gainera, α adierazgarritasun-maila izanik, proportzioak $1 - \alpha$ ingurukoak izatea espero dugu.

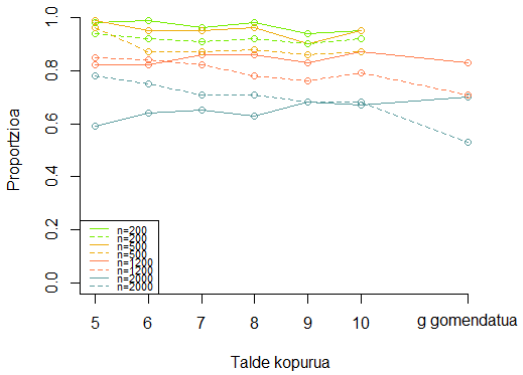
3.3 eredua doitzerakoan, HL testak doikuntza egokiko hipotesia errefusatu ez dueneko proportzioak 3.6. taulan adierazi ditugu. 3.9. irudian lortutako

emaitzak grafikoki azaldu ditugu.

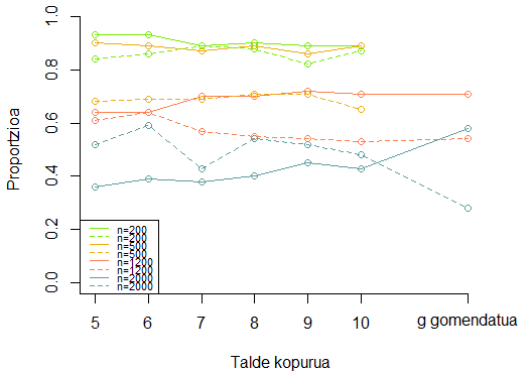
Orokorrean, espero duguna betetzen da: $1-\alpha$ inguruko proportzioak lortu ditugu $\alpha = 0.01, 0.05, 0.1$ denean. Ordea, lehenago aipatu den arazo berdina dugu: kasu batzuetan, talde kopuru gomendatua ez den talde kopuru batekin $1-\alpha$ baliora gehiago hurbiltzen diren proportzioak lortu ditugu. Honetaz gain, α-ren balioak handitzerakoan, proportzio baxuagoak lortzen ditugu eta haien sakabanapena handitzen da.

α	Lagin tamaina	Prebalentzia	Talde kopurua (g)							
			5	**6**	**7**	**8**	**9**	**10**	**14**	**34**
0.01	200	0.5	0.98	0.99	0.96	0.98	0.94	0.95		
	500		0.99	0.95	0.95	0.96	0.90	0.95		
	1200		0.82	0.82	0.86	0.86	0.83	0.87	0.83	
	2000		0.59	0.64	0.65	0.63	0.68	0.67		0.70
	200	0.9	0.94	0.92	0.91	0.92	0.90	0.92		
	500		0.96	0.87	0.87	0.88	0.86	0.87		
	1200		0.85	0.84	0.82	0.78	0.76	0.79	0.71	
	2000		0.78	0.75	0.71	0.71	0.68	0.68		0.53
0.05	200	0.5	0.93	0.93	0.89	0.90	0.89	0.89		
	500		0.90	0.89	0.87	0.89	0.86	0.89		
	1200		0.64	0.64	0.70	0.70	0.72	0.71	0.71	
	2000		0.36	0.39	0.38	0.40	0.45	0.43		0.58
	200	0.9	0.84	0.86	0.89	0.88	0.82	0.87		
	500		0.68	0.69	0.69	0.71	0.71	0.65		
	1200		0.61	0.64	0.57	0.55	0.54	0.53	0.54	
	2000		0.52	0.59	0.43	0.54	0.52	0.48		0.28
0.01	200	0.5	0.74	0.80	0.85	0.88	0.83	0.86		
	500		0.44	0.47	0.48	0.50	0.57	0.54		
	1200		0.06	0.13	0.07	0.07	0.09	0.10	0.12	
	2000		0	0	0	0	0.01	0		0.03
	200	0.9	0.69	0.70	0.62	0.66	0.63	0.63		
	500		0.60	0.56	0.51	0.49	0.51	0.47		
	1200		0.35	0.33	0.29	0.34	0.32	0.30	0.24	
	2000		0.18	0.18	0.16	0.12	0.14	0.09		0.09

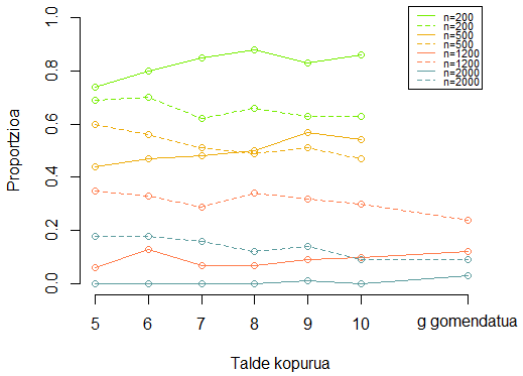
3.5. Taula: E4.1 eszenarioan, HL testak doikuntza egokiko hipotesia errefusatu ez dueneko proportzioa (II motako errorearen probabilitatea). Urdinez irudikatu ditugu talde kopuru gomendatuari dagozkion proportzioak.



(a) $\alpha = 0.01$.



(b) $\alpha = 0.05$.

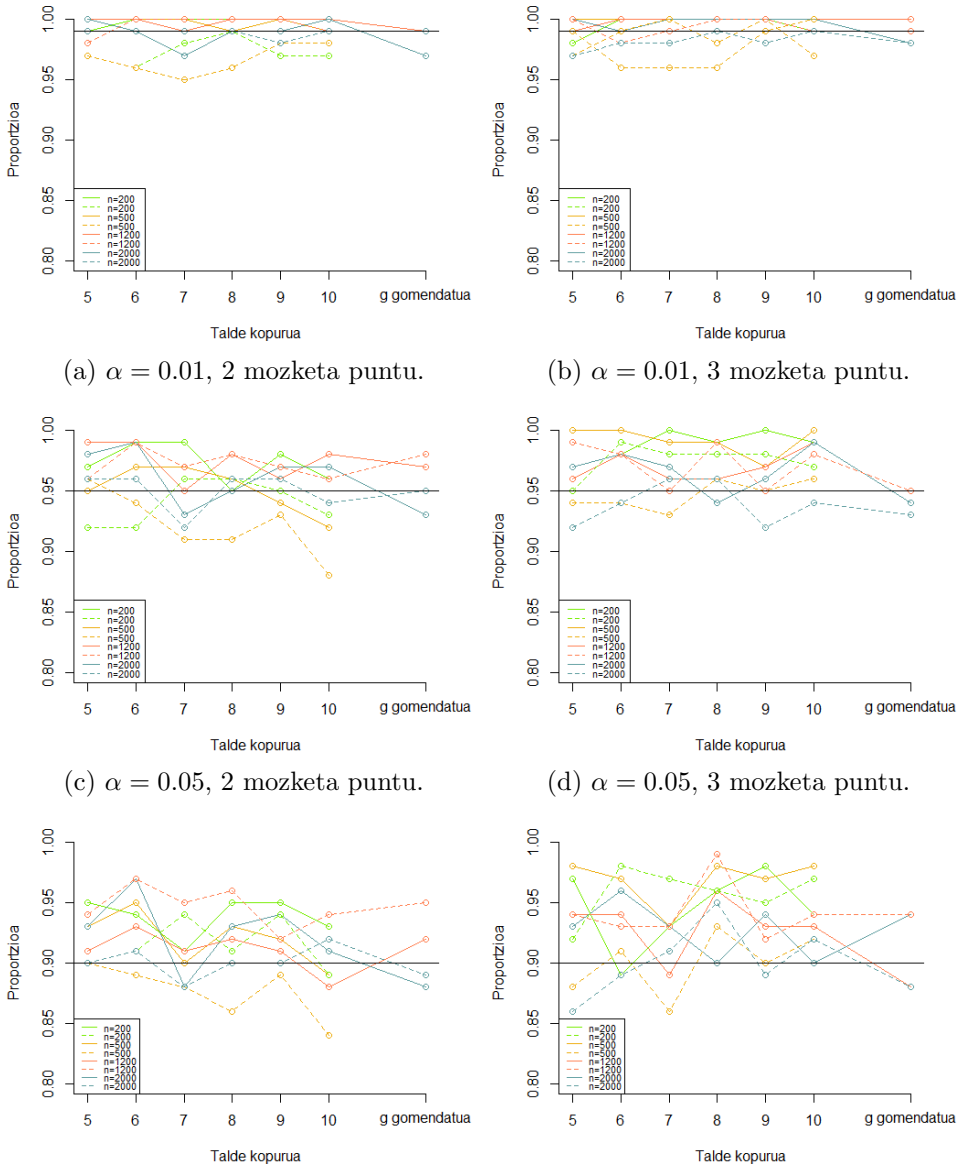


(c) $\alpha = 0.1$.

3.8. Irudia: E4.1 eszenarioan, HL testak doikuntza egokiko hipotesia errefusatu ez dueneko proportzioa (II motako errorearen probabilitatea). Lerro jarraituak prebalentzia 0.5 adierazten du eta etenak, 0.9.

α	Lagin tamaina	Prebalentzia	Mozketa puntuak	Talde kopurua (g)							
				5	**6**	**7**	**8**	**9**	**10**	**14**	**34**
0.01	200	0.5	2	0.99	1	1	1	1	1		
	500			1	1	1	0.99	1	0.99		
	1200			1	1	0.99	1	1	1	0.99	
	2000			1	0.99	0.97	0.99	0.99	1		0.97
	200		3	0.98	1	1	1	1	0.99		
	500			1	1	1	1	1	1		
	1200			0.99	1	1	1	1	1	1	
	2000			1	0.99	1	1	1	1		0.98
	200	0.9	2	0.97	0.96	0.98	0.99	0.97	0.97		
	500			0.97	0.96	0.95	0.96	0.98	0.98		
	1200			0.98	1	1	1	1	0.99	0.99	
	2000			0.99	0.99	0.99	0.99	0.98	0.99		0.99
	200		3	0.97	0.99	1	0.98	1	0.97		
	500			0.99	0.96	0.96	0.96	0.99	1		
	1200			1	0.98	0.99	1	1	0.99	0.99	
	2000			0.97	0.98	0.98	0.99	0.98	0.99		0.98
0.05	200	0.5	2	0.97	0.99	0.99	0.95	0.98	0.96		
	500			0.95	0.97	0.97	0.96	0.94	0.92		
	1200			0.99	0.99	0.95	0.98	0.96	0.98	0.97	
	2000			0.98	0.99	0.93	0.95	0.97	0.97		0.93
	200		3	0.97	0.98	1	0.99	1	0.99		
	500			1	1	0.99	0.99	0.97	1		
	1200			0.96	0.98	0.96	0.96	0.97	0.99	0.94	
	2000			0.97	0.98	0.97	0.94	0.96	0.99		0.94
	200	0.9	2	0.92	0.92	0.96	0.96	0.95	0.93		
	500			0.96	0.94	0.91	0.91	0.93	0.88		
	1200			0.96	0.99	0.97	0.98	0.97	0.96	0.98	
	2000			0.96	0.96	0.92	0.96	0.96	0.94		0.95
	200		3	0.95	0.99	0.98	0.98	0.98	0.97		
	500			0.94	0.94	0.93	0.96	0.95	0.96		
	1200			0.99	0.98	0.95	0.99	0.95	0.98	0.95	
	2000			0.92	0.94	0.96	0.96	0.92	0.94		0.93
0.1	200	0.5	2	0.95	0.94	0.91	0.95	0.95	0.93		
	500			0.93	0.95	0.90	0.93	0.92	0.89		
	1200			0.91	0.93	0.91	0.92	0.91	0.88	0.92	
	2000			0.93	0.97	0.88	0.93	0.94	0.91		0.88
	200		3	0.97	0.89	0.93	0.96	0.98	0.94		
	500			0.98	0.97	0.93	0.98	0.97	0.98		
	1200			0.94	0.94	0.89	0.96	0.93	0.93	0.88	
	2000			0.93	0.96	0.93	0.90	0.94	0.90		0.94
	200	0.9	2	0.90	0.91	0.94	0.91	0.94	0.89		
	500			0.90	0.89	0.88	0.86	0.89	0.84		
	1200			0.94	0.97	0.95	0.96	0.92	0.94	0.95	
	2000			0.90	0.91	0.88	0.90	0.90	0.92		0.89
	200		3	0.92	0.98	0.97	0.96	0.95	0.97		
	500			0.88	0.91	0.86	0.93	0.90	0.92		
	1200			0.94	0.93	0.93	0.99	0.92	0.94	0.94	
	2000			0.86	0.89	0.91	0.95	0.89	0.92		0.88

3.6. Taula: E4.2 eszenarioan, HL testak doikuntza egokiko hipotesia errefusatu ez dueneko proportzioa (konfiantza-maila). Urdinez irudikatu ditugu talde kopuru gomendatuari dagozkion proportzioak.



(a) $\alpha = 0.01$, 2 mozketa puntu.

(b) $\alpha = 0.01$, 3 mozketa puntu.

(c) $\alpha = 0.05$, 2 mozketa puntu.

(d) $\alpha = 0.05$, 3 mozketa puntu.

(e) $\alpha = 0.1$, 2 mozketa puntu.

(f) $\alpha = 0.1$, 3 mozketa puntu.

3.9. Irudia: E4.2 eszenarioan, HL testak doikuntza egokiko hipotesia errefusatu ez dueneko proportzioa (konfiantza-maila). Lerro jarraituak prebalentzia 0.5 adierazten du eta etenak, 0.9.

4. Kapitulua

Ondorioak

Lan honetan HL testaren erabakien aldaketa aztertu dugu g talde kopuruaren arabera. Simulazioetan lagin tamaina, prebalentzia, HL testean erabilitako talde kopurua, adierazgarritasun-maila, aldagai azaltzaileen kopurua, hauen bariantzak aldatu ditugu. Populazio osasuntsu eta gaixoan, aldagai azaltzaileek banaketa normalari darraie. Izan ere, kasu honetan, hauen eta $logit(p)$-ren arteko erlazio teorikoa ezaguna da. Orain, lortutako ondorio nagusiak laburbilduko ditugu.

Alde batetik, ereduaren doikuntza egokia denean, HL testaren errendimendua ona da eta ez da lagin tamainaren araberakoa. Bestalde, prebalentziak emaitzetan eragina izan du, baina proportzioek ez dute joera zehatzik jarraitu. Aldiz, ez dago desberdintasun nabarmenik talde kopuru gomendatuari dagokionez. Gainera, emaitzen aldakortasuna testaren adierazgarritasun-mailarekin handitu da.

Bestetik, ereduaren doikuntza ezegokia denean, HL testa lagin tamainarekiko sentikorra da eta bere errendimendua lagin txikietan eskasa da. Izan ere, testaren ahalmena lagin txikietan oso baxua dela ikusi dugu. Kasu honetan, lagin handietan talde kopuru gomendatuaren erabilerak eragina du. Gainera, lagin tamaina finkatuta, prebalentziaren arabera desberdintasun nabariak gertatzen dira, baina ez dugu joera zehatzik behatu. Ereduan aldagai azaltzaile bakarra erabilita, orekatu gabeko datuetan emaitza okerragoak lortu ditugu orekatuetan baino. Ordea, bi aldagai azaltzaile erabilita, datu ez-orekatuetan testaren errendimendua hobea izan da. Beraz, interesgarria izango litzateke etorkizunean testaren erabakien aldaketa aztertzea prebalentziaren arabera.

Laburbilduz, hasiera batean, talde kopuruak eragindako aldaketak aztertu nahi genituen. Ordea, guk proposatutako eszenarioetan, prebalentziak eta lagin tamainak eragin handiagoa dutela ikusi dugu. Lortutako ondo-

rioak orain arteko egindako ikerketekin bat datoz (ikusi [12], [13]). Honetaz gain, HL pertzentilen testa eta test eraldatua erabili ditugu eta bi testekin lortutako emaitzak berdinak dira (ikusi [2]), lanean ez ditugu aurkeztu.

Simulazioetan, gaixo eta osasuntsuen bariantzan desberdintasun txiki bat dagoenean, HL testean lortutako ondorioen aldaketa aztertu nahi izan dugu. Hala ere, beste eszenario batzuetan simulazioak egin ditugu eta emaitzak ezberdinak izan dira, ez ditugu aurkeztu. Etorkizunean, bariantzen arteko ezberdintasunaren eragina sakonago aztertu beharko genukeela uste dugu.

Bibliografia

[1] Paul, P., Pennell, M. L., eta Lemeshow, S. (2013). Standardizing the power of the Hosmer–Lemeshow goodness of fit test in large data sets. *Statistics in Medicine, 32*(1), 67-80. https://doi.org/10.1002/sim.5525

[2] Nattino, G., Pennell, M. L., eta Lemeshow, S. (2020). Assessing the goodness of fit of logistic regression models in large samples: A modification of the Hosmer-Lemeshow test. *Biometrics, 76*(2), 549–560. https://doi.org/10.1111/biom.13249

[3] Yu, W., Xu, W., eta Zhu, L. (2017). A modified Hosmer–Lemeshow test for large data sets. *Communications in Statistics - Theory and Methods, 46*(23), 11813–11825. https://doi.org/10.1080/03610926.2017.1285922

[4] Dimitriadis, T., Henzi, A., Puke, M., eta Ziegel, J. (2022). A safe Hosmer-Lemeshow test. https://doi.org/10.48550/arXiv.2203.00426

[5] David W. Hosmer, J., Lemeshow, S., eta Sturdivant, R. X. (2013). *Applied logistic regression.* Wiley.

[6] Pawitan, Y. (2001). *In all likelihood: Statistical modelling and inference using likelihood.* OUP Oxford.

[7] Hosmer, D. W., Lemeshow, S., eta Klar, J. (1988). Goodness-of-Fit Testing for the Logistic Regression Model when the Estimated Probabilities are Small. *Biometrical Journal, 30*(8), 911–924. https://doi.org/10.1002/bimj.4710300805

[8] Moore, D. S., eta Spruill, M. C. (1975). Unified Large-Sample Theory of General Chi-Squared Statistics for Tests of Fit. *The Annals of Statistics, 3*(3). https://doi.org/10.1214/aos/1176343125

[9] Saxena, K. M. L., eta Alam, K. (1982). Estimation of the Non-Centrality Parameter of a Chi Squared Distribution. *The Annals of Statistics, 10*(3). https://doi.org/10.1214/aos/1176345892

[10] Iparragirre, A. (2017). *On the optimism correction of the performance of prediction models* (MAL).

[11] Iparragirre, A. (2016). *Eredu aurresaleen balidazio tekniken konparaketa eta inplementazioa* (GrAL).

[12] Kramer, A. A., eta Zimmerman, J. E. (2007). Assessing the calibration of mortality benchmarks in critical care: The Hosmer-Lemeshow test revisited*. *Critical Care Medicine, 35*(9), 2052–2056. https://doi.org/10.1097/01.ccm.0000275267.64078.b0

[13] Hosmer, D. W., Hosmer, T., Le Cessie, S. eta Lemeshow, S. (1997). A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in Medicine, 16*(9), 965-980. https://doi.org/10.1002/(SICI)1097-0258(19970515)16:9¡965::AID-SIM509¿3.0.CO;2-O

A. Eranskina

R kodea

```
#PAKETEAK

library(MASS)
library(ResourceSelection)
library(CatPredi)
library(dplyr)

#LAGINA SORTZEKO

n<-50000
g1<-5
g2<-10
by<-1
n_simulazio<-100
#prop_0, prop_G, mu_0, mu_G, sigma_0, sigma_G kasuaren arabera
    zehaztu. Behin zehaztuta:
set.seed(1234)
X0<-mvrnorm(n*prop_0,mu=mu_0,Sigma=sigma_0)
XG<-mvrnorm(n*prop_G,mu=mu_G,Sigma=sigma_G)
datuak_0<-data.frame(X0)
datuak_G<-data.frame(XG)

#FUNTZIOAK

#FUNTZIO OROKORRAK

#Talde kopuru gomendatua kalkulatu
g_gomendatua<-function(n_azpilagina,prop_0){
if(n_azpilagina>=50 && n_azpilagina<1000) return (10)
else if(n_azpilagina>=1000 && n_azpilagina<=25000){
return (round(max(10,
               min(n_azpilagina*prop_0/2,
               (n_azpilagina-n_azpilagina*prop_0)/2,
               2+8*(n_azpilagina/1000)^2))))
                                                  }
return (0)
}
```

```
#Azpilagina sortu
azpilagina_sortu<-function(seed,prop_0,prop_G,n_azpilagina){
set.seed(seed)
azpilagina_datuak0<-datuak_0[sample(nrow(datuak_0),
                              n_azpilagina*prop_0,replace=F),]
azpilagina_datuakG<-datuak_G[sample(nrow(datuak_G),
                              n_azpilagina*prop_G,replace=F),]
Y<-c(rep(0,n_azpilagina*prop_0), rep(1,n_azpilagina*prop_G))
datuak<-data.frame(Y, rbind(azpilagina_datuak0,
                            azpilagina_datuakG))
return (datuak)
}

#ALDAGAIREN BAT JARRAITUA

#Eredua doitu
eredua_azpilagina<-function(seed,prop_0,prop_G,
                            n_azpilagina){
datuak<-azpilagina_sortu(seed,prop_0,prop_G,n_azpilagina)
m<-glm(Y~X1+X2,data=datuak,family='binomial',maxit = 150)
return(m)
}

#HL testa aplikatu
hl_taldeak_aldatuz<-function(n_simulazio=100,
                             prop_0,prop_G,
                             n_azpilagina,g1,g2,by){

seeds<-seq(1:n_simulazio)*100000
m<-eredua_azpilagina(seeds[1],prop_0,prop_G,n_azpilagina)
testa<-hoslem.test(m$y,m$fitted.values,g=g1)
emaitza<-data.frame("Simulazioa"=1,
                    "Tamaina"=n_azpilagina,
                    "Prebalentzia"=prop_0,
                    "Talde_kopurua"=g1,
                    "Chi_Karratu"=testa$statistic,
                    "p.balioa"=testa$p.value)
for (k in seq(from=g1+1,to=g2,by=by)){
testa<-hoslem.test(m$y,m$fitted.values,g=k)
emaitza<-rbind(emaitza,
               data.frame("Simulazioa"=1,
                          "Tamaina"=n_azpilagina,
                          "Prebalentzia"=prop_0,
                          "Talde_kopurua"=k,
                          "Chi_Karratu"=testa$statistic,
                          "p.balioa"=testa$p.value))
                                    }
if (g_gomendatua(n_azpilagina,prop_0)>0 &&
   (g_gomendatua(n_azpilagina,prop_0)>g2 ||
    g_gomendatua(n_azpilagina,prop_0)<g1)){
testa<-hoslem.test(m$y,m$fitted.values,
                   g_gomendatua(n_azpilagina,prop_0))
emaitza<-rbind(emaitza,
               data.frame("Simulazioa"=1,
```

```
91                          "Tamaina"=n_azpilagina,
92                          "Prebalentzia"=prop_0,
93                          "Talde_kopurua"=g_gomendatua(n_
    azpilagina,prop_0),
94                          "Chi_Karratu"=testa$statistic,
95                          "p.balioa"=testa$p.value))
96                                          }
97 for (r in 2:n_simulazio){
98 m<-eredua_azpilagina(seeds[r],prop_0,prop_G,n_azpilagina)
99 for (k in seq(from=g1,to=g2,by=by)){
100 testa<-hoslem.test(m$y,m$fitted.values,g=k)
101 emaitza<-rbind(emaitza,
102 data.frame("Simulazioa"=r,
103            "Tamaina"=n_azpilagina,
104            "Prebalentzia"=prop_0,
105            "Talde_kopurua"=k,
106            "Chi_Karratu"=testa$statistic,
107            "p.balioa"=testa$p.value))
108                                     }
109 if (g_gomendatua(n_azpilagina,prop_0)>0 &&
110    (g_gomendatua(n_azpilagina,prop_0)>g2 ||
111     g_gomendatua(n_azpilagina,prop_0)<g1)){
112 testa<-hoslem.test(m$y,m$fitted.values,
113                    g_gomendatua(n_azpilagina,prop_0))
114 emaitza<-rbind(emaitza,
115                data.frame("Simulazioa"=r,
116                           "Tamaina"=n_azpilagina,
117                           "Prebalentzia"=prop_0,
118                           "Talde_kopurua"=g_gomendatua(n_
    azpilagina,prop_0),
119                           "Chi_Karratu"=testa$statistic,
120                           "p.balioa"=testa$p.value))
121                                          }
122                        }
123 return(emaitza)
124 }
125
126 #Proportzioak kalkulatu
127 proportzioak<-function(n_simulazio,prop_0,prop_G,
128                        n_azpilagina,g1,g2,by,alpha){
129 datuak<-hl_taldeak_aldatuz(n_simulazio,prop_0,
130                            prop_G,n_azpilagina,g1,
131                            g2,by)  %>% arrange(Talde_kopurua)
132 pbalioak<-select(datuak,p.balioa)
133 pbalioak<-pbalioak[,1]
134 batura<-sum(pbalioak[1:n_simulazio]>alpha)
135 emaitza<-data.frame("Tamaina"=n_azpilagina,
136                     "Talde_kopurua"=g1,
137                     "Proportzioa"=batura/n_simulazio)
138 k=1
139 while (g1+k<=g2){
140 batura<-sum(pbalioak[seq(n_simulazio*(k)+1,
141                          n_simulazio*(k+1))]>alpha)
142 emaitza<-rbind(emaitza,
```

```
143                 data.frame("Tamaina"=n_azpilagina,
144                            "Talde_kopurua"=g1+k,
145                            "Proportzioa"=batura/n_simulazio))
146 k=k+1
147                  }
148 if (g_gomendatua(n_azpilagina,prop_0)>0 &&
149    (g_gomendatua(n_azpilagina,prop_0)>g2 ||
150     g_gomendatua(n_azpilagina,prop_0)<g1)){
151 batura<-sum(pbalioak[seq(n_simulazio*(g2-g1+1)+1,
152                          n_simulazio*(g2-g1+2))]>alpha)
153 emaitza<-rbind(emaitza,
154                 data.frame("Tamaina"=n_azpilagina,
155                            "Talde_kopurua"=g_gomendatua(n_
    azpilagina, prop_0),
156                            "Proportzioa"=batura/n_simulazio))
157                                         }
158 return(emaitza)
159 }
160
161 #ALDAGAIREN BAT KATEGORIKOA
162
163 #Azpilagineko datuak kategorizatu: X_1
164 catpredi_X1<-function(seed,prop_0,prop_G,n_azpilagina,
165                       mozketa_kopurua){
166 datuak<-azpilagina_sortu(seed,prop_0,prop_G,n_azpilagina)
167 emaitza<-catpredi(formula = Y~X2, cat.var = "X1",
168                   cat.points = mozketa_kopurua,
169                   data = datuak, method = "addfor",
170                   range=NULL, correct.AUC=FALSE)
171 mozketa_puntuak<-c()
172 for (i in 1:mozketa_kopurua){
173 mozketa_puntuak<-c(mozketa_puntuak,
174                    emaitza$results$cutpoints[i])
175                             }
176 mozketa_puntuak<-c(mozketa_puntuak,
177                    min(datuak$X1)-0.1,
178                    max(datuak$X1)+0.1)
179 datuak$X1=cut(datuak$X1,
180               breaks=sort(mozketa_puntuak))
181 return(datuak)
182 }
183
184 #Eredua doitu
185 eredua_azpilagina_catpredi<-function(seed,prop_0,prop_G,
186                                      n_azpilagina,
187                                      mozketa_kopurua){
188 datuak<-catpredi_X1(seed,prop_0,prop_G,n_azpilagina,
189                     mozketa_kopurua)
190 m<-glm(Y~X1+X2,data=datuak,family='binomial',maxit = 150)
191 return(m)
192 }
193
194
195
```

```
#HL testa aplikatu
hl_catpredi_taldeak_aldatuz<-function(n_simulazio,prop_0,
                                      prop_G,n_azpilagina,g1,
                                      g2,mozketa_kopurua){
seeds<-seq(1:n_simulazio)*100000
datuak<-catpredi_X1kategorizatuta(seeds[1],prop_0,prop_G,
                                  n_azpilagina,mozketa_kopurua)
m<-eredua_azpilagina_catpredi(seed,prop_0,prop_G,
                              n_azpilagina,mozketa_kopurua)
testa<-hoslem.test(m$y,m$fitted.values,g=g1)
emaitza<-data.frame("Simulazioa"=1,
                    "Tamaina"=n_azpilagina,
                    "Prebalentzia"=prop_0,
                    "Mozketa_kopurua"=mozketa_kopurua,
                    "Talde_kopurua"=g1,
                    "Chi_Karratu"=testa$statistic,
                    "p.balioa"=testa$p.value)

for (k in seq(from=g1+1,to=g2,by=by)){
testa<-hoslem.test(m$y,m$fitted.values,g=k)
emaitza<-rbind(emaitza,
               data.frame("Simulazioa"=1,
                          "Tamaina"=n_azpilagina,
                          "Prebalentzia"=prop_0,
                          "Mozketa_kopurua"=mozketa_kopurua,
                          "Talde_kopurua"=k,
                          "Chi_Karratu"=emaitza$statistic,
                          "p.balioa"=emaitza$p.value))
                                    }
if (g_gomendatua(n_azpilagina,prop_0)>0 &&
   (g_gomendatua(n_azpilagina,prop_0)>g2 ||
    g_gomendatua(n_azpilagina,prop_0)<g1)){
emaitza<-hoslem.test(m$y,m$fitted.values,
                     g=g_gomendatua(n_azpilagina,prop_0))
emaitza<-rbind(emaitza,
               data.frame("Simulazioa"=1,
                          "Tamaina"=n_azpilagina,
                          "Prebalentzia"=prop_0,
                          "Mozketa_kopurua"=mozketa_kopurua,
                          "Talde_kopurua"=g_gomendatua(n_
    azpilagina,prop_0),
                          "Chi_Karratu"=testa$statistic,
                          "p.balioa"=testa$p.value))
                                       }
for (r in 2:n_simulazio){
datuak<-catpredi_X1kategorizatuta(seeds[r],prop_0,prop_G,
                                  n_azpilagina,mozketa_kopurua)
m<-glm(Y~X1+X2, data = datuak, family='binomial', maxit = 150)
for (k in seq(from=g1,to=g2,by=by)){
testa<-hoslem.test(m$y,m$fitted.values,g=k)
emaitza<-rbind(emaitza,
               data.frame("Simulazioa"=r,
                          "Tamaina"=n_azpilagina,
                          "Prebalentzia"=prop_0,
```

```
249                           "Mozketa_kopurua"=mozketa_kopurua,
250                           "Talde_kopurua"=k,
251                           "Chi_Karratu"=testa$statistic,
252                           "p.balioa"=testa$p.value))
253                                   }
254 if (g_gomendatua(n_azpilagina,prop_0)>0 &&
255    (g_gomendatua(n_azpilagina,prop_0)>g2 ||
256     g_gomendatua(n_azpilagina,prop_0)<g1)){
257 testa<-hoslem.test(m$y,m$fitted.values,
258          g=g_gomendatua(n_azpilagina,prop_0))
259 emaitza<-rbind(emaitza,
260                data.frame("Simulazioa"=r,
261                "Tamaina"=n_azpilagina,
262                "Prebalentzia"=prop_0,
263                "Mozketa_kopurua"=mozketa_kopurua,
264                "Talde_kopurua"=g_gomendatua(n_azpilagina,
265                                             prop_0),
266                "Chi_Karratu"=testa$statistic,
267                "p.balioa"=testa$p.value))
268                                          }
269                        }
270 return(emaitza)
271 }
272
273 #Proportzioak kalkulatu
274 proportzioak_catpredi<-function(n_simulazio,prop_0,prop_G,
275                                 n_azpilagina,g1,g2,by,alpha){
276 datuak<-hl_catpredi_taldeak_aldatuz(n_simulazio,prop_0,
277                                     prop_G,
278                                     n_azpilagina,g1,g2,
279                                     by)  %>% arrange(Talde_
    kopurua)
280 pbalioak<-select(datuak,p.balioa)
281 pbalioak<-pbalioak[,1]
282 batura<-sum(pbalioak[1:n_simulazio]>alpha)
283 emaitza<-data.frame("Tamaina"=n_azpilagina,
284                     "Talde_kopurua"=g1,
285                     "Proportzioa"=batura/n_simulazio)
286 k=1
287 while (g1+k<=g2){
288 batura<-sum(pbalioak[seq(n_simulazio*(k)+1,
289                          n_simulazio*(k+1))]>alpha)
290 emaitza<-rbind(emaitza,
291                data.frame("Tamaina"=n_azpilagina,
292                           "Talde_kopurua"=g1+k,
293                           "Proportzioa"=batura/n_simulazio))
294 k=k+1
295                  }
296 if (g_gomendatua(n_azpilagina,prop_0)>0 &&
297    (g_gomendatua(n_azpilagina,prop_0)>g2 ||
298     g_gomendatua(n_azpilagina,prop_0)<g1)){
299 batura<-sum(pbalioak[seq(n_simulazio*(g2-g1+1)+1,
300                          n_simulazio*(g2-g1+2))]>alpha)
301 emaitza<-rbind(emaitza,
```

```
302                 data.frame("Tamaina"=n_azpilagina,
303                            "Talde_kopurua"=g_gomendatua(n_
    azpilagina,prop_0),
304                            "Proportzioa"=batura/n_simulazio))
305                                          }
306 return(emaitza)
307 }
308
```