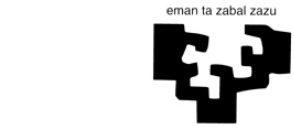


ZTF-FCT
Zientzia eta Teknologia Fakultatea
Facultad de Ciencia y Tecnología

Erregresio logistikoa eta mozketa puntuak: bihotz-gutxiegitasunaren hilkortasunaren azterketa

Gradu Amaierako Lana
Matematikako Gradua

Maialen Murua Etxeberria

Irantzu Barrio Beraza
Irakasleak zuzendutako lana

Leioa, 2014ko uztaila

Aurkibidea

Laburpena

Memoria honetan bihotz gutxiegitasun desorekatu larria (BGDL) duten pazienteen hilkortasunaren azterketa egiten da NT-pro-BNP parametro klinikoa erabiliz. Lanaren helburua, NT-pro-BNP-rentzat mozketa puntu optimoa kalkulatzea izango da, hau da, aldagaiaren zein baliotik aurrera hiltzeko probabilitatea handitzen den ikusiko dugu eta horrekin medikuei pazientea ingresatu edo ez erabaki prozesuan lagundu. Horretarako erregresio logistikoko eredu estatistiko bat eraikitzen da eta mozketa puntuak kalkulatzeko teknikak aplikatzen zaizkio ereduari.

Erregresio logistikoa, erregresio mota berezi bat da, menpeko aldagaia Y, dikotomikoa den kasurako erabiltzen dena. Azterketa honetan $Y = 1$ pazientea hil den kasua eta $Y = 0$ pazientea hil ez den kasua ditugu; arrakasta eta porrota hurrenez hurren. Erregresio logistikoa aplikatzeko arrakastarako probabilitatea zenbaki erreal bihutzen duen esteka funtzioa beharrezkoa da. Kasu honetan *logit* funtzioa erabili dugu eta, p arrakastarako probabilitatearen eta NT-pro-BNP-ren arteko erlazioa ahalbideratuko digu. Koefizienteak estimatu eta ereduaren adierazgarritasuna eta doikuntz egokitasuna aztertuko ditugu R softwarea erabiliz.

Hilkortasunaren eta NT-pro-BNP-ren arteko erlazioa adierazten duen erregresio logistiko sinpleko eredua sortu ondoren, NT-pro-BNP-ren mozketa puntu optimoa kalkulatzeko teknikak aplikatuko ditugu. Hau da, estimatutako p horiek arrakasta edo porrota diren erabakitzeko mozketa puntu optimoa, c kalkulatuko dugu. Honek, NT-pro-BNP-ren mozketa puntua emango digu. Atal honetan sailkapen taulak eta ROC kurbak erabiliko ditugu eta mozketa puntu optimoak metodo ezberdinak erabiliz kalkulatzeko R-ren OptimalCutpoints libreria erabiliko dugu.

Mozketa puntuak kalkulatzeko hiru metodo orokor erabili dira: Young-en metodoa ROC01 metodoa eta SeDSp metodoa. Lehenengo bi metodoekin aldagaiaren mozketa puntu bera lortzen dugu 6875 eta SeDSp-rekin berriz 5814.

Ikuspegi matematikotik Young eta ROC01 metodoekin emaitza hobeak lortzen ditugula esan dezakegu sailkapen egokiko proportzio global altuagoa dutelako eta fatsu positibo gutxiago ditugulako SeDSp metodoarekin baino. Horren aurrean ordea SeDSp metodoa erabiliz sentikortasun altuagoa lortzen dugu eta baita faltsu negatibo gutxiago ere. Hori dela eta azken erabakia medikuen esku utziko genuke.

Esker onak

Nire eskerrik onenak Galdakao-Usansolo ospitaleko ikerkuntza unitateari, memoria hau aurrera ateratzeko beharrezko genituen datuak ahalbideratzeagatik. Bereziki 2011111045 proiektuko ikertzaile nagusia den Susana Garcia Gutierrez-i.

1. Kapitulua

Erregresio logistikoa

Erregresio metodoak edozein datu baseren azterketan garrantzitsuak bihurtu diren metodoak dira menpeko aldagai baten, Y, eta behatutako aldagaien arteko erlazioa aztertzerakoan. Kasurik sinpleena Y aldagai jarraitua eta normala denekoa da, erregresio lineala hain zuzen. Hala ere, Y aldagaiaren banaketa normala ez denean, beste eredu batzuk kontsideratzen ditugu: eredu lineal orokortuak. Eredu lineal orokortu ezberdinak plantea ditzakegu, Y erantzun aldagaiaren banaketaren arabera:

- Y dikotomikoa denean datu bitarrak ditugu.
- Y diskretua eta kualitatiboa bada datu multinomialak.
- Y diskretua eta kuantitatibo finitua denean datu kuantitatiboak.
- Y osoa denean berriz, Poisson-en datuak.

Lan honetan Y dikotomikoa deneko kasua aztertuko dugu. Kasu honetan ereduak izen berezi bat hartzen du, erregresio logistikoa hain zuzen ere.

Erregresio logistikoan Y aldagaiak bi balio hartuko ditu $Y = 1$ eta $Y = 0$ arrakasta eta porrota izendatuko ditugunak hurrenez hurren. Hau da Y banaketa binomialari darraion aldagaia izango da, bi balio hartuko dituenez aldagai dikotomikoa edo bitarra dela esango dugu.

1.1 Erregresio logistiko sinplea

1 Sarrera

Erregresio logistiko sinplea X, aldagai aske bakarra dugun kasuari deritzo. Helburua Y menpeko aldagaiaren eta X-ren arteko erlazio bat aurkitzea da. Y esan bezala dikotomikoa izango eta X jarraitua, diskretua edota kategorikoa izan daiteke.

$E(Y|X)$, balioa Y-ren batez bestekoa da, termino horrek Y-ren itxarotako balioa adierazten du X-ren menpe.
Erregresio logistikoaren kasuan ordea, Y aldagaia banaketa binomialari darraion aldagaia da, $Y : Bin(1, p)$ non $p = P(arrakasta)$ den. $E(Y|X) = p$ izango da eta 0 eta 1 bitarteko balioak hartuko ditu.
Hurrengo pausua g funtzio bat bilatzea izango da, p X-ren menpe adierazteko, funtzio honi esteka funtzioa deituko diogu. g funtzio hau, p jarraitua bihurtzeko eraikitzen dugu. Probabilitateak $[0, 1]$ bitarteko balioak hartuko ditu $g(p)$-k ordea $(-\infty, \infty)$ tartekoak.

$$g : \mathbb{R} \longrightarrow \mathbb{R}$$
$$g(E(Y)) = g(p) = \beta_0 + \beta_1 X \tag{1.1}$$

Aukeratuko dugun g funtzioa ondokoa da:

$$g(p) = logit(p) = ln\frac{p}{1-p} \tag{1.2}$$

Logit eredua aukeratzeko arrazoiak bi dira, batetik ikuspuntu matematikotik erraz erabili daitekeen funtzioa da eta bestetik klinikoki esanguratsua den interpretazioa du kapitulu honetako 4. atalean ikusiko dugun bezala.

Suposa dezagun (x_i, y_i) $i = 1, ..., n$ n tamainako zorizko lagin bakunetik behatutako balioak direla orduan,

$$logit(p_i) = ln\frac{p_i}{1-p_i} = \beta_0 + \beta_1 x_i \qquad i = 1, ..., n$$

(logit-aren) erregresio logistiko sinpleko eredua dela esaten dugu. Non arrakastarako probabilitatea ondokoa den

$$p_i = \frac{e^{\beta_0+\beta_1 x_i}}{1 + e^{\beta_0+\beta_1 x_i}}$$

2 Parametroen estimazioa

Eredua eraikitzeko β_0 eta β_1 parametro ezezagunen balioak estimatu behar dira, horretarako egiantz handieneko metodoa erabiliko dugu. Metodo honek, parametro ezezagunen balioak estimatzen ditu behatutako balioak lortzeko probabilitatea maximizatuz [1].
Hori posible izateko egiantz handieneko funtzioa definituko dugu. Funtzioak, aztertutako datuen probabilitateak adierazten ditu parametro ezezagunen menpe.

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

β_0 eta β_1 parametroen estimatzaileak (1.3) ekuazioa maximizatzen duten balioak dira.

Matematikoki logaritmoarekin lan egitea errazagoa denez, egiantz handieneko logaritmoa definituko dugu:

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

Adierazpen hau β_0 eta β_1 ekiko deribatuz eta zerorekin berdinduz egiantz handieneko ekuazioak lortzen ditugu:

$$\sum_{i=1}^{n} y_i - p(x_i) = 0 \tag{1.4}$$

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

(1.4) eta (1.5) ekuazioak ebazteko, linealak ez direnez, zenbakizko metodoak behar dira. Ohikoena Newton-Raphson-en metodoa da [2].
(1.4) eta (1.5) ekuazioetatik lortzen ditugun β_0 eta β_1 balioak egiantz handieneko estimatzaileak deitzen dira eta $\hat{\beta}_0$ eta $\hat{\beta}_1$ adieraziko ditugu.
p-ren estimatzailea berriz ondokoa da:

$$\hat{p} = \frac{e^{\hat{\beta}_0 + \hat{\beta_1} X}}{1 + e^{\hat{\beta}_0 + \hat{\beta_1} X}}$$

3 X aldagaiaren adierazgarritasuna

Koefizienteak estimatu eta gero X-ren adierazgarritasuna aztertuko dugu, horretarako galdera hau planteatzen dugu: X aldagaia barne hartzen duen eredua hobea al da barne hartzen ez duen eredua baino?

Galdera honi erantzuteko, behatutako balioak bi ereduetan estimatutako balioekin konparatu behar ditugu; lehenengoa X aldagaia barne hartzen duenarekin eta bigarrena X barne hartzen ez duenarekin. X barne hartzen duen ereduan itxarotako balioak hobeak badira beste ereduan baino aldagaia adierazgarria izango da.

Egiantz handieneko testa

Bi ereduak konparatzeko (1.3) ekuazioan definitutako egiantz handieneko logaritmoa erabiliko dugu. X barne hartzen ez duen eredua 1E izendatuko dugu eta aldagaia barne hartzen duena 2E. Konparazioa hobeto ulertzeko lagungarria da erantzun aldagaian behatutako balioak eredu asearen balioak balira bezala ulertzea. Eredu asea parametro kopuru eta behatutako balio kopuru berdina dituen eredua da.

Ondoren datorren adierazpena *"Deviance"* izendatuko dugu:

$$D = -2ln\left[\frac{\text{estimatutako ereduaren egiantz handiena}}{\text{eredu asearen egiantz handiena}}\right] \tag{1.6}$$

(1.3) ekuazioa erabiliz ondokoa lortzen dugu

$$D = -2\sum_{i=1}^{n}\left[y_i ln\left(\frac{\hat{p_i}}{y_i}\right) + (1-y_i)ln\left(\frac{1-\hat{p_i}}{1-\hat{y_i}}\right)\right]$$

X-ren adierazgarritasuna aztertzeko bi ereduen D hartzen dugu eta G estatistikoa definitzen dugu:

$$G = D(1\text{E}) - D(2\text{E})$$

$$G = -2ln\left[\frac{\text{Egiantz handiena(1E)}}{\text{Egiantz handiena(2E)}}\right] \tag{1.7}$$

G estatistikoak 1 askatasun graduko Khi karratu banaketa jarraitzen du. X-ren adierazgarritasuna aztertzeko ondoko hipotesi kontrastea planteatzen dugu:

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$$

p-balioa: $p = P(\chi_1^2 > G)$ delarik.

Walden testa

X-ren adierazgarritasuna aztertzeko beste testa Wald-ena da. Parametroaren egiantz handieneko estimatzailea, $\hat{\beta}_1$, bere estimatutako errore estandarrarekin konparatzen du. Zatidura horrek $\beta_1 = 0$ hipotesipean banaketa normal estandarizatuari darraikio.

$$\hat{\beta}_1 \approx N(\beta_1, \hat{\text{es}})$$

non $\hat{\text{es}}$ $\hat{\beta}_1$-en errore estandarra den.
Honako kontrastea definitzen dugu:

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$$

H_0 egia izatekotan estatistikoaren balioa honakoa da

$$W_p = \frac{\hat{\beta}_1}{\hat{\text{es}}(\hat{\beta}_1)}$$

eta p-balioa $p = 2P(Z > |w_p|)$ da.

Wald-en testan oinarritzen den konfidantza tartea ondokoa da, non konfidantza tartearen bidez ere X aldagaiaren adierazgarritasuna aztertu daitekeen.

$$I_{\beta_1}^{1-\alpha} = \left(\hat{\beta}_1 - z_{\alpha/2}\ \hat{\text{es}}(\hat{\beta}_1), \hat{\beta}_1 + z_{\alpha/2}\ \hat{\text{es}}(\hat{\beta}_1)\right)$$

4 Parametroen interpretazioa

Eredua eraiki, parametroen estimazioa egin eta aldagaiaren adierazgarritasuna aztertu ondoren estimatutako koefizienteen esanahia ulertu behar da horiek interpretatzeko.

Erregresio linealean β_1 koefizienteak menpeko aldagaiaren balioen arteko diferentzia azaltzen du $x+1$ eta x balioetan. Adibidez, $y(x) = \beta_0 + \beta_1 x$ badugu orduan $\beta_1 = y(x+1) - y(x)$ eta koefizientearen interpretazioa oso erraz ikus daiteke, dagokion eskalan erantzun aldagaiaren hazkundea aldagai independentea unitate bat handitzean. Erregresio logistikoaren kasuan ordea β_1 aldagaiak logit-aren hazkundea adierazten du aldagai independentea unitate bat handitzean. $\beta_1 = g(x+1) - g(x)$. Hurrengo lerroetan kasuz kasu aztertzen joango gara erregresio logistiko sinplerako.

X dikotomoa den kasua (0 vs.1)

Erregresio logistikoko koefizienteen interpretazioarekin hasiko gara X aldagaiak bi balio hartzen dituen kasurako. Logit-aren diferentzia $X = 1$ duen indibiduo batentzat eta $X = 0$ duen indibiduo batentzat ondokoa da:

$$g(1) - g(0) = [\beta_0 + \beta_1] - [\beta_0] = \beta_1$$

Emaitza hau interpretatzeko *Odds ratio* terminoa definitu behar dugu. Arrakastarako probabilitatea X dagoenean ($X = 1$), $p(1)$ bezala definituko dugu eta arrakastarako probabilitatea X ez dagoenean ($X = 0$), $p(0)$ bezala. Hau da,

$$p(1) = P(Y = 1|X = 1) \Rightarrow 1 - p(1) = P(Y = 0|X = 1)$$
$$p(0) = P(Y = 1|X = 0) \Rightarrow 1 - p(0) = P(Y = 0|X = 0)$$

Modu honetan, arrakastarako *Odds*-a definitu daiteke X dagoenean, $o(Odds) = \frac{p(1)}{1-p(1)}$ eta X ez dagoenean, $o(Odds) = \frac{p(0)}{1-p(0)}$.

Bi termino hauen zatiduratik berriz, *Odds ratio*(OR) lortzen dugu.

$$OR = \frac{p(1)/(1-p(1))}{p(0)/(1-p(0))}$$

Balio hauek ordezkatzen baditugu,

$$OR = \frac{\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}} / \frac{1}{1+e^{\beta_0+\beta_1}}}{\frac{e^{\beta_0}}{1+e^{\beta_0}} / \frac{1}{1+e^{\beta_0}}} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{(\beta_0+\beta_1)-\beta_0} = e^{\beta_1}$$

$OR = e^{\beta_1}$ honela interpretatzen da: zenbat aldiz handiagoa den arrakastarako probabilitatea porroterako probabilitatea baino $X = 0$ eta $X = 1$ kasuen artean.

OR=1 denean, arrakastarako eta porroterako probabilitatea berdina da, X-k ez du eraginik arrakastarako probabilitatean. $OR > 1$ denean, X egoteak arrakastarako probabilitatea handitzen du eta $OR < 1$ denean, X egoteak arrakastarako probabilitatea txikitzen du. Koefizienteen interpretazioa dela eta erabiltzen da logitaren eredua kasu klinikoetan eta epidemiologian.

Adibidea 1. Demagun bihotzeko gaixotasuna edukitzea edo ez edukitzea aztertzen ari garela eta X aldagaiak indibiduoak kirola egiten duen edo ez adierazten duela. $OR = 0.5$ bada, 0.5 aldiz bihotzeko gaixotasuna izateko probabilitatea du kirola praktikatzen duen pertsonak praktikatzen ez duenak baino.

	Aldagai askea (X)	
Menpeko aldagaia (Y)	X=0	X=1
Y=0	$p(1) = \frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}$	$p(0) = \frac{e^{\beta_0}}{1+e^{\beta_0}}$
Y=1	$1-p(1) = \frac{1}{1+e^{\beta_0+\beta_1}}$	$1-p(0) = \frac{1}{1+e^{\beta_0}}$
Guztira	1.0	1.0

1.1. Taula. Erregresio logistiko sinplearen balioak X dikotomikoa denean

X jarraitua den kasua

Arrakastarako probabilitatea $X = x$ denean $p(x)$ bezala definituko dugu eta arrakastarako probabilitatea $X = x + 1$ denean $p(x + 1)$ bezala. Hau da,

$$p(x) = P(Y = 1|X = x) \Rightarrow 1 - p(x) = P(Y = 0|X = x)$$
$$p(x + 1) = P(Y = 1|X = x + 1) \Rightarrow 1 - p(x + 1) = P(Y = 0|X = x + 1)$$

Arrakastarako *Odds*-a $X = x$-rako, $o(Odds) = \frac{p(x)}{1-p(x)}$ eta arrakastarako *Odds*-a $X = x + 1$-erako, $o(Odds) = \frac{p(x+1)}{1-p(x+1)}$ definitzen dira.

Odds ratio-a honela definitzen da aldagaia jarraitua den kasurako

$$\text{OR} = \frac{p(x + 1)/(1 - p(x + 1))}{p(x)/(1 - p(x))}$$

$OR = e^{\beta_1}$ honela interpretatzen da: Zenbat aldiz handiagoa den arrakastarako probabilitatea porroterakoa baino X unitate bat handitzen denean.

$OR = 1$ bada, ($\beta = 0$) arrakastarako eta porroterako probabilitatea berdina da X-ren unitate bateko hazkunderako. $OR > 1$ bada ($\beta > 0$) arrakastarako probabilitatea altuagoa da eta $OR < 1$ bada ($\beta < 0$) arrakastarako probabilitatea bajuagoa da porroterakoa baino.

1.2 Erregresio logistiko anizkoitza

1 Sarrera

Aurreko atalean erregresio logistikoa azaldu dugu baina aldagai bakarrarentzako, orain hori bera orokortuko dugu aldagai bat baino gehiago dugun kasurako. Jarraituko dugun prozesua aurreko atalekoaren berdina izango da.
Demagun $\mathbf{X} = (X_1, X_2, ..., X_p)$ p aldagai aske ditugula, non
$E(Y|\mathbf{X}) = p(\mathbf{X})$ den. Erregresio logistiko anizkoitzerako logit-aren eredua ondokoa da:

$$g(\mathbf{X}) = \ln\frac{p}{1-p} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p \tag{1.8}$$

non

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

Demagun behatutako n balio aske ditugula $(\mathbf{x}_i, y_i)$ $i = 1, ..., n$.
Aldagai bakarrarekin gertatzen zen bezala $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)$ bektorea estimatu beharko dugu.

2 Parametroen estimazioa

Erregresio sinplean bezala parametroen estimaziorako egiantz handieneko metodoa erabiliko dugu. Egiantz handieneko funtzioa (1.3) adierazpenaren antzekoa da, desberdintasun bakarra $p(\mathbf{X})$-n dago. Egiantz handieneko $p + 1$ ekuazio edukiko ditugu egiantz handieneko logaritmoa deribatuz lortuko direnak.
Hauek dira lortzen ditugun egiantz handieneko ekuazioak:

$$\sum_{i=1}^{n}[y_i - p(\mathbf{x_i})] = 0 \tag{1.9}$$

$$\sum_{i=1}^{n} x_{ij}[y_i - p(\mathbf{x_i})] = 0 \qquad j = 1, ..., p \tag{1.10}$$

Hauek ebazteko zenbakizko metodoak behar dira, ohikoenak Newton-Raphsonen metodoa da [2]. (1.9) eta (1.10) ekuazioen soluzioa $\boldsymbol{\beta}$ izendatuko dugu.

Orain estimatutako balioen desbiderapen estandarrak aztertuko ditugu. Estimatutako balioen bariantza eta kobariantza estimatzeko metodoa egiantz handieneko teoriatik hartuko dugu. Estimatzaileak egiantz handieneko logaritmoaren bigarren deribatu partzialez osatutako matrizetik hartuko ditugu. Deribatu partzialen itxura ondokoa da:

$$\frac{\partial L^2(\beta)}{\partial \beta_j^2} = -\sum_{i=1}^{n} x_{ij}^2 p_i(1-p_i) \tag{1.11}$$

$$\frac{\partial L^2(\beta)}{\partial \beta_j \partial \beta_l} = -\sum_{i=1}^{n} x_{ij} x_{il} p_i(1-p_i) \tag{1.12}$$

non $j, l = 1, ..., p$ eta p_i $p(\mathbf{x_i})$ den.

(1.11) eta (1.12) adierazpenetan ditugun terminoek $(p+1)$x$(p+1)$ tamainako matrizea osatzen dute $\mathbf{I}(\boldsymbol{\beta})$ izendatuko dugu. Estimatzaileen bariantza eta kobariantzak matrize honen alderantzizkoaren bidez lortuko ditugu $\mathrm{Var}(\boldsymbol{\beta}) = \mathbf{I^{-1}}(\beta)$. $\mathrm{Var}(\beta_j)$ j. diagonaleko elementua izango da $\hat{\beta}_j$-ren bariantza izango dena eta $\mathrm{Cov}(\beta_j, \beta_l)$ diagonaletik kanpoko elementua $\hat{\beta}_j$ eta $\hat{\beta}_l$ elementuen kobariantza. Ondorioz, $\hat{\mathrm{es}}(\hat{\beta}_j) = \sqrt{\hat{\mathrm{Var}}(\hat{\beta}_j)}$ termino hauek hurrengo atalean erabiliko ditugu.

3 Ereduaren adierazgarritasuna

Behin eredua eraiki eta gero aurreko atalean bezala $\mathbf{X} = (X_1, ..., X_p)$-ren adierazgarritasuna aztertu behar da.

Egiantz handieneko testa

Egiantz handieneko testaren bidez kontraste globala egin dezakegu, hau da, p aldagaien adierazgarritasuna aztertu

$$\begin{cases} H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0 \\ H_1 : \exists j \in (1, 2, ..., p)/\beta_j \neq 0 \end{cases}$$

$\beta_1, \beta_2, ..., \beta_p$ nuluak dituen eredua, 1 Eredua izendatuko dugu eta bestea berriz 2 Eredua

$$G = D(1\mathrm{E}) - D(2\mathrm{E}) = -2ln\left[\frac{\text{1E-ren egiantz handiena}}{\text{2E-ren egiantz handiena}}\right] \approx \chi_p^2$$

G-k p askatasun graduko Khi karratu banaketa jarraitzen du.

p-balioa: $p = P(\chi_p^2 > G)$
Edota bi eredu konpara ditzakegu $1 < q < p$ kasurako

$$\begin{cases} H_0 : \beta_q + 1 = ... = \beta_p = 0 \\ H_1 : \exists j \in (q+1, ..., p)/\beta_j \neq 0 \end{cases}$$

H_0 1 Eredua izendatuko dugu eta H_1 2 Eredua

$$G = D(1\text{E}) - D(2\text{E}) = -2ln\left[\frac{\text{1E-ren egiantz handiena}}{\text{2E-ren egiantz handiena}}\right] \approx \chi^2_{p-q}$$

Aldagai aske gutxiago dituen eredua aldagai aske gehiago dituen ereduarekin konparatzen dugu adierazgarriagoa zein den ikusteko. G estatistikoak $p-q$ askatasun graduko Khi banaketa jarraitzen du.

p-balioa: $p = P(\chi^2_{p-q} > G)$

Wald-en testa

Aldagai bakoitzaren adierazgarritasuna aztertzeko Wald-en testa erabili dezakegu.

$$\begin{cases} H_0 : \beta_j = 0 \\ H_1 : \beta_j \neq 0 \end{cases}$$

H_0 egia izatekotan estatistikoaren balioa

$$W_p = \frac{\hat{\beta}_j}{\sqrt{\left|I^{-1}(\hat{\beta})\right|_{jj}}} \approx N(\beta_j, \left|I^{-1}(\hat{\beta})\right|_{jj})$$

p-balioa: $p = 2P(Z > |w_p|)$

Honakoa da β_j-rentzat konfidantza tartea

$$I^{1-\alpha}_{\beta_j} = \left(\hat{\beta}_j - z_{\alpha/2}\sqrt{\left|I^{-1}(\hat{\beta})\right|_{jj}}, \hat{\beta}_j + z_{\alpha/2}\sqrt{\left|I^{-1}(\hat{\beta})\right|_{jj}}\right)$$

4 Parametroen interpretazioa

Erregresio logistiko sinplean bezala interpretatzen dira.
$OR = e^{\beta_j}$-k $P(Y = 1)$, hau da, arrakastarako probabilitatea nola handitzen den adierazten digu x_j aldagaia unitate bat handitzean (suposatuz x_j aldagai jarraitua dela bestela kasuan kasukoa) gainerako aldagai independienteak konstante hartuta.

1.3 Doikuntz egokitasuna

Atal honetan, eraiki dugun ereduak Y erantzun aldagaia zenbateraino azaltzen duen jakin nahi dugu eta aukeratu dugun eredua ea hoberena den. Hori da hain zuzen doikuntz egokitasuna.
Horretarako bi gauza izan behar ditugu kontuan batetik behatutako balioen $\mathbf{y} = (y_1, y_2, ..., y_n)$ eta estimatutakoen edo doitutakoen $\hat{\mathbf{y}} = (\hat{y_1}, \hat{y_2}, ..., \hat{y_n})$

arteko distantzia txikia izatea eta bestetik errorearen balioa behatutako balioen menpekoa ez izatea.
Suposa dezagun doitutako ereduak p aldagai aske dituela, eta n banako kopurua dela.Termino batzuk definituko ditugu:
J: $\mathbf{X} = (X_1, X_2, ..., X_p)$-ren behatutako balio kopurua. Normalean $J < n$
m_j: $\mathbf{X} = x_j$ duten indibiduo kopurua.$j = 1, ..., J \longrightarrow \sum_{j=1}^{J} m_j = n$
y_j: $\mathbf{X} = x_j$ indibiduoen arrakasta kopurua

1 Pearson-en Khi-karratua eta Deviance

Erregresio logistikoan modu bat baino gehiago dago doitutako eta behatutako balioen arteko desberdintasuna neurtzeko. Doitutako j. aldagaiaren patroia horrela definitzen da:

$$\hat{y_j} = m_j p_j = mj \frac{e^{g(\hat{x}_j)}}{1 + g(\hat{x}_j)}$$

non

$$g(\hat{x}_j) = \hat{logit}(p(x_j))$$

Doitutako eta behatutako balioen desberdintasunak neurtzeko bi modu daude, Pearson-en ondarrak eta Deviance ondarrak. Aldagai jakin baterako Pearsonen ondarra honela definitzen da:

$$r(y_j, g(x_j)) = \frac{(y_j - m_j g(x_j))}{\sqrt{m_j g(x_j)(1 - g(x_j))}}$$

Pearson-en ondarretan oinarritzen den estatistikoa berriz Pearson-en Khi-karratu estatistikoa da

$$\chi^2 = \sum_{j=1}^{J} r(y_j, g(x_j))^2$$

Deviance ondarra definituko dugu

$$\mathrm{d}(y_j, g(x_j)) = \left\{ \left[y_j \ln \left(\frac{y_j}{m_j g(x_j)} \right) + (m_j - y_j) \ln \left(\frac{(m_j - y_j)}{m_j (1 - g(x_j))} \right) \right] \right\}^{\frac{1}{2}}$$

$y_j = 0$ duten aldagaientzat deviance ondarra:

$$d(y_j, g(p_j)) = -\sqrt{2m_j |\ln(1 - g(p_j))|}$$

eta $y_j = m_j$ rako berriz

$$d(y_j, g(p_j)) = \sqrt{2m_j |\ln(g(p_j))|}$$

Deviance ondarretan oinarritzen den estatistikoa:

$$D = \sum_{j=1}^{J} \mathrm{d}(y_j, g(p_j))^2$$

Doikuntz egokitasuna ona den hipotesipean $J - (p + 1)$ askatasun graduko Khi-karratu banaketa jarraitzen dute bi estatistikoek.

2 Hosmer-Lemeshow testa

Suposa dezagun $J = n$ dela. n zutabe eraikitzen dira bakoitza estimatutako probabilitateari dagokiona eta txikienetik handienera ordenatzen dira. Zutabe horiek g taldetan banatzen dira bi irizpide hauek erabiliz:

- estimatutako probabilitateen pertzentilak
- estimatutako probabilitateen balio finkoak

Normalean g=10 erabiltzen da. Hosmer eta Lemeshow-en estatistikoa horrela definitzen da:

$$\hat{c} = \sum_{k=1}^{g} \frac{(O_k - n_k \bar{p_k})^2}{n_k \bar{p_k}(1 - \bar{p_k})}$$

non $n_k \equiv k$ taldeko aldagaien konbinazio kopurua.
$O_k = \sum_{j=1}^{n_k} y_j$ n_k taldeko arrakasta kopurua den eta $\bar{p_k} = \sum_{j=1}^{n_k} \frac{m_j \bar{p_j}}{n_k}$ estimatutako batez besteko probabilitatea.
Doikuntza egokia den hipotesipean $\hat{c}$ estatistikoak $g - 2$ askatasun graduko Khi karratu banaketa jarraitzen du $J = n$-rako. Kasu honetan p-balioa horrela kalkulatzen da: $p = P(\chi^2_{g-2} > \hat{c})$

AKAIKE irizpidea (AIC)

Ereduak konparatzea ahalbideratzen du. Deviance eta ereduaren parametro kopuruarekin lotua dago.
$AIC = -2ln(\text{Ereduaren egiantz handiena}) + 2p$

Doikuntz egokitasunerako erabiltzen diren beste bi tresna sailkapen taulak eta ROC kurba dira, hurrengo kapituluan azalduko ditugu mozketa puntuekin lotura dutelako.

2. Kapitulua

Mozketa puntuak kalkulatzeko teknikak

2.1 Motibazioa

Mozketa puntu optimoak kalkulatzea izango da atal honen helburua.
1. Kapituluan arrakastarako probabilitateak estimatu ditugu erregresio logistikoaren bidez baina Y aldagai dikotomikoa denez, sailkapen bat egin beharko dugu estimatutako probabilitateak arrakasta edo porrota izango diren erabakitzeko.

Sailkapen hori c mozketa puntu baten arabera egingo dugu $\hat{p}(x) \geq c$ bada $\hat{Y} = 1$ bezala sailkatuko dugu eta $\hat{p}(x) < c$ bada $\hat{Y} = 0$ bezala. Garrantzitsua da beraz, c puntu hori egoki aukeratzea sailkapena behatutako datuen ahalik eta antzekoena izateko.

Ereduaren arabera aldagai batzuk ahalmen handiagoa dute sailkapenerako, hau da arrakastak eta porrotak sailkatzean behatutako arrakasta eta porroten kopuru oso antzekoak lortzen dituzte. Hori neurtzeko sailkapen taulak eta ROC kurba ditugu eta gure aldagaiak arrakastak eta porrotak ahalik eta gehien bereiztea interesatuko zaigu. Ideia honi aldagaiaren diskriminazio ahalmena esaten zaio.

1 Sailkapen taulak

Doitutako erregresio logistikoko eredu bat laburtzeko modu bat sailkapen taula erabiltzea da, demagun c aukeratu dugun mozketa puntua dela. Estimatutako probabilitatea c balio hori baino altuagoa bada $\hat{Y} = 1$ bezala sailkatuko dugu eta txikiagoa bada $\hat{Y} = 0$.

Guztira n indibiduo ditugu eta behatutako porrot kopurua $a+b$ eta behatutako arrakasta kopurua $c + d$ izendatuko ditugu (**2.1 Taula**). Behatutako balioak finkoak dira estimatutakoak aldiz c mozketa puntuaren arabera aldatzen joango dira.

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

2.1. Taula. Sailkapen taula, sentikortasuna eta espezifizitatea

Sailkapen taula batetik honako informazioa atera daiteke:

Egoki sailkatutako arrakasta proportzioa adierazpen honetan datza $\frac{d}{c+d}$x100 eta zatidura honi $\frac{d}{c+d}$ **sentikortasuna** esaten zaio.

Egoki sailkatutako porrot proportzioa adierazpen honetan datza $\frac{a}{a+b}$x100 eta zatidura honi $\frac{a}{a+b}$ **espezifizitate** esaten zaio

Sailkatze egokiko proportzio globala ondoko adierazpenean datza $\frac{a+d}{n}$x100

Sentikortasuna eta espezifizitatea beste modu honetan ere uler daitezke:
Asmatzeko probabilitatea

- sentikortasuna $\Rightarrow Se(c) = \mathrm{P}(\hat{p}(x) \geq c|Y = 1)$
  Jakinda arrakasta ematen dela, estimatutako arrakastarako probabilitateak c baino handiagoak edo berdinak izateko probabilitatea.

- espezifizitatea $\Rightarrow Sp(c) = \mathrm{P}(\hat{p}(x) < c|Y = 0)$
  Jakinda porrota ematen dela, estimatutako arrakastarako probabilitateak c baino txikiagoak izateko probabilitatea.

Akatsa egiteko probabilitatea

- 1- sentikortasuna $\Rightarrow 1 - Se(c) = \mathrm{P}(\hat{p}(x) < c|Y = 1)$
  Jakinda arrakasta ematen dela estimatutako arrakastarako probabilitateak c baino txikiagoak izateko probabilitatea.

- 1- espezifizitatea $\Rightarrow 1 - Sp(c) = \mathrm{P}(\hat{p}(x) \geq c|Y = 0)$
  Jakinda porrota ematen dela estimatutako arrakastarako probabilitateak c baino handiagoak edo berdinak izateko probabilitatea.

Sailkapen taulako balioak ulertzeko beste modu bat ondokoa da (**2.2 Taula**)

- EP (Egiazko positibo): Egiazkoa dela esaten dugu sailkapena zuzen egin delako, hau da, behatutako arrakastetatik $\hat{Y} = 1$ bezala sailkatu diren indibiduo kopurua da.

	Estimatutakoak ($\hat{Y}$)		
Behatutakoak (Y)	0	1	
0	EN	FP	EN+FP
1	FN	EP	FN+EP
	EN+FN	FP+EP	n

2.2. Taula. Sailkapen taula, egiazkoak eta faltsuak

- EN (Egiazko negatibo): Aurreko kasuan bezala sailkapena zuzen egin da, behatutako porrotetatik $\hat{Y} = 0$ bezala sailkatu diren indibiduo kopurua izango da.
- FN (Faltsu negatibo): Faltsua dela esaten dugu sailkapena oker egin delako, behatutako arrakastak $\hat{Y} = 0$ bezala sailkatu diren indibiduo kopurua adierazten du (horregatik negatibo).
- FP (Faltsu positibo): Aurreko kasuan bezala sailkapena oker egin da, behatutako porrotak $\hat{Y} = 1$ bezala sailkatu diren indibiduo kopurua da (horregatik positibo).

2 ROC kurba eta AUC

Receiver operating characteristic, ROC, kurba bat da non mozketa puntua aldatu ahala (1-espezifizitatea) sentikortasunarekiko adierazten duen.

$$\text{ROC}(.) = \{((1 - Sp(c)), Se(c)), c \in (-\infty, \infty)\} \tag{2.1}$$

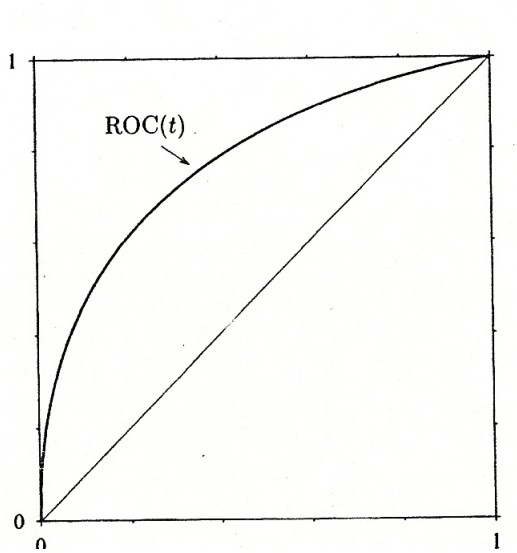
c puntua handitu ahala, $(1-Sp(c))$ eta $Se(c)$-ren balioak txikitu egiten dira. Hau da, $Sp(c)$-ren balioa handitu eta $Se(c)$-rena txikitu [3].

$c = \infty$ denean, $\lim_{c\to\infty} Se(c) = 0$ eta $\lim_{c\to\infty} 1 - Sp(c) = 0$ ditugu eta
$c = -\infty$ denean berriz, $\lim_{c\to-\infty} Se(c) = 1$ eta $\lim_{c\to-\infty} 1 - Sp(c) = 1$.

Ikusi dugun bezala ROC, funtzio monotonoa da eta gorakorra $[0, 1]$x$[0, 1]$ tartean definitua. Beste modu honetan ere idatz dezakegu:

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

2.1 irudian ikus dezakegu deskribatu dugun kurbaren itxura.



2.1. irudia. ROC kurbaren adibide bat.

Informaziorik gehitzen ez duen aldagaia, indibiduoak arrakasta eta porrot bezala bereizten ez dituen aldagaia da.
Edozein c punturako $(1 - Sp(c)) = Se(c)$ betetzen da eta X aldagaia erabilera gabekoa dela esango dugu. Kasu horretan $\mathrm{ROC}(t) = t$ 1 maldako funtzioa dugu.

X aldagai perfektua indibiduoak erabat bereizten dituena da, c mozketa puntu baterako $(1 - Sp(c)) = 0$ eta $Se = 1$ dituena. Unitate bateko koadrante positiboan ezkerraldeko goi ertzetik zehar pasatzen dena.

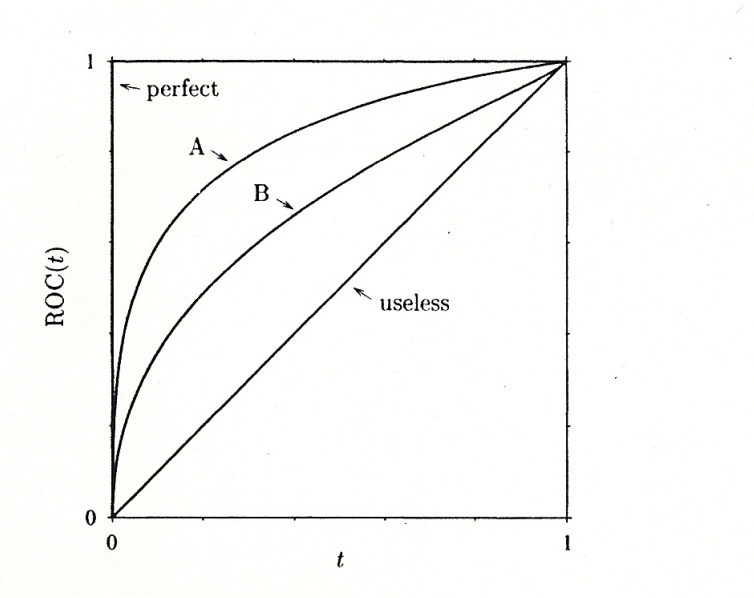
Aldagai gehienen ROC kurbak orain aipatu ditugun aldagai horien(erabilerarik gabeko aldagaia eta aldagai perfektu) kurben artean egoten dira. Bi aldagaien ROC kurbak konparatu nahi baditugu ezkerreko ertzera gehien hurbiltzen dena izango da hobea **2.2 irudian** ikus daiteken bezala. Kasu honetan A aldagaia hobea da $(1 - Sp(c))$ bakoitzeko $Se(c)$ handiagoa delako.

AUC

Gehien zabaldua dagoen ROC kurbaren neurria bere azalera da. Honela definitzen da:

$$\mathrm{AUC} = \int_0^1 \mathrm{ROC(r)}\,\mathrm{d}r \tag{2.2}$$

Aldagai perfektua, ROC kurba perfektua duen aldagaia da, AUC = 1.0 duena. Informaziorik gehitzen ez duen aldagaiak berriz, $\mathrm{ROC}(t) = t$
AUC = 0.5 du. Lehen gertatzen zen bezala aldagai gehienen AUC-ea balio hauen artean egongo da.



2.2. irudia. ROC kurba bi aldagaientzat, A eta B, non A aldagaia hobea den.

Kurbaren azpiko azalera geroz eta handiagoa izan diskriminazioa handiagoa izango da, hau da, azalera hori 1 baliotik geroz eta gertuago egon diskriminazio handiagoa izango du (0,1) puntutik gertuago egongo delako. A aldagaia B baino hobea bada:

$$\mathrm{ROC}_A(t) \geq \mathrm{ROC}_B(t) \qquad \forall t \in (0,1)$$

Orduan, AUC-en ordena ere mantentzen da, nahiz eta kontrakoa betetzen ez den.

$$\mathrm{AUC}_A \geq \mathrm{AUC}_B.$$

2.2 Mozketa puntu optimoak lortzeko metodoak

Orain arte behin eta berriz aipatu ditugu mozketa puntua eta honek sailkapenerako daukan garrantzia. Beraz, mozketa puntua nola aukeratu aztertzen hasiko gara. Behin c hori aukeratuta sentikortasun eta espezifizitatearen bidez sailkapenaren egokitasuna ikusiko dugu eta AUC-eak aldagaiak sailkapenerako duen gaitasuna emango digu.

Metodo batzuk aztertuko ditugu mozketa puntu hauek optimizatzen dituztenak, kasu honetan sentikortasun eta espezifizitate terminoetan oinarrituko gara, baina beste metodo batzuk ere badaude beste irizpide batzuk oinarritzat dituztenak gaixotasunaren kostu-irabaziak esate baterako.

- **SeDSp**: Metodo honek sentikortasun eta espezifizitatearen arteko diferentzia ahalik eta txikiena izateko c-ren balioa zein den kalkulatzen du. Ondokoa da formula:

$$c : \min\{|Se(c) - Sp(c)|\} \tag{2.3}$$

- **Youden**: Youden-en indizean oinarritzen da [4]. Esan dugu $\frac{a}{a+b}$ egoki sailkatutako arrakasta kopurua dela eta berdin esan dezakegu, $\frac{b}{a+b}$ oker sailkatutako arrakasta kopurua dela. Beraz egoki sailkatutako arrakastaren neurria $\frac{a-b}{a+b}$ da. Modu berean definitu daiteke egoki sailkatutako porrotaren neurria $\frac{d-c}{c+d}$. Bi termino hauen batez bestekoa hartuz faltsu positibo eta faltsu negatiboak orekatzen ditugu.

$$J = \frac{1}{2}\left[\frac{a-b}{a+b} + \frac{d-c}{c+d}\right] = \frac{a}{a+b} + \frac{d}{c+d} - 1 = Se + Sp - 1$$

  J-ri Youden-en indizea esaten diogu. Indizeak 0 eta 1 bitarteko balioak hartzen ditu. a=b edo c=d bada, aldagaiak ez du diskriminazio ahalmenik arrakastak sailkatzeko edo porrotak sailkatzeko eta J=0 izango litzateke. J=1 izateko faltsu positibo eta negatiboak biak, 0 izan behar dute.

  Indize hau ahalik eta altuena izatea komeni zaigu egoki sailkatutako arrakasta eta porrotak maximoak izateko. Beraz Youden-en mozketa puntu optimoa honela kalkulatzen da:

$$c : \max\{Se(c) + Sp(c) - 1\} \tag{2.4}$$

- **ROC01**: ROC kurban $(0, 1)$ puntutik hurbilen dagoen koordenatua lortzeko c-ren balioa zein izan behar duen kalkulatzen du [5].

  Lehenago aipatu dugu aldagai perfektuak $Se = 1$ eta $Sp = 1$ dituela, hau da ROC kurba (0,1) puntutik igarotzen da. Orokorrean ordea, $Se < 1$ eta $Sp < 1$ dira eta (0,1) puntutik gertuen dagoen kurbaren puntua izango da c hori. $(1-Sp(c), Se(c))$ (0,1) puntutik gertu egotea, (Sp(c),Se(c)) (1,1) puntutik gertu egotearen berdina da. Hau da, bi puntu horien distantzia minimoa izatea.

$$min\{\sqrt{(1 - Se(c))^2 + (1 - Sp)^2}\}$$

  edo distantzia hori minimizatzen duen mozketa puntu optimoa

$$c : \min\{(Se(c) - 1)^2 + (Sp(c) - 1)^2\} \tag{2.5}$$

2.3 Softwarea

Erabili dugun softwarea R izan da [6]. Erregresio logistikoa garatzeko eta mozketa puntu optimoak topatzeko, besteak beste, ondoko libreriak erabili ditugu:
stats libreriako glm funtzioa eredu lineal orokorrak doitzeko erabiliko dugu. Libreria bereko deviance eta pchisq funtzioak ereduaren adierazgarritasuna neurtzeko erabili ditugu. **MKmisc** libreriarekin doikuntz egokitasuna aztertu dugu, HLgof.test funtzioaz baliatuz Hosmer-Lemeshow-ren testa gure ereduari aplikatzeko.

2.2 ataleko metodoak aplikatzeko R-ren **OptimalCutpoints** libreria erabiliko dugu[7].
Libreria horretako optimal.cutpoints funtzioan ondoko aukera hauek zehaztu beharko ditugu:

- *X*: Menpeko aldagaiaren izena.
- *status*: Gaixo dagoen pazientea osasuntsu dagoen pazienteaz bereizten duen aldagaia. (Orokortu daiteke hil/ez hil, ingresatu/ez ingresatu kasuetara).
- *tag.healthy*: status aldagaian osasuntsu bezala kontsideratu dugun balioa.
- *method*: Mozketa puntuak optimizatzeko paketearen metodoa, hainbat aukera daude baina goian adierazitakoak erabiliko ditugu:(2.3), (2.4) eta (2.5)
- *data*: Aztertu nahi dugun datu basea.
- *direction*: ROC kurbaren norabidea adierazten du. Defektuz mozketa puntutik behera dauden probabilitateak porrota bezala sailkatuz.

3. Kapitulua

Aplikazioa

Kapitulu honetan aurreko bien aplikazioa egingo dugu datu base bat erabiliz. Datuak Galdakao-Usansolo Ospitaleko ikerkuntza unitateak ahalbideratutakoak dira eta bihotz-gutxiegitasun desorekatu larria (BGDL) duten pazienteei dagokie.

BGDL-a gaixotasun larri bat da, non bihotzak ez duen behar bezala organismoak behar duen odola ponpeatzen. Hau da, odola ez da gai organismoak behar duen oxigenoa eta elikagaiak garraitzeko. Gaixotasun hau edozein adin tartetan azaleratu daiteke baino ohikoena 65 urtetik gorako pertsonek pairatzea da.

Datu basea, Euskal Autonomia Erkidegoko hiru ospitale publikoko 252 pazienteren datuek osatzen dute. Erantzun aldagaiak pazientea ospitalera iritsi eta 90 egunera hil egin den edo ez adierazten digu ($Y = 1$ pazientea hil egin da, $Y = 0$ pazientea ez da hil) ingresoan hil diren pazienteak kontuan hartu gabe. Aldagai askea, kasu honetan jarraitua dena NT-pro-BNP izango da.

NT-pro-BPN, BNP-ren markadorea da odolean, hormona honen balioa handitu egiten da luzapen kardiakoa edo estresa handitzean. Bihotzeko horma zabalago dagoenean gehiegizko odolaren bolumenagatik edo hondatuta bihotzerako odol fluxu faltagatik, BNP-a handitu egiten da eta berarekin NT-pro-BNP-a.

Azterketa honekin medikuei pazientea ingresatu edo ez erabakiaren aurrean laguntzea lortu nahi dugu NT-pro-BNP erabiliz. Horretarako hormonaren zein baliotik aurrera hiltzeko probabilitatea handiagoa den ikusi behar dugu, hau da NT-pro-BNP-ren mozketa puntu optimoa kalkulatu.

Erregresio logistiko sinpleko eredu bat erabiliko dugu, Y aldagai dikotomikoa delako eta honen menpeko aldagai bakarra dugulako jarraitua dena.
252 pazientetatik 221 dira hiltzen ez direnak eta 31 hiltzen direnak NT-pro-BNP aldagaiak 27 balio galdu ditu horrenbestez 225 daturekin egingo dugu lan.

Aldagai askearen taula deskribatzailea aztertzen badugu (**3.1 Taula**) balio altuak hartzen dituela ikusten dugu. Aldagaiaren balio minimotik maximora dagoen diferentzia $37195,8$-koa da. Pazienteen %50ak 4059 balioa baino bajuagoa dute NT-pro-BNP-ren balioa eta beste %50ak berriz altuagoa.

Hiltzeko probabilitatearen eta NT-pro-BNP-ren arteko erlazioa aztertzeko aldagai jarraitua kuartiletan banatuko dugu eta maiztasun taula aztertuko dugu. (**3.2 Taula**)

	NT-pro-BNP
Minimoa	100.2
1. Kuartila	2055
Mediana	4059
Batez bestekoa	6126.5
3. Kuartila	7933
Maximoa	37296

3.1. Taula. Taula deskribatzailea

	BGDL			
NT-pro-BNP	n	Arrakasta	Porrota	p
100.2	1	0	1	0
2055	42	3	39	0.071
4059	70	5	65	0.071
7933	44	4	40	0.09
37296	68	16	52	0.235
Guztira	225	28	197	0.467

3.2. Taula. Maiztasun taula

Aldagaiaren eta arrakastarako probabilitatearen arteko erlazioa lineala izatea nahi dugu erregresio logistiko sinpleko eredua eraikitzeko. Maiztasun taulan ikus dezakegu NT-pro-BNP-ren balioa haunditu ahala p handituz doala.

Honakoa da lortu dugun logit-aren eredua:

$$logit(p) = -2.68 + 9.71\text{x}10^{-5}\text{NT-pro-BNP} \tag{3.1}$$

non BGDL duen pazientea hiltzeko probabilitatea honakoa den

$$p = \frac{e^{-2.68+9.71\text{x}10^{-5}\text{NT-pro-BNP}}}{1 + e^{-2.68+9.71\text{x}10^{-5}\text{NT-pro-BNP}}}$$

Aldagaia	$\beta^\star$	$OR^\star$	$IC_{OR^\star}$	$\hat{p}$	AUC
NT-pro-BNP	$9,71\text{x}10^{-3}$	1,01	(1,004-1.02)	$< 0,001$	$0,69$

3.3. Taula.

1.kapituluan azaldutako prozesua jarraitu dugu, eredua eraiki ondoren, koefizienteak interpretatu eta aldagaia adierazgarria den aztertuko dugu. Informazio hori **3.3 Taulan** dago laburtuta.

Ereduaren interpretazioa errezago egiteko asmoz, NT-pro-BNP aldagaiaren aldakuntza 100 unitatekoa hartu dugu. Beraz, $\beta^\star = 100\ \beta$ orduan $OR^\star = e^{100\ \beta}$ izango da.

Aipatu bezala *Odds Ratioa*-ren bidez oso erraz interpreta dezakegu aldagaia: 1.01 aldiz handiagoa da hiltzeko probabilitatea NT-pro-BNP hormona 100 unitate handitzean.
Beraz, 1. Kapituluan ikusi dugun bezala NT-pro-BNP aldagaiaren 100 unitateko hazkundeak, 1,01 aldiz probablegoa egiten du hiltzeko arriskua, bizirik irautekoa baino.

Aldagaia adierazgarria dela OR-ren konfidantza tartetik eta p-baliotik ondorioztatzen dugu. **3.3 Taula**-n agertzen den p-balioa Wald-en testetik lortutakoa da. H_0 errefusatzen dugu, $p < 0,05$ izateagatik eta beraz, NT-pro-BNP-ek hiltzeko hiltzeko arriskuan duen eragina adierazgarria da.

Azkenik doikuntz egokitasuna egingo dugu eraikitako eredua egokia den ikusteko. Hosmer-Lemeshow-ren testa erabiliz lortzen dugun estatistikoaren balioa ondokoa da: 0.4984. Ondorioz ereduaren doikuntza egokia dela esan dezakegu. Lortutako AUC-aren balioa 0,69-koa da.

Behin eredua eraiki eta gero gure helburuarekin hasiko gara, mozketa puntu desberdinak aukeratuta ROC kurba eta sailkapen taulak eraikiko ditugu

	Estimatutakoak		
Behatutakoak	0	1	
0	194	3	197
1	26	2	28
	220	5	225

3.4. Taula. $c = 0,5$

eta kasu bakoitzean NT-pro-BNP non moztu ikusiko dugu. Hasiera batean $c = 0,5$ hartuko dugu zer gertatzen den ikusteko **3.4 Taula**.

Sentikortasun eta espezifizitaterako $Se = 0,07$ eta $Sp = 0,98$ dira lortzen ditugun balioak. Sentikortasunerako balioa oso bajua da, hau da paziente bat hil eta guk horrela sailkatzeko probabilitatea oso txikia da. Garbi geratzen da ez dela mozketa punturik egokiena sentikortasuna altuagoa izatea interesatzen zaigulako.

Jarraian 2.kapituluan azaldutako metodoak aplikatuko ditugu

Metodoa	c	Aldagaiaren mozketa puntua	Sentikortasuna	Espezifizitatea
Youden	0,118	6875	0,607	0,731
SeDSp	0,108	5814	0,643	0,645
ROC01	0,118	6875	0,607	0,731

3.5. Taula. Mozketa puntu optimoa lortzeko hiru metodoak

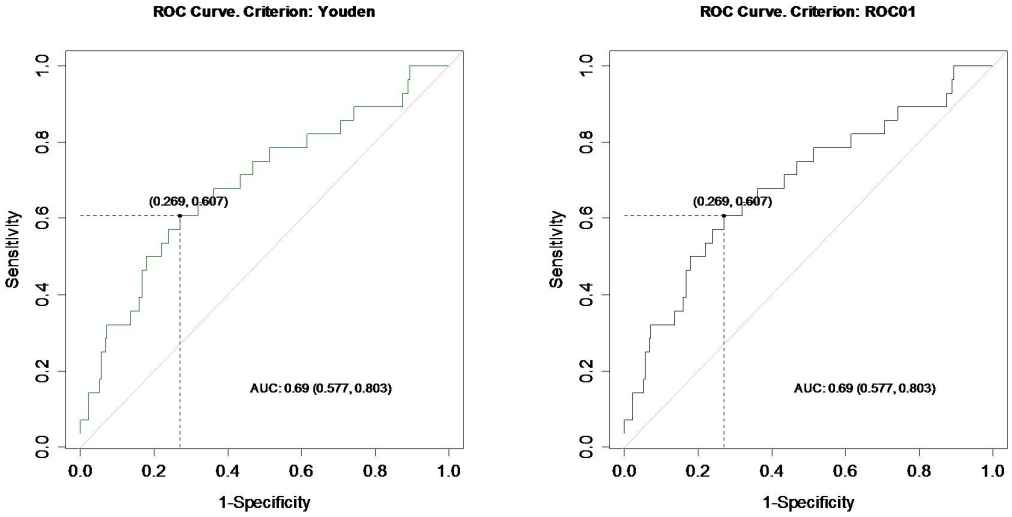
Ikus dezakegun bezala, Youden eta ROC01 metodoek emaitza berdinak ematen dizkigute (**3.5 Taula**). Kasu honetan NT-pro-BNP aldagaiaren mozketa puntua 6875 dela ikus dezakegu ($c = 0,118$).

Balio hori baino NT-pro-BNP altuagoa duten pazienteak $\hat{Y} = 1$ bezala sailkatuko ditugu eta besteak $\hat{Y} = 0$ bezala. Bi metodo hauek erabiliz NT-pro-BNP 6875 baliotik gora duten pazienteak ingresatzea komeniko da hiltzeko probabilitatea handitzen delako.

Sailkapen taula kalkulatzen badugu $c = 0,118$ baliorako (**3.6 Taula**) honako sentikortasun eta espezifizitate balioak lortzen ditugu:
$Se = 0,607$ eta $Sp = 0,731$ eta sailkatze egokiko proportzio globala ondokoa da: $\%71,6$.

	Estimatutakoak		
Behatutakoak	0	1	
0	144	53	197
1	11	17	28
	155	70	225

3.6. Taula. Youden eta ROC01 metodoekin lortutako mozketa puntu optimoarekin ($c = 0,118$) lortzen den sailkapen taula.



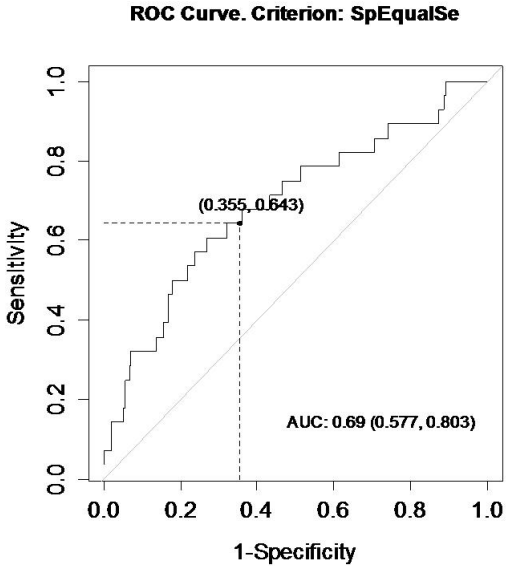
3.1. irudia. ROC kurba Youden eta ROC01 metodoak erabiliz

Orain SeDSp metodoa aztertuko dugu kasu honetan $c = 0,108$ da (**3.5 Taula**) eta NT-pro-BNP aldagaia 5814 puntuan moztu beharko genuke.

	Estimatutakoak		
Behatutakoak	0	1	
0	128	69	197
1	10	18	28
	139	86	225

3.7. Taula. SeDSp metodoarekin lortutako mozketa puntu optimoarekin ($c = 0,108$) lortzen den sailkapen taula.

$c = 0,108$-rentzat sailkapen taula kalkulatuz $Se = 0,643$ eta $Sp = 0,649$ balioak lortzen ditugu (**3.7 Taula**).



3.2. irudia. ROC kurba SpDSe metodoa erabiliz

Sailkapen egokiko proportzio globala %64, 9 da.
Bi kasuak konparatzen baditugu, sailkapen egokiko proportzio global altuagoa dute Young eta ROC01 metodoek SeDSp-k baino. Faltsu positiboak bajuagoak dira lehenengo bi metodoetan hirugarrenean baino 16 indibiduoko diferentzia batekin.

Sentikortasuna altuagoa da lehenengo kasuan bigarrenean baino, nahiz eta espezifizitatea SeDSp metodoarekin altuagoa izan. SeDSp metodoa aplikatuz faltsu negatibo bat gutxiago dugu aurreko kasuetan baino.

4. Kapitulua

Ondorioak

Koefizientearen interpretaziotik hasiko gara, lehenago aipatu dugu BGDL duen pazientea hiltzeko probabilitatea 1,01 aldiz hazten dela NT-pro-BNP-ren balioa 100 unitate handitzean.

Aipatu dugun bezala, AUC-a ereduaren diskriminazio ahalmena neurtzeko erabiltzen den parametroa da. Gure erregresio logistiko sinpleko ereduarekin lortutako AUC-a 0,69 da. Horrek esan nahi du, eredu hau simpleegia dela heriotza aurresateko, hau da, NT- pro-BNP aldagai aurresalea izan arren (aldagai adierazgarria dela lortu dugu), ez litzateke nahikoa izango heriotza aurresateko eredu bat garatu nahi izango bagenu. Eredu anizkoitz bat doitu beharko genuke, beste hainbat parametro kliniko kontuan harturik, baina hori lan honen helburutik kanpo geratzen da.

Gure helburua, NT-pro-BNP parametro klinikoarentzako mozketa puntu optimoa lortzea zen, horretarako erregresio logistiko simplea erabiliz. Behin hori lortuta, NT-pro-BNP lortutako mozketa puntua erabiliz bi kategorietan kategorizatuko genuke eta heriotzarentzako eredu aurresale batean aldagai aurresale bezala erabili, aztertu beharreko beste hainbat parametro edo aldagaiekin batera noski, hala nola adinarekin adibidez.

Aztertu diren mozketa puntuak kalkulatzeko teknikak orokorrak dira. Badaude beste teknika batzuk sentikortasuna eta espezifizitatea maximizatu edo minimizatzen dituztenak (bakoitza bere aldetik) baita horien balio jakin baterako mozketa puntuak kalkulatzen dituztenak ere. Horrelakoetan ordea, gai honetan aditua den norbaiten iritzia beharrezkoa da.

Metodo hauen arteko desberdintasunak aztertzen baditugu, ikuspuntu matematikotik ROC01 eta Young-en metodoarekin geratuko ginateke. Batetik sailkatze egokiko proportzioa altuagoa da, SeDSp metodoarekin lortzen duguna baino eta baita espezifizitatea ere.

Metodo hauekin FN=11 eta FP=53 ditugu, SeDSp metodoarekin aldiz, FN=10 eta FP=69. Faltsu negatiboen artean ez dago diferentzia handirik

baina faltsu positiboetan bai, 16 indibiduokoa hain zuzen. Irizpide hau jarraituz ere, Young eta ROC01 metodoekin geratuko ginateke akatsa egiteko probabilitatea ahalik eta bajuena izatea nahi dugulako.

Faltsu positibo eta negatiboek oker sailkatutako indibiduo kopurua adierazten dute. Lehenengo kasuan $\hat{Y} = 1$ bezala sailkatu da porrota eta beste kasuan $\hat{Y} = 0$ bezala arrakasta. Nahiz eta 16 indibiduoko aldea handia iruditu, baliteke medikuei bietarikoren bat altuagoa edo bajuagoa izatea interesatzea.

Nahiz eta faltsu negatiboen diferentzia indibiduo batekoa izan oso paziente gutxi hiltzen direnez, 252tik 31,hiltzen direnak heriotza bezala sailkatzea eta sentikortasuna ahalik eta altuena izatea komeni zaigu. Kasu honetan SeDSp metodoa aukeratuko genuke. Hala ere, azken erabakia medikuen esku uzten dugu.

A. eranskina

R-ko kodea

```
#####################################################
#####################################################
##      GRADU AMAIERAKO LANA:APLIKAZIOA            ##
#####################################################
#####################################################
############ Erregresio logistiko sinplea R-ren bidez

##Bihotz-gutxiegitasuna duten pazienteen heriotzaren aurreikuspena 90 egunera.

#Datuak inportatu eta laburpena egin

Datuak<-read.table("C:/Users/user/Desktop/datuak maialen icc.csv",header=TRUE,
sep=",",na.strings="NA",dec=".",strip.white=TRUE)
library(relimp,pos=4)

Datuak$LAB5<-as.numeric(Datuak$LAB5) ##x aldagaia jarraitu moduan irakurtzeko
Datuak$exitus_90dias<-as.factor(Datuak$exitus_90dias)

summary(Datuak)##taula deskribatzailea
quantile(Datuak$LAB5,na.rm = TRUE)##kuartilak

 #Eredua eraiki (parametroen estimazioa)

mod <-glm(exitus_90dias ~ LAB5,data=Datuak,family=binomial)
summary(mod)

## X-ren adierazgarritasuna (beta_1 hipotesi kontrastea)

# Egiantz handieneko metodoa erabiliz
```

```
pchisq(mod$null.deviance-deviance(mod), df=1, lower=FALSE)
## Hipotesi nulua errefusatu X aldagaia adierazgarria.

#Wald-en testa erabiliz
## summary-ko p-balioak dira.

####beta1-en konfidantza tartea
confint.default(mod)

##Odds ratioa eta bere konfindantza tartea
exp(coefficients(mod))
exp(confint.default(mod)) ###### Odds ratio-aren konfidantza tartea

##Doikuntz egokitasuna
###Hosmer Lemeshow-en testa
library(MKmisc)
HLgof.test(mod$fitted,mod$y, ngr = 3)
## hipotesi nulua ez errefusatu, doikuntza egokia da.

### arrakastarako probabilitatea sailkatu. <0.5 edo >0.5
phat<-fitted(mod)
yhat<-ifelse (phat < 0.5, 0, 1)

####sailkapen taula
taula<-xtabs(~mod$y+yhat, data=Datuak)
taula

###AUC
library(pROC)
roc(mod$y,mod$fitted)
roc_mod<-roc(mod$y,mod$fitted)

######### MOZKETA PUNTUAK KALKULATZEKO TEKNIKAK
#############

################################################
# Youden-en metodoa erabiliz ("Youden")
################################################

library(OptimalCutpoints)
optimal.cutpoint.Youden <- optimal.cutpoints(X = "LAB5", status = "exitus_90dias",
tag.healthy = 0, methods = "Youden", data = Datuak, pop.prev = NULL,
control = control.cutpoints(), ci.fit = FALSE, conf.level = 0.95, trace =FALSE)
```

```
summary(optimal.cutpoint.Youden)
plot(optimal.cutpoint.Youden)
#########################################################################
#   Sentikortasun eta espezifizitaterako berdintasuna ("SpEqualSe")
#########################################################################

library(OptimalCutpoints)
optimal.cutpoint.SpEqualSe <- optimal.cutpoints(X = "LAB5", status ="exitus_90dias"
tag.healthy = 0, methods = "SpEqualSe", data = Datuak, pop.prev = NULL,
control = control.cutpoints(), ci.fit = FALSE, conf.level = 0.95, trace = FALSE)

summary(optimal.cutpoint.SpEqualSe)
plot(optimal.cutpoint.SpEqualSe)

#########################################################################
# ROC kurban (0,1) puntutik hurbilen dagoen puntua ("ROC01")
#########################################################################

library(OptimalCutpoints)
optimal.cutpoint.ROC01 <- optimal.cutpoints(X = "LAB5", status = "exitus_90dias",
tag.healthy = 0, methods = "ROC01", data = Datuak, pop.prev = NULL,
control = control.cutpoints(), ci.fit = FALSE, conf.level = 0.95, trace = FALSE)

summary(optimal.cutpoint.ROC01)
plot(optimal.cutpoint.ROC01)

########################################################## Youden/ROC01
##sailkapen taula (c=0.1178)

phat<-fitted(mod)
yhat<-ifelse (phat < 0.1178, 0, 1)

taula<-xtabs(~mod$y+yhat, data=Datuak)
taula

############################################################## SeDSp
##sailkapen taula (c=0.1076)

phat<-fitted(mod)
yhat<-ifelse (phat < 0.1075864419, 0, 1)

taula<-xtabs(~mod$y+yhat, data=Datuak)
taula
```

Bibliografia

[1] David W.Hosmer, Stanley Lemeshow, Applied Logistic Regression, (2000), 1-90.

[2] Mc Cullagh and Nelder, Generalized Linear Models, (1989).

[3] Margaret Sullivan Pepe, The Statistical Evaluation of Medical Tests for Classification and Prediction, (2003), 66-81.

[4] W. J. Youden, Index for Rating Diagnostic Tests, Cancer **3**, (1950), 32-35.

[5] Neil J. Perkins, Enrique F. Schisterman, The Inconsistency of "Optimal" Cut-points Using Two ROC Based Criteria, Am J Epidemiol. **163**(**7**), (2006), 670-675.

[6] R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing 2013. URL http://www.R-project.org/.

[7] Monica Lopez-Raton, Maria Xose Rodriguez-Alvarez, Computing optimal cutpoints in diagnostic tests, (2013).