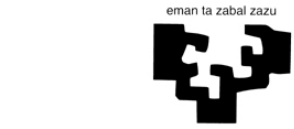


ZTF-FCT
Zientzia eta Teknologia Fakultatea
Facultad de Ciencia y Tecnología

Eredu aurresaleen balidazio tekniken konparaketa eta inplementazioa

Gradu Amaierako Lana
Matematikako Gradua

Amaia Iparragirre Letamendia

Irantzu Barrio Beraza
Irakasleak zuzendutako lana

Leioan, 2016ko ekainaren 27an

Aurkibidea

Sarrera

Gaur egun, eredu aurresaleak gero eta indar handiagoa hartzen ari dira. Etorkizuneko egoerak aurresan ahal izateko erabiltzen dira, eta adituei egoera horiek kontuan izanda, hartu beharreko neurriak hartzen laguntzeko balio diete. Esaterako, medikuntzan eredu mota hauek asko erabiltzen dira, eta medikuei gaixoen egoera, epe batean, nolakoa izango den aurreikusten laguntzen diete. Hortaz, duten garrantzia ikusita, esan beharrik ez dago behar-beharrezkoa dela eredu aurresaleak fidagarriak izatea eta aurresaten dutenaren eta benetan gertatuko denaren artean ahalik eta ezberdintasun gutxien egotea. Horretarako, ereduen balidazioa egiten da.

Lan honetan, eredu aurresaleen balidazioa egiteko metodo ezberdinak landu eta R pakete estatistikoan programatu dira. Guk eredu baten aurresateko gaitasuna AUCren bidez neurtu dugu eta, helburua da, horretarako balidazio metodorik egokiena zein den zehaztea.

Praktikan, ikerkuntza medikoan ereduen garapena egitean, split validation edo zatikako balidazioa erabiltzen da. Balidazio metodo hau beste metodo batzuekin alderatu nahi izan da, ikusteko benetan aurresateko gaitasuna neurtzeko metodorik onena erabiltzen den edo lortzen dituzten emaitzak hobetu daitezkeen beste balidazio metodo bat erabilita.

Horretarako, simulazio bidezko ikerketa bat egin da. Eszenario ezberdinak planteatu dira eta hauetako bakoitzean, metodo desberdinak erabiliz (split validation edo zatikako balidazioa, eta bootstrap balidazio metodoa, besteak beste) lortutako emaitzak AUC teorikoarekin konparatu dira.

Horrez gain, landutako balidazio metodoak datu-base erreal bati aplikatu dizkiogu. Hain zuzen ere, Galdakaoko Ospitalean egiten ari diren ikerketa bateko datuak dira, bihotzaren kongestio-gutxiegitasunari buruzko ikerketa batekoak, hain justu.

Lan hau ondoko eran antolatuta dago: 1. kapituluan, erregresio logistikoaren inguruko informazioa agertzen da. Aurresateko gaitasuna definitu eta AUC teorikoa zer den azaltzen da. 2. kapituluan, erregresio logistikoaren

balidazioaren garrantziaz hitz egiten da eta balidazio metodo ezberdinak azaltzen dira pausoz pauso. 3. kapituluan, egindako simulazioak eta lortutako emaitzak azaltzen dira. 4. kapituluan, landutakoa praktikara eramanda, datu errealak erabiliz egindako aplikazioa dago. 5. kapituluan, bildutako informazio guztia erabiliz ateratako ondorioak irakur daitezke. Amaitzeko, bibliografia eta, azkenik, eranskinean, balidazio metodo ezberdinak programatzeko garatu den kodea dago.

1. kapitulua

Erregresio logistikoa

1.1 Sarrera

Erregresio metodoek, erantzun aldagai baten eta aldagai azaltzaile bat edo gehiagoren arteko erlazioa azaltzen dute. Askotan, erantzun aldagaiaren banaketa binomiala izaten da eta, kasu honetan, erregresio logistikoa da analisia burutzeko metodo ohikoena.

Erregresio logistikoaren helburua da, biologikoki zentzua duen doikuntza egokiena lortzea, erantzun aldagaiaren eta aldagai azaltzaileen arteko erlazioa deskribatzeko. Bere berezitasuna da, erantzun aldagaia bitarra edo dikotomikoa dela.

1.2 Ereduaren doikuntza eta estimazioa

Izan bitez, $X_1, X_2, \ldots, X_p$ p aldagai aske non $\mathbf{X} = (X_1, X_2, \ldots, X_p)^t$ eta Y erantzun aldagai dikotomikoa diren. Izan bedi, $P(Y = 1|\mathbf{X}) = p(\mathbf{X})$ arrakasta izatearen probabilitate baldintzatua. Erregresio logistikoaren eredua hau da:

$$p(\mathbf{X}) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}} \in (0, 1)$$

Logit transformazioa deituko diogu $p(\mathbf{X})$-ren transformazio honi:

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

$g(\mathbf{X})$ funtzioak erregresio lineal orokorraren ereduaren propietate batzuk betetzen ditu: funtzio lineala da, jarraitua izan daiteke eta $(-\infty, +\infty)$ balioak har ditzake.

$\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^t$ parametroak estimatzeko egiantz handieneko metodoa erabiltzen da. Demagun n tamainako lagin bat dugula eta izan bedi $(\mathbf{x}_i, y_i), i =$

$1, 2, \ldots, n$ bikotea, non y_i erantzun aldagaiaren balio dikotomikoa den eta $\mathbf{x}_i$ aldagai askeen bektorearen balioa i.gaian. Ondokoa da egiantz-funtzioa:

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

Izan bedi $L(\boldsymbol{\beta}) = ln[l(\boldsymbol{\beta})]$. $L(\boldsymbol{\beta})$ maximizatzen duen $\boldsymbol{\beta}$ aurkitzeko, $L(\boldsymbol{\beta})$ diferentziatuko dugu $\beta_j, j = 0, 1, \ldots, p$ bakoitzerako. Lortzen diren ekuazioei, egiantz-ekuazioak deitzen zaie. $p+1$ egiantz-ekuazio lortuko ditugu guztira, $p + 1$ koefizienteei dagozkienak, hain zuzen ere. Honako hauek dira:

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

eta

$$\sum_{i=1}^{n} x_{ij}(y_i - p(\mathbf{x}_i)) = 0, \quad \forall j = 1, 2, \ldots, p.$$

Ekuazio hauek Newton-Raphson-en metodoaren bidez ebazten dira[1].

$\boldsymbol{\beta}$-rentzako estimatutako erregresio koefizienteak adierazteko $\hat{\boldsymbol{\beta}}$ erabiliko dugu.

1.3 Aurresateko gaitasuna

Doitutako ereduan lortutako emaitzak laburbiltzeko modu bat, sailkapen-taulak erabiltzea da. Izan bedi

$$\hat{p}(\mathbf{x}) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x_1+\cdots+\hat{\beta}_p x_p}}{1 + e^{\hat{\beta}_0+\hat{\beta}_1 x_1+\cdots+\hat{\beta}_p x_p}},$$

indibiduo bakoitzari dagokion arrakastarako probabilitatearen estimazioa. Indibiduo hauek arrakasta edo porrota bezala sailkatzeko, c mozketa-puntu bat aukeratzen da. Estimatutako probabilitatea mozketa-puntu hau baino handiagoa bada, indibiduo hori arrakasta bezala sailkatzen da eta 1 balioa emango diogu. Txikiagoa bada aldiz, porrot bezala sailkatuko dugu eta 0 balioa izango du. Orduan, itxura honetako 2×2-ko taula bat sor dezakegu, erantzun aldagaiaren behatutako (Y) eta estimatutako balioekin ($\hat{Y}$):

	Estimatutakoak ($\hat{Y}$)		
Behatutakoak (Y)	0	1	
0	a (EN)	b(PF)	a+b
1	c (NF)	d (EP)	c+d
	a+c	b+d	n

1.1. Taula. EN≡ egiazko negatiboak, EP≡ egiazko positiboak, NF≡ negatibo faltsuak eta PF≡ positibo faltsuak.

Ikus dezakegunez,

- egoki sailkatutako proportzio globala $\frac{a+d}{n} \times 100$ da.
- egoki sailkatutako arrakasta proportzioa $\frac{d}{c+d} \times 100$ da eta $\frac{d}{c+d}$ zatiketari sentikortasuna deritzo.
- egoki sailkatutako porrot proportzioa $\frac{a}{a+b} \times 100$ da eta $\frac{a}{a+b}$ zatiketari espezifikotasuna deritzo.

Beraz, sentikortasuna eta espezifikotasuna aukeratutako mozketa-puntuaren araberakoak dira. Askotan, $c = 0.5$ mozketa-puntua aukeratzen da. Hala ere, arrakasta izateko probabilitatea oso txikia (edo oso handia) bada, gerta daiteke indibiduo guztiak talde berean sailkatzea, hau da, guztiak porrot (edo arrakasta) bezala sailkatzea. Ondorioz, ereduak ez luke bereiziko arrakasta eta porrotaren artean.

Mozketa-puntu ezberdinentzako emaitzak, denak batera, kontuan hartzen dituen parametro bat, AUC (*area under the ROC curve*) da. ROC (*Receiver Operating Characteristic*) kurba, sentikortasuna eta $1-$ espezifikotasuna batera irudikatzen dituen grafikoari deitzen zaio. AUC, ROC kurbaren azpiko azalera da eta 0.5etik 1erako balioak hartzen ditu. AUCak doitutako ereduaren aurresateko gaitasuna neurtzen du. Beste modu batera esanda, ROC kurbaren azpiko azalerak, ereduak duen bereizteko gaitasuna neurtzen du, arrakasta duen indibiduo baten eta arrakastarik ez duenaren artean.

AUC zenbat eta handiagoa izan, orduan eta hobeagoa da ereduaren aurresateko gaitasuna. $AUC = 0.5$ bada, eredua ez da gai bereizketarik egiteko, txanpon bat airera botatzearen parekoa izango litzateke bere aurresateko gaitasuna. $0.7 \leq AUC \leq 0.8$ bada, ereduaren aurresateko gaitasuna egokia dela esaten da. $0.8 \leq AUC \leq 0.9$ bada, ereduaren aurresateko gaitasuna bikaina da eta $AUC \geq 0.9$ denean are hobea, hau praktikan oso gutxitan gertatzen bada ere [1].

Eman dezagun orain aurresateko gaitasunaren beste definizio bat. Izan bitez $TPF(c) = P(p(\mathbf{x}) \geq c|Y = 1)$ non TPF-k adierazten duen egiazko positiboen arrazoia, sentikortasuna alegia, eta $FPF(c) = P(p(\mathbf{x}) \geq c|Y = 0)$ non FPF-k adierazten duen positibo faltsuen arrazoia, 1-espezifikotasuna, hain justu. Definizioz, ondokoa da ROC kurba:

$$ROC(\cdot) = \{(FPF(c), TPF(c)), c \in (0, 1)\}.$$

ROC kurba, lehen koadranteko funtzio monotono gorakorra da eta beste modu honetan ere idatz daiteke:

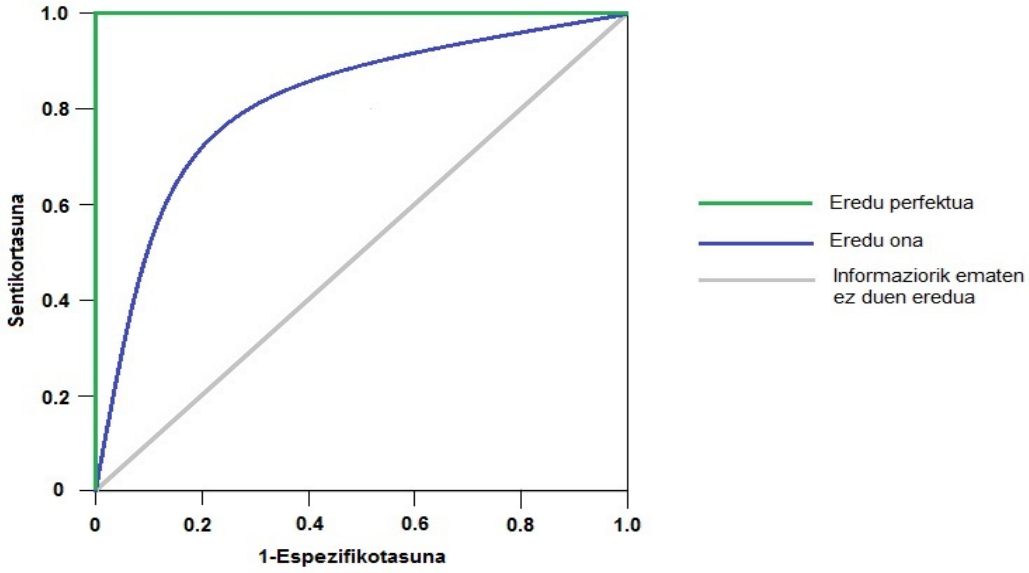
$$ROC(\cdot) = \{(t, ROC(t)), t \in (0, 1)\}$$

non $ROC(t) = TPF(c)$ den, c $FPF(c) = t$ egiten duen mozketa-puntua izanik.

Edozein mozketa-punturentzat $TPF(c) = FPF(c)$ gertatzen bada, ereduak informaziorik ez digula ematen ondorioztatuko dugu. Erabilgarria ez den eredu baten ROC kurba, beraz, $ROC(t) = t$ da, malda 1 duen zuzena, alegia. Aurresateko gaitasuna, esan bezala, ROC kurbaren azpiko azalera da. Ondorioz, honela lor dezakegu:

$$AUC = \int_0^1 ROC(t)dt.$$

1.1. irudian ikus daitezke eredu mota ezberdinen ROC kurbak:

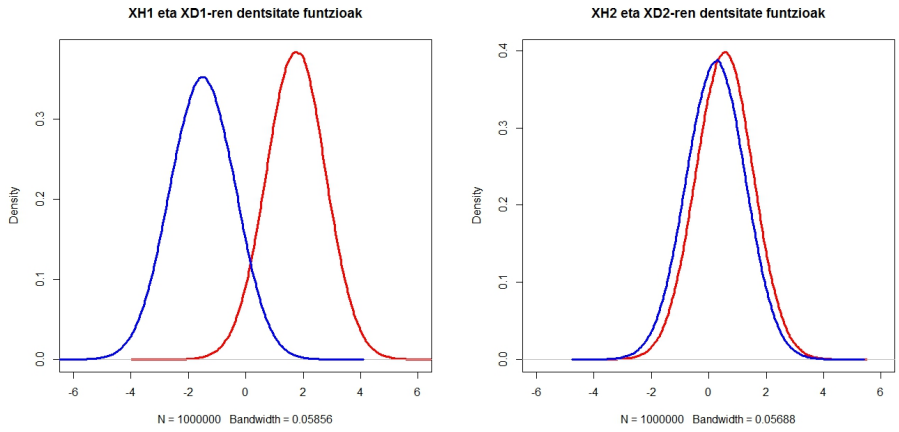


1.1. irudia. Eredu mota ezberdinen ROC kurbak

1.4 AUC teorikoa

Azaldu dugun moduan, AUCak ematen digu eredu batek duen gaitasuna gaixo eta osasuntsu baten arteko bereizketa egiteko. AUC teorikoaren ideia intuitiboa ulertzeko adibide bat ikusiko dugu.

Demagun bi eredu ditugula, bakoitza aldagai azaltzaile batekin, X_1 eta X_2, hurrenez hurren. Irudika ditzagun aldagai hauen dentsitate funtzioak gaixoetan eta osansuntsuetan bakoitza bere aldetik 1.2 irudian agertzen den bezala.



(a) (b)

1.2. irudia. Ezkerrean (a): X_1 aldagaiaren dentsitate funtzioa gaixoetan (X_{D1}) (lerro gorria) eta osasuntsuetan (X_{H1}) (lerro urdina). Eskuinean (b): X_2 aldagaiaren dentsitate funtzioa gaixoetan (X_{D2}) (lerro gorria) eta osasuntsuetan (X_{H2}) (lerro urdina).

Izan bitez X_{H1} eta X_{D1} X_1 aldagaia osasuntsuen eta gaixoen azpimultzoetan hurrenez hurren (gauza bera X_2 aldagaiarekin). 1.2. irudiari begiratuta, garbi dago, ezkerreko azpirudian gaixo eta osasuntsuen aldagaien banaketa ezberdinagoa dela. Beste modu batera esanda, X_{H1} eta X_{D1}-ren dentsitate funtzioen arteko eremua, txikiagoa da X_{H2} eta X_{D2}-ren artekoa baino. X_{H2} eta X_{D2} ia elkarren gainean daude eta, kasu honetan, zailagoa da bereiztea indibiduo bat gaixo edo osasuntsu dagoen. Lehen ereduak ordea, errazago bereiziko du indibiduo bat gaixoa edo osasuntsua den, euren dentsitate funtzioak oso banatuta daudelako bata bestearengatik. Eta hau da, hain zuzen ere, AUC teorikoari buruz hitz egitean esan nahi duguna. Esan bezala, AUC teorikoa da doitu dugun ereduak duen bereizteko gaitasuna gaixo eta osasuntsuen artean. Argi dago, beraz, lehen ereduaren AUC teorikoa altuagoa izango dela bigarren ereduaren AUC teorikoa baino, dentsitate funtzioak begiratuz ondorio horretara iritsi garelako.

Aldagai bakarrarekin azaldu dugun kontzeptu hau, p aldagaitara ere heda daiteke. Hala ere, AUC teoriko honen balio zehatza kasu gutxi batzuetan bakarrik lortu ahal izango dugu. Izan bedi, Y erantzun aldagaia, $Y = 1$ gaixo eta $Y = 0$ osasuntsu izanik. Demagun, p aldagai azaltzaile jarraitu ditugula non $\mathbf{X} = (X_1, X_2, \ldots, X_p)$ den. Izan bedi $\mathbf{X}_D$ aldagai azaltzaileen bektorea gaixoetan eta $\mathbf{X}_H$ osasuntsuetan. Baldin $\mathbf{X}_D$-ren banaketa $\mathbf{X}_D \sim N(\boldsymbol{\mu}_D, \boldsymbol{\Sigma}_D)$ bada eta $\mathbf{X}_H$-rena $\mathbf{X}_H \sim N(\boldsymbol{\mu}_H, \boldsymbol{\Sigma}_H)$, Su eta Liu, 1993-

ren korolarioaren arabera AUC teorikoa ondoko erara kalkula daiteke [3]:

$$A = \Phi\left(\sqrt{\boldsymbol{\mu}^T(\boldsymbol{\Sigma}_D + \boldsymbol{\Sigma}_H)^{-1}\boldsymbol{\mu}}\right) \tag{1.1}$$

$\boldsymbol{\mu} = \boldsymbol{\mu}_D - \boldsymbol{\mu}_H$ eta Φ banaketa normal estandarizatuaren banaketa-funtzioa izanik.

Beraz, gaixo eta osasuntsuen banaketak ezagutzen ditugunean, eta hauek normalak direnean, AUC teorikoa kalkulatu dezakegu 1.1 formula erabiliz.

2. kapitulua

Erregresio logistikoaren aurresateko gaitasunaren balidazioa

2.1 Erregresio logistikoaren aurresateko gaitasuna

Eredu aurresaleak oso tresna garrantzitsua dira etorkizuneko gertakariak aurresan ahal izateko. Eredu hauek erabiltzeko euren aurresateko gaitasuna neurtzea behar-beharrezkoa dugu. Ez du ezertarako balio eredu bat doitzeak, gero bere aurresateko gaitasuna txanpon bat airera botatzearen parekoa bada. Erregresio logistikoaren aurresateko gaitasuna neurtzeko maizen erabiltzen den parametroa AUCa da. Dena dela, AUCa estimatzeko eredua doitzeko erabili den lagin berdina erabiltzen baldin bada, AUC hori gainestimatuta egon daiteke, eta ondorioz garrantzitsua da lortutako AUC hori balidatzea.

2.2 Balidazio metodoak

Eredu aurresale baten helburua, etorkizunean gerta daitezkeen egoerak aurresan ahal izatea da. Esan bezala, oso garrantzitsua da eredu baten estimatutako aurresateko gaitasuna balidatzea. Bi motatako balidazio teknikak aurki ditzakegu: barne-balidazioko metodoak eta kanpo-balidazioko metodoak.

Lagin bat hartu eta eredu bat doitzen badugu, lagin honetako indibiduoek ezaugarri komun batzuk dituzte (adibidez, ospitale berera joan diren gaixoak dira guztiak). Gutxienez, populazio honentzako, gure ereduaren aurresateko gaitasuna neurtu beharko genuke. Barne-balidazioa deitzen zaio honi. Barne-balidazioaren bidez, gure datu-basea sortu den inguruneko balidazioa lor dezakegu.

Kanpo-balidazioa, "nahiko erlazionatuta" dauden populazioen bidezko balidazioa da. Populazioak euren artean "nahiko erlazionatuta" egoteak esan nahi du, adibidez, gaixotasun bera duten indibiduoekin ari garela lanean, baina gaixo hauek ospitale desberdinetan edo garai desberdinetan artatuak izan direla. Orokortasuna lor daitekeen edo ez ikus dezakegu teknika honen bidez eta hau oso interesgarria da, bai zientifikoki (hipotesi eta teoriek indarra hartzen dutelako), eta baita praktikan ere (eredu bat gure beharrizanetan erabili ahal izateko aukera ematen baitigu). Zenbat eta gehiagotan burutu kanpo-balidazioa eta populazio hauek zenbat eta desberdinagoak izan, orduan eta egonkortasun handiagoa eta ondorio orokorragoak lortu ahal izango ditugu.

Lan honetan, barne-balidazioko teknikak landuko ditugu.

1 Apparent Validation edo Balidazio Behagarria

Balidazio behagarriak, doitutako ereduaren balidazioa jatorrizko laginean egiten du, hau da, eredua doitzeko erabili den laginean bertan. Honek estimazioa baikorregia izatea dakar, ereduaren parametroak lagin horrentzat egokienak izan daitezen doitu dugulako eredu hori. Dena den, datu-base osoa erabiltzen da, bai eredua doitzeko, baita balidazioa egiteko ere. Beraz, estimazio baikorrak baina egonkorrak lortzen dira.

2 Split-Sample Validation edo Zatikako Balidazioa

Zatikako balidazioa burutzeko, lagina bi zatitan banatzen da zoriz. Balidazio mota hau, kanpo-balidazioaren ideiatik dator. Zoriz aukeratutako bi talde hauetako batean eredua doitzen da, eta bestean, balidatu egiten da aurrez doitutako eredu hori. Laginaren zatiketa ohikoenak $1/2 : 1/2$ edo $2/3 : 1/3$ dira ($2/3$ eredua doitzeko, $1/3$ balidatzeko).

Lagina zatitzerako orduan, kontuan izan behar dugu hainbat arazo sor daitezkeela. Hasteko, laginaren zatiketa guztiz zorizkoa bada, desoreka egon daiteke doikuntza-lagin eta balidazio-laginen artean. Ezer egin baino lehen, ziurtatu behar da baliokideak direla bi laginak.

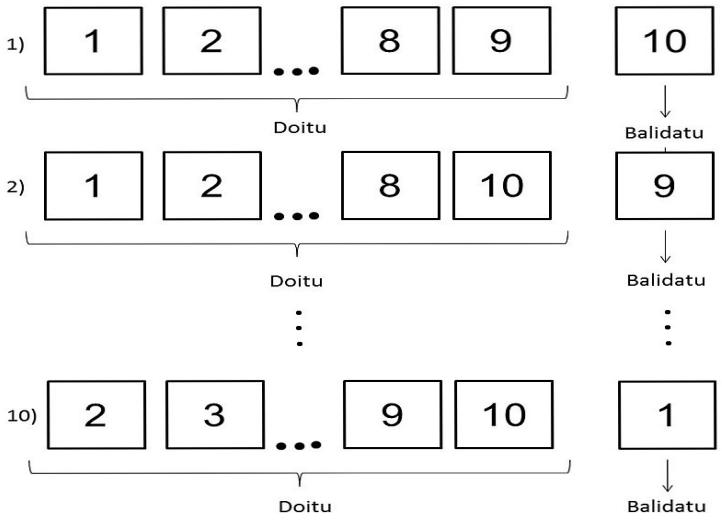
Metodo honek desabantaila asko ditu. Datuen zati bat bakarrik erabiltzen da eredua doitzeko. Ondorioz, lortutako emaitzak ezegonkorragoak dira, datu guztiak erabiliz lortutakoak baino. Horrez gain, balidazioaren lagin zatia ere nahiko txikia da, eta honek fidagarritasuna ahultzen du. Gainera ikertzaileak zorte txarra izan dezake laginaren zatiketan. Lortzen dugun balidazioaren emaitza lagin zati bati dago lotuta, eta guk lagin osoaren gaineko emaitzak lortu nahi ditugu.

Laburtuz, zatikako balidazioa ez da oso eraginkorra, oso erabilia izan arren. Simulazioen bidez frogatu da tamaina handiko laginak behar direla metodo honen bidez emaitza esangarriak lortzeko [2]. Beraz, balidazio-metodo hau, balidatzea beharrezkoa ez den kasuan bakarrik da egokia.

3 Cross-Validation edo Balidazio Gurutzatua

Balidazio gurutzatua, zatikako balidazioaren hedapena baino ez da eta bere helburua da egonkortasun handiagoa lortzea. Berriro ere, eredua zoriz aukeratutako zati batean balidatzen da, eta laginaren gainerako zatian doitzen da. Baina prozesu hau hainbat aldiz errepikatzen da. Adibidez, lagina hamar azpilaginetan bana daiteke. Eredua hamar zati hauetako bederatzitan doitzen da eta hamargarrenean balidatu egiten da. Prozesu hau hamar aldiz errepikatzen da, eta aldi bakoitzean multzo ezberdina erabiltzen da balidaziorako.

Zatikako balidazioarekin konparatzen badugu, metodo honek lagin zati handiagoa erabiltzen du eredua doitzeko eta hau abaintaila bat da, egonkorragoa delako. Balidazio gurutzatua muturreraino eramanda, jack-knife izeneko prozedura dugu. Metodo honek indibiduo bakoitza behin erabiltzen du ereduaren balidaziorako. Lagina handia bada ez da eraginkorra.



2.1. irudia. Balidazio gurutzatua

4 Bootstrap Balidazioa

Bootstrap balidazio metodoa, hasierako laginean oinarrituta behin eta berriz laginketak egitean datza. Laginketa hauetan errepikapenak onartzen dira eta jatorrizko laginaren indibiduo kopuru bera izan behar dute lorturiko bootstrap lagin berri hauek. Orokorrean, eredu aurresale baten balidazioa egitean, 100 edo 200 bootstrap nahikoak izan ohi dira estimazio egonkorrak lortzeko[2]. Gaur egun, teknika hau da erabiliena eredu baten balidazioa egiteko.

Bootstrap metodoa aplikatzeko honako pauso hauek jarraitu behar dira:

1. Doitu erregresio eredua jatorrizko laginean eta egin balidazioa lagin horretan bertan. Balidazio behagarria da hau eta AUC_{app} deituko diogu parametro honi.

2. $b = 1, 2, \ldots, B$ bakoitzerako, sortu bootstrap lagin berri bat zoriz, jatorrizko laginean oinarrituta, indibiduo kopuru bera duena eta errepikapenak onartzen direlarik. Doitu eredua bootstrap laginarekin, estimatu $\hat{\beta}_0^b, \hat{\beta}_1^b, \ldots, \hat{\beta}_p^b$ koefizienteak (p aldagai azaltzaile kopurua izanik), eta AUCa estimatu doitutako bootstrap laginean: AUC_{boot}^b.

3. Aplikatu 2. pausoan lortutako eredua jatorrizko laginean eta deitu AUC_o^b parametro honen balioari, $b = 1, 2, \ldots, B$ bakoitzerako.

Prozesu hau burutu ondoren, aurresateko gaitasunaren jatorrizko parametroaren optimismoa bi modutara kalkula daiteke:

$$O = \frac{1}{B} \sum_{b=1}^{B} (AUC_{boot}^b - AUC_o^b)$$

edo

$$O = \frac{1}{B} \sum_{b=1}^{B} \left| AUC_{boot}^b - AUC_o^b \right|.$$

Ondorioz, zuzendutako aurresateko gaitasunaren parametroaren balioa honela kalkulatzen da:

$$AUC_{zuzendua} = AUC_{app} - O$$

Bootstrap balidazioak hainbat abantaila ditu. Eredua doitu eta balidatzeko tamaina bereko laginak erabiltzen direnez, metodoa egonkorra da, balidazio behagarrian gertatzen zen bezala. Zatikako balidazioa eta balidazio gurutzatuarekin konparatuta, hau abantaila bat da. Balidazio behagarriarekin

alderatuta, optimismoa estimatu beharrak zenbait zalantza sor ditzakeen arren, nahikoa bootstrap laginketa eginez gero, zalantza hauek ez dira kontuan hartzeko modukoak.

3. kapitulua

Simulazioak

3.1 Sarrera

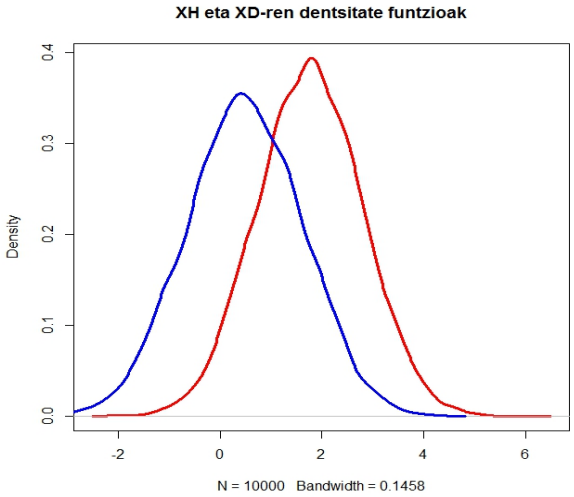
2. kapituluan eredu aurresale baten balidazioa egiteko hainbat metodo aipatu dira. Lan honen helburua da, egoera jakin batzuetan, balidazio metodo hauen guztien bidez lortutako AUCen arteko konparazioa egin eta AUC teorikora gehien hurbiltzen den metodoa aurkitzea. Beste modu batera esanda, baldintza batzuen menpe, metodo horien guztien artetik, egokiena zein den zehaztea, eredu aurresale baten aurresateko gaitasuna neurtzeko. Horretarako, hainbat simulazio egin ditugu, egoera ezberdinak planteatuz.

Bi aldagai azaltzaile eraiki nahi ditugu: X_1 eta X_2. Aldagai hauek jarraituak izatea nahi dugu, eta doitutako ereduak zentzua izateko *logit* p-rekiko erlazio lineala bete behar dute. Itxura hau izango dute doituko ditugun ereduek:

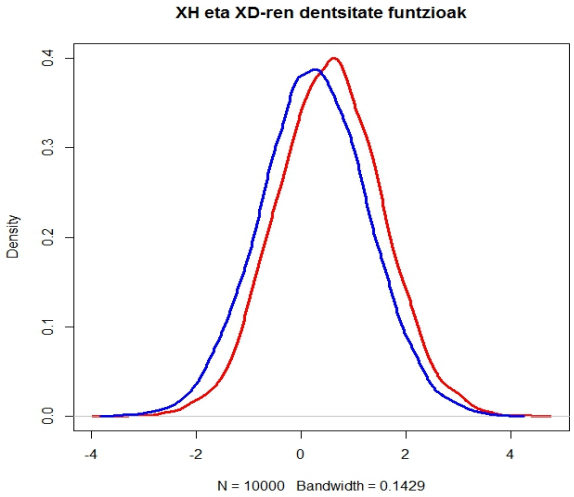
$$logit\ p = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

Izan bedi $\mathbf{X} = (X_1, X_2)$ aldagai azaltzaileen bektorea. Lehenik eta behin, bektore honen gaixo eta osasuntsuen bektoreak sortu ditugu. Bi bektoreek banaketa normalari darraie, $\mathbf{X}_H \sim N(\boldsymbol{\mu}_H, \boldsymbol{\Sigma}_H)$ eta $\mathbf{X}_D \sim N(\boldsymbol{\mu}_D, \boldsymbol{\Sigma}_D)$. Bi kasu bereiztu ditugu:

- Alde batetik, $\boldsymbol{\mu}_H = (0,1), \boldsymbol{\mu}_D = (1.5,2), \boldsymbol{\Sigma}_H = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \boldsymbol{\Sigma}_D$ definitu ditugu. 1.1 formula aplikatuz, $AUC_{teorikoa} = 0.8988$ dela kalkula dezakegu. 3.1. irudian ikus ditzakegu $\boldsymbol{X}_H$ eta $\boldsymbol{X}_D$-ren dentsitate funtzioak.
- Bestetik, $\boldsymbol{\mu}_H = (0, 0.5), \boldsymbol{\mu}_D = (0.5, 0.6), \boldsymbol{\Sigma}_H = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \boldsymbol{\Sigma}_D$ definitu ditugu. Kasu honetan, 1.1 formula erabiliz berriro ere, $AUC_{teorikoa} = 0.6408$ da. 3.2. irudian daude $\boldsymbol{X}_H$ eta $\boldsymbol{X}_D$-ren dentsitate funtzioak irudikatuta.

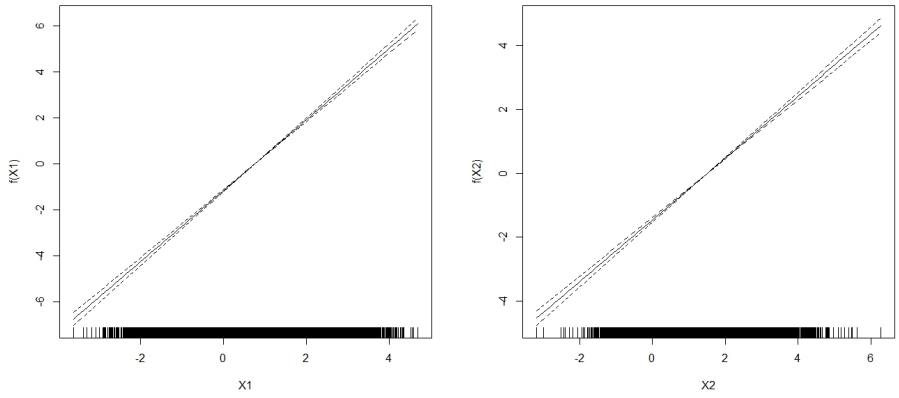


3.1. irudia. $AUC_{teorikoa} = 0.8988$. X_H-ren dentsitate funtzioa lerro urdina eta X_D-rena lerro gorria.



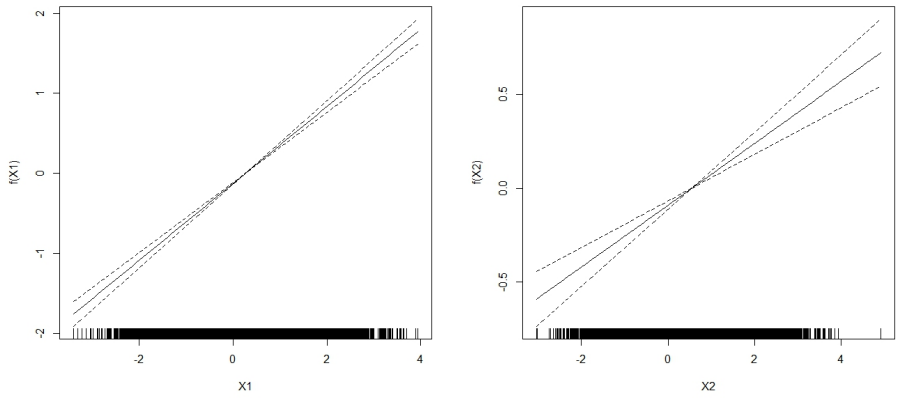
3.2. irudia. $AUC_{teorikoa} = 0.6408$. $\boldsymbol{X}_H$-ren dentsitate funtzioa lerro urdina eta $\boldsymbol{X}_D$-rena lerro gorria.

Bi egoeretan 10.000 indibiduoko lagina sortu dugu, indibiduo hauen %50 gaixoa eta %50 osasuntsua izanik. Bi egoeretan Σ_H eta Σ_D bariantzak identitate matrize bezala definitu ditugu, frogatuta dagoelako, honela X_1 eta X_2 aldagaiek *logit* p-rekiko linealtasuna beteko dutela [5]. 3.3. eta 3.4. irudietan ikus daiteke, linealtasun hau betetzen dela.



(a) (b)

3.3. irudia. $AUC_{teorikoa} = 0.8988$. Ezkerrean X1 aldagaiaren linealtasuna. Eskuinean X2 aldagaiaren linealtasuna.



(a) (b)

3.4. irudia. $AUC_{teorikoa} = 0.6408$. Ezkerrean X1 aldagaiaren linealtasuna. Eskuinean X2 aldagaiaren linealtasuna.

Ondoren, simulazioak egiteko, 10.000 indibiduoko lagin hauetatik azpilagin desberdinak hartu ditugu, bi AUC teoriko ezberdinentzat. Sortu ditugun azpilaginek 200, 500 eta 1000 indibiduo dituzte, eta bi modutan eraiki ditugu:

- Hasteko, indibiduoen %50 osasuntsu eta %50 gaixo dituzten azpilaginak sortu ditugu. Hau da, $N = 200$ tamainako lagina sortzeko, 100

osasuntsu aukeratu ditugu zoriz, X_H bektoretik, eta beste 100 gaixo, X_D bektoretik. Modu berean sortu ditugu $N = 500$ (250 osasuntsu eta 250 gaixo) eta $N = 1000$ (500 osasuntsu eta 500 gaixo) tamainako azpilaginak.

- Horrez gain, indibiduoen %90 osasuntsu eta %10 gaixo dituzten azpilaginak ere sortu ditugu. Horretarako, $N = 200$ indibiduoko azpilagina sortzeko, $200 \times 0.9 = 180$ osasuntsu eta $200 \times 0.1 = 20$ gaixo hartu ditugu zoriz X_H eta X_D bektoreetatik. Modu berean, $N = 500$ (450 osasuntsu eta 50 gaixo) eta $N = 1000$ (900 osasuntsu eta 100 gaixo) indibiduoko azpilaginak ere sortu ditugu.

Ondorengo eskeman ikus daiteke laburtuta, ditugun eszenario guztien multzoa.

$$
\begin{cases}
AUC_{teo.} = 0.8988 \begin{cases}
N = 200 \begin{cases} \%50 - \%50 : 100H - 100D & \text{Eszenario 1} \\ \%90 - \%10 : 180H - 20D & \text{Eszenario 2} \end{cases} \\
N = 500 \begin{cases} \%50 - \%50 : 250H - 250D & \text{Eszenario 3} \\ \%90 - \%10 : 450H - 50D & \text{Eszenario 4} \end{cases} \\
N = 1000 \begin{cases} \%50 - \%50 : 500H - 500D & \text{Eszenario 5} \\ \%90 - \%10 : 900H - 100D & \text{Eszenario 6} \end{cases}
\end{cases} \\
AUC_{teo.} = 0.6408 \begin{cases}
N = 200 \begin{cases} \%50 - \%50 : 100H - 100D & \text{Eszenario 7} \\ \%90 - \%10 : 180H - 20D & \text{Eszenario 8} \end{cases} \\
N = 500 \begin{cases} \%50 - \%50 : 250H - 250D & \text{Eszenario 9} \\ \%90 - \%10 : 450H - 50D & \text{Eszenario 10} \end{cases} \\
N = 1000 \begin{cases} \%50 - \%50 : 500H - 500D & \text{Eszenario 11} \\ \%90 - \%10 : 900H - 100D & \text{Eszenario 12} \end{cases}
\end{cases}
\end{cases}
$$

Ikus dezakegunez, guztira hamabi eszenario ditugu. Eszenario guztietan 2. kapituluan aipaturiko metodoak aplikatu ditugu.

Lehenik eta behin, balidazio behagarria egin dugu.

Ondoren, zatikako balidazioa egin dugu. Horretarako, N tamainako lagina $N/2$ tamainako bi azpilaginetan banatu dugu zoriz. Horietako batean, eredua doitu dugu, eta bestean balidatu.

Balidazio gurutzatua egiteko, N luzerako lagina $N/10$ luzerako hamar azpilaginetan banatu da. Bederatzi azpilagin erabili dira eredua doitzeko, eta azpilagin bat doitutako ereduaren balidaziorako. Prozesu hau hamar aldiz errepikatu da, aldi bakoitzean azpilagin ezberdin bat utzita balidaziorako.

Amaitzeko, bootstrap balidazioa burutzeko, N tamainako 100 bootstrap laginketa egin ditugu. Optimismoa bi modutara kalkulatu dugu, balio absoluturik gabe lehendabizi, eta balio absolutuarekin ondoren.

Balidazio metodo bakoitzetik 500 simulazio egin ditugu, eszenario bakoitzean. Simulazio hauek Steyerberg, 2001 artikuluan proposatutakoak dira [4].

3.2 Emaitzak

Lortutako emaitzak lau multzotan sailkatu ditugu AUC teorikoaren eta lagin bakoitzeko osasuntsu eta gaixo proportzioen arabera. Multzo bakoitzari dagozkion kutxa-diagramak eta laburpen-taula bana egin ditugu. 3.1., 3.2., 3.3. eta 3.4. tauletan eta 3.5., 3.6., 3.7. eta 3.8. irudietan ikus ditzakegu emaitzak. Irudiak interpretatzerako orduan, kontuan hartu behar dugu, ardatzak multzo batetik bestera aldatu egiten direla. Kutxa-diagrama guztiak ondo ikusteko antolatu ditugu modu honetan.

Egoera bakoitza banan-banan aztertu dugu. Lehenik eta behin, multzo guztiak kontuan hartuz, ikus dezakegu lagin tamaina handitu ahala, metodo guztiek hobeto funtzionatzen dutela. Hau da, lagin kopuru txikia denean ditugu arazo gehien, esperogarria zen moduan.

$AUC_{teorikoa} = 0.8988$ kontsideratu dugunean, ikusi dugu balidazio behagarria eta bootstrap balidazioa direla ondoen funtzionatzen duten metodoak. Balidazio behagarria erabilita ez dago gainestimaziorik eta bootstrap metodoaren bidez lortutako emaitzak ere oso hurbil daude AUC teorikotik. Bi metodoen desbiderapen estandarra eta alborapena oso antzekoak dira eta biak dira baxuak.

AUC teorikotik gehien urruntzen den metodoa balio absolutuaren bidez egindako bootstrap metodoa da. Metodo honen bidez AUCa gehiegi zuzentzen da. Antzeko zerbait gertatzen da balidazio gurutzatuarekin, lagina txikia denean. Balidazio gurutzatuaren bidez lortutako emaitzak AUC teorikoaren oso azpitik gelditzen dira. 3.1. eta 3.2. tauletan ikus dezakegunez, metodo honen alborapena handia da, batez ere, lagina txikia denean, lagina handitu ahala alborapena txikitzen doan arren. Horrez gain, osasuntsuen proportzioa %50 izatetik %90 izatera pasatzean, alborapena bikoiztu egin da.

Zatikako balidazioari dagokionez, aipagarria da, alborapena ez dela hain handia, baina oso sakabanapen handia duela, batez ere lagina txikia denean. Kasu honetan ere, metodoak askoz okerrago funtzionatzen du osasuntsuen proportzioa %90 denean %50 denean baino. Desberdintasun hauek oso nabariak dira 3.5. eta 3.6. irudiei begiratuta.

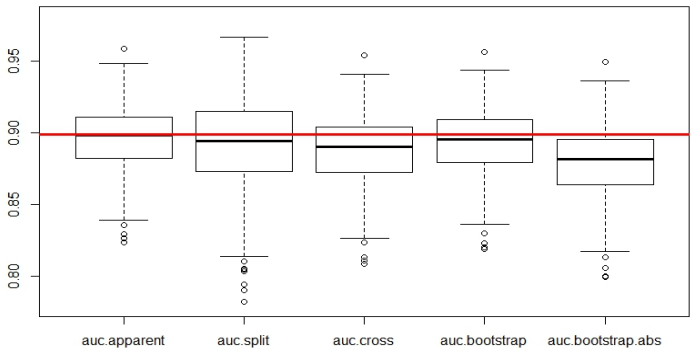
$AUC_{teorikoa} = 0.6408$ denean, balidazio behagarria gainestimatuta agertzen da, proposatutako eszenario ezberdin guztietan. Gainestimazio hau nabariagoa da osasuntsuen proportzioa %90 denean. Berriro ere, bistakoa da balio absolutu bidezko bootstrap metodoaren AUCaren gehiegizko zuzenketa eta gauza bera esan daiteke balidazio gurutzatuaren kasuan lagina oso handia ez denean. Gehiegizko zuzenketa hau are argiago ikusten da osasuntsuak indibiduoen %90 direnean.

Orokorrean, $AUC_{teorikoa} = 0.6408$ dugunean, metodo guztietan desbiderapena handiagoa da $AUC_{teorikoa} = 0.8988$ denean baino. Baina, desbiderapen hau zatikako balidazioan da handiena zalantzarik gabe, 3.3. eta 3.4. taulek erakusten duten bezala. 3.7. eta 3.8. irudietan ere nabarmena da zatikako balidazioa dela sakabanatuena.

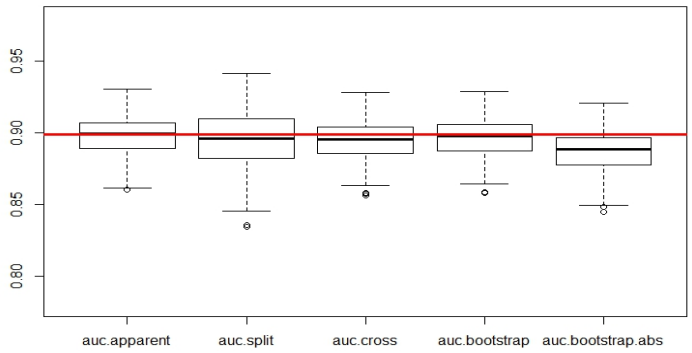
3.7. eta 3.8. irudietan ikus daitekeenez, egoera honetan AUC teorikora gehien hurbiltzen den metodoa bootstrap balidazioa dela esan dezakegu. 3.3. eta 3.4. taulek erakusten digute alborapena txikia dela kasu guztietan eta zatikako balidazioarekin alderatuz, desbiderapena txikia dela esan daiteke. Horrez gain, balidazio behagarriaren gainestimazioa zuzentzen du.

3.1. Taula. $AUC_{teorikoa} = 0.8988$ eta H-D: $\%50 - \%50$

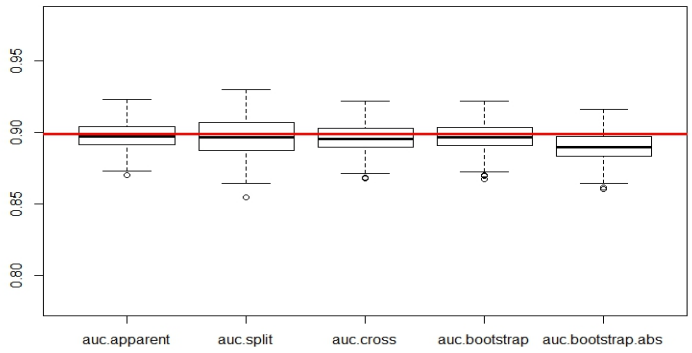
Metodoa	**Lagin tamaina**	**Batezbestekoa (sd)**	**Mediana**	**Alborapena**
Behagarria	N=200	0.8964 (0.0227)	0.8984	-0.0024
	N=500	0.8978 (0.0134)	0.8993	-0.0010
	N=1000	0.8976 (0.0095)	0.8973	-0.0012
Zatikakoa	N=200	0.8920 (0.0321)	0.8946	-0.0068
	N=500	0.8955 (0.0197)	0.8958	-0.0033
	N=1000	0.8966 (0.0138)	0.8964	-0.0022
Gurutzatua	N=200	0.8881 (0.0241)	0.8904	-0.0107
	N=500	0.8945 (0.0137)	0.8956	-0.0043
	N=1000	0.8960 (0.0096)	0.8957	-0.0028
Bootstrap	N=200	0.8937 (0.0232)	0.8956	-0.0051
	N=500	0.8967 (0.0135)	0.8980	-0.0021
	N=1000	0.8972 (0.0096)	0.8969	-0.0016
Balio abs.	N=200	0.8790 (0.0248)	0.8816	-0.0198
	N=500	0.8869 (0.0143)	0.8884	-0.0119
	N=1000	0.8899 (0.0099)	0.8898	-0.0089



(a)



(b)

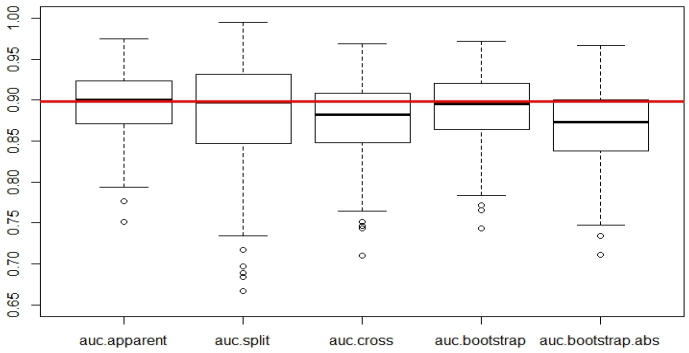


(c)

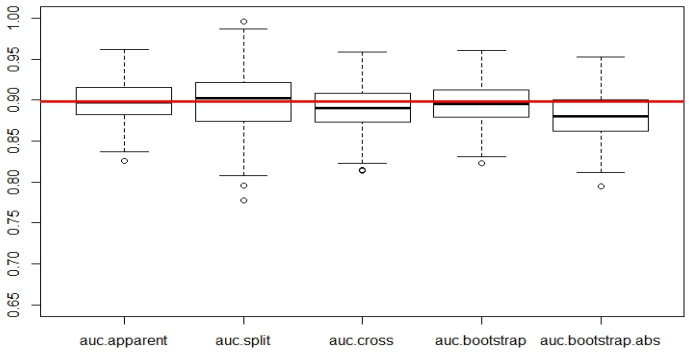
3.5. irudia. $AUC_{teorikoa} = 0.8988$. %50 osasuntsu - %50 gaixo. Lagin tamainak: (a) N=200; (b) N=500; (c) N=1000;

3.2. Taula. $AUC_{teorikoa} = 0.8988$ eta H-D: $\%90 - \%10$

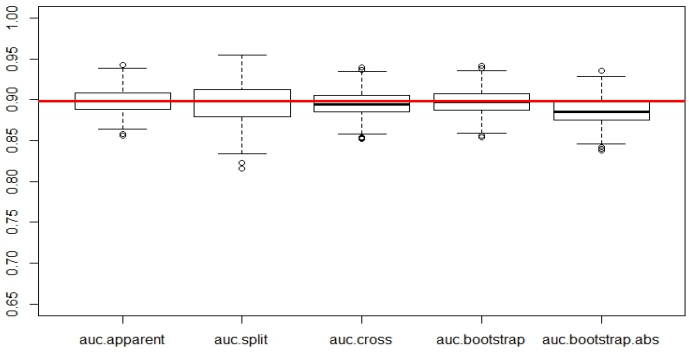
Metodoa	**Lagin tamaina**	**Batezbestekoa (sd)**	**Mediana**	**Alborapena**
Behagarria	N=200	0.8974 (0.0375)	0.9008	-0.0014
	N=500	0.8984 (0.0224)	0.8975	-0.0004
	N=1000	0.8980 (0.0155)	0.8984	-0.0008
Zatikakoa	N=200	0.8867 (0.0587)	0.8972	-0.0121
	N=500	0.8975 (0.0348)	0.9024	-0.0013
	N=1000	0.8970 (0.0235)	0.8988	-0.0018
Gurutzatua	N=200	0.8763 (0.0434)	0.8819	-0.0225
	N=500	0.8906 (0.0241)	0.8904	-0.0082
	N=1000	0.8943 (0.0159)	0.8946	-0.0045
Bootstrap	N=200	0.8914 (0.0396)	0.8957	-0.0074
	N=500	0.8962 (0.0227)	0.8959	-0.0026
	N=1000	0.8969 (0.0157)	0.8974	-0.0019
Balio abs.	N=200	0.8693 (0.0449)	0.8737	-0.0295
	N=500	0.8806 (0.0256)	0.8804	-0.0182
	N=1000	0.8853 (0.0168)	0.8854	-0.0135



(a)



(b)

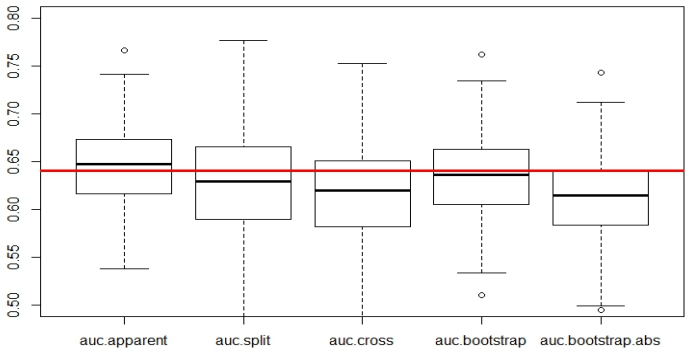


(c)

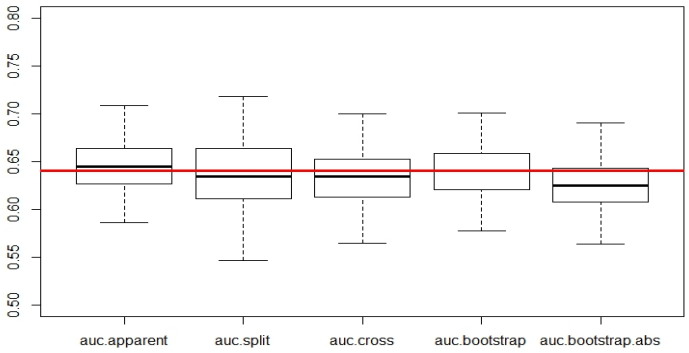
3.6. irudia. $AUC_{teorikoa} = 0.8988$. %90 osasuntsu - %10 gaixo. Lagin tamainak: (a) N=200; (b) N=500; (c) N=1000;

3.3. Taula. $AUC_{teorikoa} = 0.6408$ eta H-D: $\%50 - \%50$

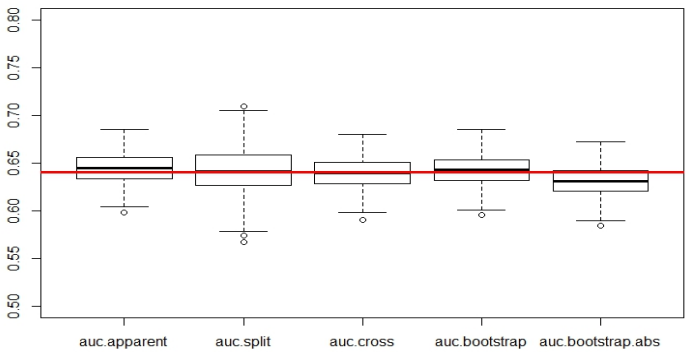
Metodoa	**Lagin tamaina**	**Batezbestekoa (sd)**	**Mediana**	**Alborapena**
Behagarria	N=200	0.6447 (0.0404)	0.6469	0.0039
	N=500	0.6445 (0.0241)	0.6446	0.0037
	N=1000	0.6450 (0.0162)	0.6448	0.0042
Zatikakoa	N=200	0.6249 (0.0610)	0.6289	-0.0159
	N=500	0.6363 (0.0351)	0.6346	-0.0045
	N=1000	0.6422 (0.0242)	0.6413	0.0014
Gurutzatua	N=200	0.6147 (0.0480)	0.6193	-0.0261
	N=500	0.6325 (0.0264)	0.6347	-0.0083
	N=1000	0.6392 (0.0170)	0.6392	-0.0016
Bootstrap	N=200	0.6332 (0.0437)	0.6361	-0.0076
	N=500	0.6399 (0.0250)	0.6408	-0.0009
	N=1000	0.6431 (0.0166)	0.6429	0.0023
Balio abs.	N=200	0.6130 (0.0425)	0.6146	-0.0278
	N=500	0.6250 (0.0245)	0.6248	-0.0158
	N=1000	0.6312 (0.0164)	0.6311	-0.0096



(a)



(b)

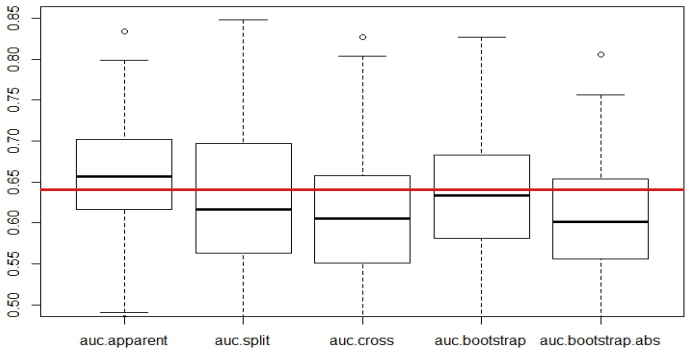


(c)

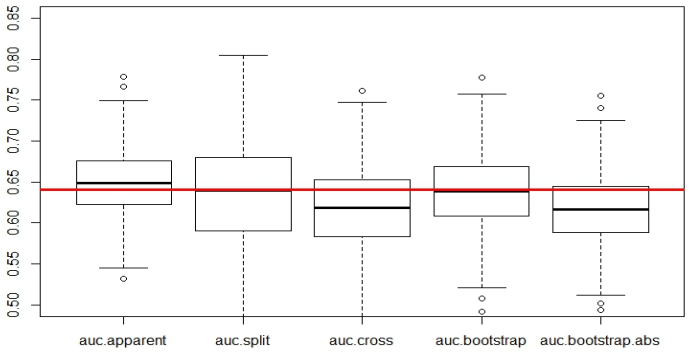
3.7. irudia. $AUC_{teorikoa} = 0.6408$. %50 osasuntsu - %50 gaixo. Lagin tamainak: (a) N=200; (b) N=500; (c) N=1000;

3.4. Taula. $AUC_{teorikoa} = 0.6408$ eta H-D: $\%90 - \%10$

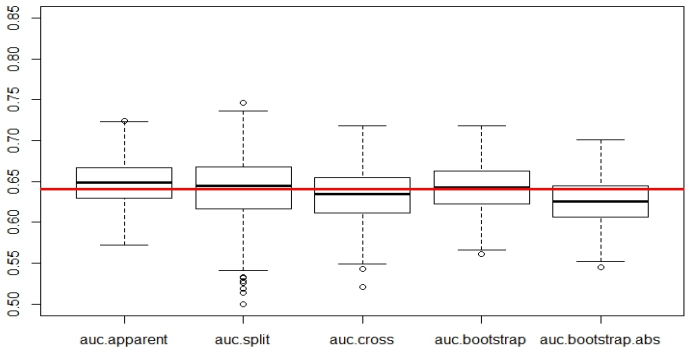
Metodoa	**Lagin tamaina**	**Batezbestekoa (sd)**	**Mediana**	**Alborapena**
Behagarria	N=200	0.6549 (0.0668)	0.6565	0.0141
	N=500	0.6496 (0.0407)	0.6490	0.0088
	N=1000	0.6484 (0.0272)	0.6488	0.0076
Zatikakoa	N=200	0.6270 (0.0908)	0.6168	-0.0138
	N=500	0.6342 (0.0679)	0.6400	-0.0066
	N=1000	0.6411 (0.0418)	0.6444	0.0003
Gurutzatua	N=200	0.6070 (0.0721)	0.6053	-0.0338
	N=500	0.6191 (0.0497)	0.6186	-0.0217
	N=1000	0.6329 (0.0307)	0.6346	-0.0079
Bootstrap	N=200	0.6250 (0.0810)	0.6335	-0.0158
	N=500	0.6381 (0.0444)	0.6381	-0.0027
	N=1000	0.6428 (0.0286)	0.6429	0.0020
Balio abs.	N=200	0.5985 (0.0768)	0.6012	-0.0423
	N=500	0.6160 (0.0436)	0.6165	-0.0248
	N=1000	0.6253 (0.0278)	0.6252	-0.0155



(a)



(b)



(c)

3.8. irudia. $AUC_{teorikoa} = 0.6408$. %90 osasuntsu - %10 gaixo. Lagin tamainak: (a) N=200; (b) N=500; (c) N=1000;

4. kapitulua

Aplikazioa

Aurreko kapituluetan lortutako emaitza teorikoak praktikan jartzeko, datubase erreal batekin lan egin dugu. Datu hauek Galdakaoko Ospitaleko ikerketa batean darabiltzaten datuak dira, bihotzaren kongestio-gutxiegitasunaren inguruko ikerketa batean, hain justu.

Lortu dugun datu-basean, erantzun aldagaia, eboluzio txarra da eta bi aldagai azaltzaile daude: bun, odoleko nitrogeno ureikoa, (X_1) eta odoleko pH-a (X_2). Bi aldagai azaltzaileak egiten ari diren ikerketan adierazgarriak direla baieztatu digute. Guztira 1168 gaixoz osatutako lagina da. Hauetatik 185 dira eboluzio txarra izan dutenak (gaixo guztien %16 gutxi gorabehera), gainerako 983 gaixoek ez dute eboluzio txarra izan (indibiduo guztien %84).

Bun eta pH aldagaien linealtasuna aztertu dugu lehenik eta behin. 4.1. irudian ikus ditzakegu. Argi eta garbi esan dezakegu bun lineala dela. pH ere bun bezain lineala ez izan arren, aldagai linealtzat har dezakegu. Beraz, zentzua du aldagai jarraitu bezala ereduan sartzeak.

Bi aldagai azaltzaile hauen eboluzio txarra izan dutenen ($\mathbf{X}_D$) eta eboluzio txarra izan ez dutenen ($\mathbf{X}_H$) bektoreen dentsitate funtzioak ere irudikatu ditugu. 4.2. irudian ikus ditzakegu.

Kasu honetan, X_H eta X_D bektoreek ez dute normaltasuna betetzen. Shapiro-Wilk-en normaltasun testa aplikatu diegu bi bektoreei eta bi kasuetan normaltasuna baztertu dugu, lortutako p-balioa < 0.05 delako.

Ondorioz, ezin izango dugu AUC teorikoa lortu, 1.1 formula aplikatzeko, nahitaezkoa baita X_H eta X_D bektoreen banaketa normala izatea.

Beraz, eredu honen behatutako balidazioa, zatikako balidazioa, balidazio gurutzatua eta bootstrap balidazioa (bai balio absoluturik gabe, bai balio

absolutuarekin) egin ditugu. 4.1. taulan daude lortu ditugun emaitzak:

Balidazio metodoa	**AUC**
Behatutakoa	0.6875395
Zatitutakoa	0.7053817
Gurutzatutakoa	0.6843914
Bootstrap	0.6851439
Bootstrap (balio absolutua)	0.6675082

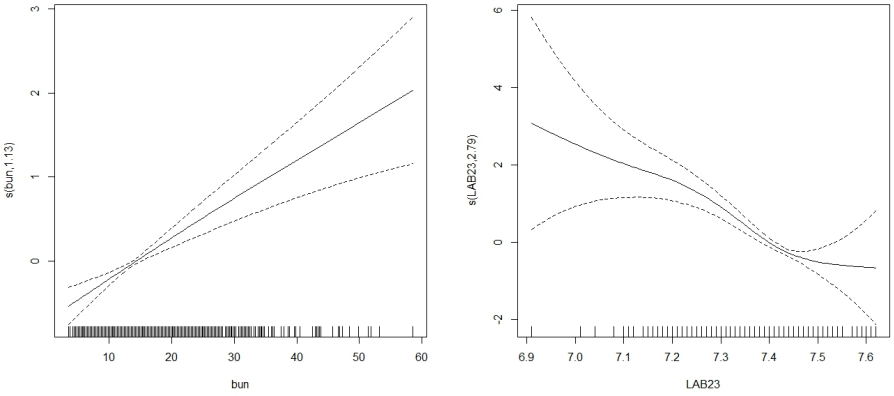
4.1. Taula. Datu errealei balidazio metodo ezberdinak aplikatuta lortutako emaitzak.

Emaitzei erreparatuta, nabaria da bootstarap balidazioak balio absolutuarekin ez duela ondo funtzionatzen, gainerako AUC guztiekin konparatuta, oso azpitik geratu baita 0.6675ko AUCarekin, simulazioetan gertatzen zen bezala.

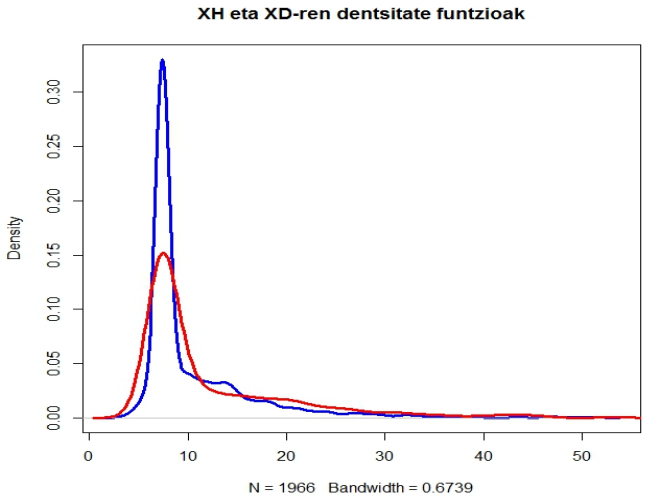
Behatutako, gurutzatutako eta bootstrap balidazioekin lortu ditugun emaitzak, oso antzekoak dira. Simulazioetan ikusi dugu behatutako balidazioak nahiko ondo funtzionatzen duela lagina handia denean baina gainestimaziorako joera duela. Bootstrap balidazioak gainestimazio hau zuzentzen du. Kasu honetan ere behatutako balidazioarekin lortu dugun AUCa (0.6875) bootstraparekin lortutakoaren gainetik dago (0.6851). Bestalde, gurutzatutako balidazioaren bidez lagin handietan emaitza hobeak lortu ditugun arren, ikusi dugu askotan gehiegizko zuzenketa egiteko joera duela, eta kasu partikular honetan ikus dezakegunez ere, bootstrap balidazioarekin lortutako AUCaren azpitik dago (0.6844).

Azkenik, zatikako balidazioaren bidez lortutako AUCaren gainestimazioa nabarmena da (0.7054). Simulazioetan ikusi dugu zatikako balidazioaren sakabanapena oso handia dela eta ondorioz emaitza ez da fidagarria, gerta baitaiteke, kasu honetan bezala, metodo honek emandako AUCa handiegia izatea.

Ondorioz, informazio hau guztia laburtuz, eredu honen aurresateko gaitasuna zein den zehazteko, bootstrap bidezko balidazioa da fidagarriena. Hau da, ereduaren AUCa 0.6851 dela esan dezakegu.



4.1. irudia. Bun aldagaiaren (ezkerrean) eta LAB23 aldagaiaren (eskuinean) linealtasuna.



4.2. irudia. X_H-ren (lerro urdina) eta X_D-ren (lerro gorria) dentsitate funtzioak.

5. kapitulua

Ondorioak

Aurreko kapituluetan jasotako informazio guztia aztertu eta hausnartu ondoren, hainbat konklusio atera ditugu. Ondorengo lerroetan daude azalduta lortu ditugun ondorioak.

Hasteko, behatutako balidazioak AUC teorikoa altua denean ondo funtzionatzen duela ikusi dugu, lortutako emaitzak egonkorrak dira eta ez dago gainestimaziorik. Hala ere, AUC teorikoa oso altua ez den kasu guztietan, begi bistakoa da gainestimazioa agertzen dela. Hau izan daiteke, AUCa altua denean behatutako balidazioa emaitza egokiak emateko prest dagoelako. Aldiz, AUCa hain altua ez denean emaitza baikorregiak lortuko ditugu. Kontuan izan behar dugu, normalean hain altuak ez diren AUCekin lan egin beharko dugula. Hau da, AUC teorikoa gehienetan ez da hain altua izango eta, horrez gain, normalean indibiduoen erdiak ez dira gaixoak izango. Beraz, behatutako balidazioarekin lan egitean gogoan izan behar dugu lortutako AUCa benetakoa baino altuagoa izan daitekeela.

Zatikako balidazioari dagokionez, aipagarria da askotan alborapena oso handia ez izan arren, oso sakabanatua dela ia kasu guztietan. Sakabanapen hau are argiago ikusten da lagina txikia denean eta osasuntsu eta gaixoen proportzioak berdinak ez direnean. Izan ere, ohar gaitezen $N = 200$ indibiduoko lagin batekin %90 osasuntsu eta %10 gaixo baldin baditugu, guztira gaixo kopurua 20 baino ez dela eta hauek bi laginetan banatu behar direla. Ondorioz, kopuru horietan gaixotasuna duen indibiduo bat gorabehera izateak diferentzia nabarmenak suposa ditzake ereduaren estimazioan. Talde hauen banaketak AUCaren neurketan eragin zuzena duenez, oso emaitza ezberdinak lor ditzakegu zatikako balidazioa erabiliz. Urrutira joan gabe, 4. kapituluko aplikazioan ikusi dugu metodo honen bidez lortutako AUCa altuegia izan dela. Hau dela eta, esan dezakegu zatikako balidazioa ez dela oso metodo fidagarria eredu baten aurresateko gaitasuna neurtzeko.

Balidazio gurutzatuari buruz hitz egitean, gehiegizko zuzenketa egiten duela azpimarratzea beharrezkoa da. Kontuan izan behar dugu, zatikako balidazioan lagina bi zatitan banatzen genuen moduan, gurutzatutako balidazioan hamar zati egiten ditugula eta baliteke gehiegizko zuzenketa hori honen ondorio izatea. Izan ere, lehen aipatu dugun moduan, talde hauen banaketak eragin handia du AUCaren kalkuluan. Hala ere, metodo honen bidez prozesua hamar aldiz errepikatzen denez multzo bakoitza behin erabiliz balidaziorako, lortzen diren emaitzak ez daude hain sakabanatuta. Baina errepikapen kopuru hau ez da nahikoa gehiegizko zuzenketa hori konpontzeko. Orduan, interesgarria izan daiteke Steyerberg, 2001 [4] artikuluan proposatutako beste simulazio batekin emaitzak hobetzen diren edo ez ikustea. Artikulu honetan, guk egindako simulazioez gain, proposatzen du gurutzatutako balidazioan egindako guztia 10 aldiz errepikatzea. Hau da, metodo honen bidez 10 errepikapen egin ordez (errepikapen bakoitzean multzo ezberdin bat utziz balidaziorako), 100 errepikapen egitea da bere proposamena.

Balio absoluturik gabeko bootstrap balidazioa da eszenario guztietan ondoen funtzionatu duen metodoa. Oso alborapen txikia eman digu kasu guztietan eta lortutako emaitzak ez dira oso sakabanatuak, balidazio behagarriaren antzeko sakabanapena du eszenario guztietan. Gogora dezagun metodo honetan sortzen ditugun lagin berriak, gure laginaren tamaina berekoak direla eta jatorrizko laginetik sortzen ditugula, errepikapenak onartuz. Ondorioz, osasuntsu eta gaixoen proportzioak antzeko mantenduko direla pentsa dezakegu, sortutako lagin berri guztietan. Hau abantaila handia da zatikako balidazioa eta balidazio gurutzatuarekin konparatuta. Horrez gain, ikus dezakegu, balidazio behagarriaren gainestimazioa ondo zuzentzen duela metodo honek, eta gainera, zuzenketa hau ez da gehiegizko zuzenketa. Balidazio behagarriaren eta bootstraparen arteko desberdintasun hau AUC oso altua ez denean ikus dezakegu batez ere, balidazio behagarriak arazoak ematen zizkigun kasuak hobetzen baitizkigu.

Bestalde, bootstrap balidazioa balio absolutuarekin erabiltzen badugu, AUCaren zuzenketa gehiegizkoa da. Sakabanapena txikia izan arren, alborapena oso handia da metodo honekin, edozein eszenariotan. Gainerako metodo guztiekin lortutako emaitzekin konparatuta, AUCa benetakoaren oso azpian gelditzen da. Argi dago beraz, aldagai jarraituak eta *logit* p-rekiko linealak ditugun kasuan, balio absolutu bidezko bootstrap balidazioak ez digula balio.

Laburtuz, gaur egun ikerkuntza medikoan zatikako balidazioa izan arren erabiliena, ikusi dugu aldagaiak jarraituak eta *logit* p-rekiko linealak diren kasuan ezin gaitezkeela fidatu metodo honek ematen dizkigun emaitzekin. Sakabanapena hain handia izanik, ohikoa da AUC teorikotik oso urrun dauden balioak lortzea metodo honen bidez eta ondorioz, doitutako ereduaren

aurresateko gaitasuna ez da izango guk uste duguna, horrek izan ditzakeen ondorio guztiekin. Hau ekiditeko, bootstrap balidazioa erabiltzea da gure proposamena, ikusi baita metodo honen bidez lortzen diren emaitzak askoz hobeak direla eta beraz, guk doitutako ereduaren balidazioa metodo honen bidez egitean, AUC teorikotik oso hurbil dauden emaitzak lortuko ditugu. Etorkizunean, interesgarria izan daiteke aldagaien eta *logit* p-ren arteko erlazioa lineala ez den kasuan zer gertatzen den ikustea ere.

Azkenik, lan hau bi kongresu hauetan onartua izan da: *XXXVI Congreso Nacional de Estadística e Investigación Operativa* eta *II Jornadas de jóvenes estudiantes de la Sociedad Española de Biometría.*

Bibliografia

[1] D.W. Hosmer & S. Lemeshow, *Applied Logistic Regression*, Wiley, 1989.

[2] Ewout W Steyerberg. *Clinical prediction models: a practical approach to development, validation, and updating.* Springer, 2008

[3] John Q. Su & Jun S. Liu."Linear Combinations of Multiple Diagnostic Markers". Journal of the American Statistical Association, December 1993, Vol. 88, No. 424, Theory and Methods.

[4] Ewout W. Steyerberg, Frank E. Harrell Jr, Gerard J. J. M. Borsboom, M. J. C. (René) Eijkemans, Yvonne Vergouwe, J. Dik F. Habbema. "Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis". Journal of Clinical Epidemology 54 (2001) 774-781.

[5] Barrio I. *Proposal and validation of methodologies for the categorisation of continuous variables in the development of prediction model.* ADDI, 2015. http://hdl.handle.net/10810/17542.

A. eranskina

R-ko kodea

```
#Kargatu beharreko paketeak
library(MASS) #mvrnorm funtzioa erabili ahal izateko
library(pROC) #roc funtzioa erabili ahal izateko



#AUC teorikoa kalkulatzeko formula:
#pnorm(sqrt(t(MU)%*%solve(sigma_H+sigma_D)%*%MU))

#Balidazio behagarria egiteko funtzioa:
apparent.val <- function(datuak){
fit.app <- glm(Y ~ X1+ X2, data=datuak, family=binomial)
auc.app<- roc(fit.app$y, fit.app$fitted)$auc[1]
auc.app
}


#Zatikako balidazioa egiteko funtzioa
split.val <- function(datuak,n){

N <- vector(l=n)
for(i in 1:n){N[i]=i}

N1 <- sample(N,n/2,replace=FALSE)
N2 <- N[-N1]

datuak.deriv <- datuak[N1,]
datuak.valid <- datuak[N2,]

fit.deriv <- glm(Y ~ X1 + X2, family = binomial,
```

```
            data=datuak.deriv)
fit.valid <- predict(fit.deriv, newdata=datuak.valid,
            type='response')
auc.split <- roc(datuak.valid$Y, fit.valid)$auc[1]
auc.split
}


#Balidazio gurutzatua egiteko funtzioa:
cross.val <- function(datuak,n){

N <- vector(l=n)
for(i in 1:n){N[i]=i}

N1 <- sample(N,n/10,replace=FALSE)
N2 <- sample(N[-N1],n/10,replace=FALSE)
N3 <- sample(N[-c(N1,N2)],n/10,replace=FALSE)
N4 <- sample(N[-c(N1,N2,N3)],n/10,replace=FALSE)
N5 <- sample(N[-c(N1,N2,N3,N4)],n/10,replace=FALSE)
N6 <- sample(N[-c(N1,N2,N3,N4,N5)],n/10,replace=FALSE)
N7 <- sample(N[-c(N1,N2,N3,N4,N5,N6)],n/10,replace=FALSE)
N8 <- sample(N[-c(N1,N2,N3,N4,N5,N6,N7)],n/10,replace=FALSE)
N9 <- sample(N[-c(N1,N2,N3,N4,N5,N6,N7,N8)],n/10,replace=FALSE)
N10 <- sample(N[-c(N1,N2,N3,N4,N5,N6,N7,N8,N9)],n/10,replace=FALSE)


datuak.c1 <- datuak[N1,]
datuak.c2 <- datuak[N2,]
datuak.c3 <- datuak[N3,]
datuak.c4 <- datuak[N4,]
datuak.c5 <- datuak[N5,]
datuak.c6 <- datuak[N6,]
datuak.c7 <- datuak[N7,]
datuak.c8 <- datuak[N8,]
datuak.c9 <- datuak[N9,]
datuak.c10 <- datuak[N10,]


fit.valid <- vector(l=n)

datuak1 <- datuak[c(N1,N2,N3,N4,N5,N6,N7,N8,N9),]
fit.deriv1 <- glm(Y ~ X1 + X2, family = binomial, data=datuak1)
fit.valid[N10] <- predict(fit.deriv1, newdata=datuak.c10,
                  type='response')
```

```
datuak2 <- datuak[c(N1,N2,N3,N4,N5,N6,N7,N8,N10),]
fit.deriv2 <- glm(Y ~ X1 + X2, family = binomial, data=datuak2)
fit.valid[N9] <- predict(fit.deriv2, newdata=datuak.c9,
                 type='response')

datuak3 <- datuak[c(N1,N2,N3,N4,N5,N6,N7,N9,N10),]
fit.deriv3 <- glm(Y ~ X1 + X2, family=binomial, data=datuak3)
fit.valid[N8] <- predict(fit.deriv3, newdata=datuak.c8,
                 type='response')

datuak4 <- datuak[c(N1,N2,N3,N4,N5,N6,N8,N9,N10),]
fit.deriv4 <- glm(Y ~ X1 + X2, family=binomial, data=datuak4)
fit.valid[N7] <- predict(fit.deriv4, newdata=datuak.c7,
                 type='response')

datuak5 <- datuak[c(N1,N2,N3,N4,N5,N7,N8,N9,N10),]
fit.deriv5 <- glm(Y ~ X1 + X2, family=binomial, data=datuak5)
fit.valid[N6] <- predict(fit.deriv5, newdata=datuak.c6,
                 type='response')

datuak6 <- datuak[c(N1,N2,N3,N4,N6,N7,N8,N9,N10),]
fit.deriv6 <- glm(Y ~ X1 + X2, family=binomial, data=datuak6)
fit.valid[N5] <- predict(fit.deriv6, newdata=datuak.c5,
                 type='response')

datuak7 <- datuak[c(N1,N2,N3,N5,N6,N7,N8,N9,N10),]
fit.deriv7 <- glm(Y ~ X1 + X2, family=binomial, data=datuak7)
fit.valid[N4] <- predict(fit.deriv7, newdata=datuak.c4,
                 type='response')

datuak8 <- datuak[c(N1,N2,N4,N5,N6,N7,N8,N9,N10),]
fit.deriv8 <- glm(Y ~ X1 + X2, family=binomial, data=datuak8)
fit.valid[N3] <- predict(fit.deriv8, newdata=datuak.c3,
                 type='response')

datuak9 <- datuak[c(N1,N3,N4,N5,N6,N7,N8,N9,N10),]
fit.deriv9 <- glm(Y ~ X1 + X2, family=binomial, data=datuak9)
fit.valid[N2] <- predict(fit.deriv9, newdata=datuak.c2,
                 type='response')

datuak10 <- datuak[c(N2,N3,N4,N5,N6,N7,N8,N9,N10),]
fit.deriv10 <- glm(Y ~ X1 + X2, family=binomial, data=datuak10)
fit.valid[N1] <- predict(fit.deriv10, newdata=datuak.c1,
```

```
                type='response')

auc.cross <- roc(datuak$Y, fit.valid)$auc[1]
auc.cross
}

##Bootstrap lagina
bootstrap.sample <- function(data, group) {
res <- data[sample(nrow(data), replace=TRUE),]
res
}

###Boostrap balidazioa egiteko funtzioa:
auc.boot <- function(datuak, AUC) {
auc.boot <- auc.original <- vector(l=100)
for (i in 1:100) {

data.b <- bootstrap.sample(data.frame(X1 = datuak$X1,
          X2 = datuak$X2, Y=datuak$Y),"Y")
# Bootstrap
fit.boot <- try(glm(Y~X1 + X2, data=data.b, family=binomial))
if("try-error" %in% class(fit.boot)){
print(cat(">>>>>>>", i, ">>>>>>>", " FIT boot " ))
auc.boot[i] <- NA
auc.original[i] <- NA

} else {
auc.boot[i] <- roc(fit.boot$y, fit.boot$fitted)$auc[1]
# Original Sample
fit.orig <- try(predict(fit.boot, newdata = datuak,
            type = "response"))
if("try-error" %in% class(fit.orig)){
print(cat(">>>>>>>", i, ">>>>>>>", " FIT orig"))
auc.original[i] <- NA
} else {
auc.original[i] <- roc(datuak$Y, fit.orig)$auc[1]
}
}
}
auc.corrected <- AUC - mean(auc.boot - auc.original, na.rm=TRUE)
auc.corrected
}
```

```
### Boostrap balio absolutuarekin egiteko funtzioa:
auc.boot.abs <- function(datuak, AUC) {
auc.boot <- auc.original <- vector(l=100)
for (i in 1:100) {

data.b <- bootstrap.sample(data.frame(X1 = datuak$X1,
          X2 = datuak$X2, Y=datuak$Y),"Y")
# Bootstrap
fit.boot <- try(glm(Y~X1 + X2, data=data.b, family=binomial))
if("try-error" %in% class(fit.boot)){

print(cat(">>>>>>>", i, ">>>>>>>", " FIT boot - abs " ))
auc.boot[i] <- NA
auc.original[i] <- NA

} else {

auc.boot[i] <- roc(fit.boot$y, fit.boot$fitted)$auc[1]
# Original Sample
fit.orig <- try(predict(fit.boot, newdata = datuak,
            type = "response"))
if("try-error" %in% class(fit.orig)){
print(cat(">>>>>>>", i, ">>>>>>>", " FIT orig - abs"))
auc.original[i] <- NA
} else {
auc.original[i] <- roc(datuak$Y, fit.orig)$auc[1]
}
}

}
auc.corrected <- AUC - mean(abs(auc.boot - auc.original), na.rm=TRUE)
auc.corrected
}

###Simulazioak
simulazioak.azpilaginak <- function(n=10000, nsim=500, nazpi,
                           prop_H=0.5, prop_D=0.5, prop_H_nazpi,
                           prop_D_nazpi, auc){

denb1 <- Sys.time()

if (auc==89){
mu_H <- c(0,1)
mu_D <- c(1.5,2)
```

```
MU <- mu_D-mu_H
sigma_D <- matrix(c(1,0,0,1), byrow=TRUE, nrow=2, ncol=2)
sigma_H <- matrix(c(1,0,0,1), byrow=TRUE, nrow=2, ncol=2)
}

if (auc==64){
mu_H <- c(0,0.5)
mu_D <- c(0.5,0.6)
MU <- mu_D-mu_H
sigma_D <- matrix(c(1,0,0,1), byrow=TRUE, nrow=2, ncol=2)
sigma_H <- matrix(c(1,0,0,1), byrow=TRUE, nrow=2, ncol=2)
}


set.seed(11235)
XH <- mvrnorm(n*prop_H,mu=mu_H,Sigma=sigma_H)
XD <- mvrnorm(n*prop_D,mu=mu_D,Sigma=sigma_D)
datuakH <- data.frame(XH)
datuakD <- data.frame(XD)


N <- vector(l=n)
for(i in 1:n){N[i]=i}

auc.apparent <- vector(l=nsim)
auc.split <- vector(l=nsim)
auc.cross <- vector(l=nsim)
auc.bootstrap <- vector(l=nsim)
auc.bootstrap.abs <- vector(l=nsim)


auc.teorikoa <- pnorm(sqrt(t(MU)%*%solve(sigma_H+sigma_D)%*%MU))

seeds <- runif(nsim)*1000


for (i in 1:nsim){
hasi <- Sys.time()

print(cat(">>>>>>>", i, ">>>>>>>"))

set.seed(seeds[i])

datuakHazpi <- datuakH[sample(nrow(datuakH),nazpi*prop_H_nazpi,
```

```
               replace=FALSE),]
datuakDazpi <- datuakD[sample(nrow(datuakD),nazpi*prop_D_nazpi,
               replace=FALSE),]
Y <- c(rep(0,nazpi*prop_H_nazpi),rep(1,nazpi*prop_D_nazpi))
datuak <- data.frame(Y,rbind(datuakHazpi,datuakDazpi))


options(warn=2)
app.val <-  try(apparent.val(datuak))
spl.val <-  try(split.val(datuak,nazpi))
crr.val <-  try(cross.val(datuak,nazpi))
boot.val <- try(auc.boot(datuak,app.val))
boot.abs.val <- try(auc.boot.abs(datuak,app.val))

if("try-error" %in% class(app.val)){
auc.apparent[i]<-NA
print("Warning atera da apparent validation-en")
} else {
auc.apparent[i] <- app.val
}

if("try-error" %in% class(spl.val)){
auc.split[i]<-NA
print("Warning atera da split validation-en")
} else {
auc.split[i] <- spl.val
}

if("try-error" %in% class(crr.val)){
auc.cross[i]<-NA
print("Warning atera da cross validation-en")
} else {
auc.cross[i] <- crr.val
}

auc.bootstrap[i] <- boot.val
auc.bootstrap.abs[i] <- boot.abs.val


bukatu <- Sys.time()
print(bukatu-hasi)
}
```

```
denb2 <- Sys.time()
print("Guztira:")
print(denb2 - denb1)

emaitza <- list(auc.teorikoa=auc.teorikoa,
           auc.apparent=auc.apparent, auc.split=auc.split,
           auc.cross=auc.cross, auc.bootstrap=auc.bootstrap,
           auc.bootstrap.abs=auc.bootstrap.abs)
emaitza
}
```