Ekaia, 2024, 45, 327-344
https://doi.org/10.1387/ekaia.24696

ekaia
ZIENTZIA eta TEKNOLOGIA ALDIZKARIA

ISSN 0214-9001 – eISSN 2444-3255

Hosmer-Lemeshow testean erabilitako talde-kopuruaren azterketa simulazioen bidez

(Analysis of the number of groups used in the Hosmer-Lemeshow test using simulations)

*Ane Moreno-Oya**[1,2], *Irantzu Barrio Beraza*[1,3]

[1] Matematika Saila, Euskal Herriko Unibertsitatea UPV/EHU
[2] Spanish National Cancer Research Centre (CNIO), Madrid, and CIBERONC, Spain
[3] Basque Center for Applied Mathematics, BCAM

LABURPENA: Egun, Hosmer-Lemeshow (HL) testa erregresio logistikoko ereduen doikuntza-egokitasuna neurtzeko maiz erabiltzen den hipotesi-kontrastea da. Ordea, lagin tamainarekin lotuta dauden hainbat eragozpen ditu eta, hori dela eta, azken urteetan eraldaketa ugari jasan ditu. Lan honetan, *g* talde kopurua aldatuta, testaren erabakien egonkortasuna aztertu dugu. Aukeratutako eredua egokia denean, HL testaren errendimendua ona dela lortu dugu. Gainera, ez da lagin tamainaren araberakoa. Bestalde, egoera honetan, proposatutako talde kopuru gomendatuaren erabilerak ez du eragin nabarmenik. Aldiz, doitutako eredua desegokia denean, HL testa lagin tamainarekiko sentikorra da eta, batez ere lagin txikietan, errendimendua eskasa da. Honetaz gain, lagin handietan bai gaixotasunaren prebalentziak bai ereduaren konplexutasunak eragina dute.

HITZ GAKOAK: Hosmer-Lemeshow, erregresio logistikoa, doikuntza-egokitasuna, lagin tamaina, talde kopurua.

ABSTRACT: *Nowadays, the Hosmer-Lemeshow (HL) test is a tool often used to measure the goodness of fit of logistic regression models. However, it has several inconveniences associated with sample size. As a result, it has undergone many modifications in recent years. In this work, by changing the number of groups, we have analyzed the stability of the test's decisions. When the model is correct, the performance of the HL test is good. What's more, it doesn't depend on the sample size. On the other hand, in this case, the use of the proposed number of groups has no significant effect on the results. When the chosen model is misspecified, the performance of the HL test is poor and it is sensitive to sample size. Above all, its performance is poor in small samples. Apart from that, in big samples, both prevalence and the model's complexity have an effect on the decisions of the test.*

KEYWORDS: Hosmer-Lemeshow, logistic regression, goodness-of-fit, sample size, number of groups.

**** Harremanetan jartzeko / Corresponding author:*** Ane Moreno-Oya. Zientzia eta Teknologia Fakultatea UPV/EHU, Sarriena auzoa, z/g (48940 Leioa-Bizkaia). – anemorenooya@gmail.com

Nola aipatu / How to cite: Moreno-Oya, Ane; Barrio Beraza, Irantzu (2024). «Hosmer-Lemeshow testean erabilitako talde-kopuruaren azterketa simulazioen bidez». *Ekaia*, 45, 2024, 327-344. (https://doi.org/10.1387/ekaia.24696).

Jasotze-data: 2023, martxoak 28; Onartze-data: 2023, ekainak 29.

ISSN 0214-9001 - eISSN 2444-3255 / © 2024 UPV/EHU

1. **SARRERA**

Gaur egun, erregresio logistikoa datu bitarrak modelizatzeko gehien erabiltzen den metodo estatistikoetako bat da.

Behin eredua doituta, aukeratutako aldagai azaltzaileak adierazgarriak direla suposatuz, ezin-bestekoa da estimatutako probabilitateek datuetan behatutako erantzun aldagaiaren balioak adierazten dituztela egiaztatzea. Eredu baten doikuntza-egokitasuna aztertzeko bi irizpide nagusi daude: diskriminazioa eta kalibrazioa. Diskriminazio ona duen eredu batek doitasunez bereizten du noiz gertatzen den interesekoa dugun gertaera. Bestalde, kalibrazio ona duen eredu batek zehaztasunez estimatzen ditu probabilitateak.

Erabakiak askotan gertaera bat jasateko arriskuan oinarritzen dira eta, beraz, ereduak estimatutako probabilitateak fidagarriak izan beharko lirateke. Gainera, kalibrazio eskasak eredu bat praktikan alferrikako bilaka dezake, baita kaltegarria ere [1]. Bestalde, kalibrazioa gainestimazioa saihesteko erabiltzen da [2].

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

testean estimatutako probabilitateak sailkatzeko g talde kopuru jakin bat erabiltzen da.

HL testak lagin tamainarekin lotutako hainbat muga ditu. Alde batetik, taldeak sortzerako orduan, behaketa kopurua txikia bada, gerta liteke ez ohartzea ereduaren doikuntza ezegokia dela [3, 4].

Bestalde, 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 [4, 5]. Izan ere, testaren ahalmena behaketa kopuruarekin handitzen da.

Azken urteetan HL testak aldaketa ugari jasan ditu azaldutako eragozpenak gainditzeko helburuarekin (ikusi [6, 7, 8, 9]).

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.

Aurkezten dugun lana lau atal nagusitan banatuta dago. Metodoen atalean, erregresio logistikoa eta HL testaren oinarri teorikoa azaltzen dira. Simulazioen atalean, testaren egonkortasuna aztertzeko planteatu ditugun egoera ezberdinak laburtzen dira. Emaitzen atalean, simulazioetan lortutako emaitzak grafikoki eta taula batean adierazita daude. Azkenik, eztabaidaren atalean, emaitzen ondorioak azaltzen dira.

2. METODOAK

Metodoen atala bi azpiataletan banatuta dago. Lehenengo azpiatalean, erregresio logistikoaren sarrera egiten da, HL testa hobeto uler dadin. Bigarren azpiatalean, HL testa eta HL test eraldatua azaltzen dira.

2.1. Erregresio logistikoa

Izan bitez $X_1, \ldots, X_p$ p zorizko aldagai aske, $\mathbf{X} = (X_1, \ldots, X_p)^t$ eta Y erantzun aldagai bitarra. Izan bedi $p(\mathbf{X}) = P(Y = 1|\mathbf{X})$ arrakastaren probabilitate baldintzatua. Orduan, erregresio logistikoko ereduaren *logit* transformazioa

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

da, non

$$p(\mathbf{X}) = \frac{e^{f(\mathbf{X})}}{1+e^{f(\mathbf{X})}} \tag{2}$$

den.

$\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^t$ bektorearen balioak estimatzeko, ($\mathbf{X}$, Y) aldagaien n tamainako lagina emanda, egiantz handieneko metodoa eta haztatutako karratu txikien metodoan oinarritutako algoritmo iteratiboak erabiltzen dira [10]. $\boldsymbol{\beta}$-ren estimatzailea $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \ldots, \hat{\beta}_p)^t$ denotatuko dugu.

2.2. Hosmer-Lemeshow pertzentilen testa

Demagun ($\mathbf{X}$, Y) aldagaien n tamainako lagina dugula, $\{(\mathrm{x}_i, y_i)\}_{i=1}^{n}$.

Izan bedi J $\mathbf{X}$ aldagai-bektorearen behatutako balio ezberdinen kopurua. Orduan, ereduak estimatutako probabilitateak txikienetik handienera ordenatu ostean, estimatutako probabilitateak pertzentilak erabilita taldekatzen dira, g talde osatuz.

$\hat{C}$ estatistikoak behatutako eta itxarotako probabilitateen arteko desberdintasuna adierazten du:

$$\hat{C} = \sum_{k=1}^{g}\left[\frac{(o_{1k} - e_{1k})^2}{e_{1k}} + \frac{(o_{0k} - e_{0k})^2}{e_{0k}}\right], \tag{3}$$

non e_{1k} eta o_{1k} k. taldeko arrakastarako ($Y = 1$) itxarotako eta behatutako maiztasunak eta e_{0k} eta o_{0k} k. taldeko porroterako ($Y = 0$) itxarotako eta behatutako maiztasunak diren, hurrenez hurren.

https://doi.org/10.1387/ekaia.24696

$J = n$ eta aukeratutako ereduaren doikuntza egokia denean, $\hat{C}$ estatistikoak χ^2_{g-2} anaketa asintotikoa du [5].

Orduan,

$$\begin{cases} H_0 : \text{Ereduaren doikuntza egokia da.} \\ H_1 : \text{Ereduaren doikuntza ezegokia da.} \end{cases} \tag{4}$$

hipotesi-kontrastea egiteko erabiliko dugun estatistikoa $\hat{C}$ da. Ondorioz, (4) testaren *p-balioa*

$$p = \int_{\hat{C}}^{\infty} \chi^2_{g-2}(z)dz \tag{5}$$

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

Subjektibitatea eta testaren gehiegizko ahalmen estatistikoa saihesteko, behaketa kopurua 1.000 eta 25.000 artean badago, g talde kopurua

$$g = \max\left(10, \min\left\{\frac{n_1}{2}, \frac{n-n_1}{2}, 2+8\left(\frac{n}{1000}\right)^2\right\}\right) \tag{6}$$

ekuazioaren arabera aukeratzea proposatu dute Paul *et al.* (2013) ikertzaileek, n_1 lagineko arrakasta kopuru totala izanik, $n_1 = \sum_{i=1}^{n} y_i$.

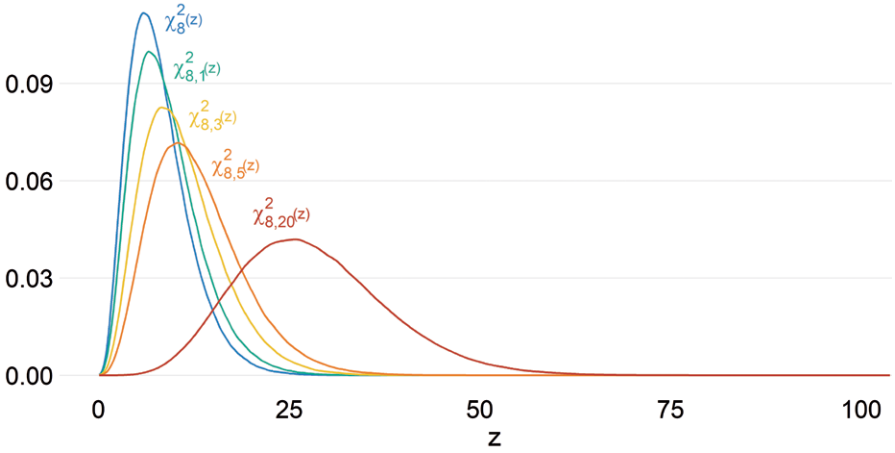
Ikertzaile horien arabera, behaketa kopurua 25.000 baino altuagoa bada, test honen erabilpena ez da gomendagarria eta kopurua 1.000 baino baxuagoa bada, $g = 10$ erabiltzea gomendatzen da. Gainera, talde bakoitzean behintzat 5 banako egotea aholkatzen da.

2.3. **Hosmer-Lemeshow testaren eraldaketa**

HL pertzentilen testaren erabilera ez da gomendatzen 25.000 behaketa baino gehiago ditugunean. Arazo hori gainditzeko, HL test eraldatua proposatu dute Nattino *et al.* ikertzaileek (2020).

Lehenago aipatu denez, ereduaren doikuntza egokia denean, (3) ekuazioan definitu dugun $\hat{C}$ estatistikoaren banaketa asintotikoa χ^2_{g-2} da. Aldiz, 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 [11]. 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.



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

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.

Defini dezagun

$$\epsilon = \sqrt{\frac{\lambda}{n}} \tag{7}$$

ez-zentralizazio parametro estandarizatua. Orduan,

$$\begin{cases} H_0 : \epsilon \leq \epsilon_0 \\ H_1 : \epsilon > \epsilon_0 \end{cases} \tag{8}$$

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

$$\epsilon_0 = \sqrt{\frac{\chi^2_{g-2,\alpha} - (g-2)}{10^6}} \tag{9}$$

erabiltzea gomendatzen da, non $\chi^2_{g-2,\alpha}$ χ^2_{g-2} banaketaren %$(1-\alpha)$100-kuantila den [7].

Bestalde, (8) testaren *p-balioa*

$$p = \int_{\hat{C}}^{\infty} \chi^2_{g-2,\epsilon_0^2 n}(z)dz \tag{10}$$

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.

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

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 doikuntza-egokitasunaren hipotesia 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 [7].

3. **SIMULAZIOAK**

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

3.1. **Eredu bakun teorikoa**

Hasteko, aldagai jarraitu bat eraiki dugu: *Z*. Gaixo ($Y = 1$) eta osasuntsuen ($Y = 0$) azpitaldetan *Z* sortu dugu, Z_G eta Z_O izendatuko ditugunak, hurrenez hurren. Bi aldagaiek banaketa normala izango 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 bereizi ditugu: $\sigma_O^2 = 1$ $(\sigma_O = 1)$ edo $\sigma_O^2 = 0.5$ $\left(\sigma_O = \sqrt{0.5}\right)$ izatea.

Orduan,

$$logit\left[p(Z)\right] = \beta_0 + \beta_1 Z \tag{11}$$

itxurako ereduak doitu ditugu.

Alde batetik, $\sigma_G = \sigma_O = 1$ direnean, Z eta $logit(p)$-ren arteko linealtasuna betetzen dela frogatuta dago [12].

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, Y) aldagaien n tamainako lagin batean, $\{z_i, y_i\}_{i=1}^{n}$, Z eta $logit(p)$-ren erlazioa

$$logit\left[p(z_i)\right] = -0.25z_i^2 + 1.5z_i + \ln\left[\frac{n_1}{\sqrt{2}(n-n_1)}\right] - 1.125 \tag{12}$$

da, non n_1 lagineko arrakasta kopuru totala den [12].

3.2. **Bi aldagai azaltzaile dituen eredu teorikoa**

X_1 eta X_2 aldagai aske jarraituak kontsideratu ditugu 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 bereizi ditugu:

$$\boldsymbol{\Sigma}_O = I_2 \text{ edo } \Sigma_O = \begin{pmatrix} 0.5 & 0 \\ 0 & 1 \end{pmatrix} \text{ izatea.}$$

Orduan, bi aldagaiak erabiliz

$$logit\left[p(\mathbf{X})\right] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \tag{13}$$

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 [12].

Bigarren kasuan, $\Sigma_O = \begin{pmatrix} 0.5 & 0 \\ 0 & 1 \end{pmatrix}$ eta $\boldsymbol{\Sigma}_G = I_2$ 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, (**X**, Y) aldagaien n tamainako lagin batean, $\{\mathrm{x}_i, y_i\}_{i=1}^{n}$,

$$logit\left[p(\mathrm{x}_i)\right] = -0.25x_{1,i}^2 + 1.5x_{1,i} + x_{2,i} + \ln\left[\frac{n_1}{\sqrt{2}(n-n_1)}\right] - 2.625 \qquad (14)$$

betetzen da non n_1 lagineko arrakasta kopuru totala den [12].

Gainera, X_1 2 eta 3 mozketa puntu erabilita kategorizatu dugu R programako CatPredi paketea erabiliz [13] eta

$$logit\left[p(\mathbf{X})\right] = \beta_0 + \sum_{l=1}^{m-1} \beta_{1_l} D_{1_l} + \beta_2 X_2 \qquad (15)$$

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

Bestalde, HL testa $J = n$ kasurako garatuta dago eta aldagai bat kategorizatzean, J txikiagoa izango da. Simulazioetan horren eragina aztertu nahi izan da. Honetaz gain, X_1 kategorizatzerakoan ez-linealtasunaren arazoa saihesten da.

Laburbilduz, 4 egoera ezberdin kontsideratu ditugu, 16. diagraman adierazita daudenak.

- Eredu bakuna
  - $\sigma_G = \sigma_O$ 1 Egoera (E1)
  - $\sigma_G \neq \sigma_O$ 2 Egoera (E2)
- Eredu anizkoitza
  - $\Sigma_G = \Sigma_O$ 3 Egoera (E3)
    - X_1 jarraitua (E3.1)
    - X_1 kategorikoa (E3.2)
  - $\Sigma_G \neq \Sigma_O$ 4 Egoera (E4)
    - X_1 jarraitua (E4.1)
    - X_1 kategorikoa (E4.2)

(16)

Lau egoeretan 0.5 eta 0.9 prebalentzia duten 50.000 behaketako laginak sortu ditugu. Lagin hauetatik 200, 500, 1.200 eta 2.000 tamainako 100 azpilagin 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 = 5tik g = 10era 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 deneko emaitzak kalkulatu ditugu. Ohartu E1, E3 eta E4.2 egoeretan H_0 egia dela eta, beraz, H_0 errefusatzeko proportzioa α inguru egotea espero dugula. Bestalde, E2 eta E4.1 egoeretan, H_1 egia da eta, ondorioz, H_0 errefusatzeko proportzioa testaren ahalmenaren estimazioa $(1-\hat{\beta})$ izango da.

4. **EMAITZAK**

Hasteko, simulazioetan erabilitako parametro kopurua handia denez, emaitza nagusiak laburtuko ditugu. Aldagai azaltzaileen eta *logit*(p)-ren arteko erlazioa lineala denean (E1, E3.1), α finkaturik, estimatutako adierazgarritasun-maila, $\hat{\alpha}$, α inguruan egon da. Antzeko emaitzak lortu ditugu n lagin tamaina ezberdinetarako (n = 200, 500, 1200, 2000). Berriz, aldagai askeen eta *logit*(p)-ren arteko erlazioa ez-lineala denean (E2, E4.1), batez ere tamaina txikiko laginetan (n = 200), ahalmenaren estimazio baxuak lortu dira. Gainera, prebalentziaren arabera oso aldakorrak diren emaitzak lortu ditugu.

Jarraian, azaldutako egoeretan lortutako emaitzak taula eta irudien bitartez adieraziko ditugu. 1. taulan eta 2. irudian, doikuntza egokia betetzen den egoeretan lortutako emaitzak adierazi ditugu. Doikuntza ezegokia betetzen den egoeretan lortutako emaitzak 2. taulan eta 3. irudian adierazi ditugu. Irudietan, $\alpha = 0.01$ eta $\alpha = 0.1$ erabilita lortutako estimazioak adierazi ditugu, antzeko emaitzak lortu dira $\alpha = 0.05$ kontsideratu denean.

4.1. **Aldagai azaltzaileen eta *logit*(p)-ren arteko erlazio teorikoa lineala**

4.1.1. *E1: eredu bakuna*

E1 egoeran, aldagai azaltzaile bakarra dugu ereduan eta $\sigma_G = \sigma_O$ betetzen da. Doitutako ereduaren itxura (11) ekuazioan ikus daiteke. Kasu honetan, frogatuta dago aldagai azaltzailearen eta *logit*(p)-ren arteko erlazioa lineala dela.

(11) eredua doitzerakoan, HL testaren I motako errorearen estimazioak, $\hat{\alpha}$, kalkulatu ditugu adierazgarritasun-maila $\alpha = 0.01, 0.05, 0.1$ kontsideratuz.

Orokorrean, espero duguna betetzen da; $\alpha = 0.01, 0.05, 0.1$ denean, α inguruko proportzioak lortu ditugu. Honetaz gain, ez dago I motako erro-

rearen estimazioen arteko desberdintasun nabarmenik lagin tamainari eta talde kopuruari erreparatzen badiogu (ikusi 2. irudia). Ordea, aipatu beharra dago egoera batzuetan talde kopuru gomendatua ez den talde kopuru batekin α baliora gehiago hurbiltzen diren emaitzak lortu ditugula. Adibidez, gertaera hau $\alpha = 0.01$, $n = 1200$ eta prebalentzia 0.9 diren kasuan ikus daiteke: $g = 10$ erabilita 0.02 lortu dugu eta $g = 14$ (talde kopuru gomendatua) erabilita, 0.07. Bestalde, kasu batzuetan, I motako errorearen estimazioetan prebalentziaren araberako α balioarekiko aldaketak eman dira. Adibidez, $\alpha = 0.05$, $n = 1200$ eta prebalentzia 0.5 direnean, α baino I motako errorearen estimazio baxuagoak edo berdinak lortu ditugu. Aldiz, prebalentzia 0.9 denean, α baino $\hat{\alpha}$ altuagoak zein baxuagoak lortu ditugu. Adierazgarritasun-maila $\alpha = 0.1$ eta $n = 1200$, 2000 denean berdina lortu dugu. Gainera, bi kasu hauetan, prebalentzia 0.9 denean, I motako errorearen estimazioak oso baxuak izan dira α-rekin alderatuta. Bestalde, α handitzerakoan, lortutako I motako errorearen estimazioak handitzen dira eta hau espero daitekeena da estimazioek konfiantza-maila islatzen dutelako. Honetaz gain, α-ren balioa handitzean, $\hat{\alpha}$ balioen sakanabatzea handitzen da.

4.1.2. *E3: eredu anizkoitza*

E3 egoeran, bi aldagai azaltzaile ditugu ereduan eta doikuntza egokia betetzen da. Doitutako ereduen itxura (13) eta (15) ekuazioetan ikus daitezke. Egoera honetan, bi aldagaiak jarraituak izanik, frogatuta dago aldagai azaltzaileen eta *logit*(p)-ren arteko erlazioa lineala dela.

X_1 **jarraitua**

(13) eredua doitzerakoan, HL testaren I motako errorearen estimazioak, $\hat{\alpha}$, kalkulatu ditugu adierazgarritasun-maila $\alpha = 0.01$, 0.05, 0.1 ezarriz.

Kasu batzuetan, prebalentziaren arabera α balioarekiko aldaketak ikusi ditugu. Adibidez, $\alpha = 0.05$ eta $n = 500$ direnean, α baino I motako errorearen estimazio altuagoak lortu ditugu prebalentzia 0.9 denean eta, prebalentzia 0.5 denean, α baino $\hat{\alpha}$ altuago zein baxuagoak lortu ditugu. $\alpha = 0.01$ eta $n = 2000$ direnean, prebalentzia 0.9 denean, α baino I motako errorearen estimazio baxuagoak lortu ditugu eta prebalentzia 0.5 denean, altuagoak zein baxuagoak. Honetaz gain, α-ren balioak handitzean, I motako errorearen estimazioak handitzen dira eta espero genuena. Gainera, α-ren balioa handitzean, $\hat{\alpha}$ balioen sakanabatzea 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 I motako errorearen estimazioa adierazten dute. Espero duguna emaitzak α balioaren ingurukoak izatea da.

(15) eredua doitzerakoan, HL testaren I motako errorearen estimazioak, $\hat{\alpha}$, kalkulatu ditugu adierazgarritasun-maila $\alpha = 0.01, 0.05, 0.1$ kontsideratuz. 2 mozketa puntu erabilita lortutako emaitzak ez ditugu aurkeztu.

Orokorrean, espero duguna betetzen da: α 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 $\hat{\alpha}$ baxuagoak lortu ditugu 2 mozketa punturekin baino, prebalentzia 0.5 zein 0.9 izanik.

Bestalde, α handitzean, emaitza sakabanatuagoak lortu ditugu.

4.2. Aldagai azaltzaileen eta *logit(p)*-ren arteko erlazio teorikoa ez-lineala

4.2.1. *E2: eredu bakuna*

E2 egoeran, aldagai azaltzaile bakarra dugu ereduan eta $\sigma_G \neq \sigma_O$ dira. Doitutako ereduaren itxura (11) ekuazioan ikus daiteke. Egoera honetan, aldagai askearen eta *logit*(p)-ren arteko erlazioa ez-lineala dela frogatuta dago.

(11) eredua doitu ostean, HL testak doikuntza egokiko hipotesia errefusatu dueneko proportzioak kalkulatu ditugu $\alpha = 0.01, 0.05, 0.1$ izanda.

Espero genuena ez da betetzen: HL testak, orokorrean, doikuntza egokia ez du errefusatzen. Hau da, testaren ahalmena oso baxua da. Bereziki, $n = 200, 500$ denean, II motako errorearen estimazio baxuak lortu ditugu. Adibidez, $\alpha = 0.01$, $n = 200$, prebalentzia 0.5 eta $g = 10$ (talde kopuru gomendatua) direnean, lortutako ahalmenaren estimazioa 0.04 da (ikusi 2. taula). Egoera honetan, HL testak erabaki zuzena 4 aldiz hartu du 100etik. Bestalde, eta espero bezala, ahalmenaren estimazioak n lagin tamainarekin handitzen dira, $n = 2000$ denean estimatutako ahalmena %85ekoa izanik. Bestalde, kasu batzuetan talde kopuru gomendatua erabilita, ahalmenaren estimazioa handitzen da; $\alpha = 0.05$, $n = 1200$ eta prebalentzia 0.5 direnean, $g = 10$ aukeratuta 0.86 lortu dugu $g = 14$rekin (talde kopuru gomendatuarekin), berriz, 0.75 lortu dugu. Honetaz gain, $\alpha = 0.01$ izanda, prebalentzia 0.9 denean ahalmenaren estimazio baxuagoak lortzen ditugu prebalentzia 0.5 denean baino eta, lagin tamaina handitzean, haien arteko diferentzia nabarmenki hazten da. $\alpha = 0.05$ eta $n = 200$ direnean, prebalentzia 0.5 denean II motako errorearen estimazioak baxuagoak dira prebalentzia 0.9 direnean baino. Aldiz, $\alpha = 0.05$ eta $n = 500, 1200, 2000$ direnean, prebalentzia 0.9 denean ahalmenaren estimazioak altuagoak dira prebalentzia 0.5 denean lortutako II motako errorearen estimazioekin alderatuta. Gainera, n handitzean, haien arteko diferentziak asko handitzen dira. Noski, α-ren balioak handitzean, ahalmenaren estimazio altuagoak lortu ditugu.

4.2.2. *E4: eredu anizkoitza*

E4 egoeran, bi aldagai azaltzaile ditugu ereduan eta $\mathbf{\Sigma}_O \neq \mathbf{\Sigma}_G$ betetzen da. Doitutako ereduen itxura (13) eta (15) ekuazioetan ikus daitezke. (13) ereduan, X_1 aldagai azaltzailearen erlazioa *logit*(p)-rekiko ez-lineala da eta X_2-rena, lineala.

X_1 **jarraitua**

(13) eredua doitu ostean, HL testean doikuntza egokiko hipotesia errefusatu deneko proportzioak kalkulatu ditugu, α = 0.01, 0.05, 0.1 ezarrita (ikusi 2. taula).

Bereziki n = 200, 500 denean, orokorrean lortutako ahalmenaren estimazioak baxuak dira. Adibidez, α = 0.01, n = 200, prebalentzia 0.5 eta g = 10 (talde kopuru gomendatua) denean, lortutako ahalmenaren estimazioa estimazioa 0.05 da. Beste era batera esanda, HL testak erabaki zuzena 5 aldiz hartu du 100etik. Honetaz gain, orokorrean, ahalmenaren estimazioak n lagin tamaina handitzean, handitu egiten dira, baina ahalmenaren estimazioa 0.8tik gora oso kasu bakarretan behatu dugu, esaterako, n = 2000, α = 0.1, prebalentzia 0.5 eta g = 34 (talde kopuru gomendatua) direnean, lortutako ahalmenaren estimazioa 0.91 da. Gainera, α = 0.01, 0.05 eta n = 1200 direnean, prebalentzia 0.5 baliotik 0.9 baliora handitzean, ahalmenaren estimazioak handitu egin dira. Adibidez, α = 0.01 eta g = 14 (talde kopuru gomendatua) erabilita, prebalentzia 0.5 denean lortutako estimazioa 0.42 da eta, prebalentzia 0.9 denean, 0.46. Ordea, α = 0.1 eta n = 1200 denean, ahalmenaren estimazioak txikitu egiten dira prebalentzia handitzean. Esate baterako, g = 14 (talde kopuru gomendatua) erabilita, prebalentzia 0.5 denean 0.12 behatu dugu eta prebalentzia 0.9 denean, 0.06. Bestalde, n = 2000 denean, α = 0.01, 0.05 izanik, 0.5 prebalentziari dagozkion estimatutako ahalmenak 0.9 prebalentziakoak baino baxuagoak dira, talde kopuru gomendatua erabiltzean izan ezik. Gainerako kasuetan, talde kopuru gomendatuaren erabilerak ez du eragin garrantzitsurik izan. α = 0.1 denean, prebalentzia 0.9 denean, ahalmenaren estimazio baxuagoak lortzen ditugu prebalentzia 0.5 denean baino. Bestalde, n = 200, 500 denean, orokorrean, prebalentzia txikitzean, ahalmenaren estimazio altuagoak lortu ditugu. Honetaz gain, α adierazgarritasun-maila handitzean, orokorrean, ahalmenaren estimazio altuagoak 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 I motako errorearen estimazioa adierazten dute. Hortaz, α finkaturik, emaitzak α ingurukoak izatea espero dugu.

(15) eredua doitzerakoan, HL testak doikuntza egokiko hipotesia errefusatu dueneko emaitzak kalkulatu ditugu, $\alpha = 0.01, 0.05, 0.1$ izanik (ikusi 1. taula).

Orokorrean, espero duguna betetzen da: α inguruko emaitzak 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 α baliora gehiago hurbiltzen diren emaitzak lortu ditugu.

5. **EZTABAIDA**

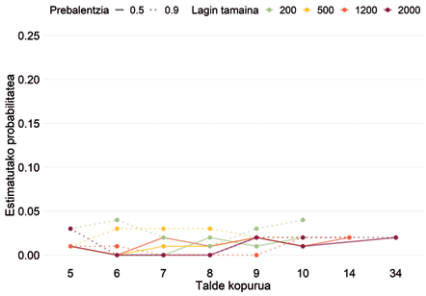
Lan honetan HL testaren erabakien aldaketa aztertu dugu *g* talde kopuruaren arabera. Simulazioetan lagin tamaina, prebalentzia, adierazgarritasun maila, testean erabilitako talde kopurua, aldagai azaltzaileen kopurua eta haien bariantzak aldatu ditugu. Populazio osasuntsu eta gaixoan, aldagai azaltzaileak banaketa normalari darraizkio. Izan ere, kasu honetan, aldagai azaltzaileen 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 ezta erabilitako talde kopuruarena ere. Bestalde, prebalentziak emaitzetan eragina izan du, baina proportzioek ez diote joera zehatz bati jarraitu.

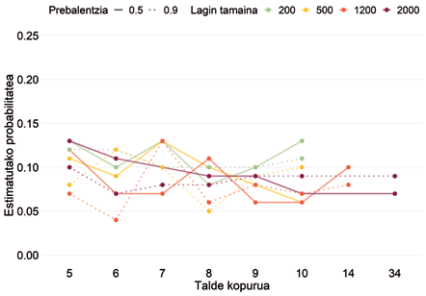
Gainera, emaitzen aldakortasuna testaren adierazgarritasun-mailarekin handitu da.

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. Erabilitako talde kopuruari dagokionez, aztertutako egoeretan, ez dugu diferentzia nabarmenik aurkitu erabilitako talde kopuruaren arabera. Doitutako eredua desegokia denean, lagin handietan bai gaixotasunaren prebalentziak bai ereduaren konplexutasunak eragina dute.

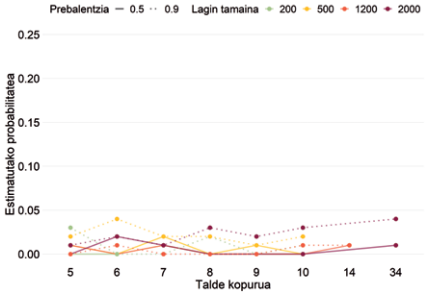
Lagin tamaina finkatuta, prebalentziaren arabera desberdintasun nabariak gertatzen dira ahalmenaren estimazioetan, baina ez dugu joera zehatzik behatu. Bestalde, orokorrean, aldagai azaltzaileen kopurua handitzean, ahalmenaren estimazioak baxuagoak izan dira. Ereduan aldagai azaltzaile bakarra erabilita, 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.



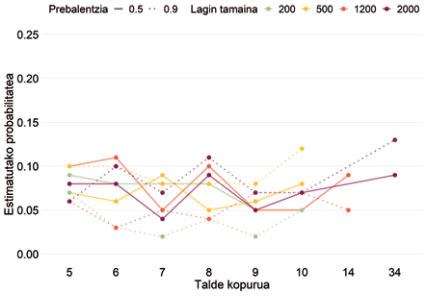
a) E3.1: $\alpha = 0.01$.



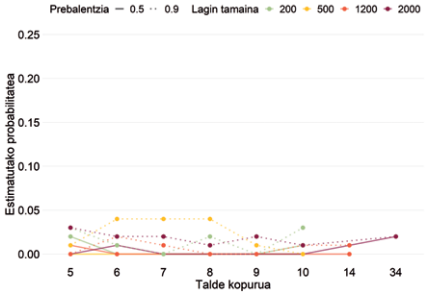
b) E3.1: $\alpha = 0.1$.



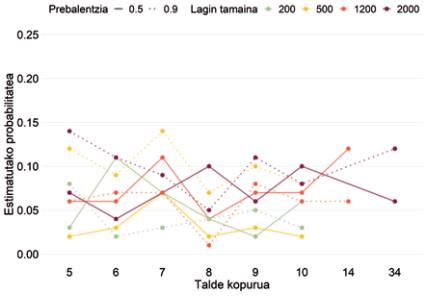
c) E3.2: $\alpha = 0.01$, 3 mozketa puntu.



d) E3.2: $\alpha = 0.1$, 3 mozketa puntu.

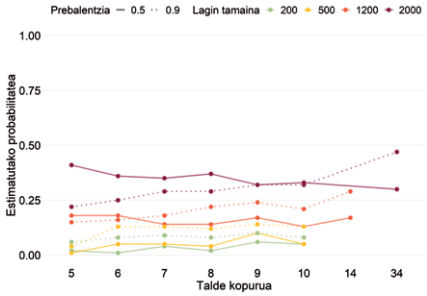


e) E4.2: $\alpha = 0.01$, 3 mozketa puntu.

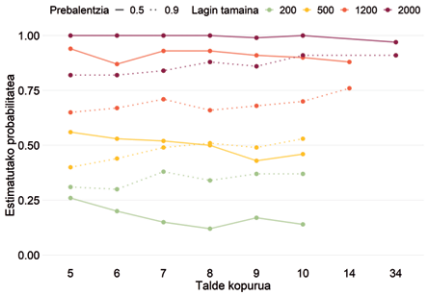


f) E4.2: $\alpha = 0.1$, 3 mozketa puntu.

2. irudia. H_0 egia den egoeretan, HL testak doikuntza egokiko hipotesia errefusatu dueneko proportzioa, I motako errorea egiteko probabilitatearen estimazioa. Lagin tamaina $n = 1200$ denean, g talde kopuru gomendatua $g = 14$ da eta $n = 2000$ denean, g talde kopuru gomendatua $g = 34$ da.



a) E4.1: $\alpha = 0.01$.



b) E4.1: $\alpha = 0.1$.

3. irudia. H_1 egia den egoeretan, HL testak doikuntza egokiko hipotesia errefusatu dueneko proportzioa, ahalmenaren estimazioa. Lagin tamaina n = 1200 denean, g talde kopuru gomendatua g = 14 da eta n = 2000 denean, g talde kopuru gomendatua g = 34 da.

1. taula. H_0 egia den egoeretan, HL testak doikuntza egokiko hipotesia errefusatu dueneko proportzioa g talde kopuru gomendatua erabiliz.

α	Lagin tamaina (*g gomendatua*)	Prebalentzia	Egoera			
			E1	E3.1	E3.2	E4.2
0.01	200	0.5	0.03	0.02	0.00	0.01
	500		0.02	0.01	0.00	0.00
	1200		0.01	0.02	0.01	0.00
	2000		0.03	0.02	0.01	0.02
	200	0.9	0.00	0.04	0.01	0.03
	500		0.03	0.02	0.02	0.00
	1200		0.07	0.02	0.01	0.01
	2000		0.02	0.02	0.04	0.02
0.05	200	0.5	0.06	0.07	0.02	0.01
	500		0.04	0.04	0.03	0.00
	1200		0.03	0.08	0.03	0.06
	2000		0.07	0.02	0.03	0.06
	200	0.9	0.06	0.07	0.05	0.03
	500		0.10	0.08	0.05	0.04
	1200		0.04	0.08	0.03	0.05
	2000		0.08	0.05	0.09	0.07
0.1	200	0.5	0.11	0.13	0.05	0.06
	500		0.09	0.06	0.08	0.12
	1200		0.06	0.10	0.09	0.12
	2000		0.10	0.07	0.09	0.06
	200	0.9	0.08	0.11	0.05	0.03
	500		0.10	0.10	0.12	0.08
	1200		0.10	0.08	0.05	0.06
	2000		0.13	0.09	0.13	0.12

2. taula. H_1 egia den egoeretan, HL testak doikuntza egokiko hipotesia errefusatu dueneko proportzioa *g* talde kopuru gomendatua erabiliz.

α	Lagin tamaina (*g gomendatua)*	Prebalentzia	Egoera	
			E2	E4.1
0.01	200	0.5	0.04	0.05
	500		0.22	0.05
	1200		0.52	0.17
	2000		0.85	0.30
	200	0.9	0.03	0.08
	500		0.16	0.13
	1200		0.26	0.29
	2000		0.46	0.47
0.05	200	0.5	0.11	0.11
	500		0.38	0.11
	1200		0.75	0.29
	2000		0.91	0.42
	200	0.9	0.23	0.23
	500		0.36	0.35
	1200		0.58	0.46
	2000		0.80	0.72
0.1	200	0.5	0.04	0.14
	500		0.46	0.46
	1200		0.88	0.88
	2000		0.97	0.97
	200	0.9	0.37	0.37
	500		0.53	0.53
	1200		0.76	0.76
	2000		0.91	0.91

Laburbilduz, hasiera batean, talde kopuruak eragindako aldaketak aztertu nahi genituen. Ordea, guk proposatutako egoeretan, prebalentziak eta lagin tamainak eragin handiagoa dutela ikusi dugu.Izan ere, doitutako ereduaren doikuntza egokia zein ezegokia denean, HL testaren errendimendua antzekoa da talde kopuru ezberdinetarako: ez dugu joera zehatzik behatu HL testaren erabakietan talde kopurua aldatzean prebalentzia eta lagin tamaina desberdinetarako. Lortutako ondorioak orain arteko egindako ikerketekin bat datoz (ikusi [14, 3]). Honetaz gain, HL pertzentilen testa eta test eraldatua erabili ditugu eta bi testekin lortutako emaitzak berdinak dira (ikusi [7]), baina 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 egoera batzuetan simulazioak egin ditugu eta emai-

tzak ezberdinak izan dira. Etorkizunean, bariantzen arteko ezberdintasuna sakonago aztertu beharko genukeela uste dugu.

ESKER ONAK

Lan honen garapena posiblea izan da honako laguntza hauei esker: Eusko Jaurlaritzako Hezkuntza, Hizkuntza Politika eta Kultura Sailaren Ikerketa Taldea MATHMODE [IT1456-22] eta Zientzia eta Berrikuntza Ministerioak BCAM Severo Ochoa kreditazioaren bidez [CEX2021-001142-S/MICIN/AEI /10.13039/501100011033) eta proiektuaren bidez [PID2020-115882RB-I00/AEI /10.13039/501100011033) Ikerketako Estatu Agentziak finantzatua eta «S3M1P4R» akronimoa duena, eta Eusko Jaurlaritzaren BERC 2022-2025 programaren bidez.

BIBLIOGRAFIA

[1] VAN CALSTER, B., MCLERNON, D. J., VAN SMEDEN, M., WYNANTS, L., eta STEYERBERG, E. W. 2019. «Calibration: the Achilles heel of predictive analytics.» *BMC Medicine*, **17**(1).

[2] HARRELL, F. E. 2015. *Regression Modeling Strategies: With Applications to Linear Models*, *Logistic and Ordinal Regression*, *and Survival Analysis*. Springer.

[3] 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.

[4] STEYERBERG, E. W. 2019. *Clinical Prediction Models.* Springer Publishing.

[5] HOSMER, D. W., LEMESHOW, S., eta STURDIVANT, R. X. 2013. *Applied logistic regression.* Wiley.

[6] 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.

[7] 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.

[8] 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.

[9] DIMITRIADIS, T., HENZI, A., PUKE, M., eta ZIEGEL, J. 2022. «A safe Hosmer-Lemeshow test.»

[10] MCCULLAGH, P., NELDER, eta JOHN, A. 1989. *Generalized Linear Models*, *Second Edition (Chapman & Hall/CRC Monographs on Statistics & Applied Probability)* Chapman and Hall/CRC.

[11] 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).

[12] IPARRAGIRRE, A., BARRIO, I. eta RODRÍGUEZ-ÁLVAREZ, M. X. 2019. «On the optimism correction of the area under the receiver operating characteristic curve in logistic prediction models.» *SORT-Statistics and Operations Research Transactions*, **43**(1), 145-162.

[13] BARRIO, I., AROSTEGUI, I., RODRÍGUEZ-ÁLVAREZ, M. X. eta QUINTANA, J. M. 2015. «A new approach to categorising continuous variables in prediction models: Proposal and validation.» *Statistical Methods in Medical Research*, **26**(6), 2586-2602.

[14] 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.