TREBALL DE FI DE GRAU

Grau en Estadística

Títol: Caracterització de les empreses seleccionades per EIT Health ajustant Model Lineal Generalitzat de resposta binaria

Autor: Pau Font Farré

Directors: Magí Lluch-Ariet, Ramon Alemany Leira

Departament: Econometria, Estadística y Economia Aplicada

Convocatòria: Juny 2021

UNIVERSITAT DE BARCELONA

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

Resum

Els models lineals generalitzats són una eina molt útil per d'explicar l'efecte d'unes variables (explicatives) sobre una variable resposta. Tenen dues utilitats principals que són: l'anàlisi estructural de cadascun dels coeficients i la realització de prediccions. Aquest treball ajusta un model per conèixer els factors més rellevants que expliquen la selecció d'empreses per EIT Health i poder realitzar prediccions sobre la seva selecció. EIT Health és una organització de la Unió Europea que és dedica a ajudar i finançar empreses i projectes relacionats amb el món de la salut. Ateses les característiques de la variable resposta s'ajustarà un model de resposta binaria. Aquest treball permetrà conèixer millor el perfil de les empreses seleccionades per EIT.

Paraules clau: *models lineals, significació, lògit, odds ràtio, startups, empreses de salut*

Classificació AMS (MSC2010): 62J12 Generalized linear models

62-07 Data analysis

62J20 Diagnostics

Agrair a la meva família i als meus amics per escoltar-me parlar del treball. I sobretot als tutors, tan al Magí com al Ramon, pels consells, recomanacions i ajuda.

Índex

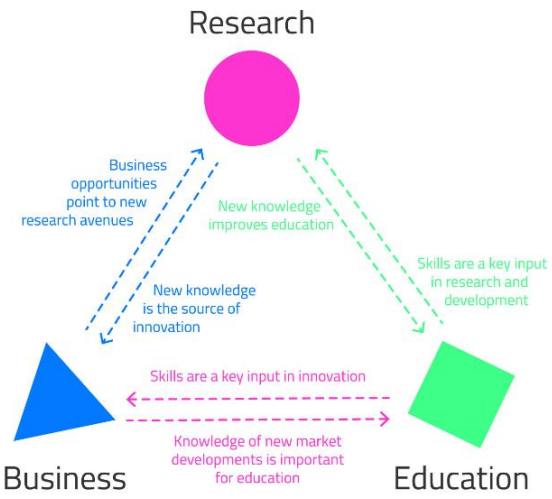
Índex de taules

Índex de gràfics

INTRODUCCIÓ

EIT Health és una organització sense ànim de lucre finançada per la Unió Europea que és dedica a finançar diferents empreses i projectes relacionats amb el món de la salut.

Va ser creada l'any 2015 i és una de les branques *EIT (European Institute of Innovation and Technology)* que pretén ser una organització puntera en el coneixement i la innovació per millorar la vida de les persones.



La idea d'EIT és que floreixi la innovació, i per fer-ho possible la millor manera és que es reuneixen les persones adequades per compartir els coneixements. L'anomenat "triangle del coneixement" és el principi que estableix que quan els experts en negocis, en investigació i en educació treballen junts, es crea un entorn òptim per a la innovació.

Les quatre branques en les que treballa *EIT Health* són:

- *Education*: Participen estudiants a través de programes educatius per permetre la innovació contínua en la sanitat europea.
- *Acceleration*: Catalitza un nou creixement empresarial per oferir productes i serveis transformadors. Treballa amb emprenedors, empreses emergents i PIMES per escurçar el temps de comercialització de productes i serveis que canvien la vida, creant al mateix temps nous llocs de treball i contribuint a una economia sanitària pròspera.
- *Innovation*: Els projectes d'innovació s'esforcen per donar resposta a alguns dels majors reptes sanitaris que s'enfronten a Europa. És aquí on pren vida l'enfocament únic d'EIT Health sobre la innovació.
- *Think Tank*: És un fòrum que reuneix els líders sanitaris per preparar el terreny per a la innovació que canviï la vida i identificar la propera oportunitat d'un canvi de pas en la manera com s'ofereix la salut. Col·laboren entre disciplines i fronteres per explorar i avaluar els temes més urgents que afecten la innovació sanitària actual.

La branca d'interès per aquest treball és la *d'accelaration*. Les bases de dades amb les que es treballarà pertanyen a aquesta branca.

L'*accelerator* treballa amb *Start-ups* i PIMES per tal d'ajudar-les i donar-les-hi l'impuls per poder créixer i entrar en el mercat de salut o per consolidar-se i expandir el mercat.

Per fer-ho crea diversos programes enfocats a diferents tipus d'empreses on aquestes apliquen omplint un formulari i se'n seleccionen un nombre determinat. Al llarg de l'any són varis els programes la selecció d'empreses dels quals rau sota el criteri d'experts en salut i empreses de salut.

A partir de les dades extretes dels formularis de cada programa s'estudiaran els factors més significatius a l'hora de seleccionar les empreses.

La motivació d'aquest treball és, després de mesos d'estar treballant amb la categorització i manteniment de la base de dades, fer-ne un us de les mateixes que generés informació rellevant. Atès que he estat aquests mesos entenent el funcionament de l'entitat i aprenent com funciona el sistema em sembla de gran interès l'anàlisi de la variable principal de la base de dades i el motiu que mou a les empreses a presentar-se, que és el fet de ser seleccionades.

OBJECTIUS

Els models lineals generalitzats són una forma molt útil d'obtenir informació molt clara i efectiva de l'efecte d'una o varies variables sobre la variable dependent, veure el pes específic de cada una i si realment aquest efecte és significatiu o no. A part de la l'anàlisi estructural de cada un dels coeficients també es poden fer prediccions. D'aquesta manera un sol procés, la creació d'un model, ens permet tenir dos utilitats.

En aquest treball s'ajustarà un model amb la variable status ,variable que informa sobre l'estat de "l'aplication" de l'empresa(rebutjada o seleccionada), com a variable resposta binaria per conèixer les variables que més afectin a la variable resposta. Es a dir, conèixer si hi ha algun factor que tingui una influencia significativa a l'hora d'acceptar o rebutjar una l'aplicació d'una empresa. A part d'això com ja s'ha dit l'altra aplicació que tenen els models lineals és la de predicció ja que tenim els coeficients per a cada variable, es pot predir els valors de la variable resposta. Per ajustar el model ho farem una part de la base de dades, que anomenarem mostrals. La resta de dades les anomenarem extramostrals i ens serviran per fer les prediccions.

Per tant els objectius són els següents:

- Estimar i seleccionar un model lineal de resposta binaria
- Validar el model
- Fer prediccions de la variable resposta a partir de les variables explicatives del millor model possible
- Conèixer les variables explicatives més determinants en la variable resposta i saber l'efecte que tenen.
- Aplicar els coneixements apresos durant el grau

ESTAT DE L'ART

La proporció de la variable resposta en la base de dades no està equilibrada, per tant, el model ajustat no serà balancejat. La proporció d'empreses seleccionades de la base de dades amb la qual es treballarà és el 25% (mostral).

En l'article (King & Zeng, 2003) s'aborda el problema de l'ajust dels models lineals de resposta binaria en casos on la proporció d'uns de la variable resposta és petita. Es planteja diversos mètodes per ajustar els models i corregir aquest desequilibri de les dades i s'analitza l'efecte d'aquesta correcció. Conclou que l'efecte és gran en mostres petites i proporcions de la variable resposta inferiors al 5%.

En l'article (Salas-Eljatib et al., 2018) a partir de d'una base de dades inicial crea bases de dades de la mateix mida amb proporcions de zeros en la variable resposta binaria de 10% a 90% de 20 en 20 ,és a dir, 5 bases de dades (10%,30%,50%,70%,90% de proporcions de zeros). Primerament, compara el biaix dels paràmetres del models, el model de la base de dades balancejada (50% de zeros) és la que té menys biaix, seguidament del de les proporcions de 30% i 70%. Els models de les bases de dades amb les proporcions de zeros més extremes de 10% i 90% són les que tenen més biaix.

Pel que fa les prediccions passa el mateix: el model balancejat és el millor model predictor seguit dels models amb les bases de dades de 30% i 70% de zeros i els models amb base de dades de 10% i 90% de zeros fan pitjors prediccions.

Com a conclusions recomana utilitzar models balancejats sempre que es pugui ja que segons l'estudi és el model amb menys baix en els paràmetres i prediccions amb més encert. En l'article afirma que alguns autors argumenten que no hi ha major efecte dels models no balancejats amb bases de dades binaries.

Després de llegir aquest dos articles es demostra que treballar amb dades no balancejades no és un problema, sempre i quan, no siguin proporcions extremes. En l'article (Salas-Eljatib et al., 2018) recomana utilitzar models balancejats sempre que es pugui, però mirant els gràfics i les taules dels articles s'observa que la diferència del baix i de les prediccions dels models amb les bases de dades de 30% i 70% de zeros són molt similars als del model balancejat. En els models amb les bases de dades 10% i 90% els resultats si que empitjoren dràsticament. Per tant, relacionat amb l'article (King & Zeng, 2003) seria un problema si la proporció d'uns fos molt inferiors, sent del 25 % no és un problema i es pot ajustar un model.

METODOLOGIA

1. BASE DE DADES

En el moment inicial l'objectiu principal del treball era la predicció del creixement de les empreses del sector de la salut. Per dur-ho a terme es va fer una neteja exhaustiva que va comportar força feina per tal d'obtenir una base de dades amb condicions. Per fer aquesta predicció s'utilitzaria com a variable resposta el creixement de la variable *revenue*, els ingressos. Per tant, s'havien de trobar empreses que haguessin participat dos cops a *EIT Health* en anys diferents i tinguessin el camp omplert. Es van trobar 84 empreses, de les quals es va seleccionar el 80% com a dades mostrals i la resta com a dades extramostrals per comprovar l'encert del model. Es van provar diferents models lineals on es mirava l'ajust i els residus. No eren mals models, però els resultats no eren els esperats: coeficients amb valors estranys, pocs coeficients significatius i prediccions dolentes. Fins aquest moment es portaven forces hores i esforç, s'arribava just a la data d'entrega i vist el moment i els resultats es va optar amb l'ajuda del tutor per canviar d'objectiu. Buscar una nova variable resposta interessant la qual tingués bastants registres per tal d'obtenir una base de dades extensa i, com a conseqüència, un model més robust.

Es va buscar una nova variable que pogués tenir una interpretació útil i que a la base de dades la tinguéssim en la majoria de registres. Es va optar per la variable *status* ja que la trobem en tots els registres i ens aporta la informació de si l'empresa ha estat seleccionada per donar-hi suport o no. Es tracta d'una variable binaria. Per tant, a diferència del treball anterior, no s'ajustarà un model de regressió lineal múltiple sinó un model lineal generalitzat amb variable resposta binaria.

La base de dades amb la que s'ajustarà el model, com ja s'havia fer al treball anterior, es dividirà en dos:

- Mostral: Selecció aleatòria del 80 % de les dades
- Extramostral: Secció aleatòria del 20 % de les dades

D'aquesta manera s'estimaran els models sobre la base de dades mostral i es faran les prediccions sobre les dades extramostrals.

2. MODELS LINEALS

Un model de regressió lineal múltiple ens permet explicar el comportament d'una variable resposta Y a partir d'unes variables explicatives X.

$$y= \beta_0+ \beta_1 x_1+ \beta_2 x_2+... +\beta_n * x_n+ \varepsilon$$

- La variable y és la variable resposta o endògena i les variables x_1,x_2, …, x_n són les variables explicatives o exògenes.
- β_0, β_1 ,β_2,...,β_n són els paràmetres o coeficients a estimar.
- ε és l'error o terme de pertorbació aleatòria

2.1. ESTIMACIÓ PER MÍNIMS QUADRATS ORDINARIS

Un cop especificat el model, el següent pas és fer l'estimació dels paràmetres, per fer-ho s'utilitza l'estimació per mínims quadrats ordinaris, aquest mètode obté els valors dels paràmetres que fan mínims els residus de l'estimació. Per obtenir els valors dels paràmetres es fa el següent càlcul de matrius:

$$\beta_{MQO} = (X'X)^{-1}X'Y$$

Matriu X:

$$\begin{bmatrix} 1 & x_{21} & \cdots & x_{k1} \\ 1 & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{2N} & \cdots & x_{kN} \end{bmatrix}$$

Matriu Y:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

On X és la matriu que conté els valors de les variables explicatives i Y és la matriu què conté els valors de la variable resposta.

Propietats dels residus de de l'estimació MQO:

- L'esperança matemàtica dels errors a de ser igual a 0.
- Homoscedasticitat: La variància dels errors ha de ser constant.
- No autocorrelació: Els residus han de ser independents entre si.
- Els residus es distribueixen segons una distribució normal.

3. MODELS LINEALS GENERALITZATS

Els models lineals són aplicables sota els supòsit de normalitat en els residus quan es distribueixen amb una altra distribució s'utilitzen els models lineals generalitzats.

Es composen de 3 parts:

- Distribució de probabilitats que pot escriure's com:

$$f(y;\theta,\phi) = e^{\frac{\theta y - b(\theta)}{a(\phi)} + c(y,\phi)}$$

  On :
  - θ és el paràmetre canònic
  - Φ és el paràmetre de dispersió

- Predictor lineal: Quantitat que incorpora la informació sobre les variables independents del model

$$\eta = X\beta$$

- Funció d'enllaç (*link*). Funció que relaciona el valor esperat amb el predictor lineal

$$g(\mu) = \eta = X\beta$$

Propietats MLG:

- L'estimació dels paràmetres es farà per màxima versemblança
- La variància no cal que sigui constant
- S'escollirà la distribució
- La funció *link* s'escollirà en funció de com sigui l'esperança del model que volem descriure.

3.1. MODELS LINEALS DE RESPOSTA BINARIA

Quan la variable resposta és qualitativa binaria, com és el cas d'estudi, la distribució serà binomial. Tenim diverses funcions d'enllaç (*link*):

- *Logit*: és la funció *link* més típica.
  El funció *link* és la següent $g(\mu) = \eta = \text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right)$
- *Probit*: Aquest *link* és la inversa de la distribució normal estàndard amb paràmetres entre 0 i 1.
  La funció *link* és la següent $g(\mu) = \eta = \phi(\pi)$
- *Cloglog*: Aquest *link* és la inversa de la funció distribució per *Gompertz*
  La funció *link* és $g(\mu) = \eta = \log(\log\left(\frac{1}{1-\pi}\right))$

3.2. MÈTODES DE BONDAT D'AJUST PER COMPARAR MODELS (AIC, BIC, LOGLIK)

- AIC (*Akaike Infromation Criteria*): proposada per *Akaike* (1974), es defineix com un *trade-off* entre la bondat d'ajust del model i el nombre de paràmetres p:
  $AIC = 2(p - \ell(\pi, y))$
- BIC (*Bayesian Information Criteria*): proposat per Schwartz (1978), a diferència del AIC té en conte la grandària la mostra.

$$BIC = p \log n - 2\ell(\pi, y)$$

Són preferibles els models amb mínim AIC i BIC

3.3. FUNCIÓ STEP

Funció del programa R del paquet *stats* que a partir d'introduir un model et selecciona un nou model amb AIC a partir d'un algoritme *Stepwise*.

```
step(object, scope, scale = 0,
    direction = c("both", "backward", "forward"),
    trace = 1, keep = NULL, steps = 1000, k = 2, …)
```

Per utilitzar-la s'usarà la funció i el model com a objecte:

>step(model)

COS DEL TREBALL

4. DESCRIPCIÓ DE LA BASE DE DADES

La base de dades amb la que s'ha treballat internament l'anomenem *historical data* ja que conté totes les aplicacions que han fet les empreses o *startups* als diferents programes *d'EIT Health*. Està estructurada en dos parts:

- La part que conté els valors originals: on tenim les dades sense codificar on hi ha noms complerts, descripcions de text lliure...
- La part estandarditzada: variables codificades segons el valor que tinguin i en alguns camps extraiem un valor a partir d'un text lliure.

La part que s'ha utilitzat és la codificada ja que ens simplifica el procés a l'hora d'analitzar les dades. També al tenir les dades estandarditzades ens permetrà treballar amb variables qualitatives com si fossin quantitatives.

Les variables de la base de dades són les següents:

- IFOPID (*Project ID*): Identificador únic per a cada aplicació feta a qualsevol programa d'EIT Health.
- IFOYEAR (*Year*): Any del programa en que l'empresa va aplicar.
- IFOCFA (*Call number):* Edició en la que és va presentar, normalment n'hi ha una i algun cop obren dos o tres terminis o edicions en que les empreses poden tornar a aplicar.
- SFO000 (*EIT Health Programme*): Programa *d'EIT Health* en el que s'ha presentat.
- SFO001 (*First Name*): Nom de la persona que fa l'aplicació.
- SFO002 (*Last Name*): Cognom de l'aplicant.
- SFO003 (*Email*): Carreu electrònic de l'aplicant.
- SFO004 (*Function*): Càrrec o rol que té dins l'empresa la persona que aplica.
- SFO005 (*Gender*): Gènere de l'aplicant.
    - 1 – *Male* (Masculí)
    - 2 – *Female* (Femaní)
    - 3 – *Other* (Altres)
    - 4 – *Undisclosed* (No contestat)
- SFO006 (*Incorporated*): Si l'empresa està registrada o no.
    - 1 – *Yes* (Sí)
    - 2 – No
- SFO007 (*Incorporation year*): Any de registre de l'empresa.
- SFO008 (*Registered Country*): País de registre de l'empresa.

- SFO009 (*Company name*): Nom de l'empresa.
- SFO010 (*Company team*): Equip o treballadors que conformen l'empresa.
- SFO011 (*Category*): Categoria o sector de l'empresa
    - 1 – *Biotech* (Biotecnologia)
    - 2 – *Medtech* (Serveis mèdics)
    - 3 – *Digital Health* (Salut digital: Softwares, aplicacions de mòbil o suport digital aplicat al sector de la salut)
    - 4 – *Service/Care Model* (Cura o atenció de les persones)
    - 5 – *Other* (altres)
- SFO012 (*TRL*): Nivells de maduresa de la tecnologia.
    - Prenent valors de TRL 1 a TRL 9 categoritzats de 1 a 9.
- SFO013 (*Elevator pitch*): Descripció de l'empresa i l'activitat que fa.
- SFO014 (*FTE*): Nombre de treballadors en jornada completa.
- SFO015 (*Project name*): Nom del projecte de l'empresa.
- SFO016 (*Project description*): Descripció del projecte.
- SFO017 (*Innovation*): Innovació que aporta l'empresa.
- SFO018 (*Funding Stage*): Etapa de finançament.
    - 1 – *Pre Seed*
    - 2 – *Seed*
    - 3 – *Series A*
    - 4 – *Series B*
    - 5 – *Series C*
- SFO019 (*Total funding*): Finançament rebut en milers d'euros
- SFO020 (*Planned spending*): Amb quins recursos i amb què tenen pensat les empreses invertir o gastar els diners del finançament que rebin.
- SFO021 (*Equity sought*): Liquiditat de l'empresa en milers d'euros.
- SFO022 (*Revenue*): Ingressos de l'empresa en milers d'euros.
- SFO023 (*Business model*): Model de negoci de l'empresa.
- SFO024 (*EIT Health partner*): *Partner* vinculat a *EIT Health* que ha rebut ajut o té contacte amb l'empresa.
- SFO025 (*CLC*): *Co-Location Centre*, regió establerta per *EIT Health* a la que pertany l'empresa, depèn del país de registre o del lloc on tingui seu l'empresa.
    - 1 – *Belgium/Netherlands*
    - 2 – *France*
    - 3 – *Germany/Switzerland*
    - 4 – *Scandinavia*
    - 5 – *Spain*

    - 6 – *UK/Ireland*
    - 7 – *Innostars (Hungary, Poland, Portugal, Italy, Czech Republic, Greece)*
- SFO026 (*Valuation*): Estimació econòmica del valor de l'empresa en milers d'euros.
    - 1 – [1,250)
    - 2 – [250,500)
    - 3 – [500,1000)
    - 4 – [1000,2000)
    - 5 – [2000,5000)
    - 6 – [5000,∞)
- IFOSTAT (*Status*): Si l'aplicació feta per l'empresa ha estat seleccionada o no.
    - 0 – *Rejected*
    - 1 – *Selected*
- IFOCMP (*Starup* ID): Identificador intern i únic que assignem a cada empresa. Una empresa que hagi participat amb més d'un programa tindrà un *project* ID diferent però el mateix *Startup* ID.

Els valors desconeguts (*UNKNOWNS*) d'alguns camps ja sigui perquè no és demanava en la inscripció del programa o perquè no han respòs, tenen el valor 0 o -1 depenent de la variable.

Per protecció de dades no s'utilitzen els camps que hi surten noms de persones, d'empreses o entitats. Tampoc es tindran en compte les variables de text lliure.

Per tant es tindran les variables categoritzades com variables categòriques. Inclús les que tenen bastantes categories i són ordinals les podrem utilitzar com numèriques.

També seran d'interès les variables quantitatives ja que es té el valor exacte de la variable en qüestió.

Malgrat no tenir els noms de les empreses tindrem l'identificador de cada aplicació (*project* ID) i el codi de cada empresa o *Start-up* (*Startup* ID).

La base de dades en la què es comença a treballar conté 5030 registres, que anomenem *applications* que són el total de sol·licituds rebudes en tots els anys en tots els programes. D'aquests 5030 registres tenim 2939 empreses diferents. Aquesta diferència ve donada pel fet que una sola empresa pot presentar-se en varis programes i en diferents anys. En el treball s'utilitzarà el nombre d'empreses enlloc *d'applications*.

Pel que fa a la variable d'estudi, IFOSTAT, que és la variable resposta del nostre model ens indica si aquella aplicació que ha fet una empresa en un programa ha estat seleccionada per donar-li ajuda o no.

5. MANIPULACIÓ DE LA BASE DE DADES

5.1. DEPURACIÓ I COMPLEMENTACIÓ DE LA BASE DE DADES

La base de dades conté tota la informació històrica de les aplicacions que han anat fent les empreses, com l'objectiu és saber el perfil de les empreses seleccionades i no el de les aplicacions s'han d'eliminar les aplicacions repetides d'una mateixa empresa.

Per tant, teníem clar que en aquest procés d'eliminació de registres no es volia perdre informació o perdre la menor possible ,per la qual cosa, a l'hora d'eliminar els registres duplicats de les empreses s'havia d'assegurar que aquella informació ja es trobava en l'aplicació que deixàvem.

Primerament, com s'ha dit a l'apartat anterior hi ha columnes que contenen variables que no ens interessen i va quedar una base de dades amb les següents columnes:

- IFOPID (*Project* ID)
- IFOYEAR (*Year*)
- SFO000 (*EIT Health Programme*)
- SFO005 (*Gender*)
- SFO006 (*Incorporated*)
- SFO007 (*Incorporation year*)
- SFO008 (*Registered Country*)
- SFO011 (*Category*)
- SFO012 (*TRL*)
- SFO014 (*FTE*)
- SFO018 (*Funding Stage*)
- SFO019 (*Total funding*)
- SFO021 (*Equity sought*)
- SFO022 (*Revenue*)
- SFO025 (*CLC*)
- SFO026 (*Valuation*)
- IFOSTAT (*Status*)
- IFOCMP (*Starup ID*)

Seguidament, es va mirar quines variables són atemporals, és a dir, seran fixes al llarg del temps hi havia registres buits. Aquestes variables són SFO006 (*Incorporated*), SFO007 (*Incorporation year*) i SFO011 (*Category*).

Es van ordenar les empreses segons l'any del programa que havien participat, de més nou a més vell, pel camp status, primer si estaven seleccionades, i per codi d'Empresa. D'aquesta

manera si una empresa estava repetida la primera era la participació més nova que estès seleccionada i si no ho estava era la participació més novella, ja que com més nova més informació es conté.

Abans de fer la eliminació dels duplicats es va complimentar per les empreses duplicades les variables que hem anomenat atemporals. Per fer-ho si els registre que no s'ha d'eliminar de l'empresa conté alguns d'aquestes variables buides es complimenta amb un dels valors de les aplicacions duplicades.

Un cop complimentada aquesta informació es va eliminar les participacions repetides com que ja estaven ordenades simplement era eliminar els valors duplicats segons el codi d'empresa d'aquesta manera s'aconseguia un codi únic per empresa.

5.2. CANVI DE TIPUS D'ALGUNES VARIABLES

Al introduir les dades al programa usat per fer l'anàlisi (*RStudio*) la majoria de variables es detectaven com numèriques, com ja s'ha dit prèviament, la majoria de les variables seleccionades per realitzar l'estudi estan codificades. Per tant, s'ha de canviar el tipus de variables per les que ens interessava. Les variables van quedar de la següent manera:

- IFOPID (*Project ID*) – Categòrica
- IFOYEAR (*Year*) – Categòrica
- SFO000 (*EIT Health Programme*) – Categòrica
- SFO005 (*Gender*) – Categòrica
- SFO006 (*Incorporated*) – Categòrica
- SFO007 (*Incorporation year*) – Numèrica
- SFO008 (*Registered Country*) – Categòrica
- SFO011 (*Category*) – Categòrica
- SFO012 (*TRL*) – Numèrica
- SFO014 (*FTE)* – Numèrica
- SFO018 (*Funding Stage*) – Categòrica
- SFO019 (*Total* funding) – Numèrica
- SFO021 (*Equity sought*) – Numèrica
- SFO022 (*Revenue*) – Numèrica
- SFO025 (*CLC*) – Categòrica
- SFO026 (*Valuation*) – Categòrica
- IFOSTAT (*Status*) – Categòrica
- IFOCMP (*Startup* ID) – Categòrica

- Antic: Anys d'antiguitat de l'empresa quan es va presentar al programa d'EIT Health. Diferencia entre SFO000 menys SFO007 – Numèrica

5.3. CREACIÓ DE SUBBASES DE DADES

Per fer l'anàlisi es volia perdre el mínim d'informació possible i es va detectar que la major part de camps buits es trobava en les variables econòmiques(les que contenen informacions monetàries de les empreses- (SFO019, SFO021, SFO022, SFO026)).

Per tant, a l'hora d'eliminar les aplicacions que tenien algun dels camps necessaris per l'estudi buits quedaven 207 empreses.

Es va optar per crear una base de dades per cada variable d'aquesta manera com que les 4 variables econòmiques són les que tenen més registres buits, es perdia la mínima informació possible.

Per evitar problemes de multicol·linealitat es va mirar la correlació entre les variables (Taula 5.2.1) econòmiques, totes les correlacions són baixes, inferiors a 0,3.

D'aquesta forma vam obtenir les següents bases de dades (Vegeu taula 5.2.1):

Taula 5.2.1 Informació de les bases de dades

Base de dades	**Camps que conté la base de dades**	**Nre d'empreses**
1	Resta de camps	823
2	Resta de camps[1] + SFO019	515
3	Resta de camps + SFO021	698
4	Resta de camps + SFO022	576
5	Resta de camps + SFO026	328
6	Tots els camps	207
7	Resta de camps + SFO019+ SFO022	228

Taula 5.2.1 Correlacions de les variables econòmiques

	SFO019	SFO021	SFO022
SFO019	1.0000000	0.1172593	0.105836
SFO021	0.1172593	1.0000000	0.260475
SFO022	0.1058360	0.2604750	1.000000

[1] Totes les variables de la base de dades exceptuant les variables econòmiques (SFO019, SFO021, SFO022, SFO026).

6. DESCRIPCIÓ UNIVARIANT I BIVARIANT DE LES VARIABLES

Per fer la descripció univariant i bivariant de les dades es va fer amb la base de dades 6 ja que conté tots els camps.

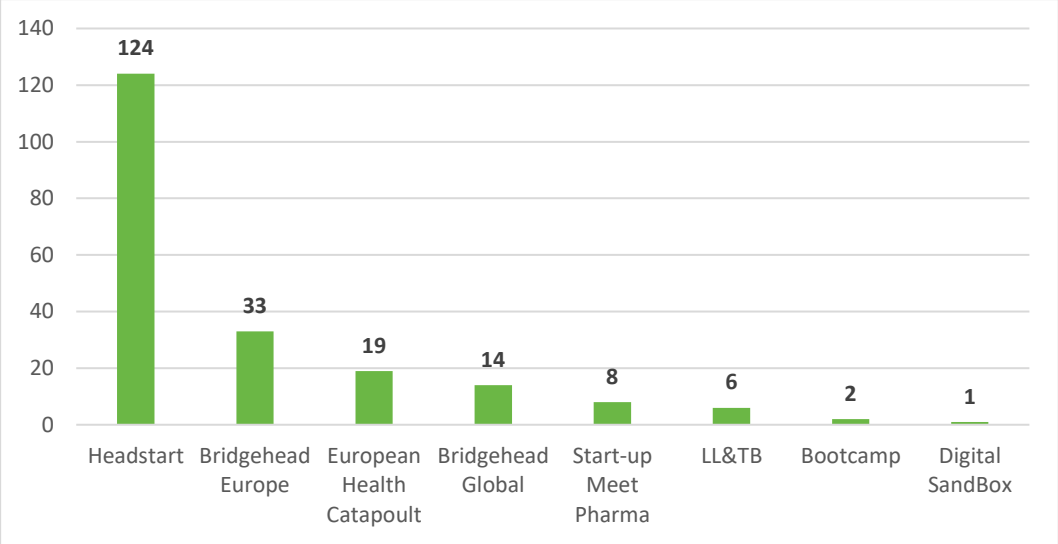
6.1. DESCRIPCIÓ UNIVARIANT VARIABLES QUALITATIVES

- SFO000 (*EIT Health Programme*)

Taula 6.1.1 Descriptiva de la variable SFO000

item	count	percent	cum_count	cum_percent
Headstart	124	0.5990338	124	0.5990338
Bridgehead Europe	33	0.1594203	157	0.7584541
European Health Catapult	19	0.0917874	176	0.8502415
Bridgehead Global	14	0.0676329	190	0.9178744
Start-up Meet Pharma	8	0.0386473	198	0.9565217
LL&TB	6	0.0289855	204	0.9855072
Bootcamp	2	0.0096618	206	0.9951691
Digital SandBox	1	0.0048309	207	1.0000000

Gràfic 6.1.1 Diagrama de barres de la variable SFO000



Per la variable SFO000 es pot veure que la majoria de d'empreses de la base de dades, pràcticament el 60%, han aplicat al programa *Heastart*. Vegeu Taula 6.1.1 i Gràfic 6.1.1.

- IFOYEAR (*Year*)

Taula 6.1.2 Descriptiva de la variable IFOYEAR

item	count	percent	cum_count	cum_percent
2020	207	1	207	1

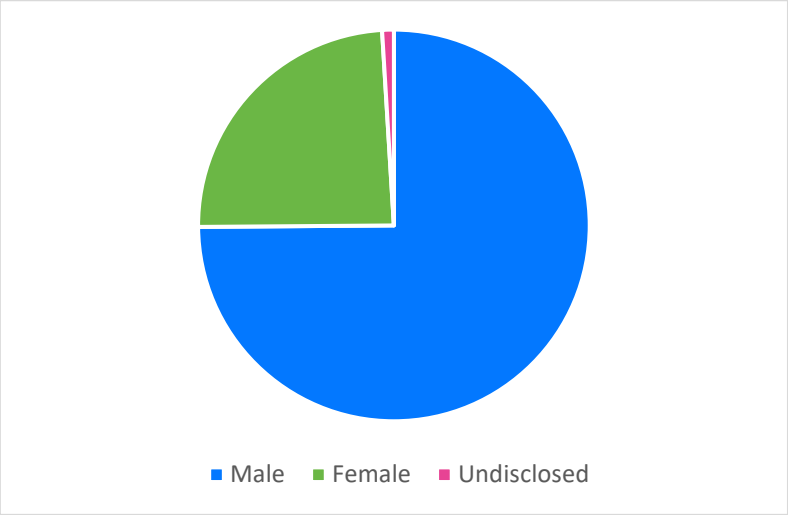
Per la variable IFOYEAR es veu a la Taula 6.1.2 que les empreses de la base de dades s'han presentat a programes del 2020.

- SFO005 (*Gender*)

Taula 6.1.3 Descriptiva de la variable SFO005

item	count	percent	cum_count	cum_percent
Male	155	0.7487923	155	0.7487923
Female	50	0.2415459	205	0.9903382
Undisclosed	2	0.0096618	207	1.0000000

Gràfic 6.1.2 Diagrama de sectors de la variable SFO005



La variable SFO005 que fa referència al gènere de l'aplicant es distribueix de manera que la majoria dels aplicants ,pràcticament 3 de cada 4, són homes. Vegeu Taula 6.1.3 i Gràfic 6.1.3.

- SFO006 (*Incorporated*)

Taula 6.1.4 Descriptiva de la variable SFO006

item	count	percent	cum_count	cum_percent
Incorporated	207	1	207	1

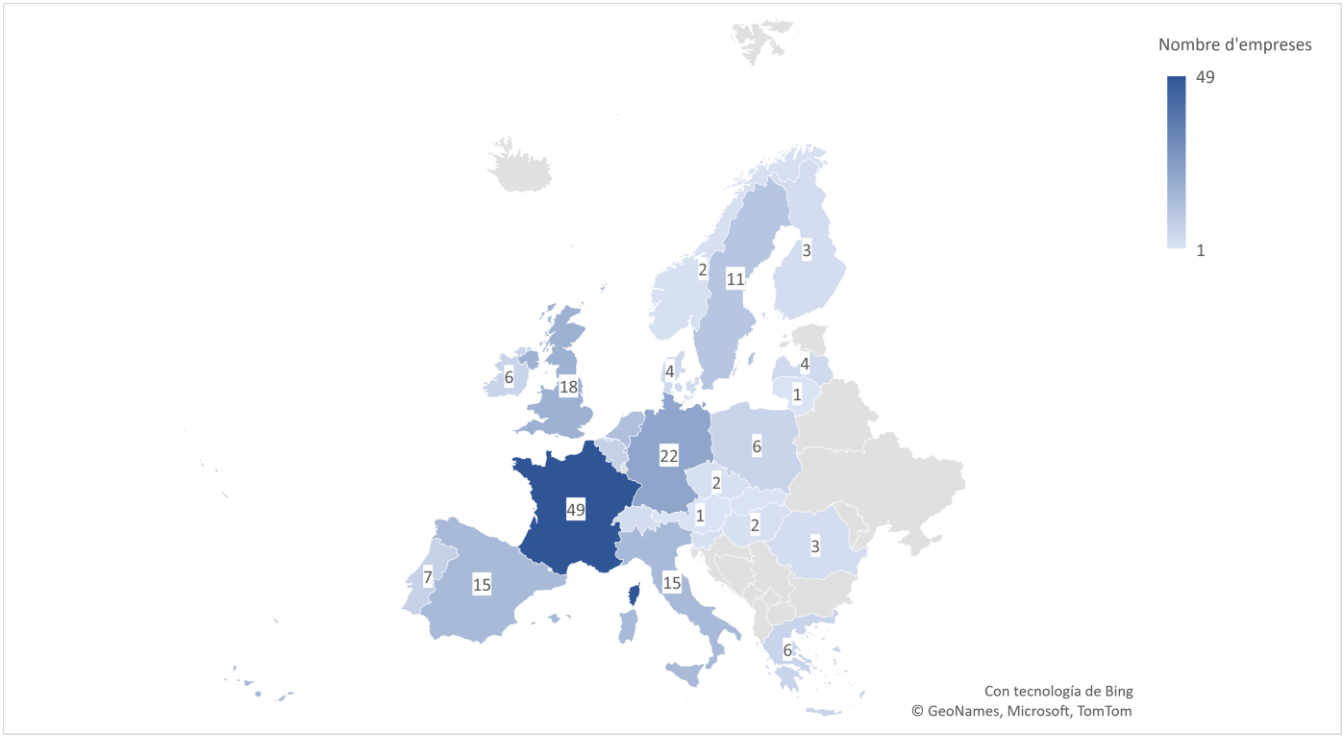
Per la variable SFO006 s'observa a la Taula 6.1.4 que totes les empreses de la base de dades estan registrades.

- SFO008 (*Registered Country*)

Taula 6.1.5 Descriptiva de la variable SFO008

item	count	percent	cum_count	cum_percent
FR	49	0.2367150	49	0.2367150
DE	22	0.1062802	71	0.3429952
GB	18	0.0869565	89	0.4299517
ES	15	0.0724638	104	0.5024155
IT	15	0.0724638	119	0.5748792
NL	13	0.0628019	132	0.6376812
SE	11	0.0531401	143	0.6908213
BE	7	0.0338164	150	0.7246377
PT	7	0.0338164	157	0.7584541
GR	6	0.0289855	163	0.7874396
IE	6	0.0289855	169	0.8164251
PL	6	0.0289855	175	0.8454106
DK	4	0.0193237	179	0.8647343
LV	4	0.0193237	183	0.8840580
FI	3	0.0144928	186	0.8985507
RO	3	0.0144928	189	0.9130435
CH	2	0.0096618	191	0.9227053
CZ	2	0.0096618	193	0.9323671
HU	2	0.0096618	195	0.9420290
IL	2	0.0096618	197	0.9516908
SI	2	0.0096618	199	0.9613527
AT	1	0.0048309	200	0.9661836
LT	1	0.0048309	201	0.9710145
NG	1	0.0048309	202	0.9758454
NO	1	0.0048309	203	0.9806763
SK	1	0.0048309	204	0.9855072
SZ	1	0.0048309	205	0.9903382
TR	1	0.0048309	206	0.9951691
US	1	0.0048309	207	1.0000000

Gràfic 6.1.3 Mapa de calor de la variable SFO008



[2]No està representat ni Israel, ni Turquia ni Estats Units. Tots tres països amb una empresa

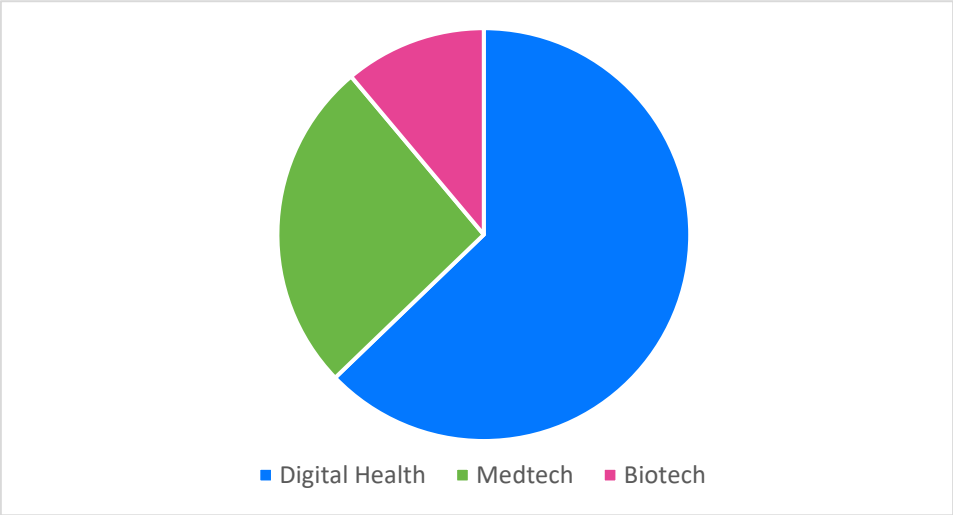
De la variable *registred country* destacaríem que hi ha un gran nombre d'empreses de França a la base de dades, molt per sobre la resta de països. Vegeu Taula 6.1.5 i Gràfic 6.1.3.

- SFO011 (Category)

Taula 6.1.6 Descriptiva de la variable SFO011

item	count	percent	cum_count	cum_percent
Digital Health	130	0.6280193	130	0.6280193
Medtech	54	0.2608696	184	0.8888889
Biotech	23	0.1111111	207	1.0000000

Gràfic 6.1.4 Diagrama de sectors de la variable SFO011



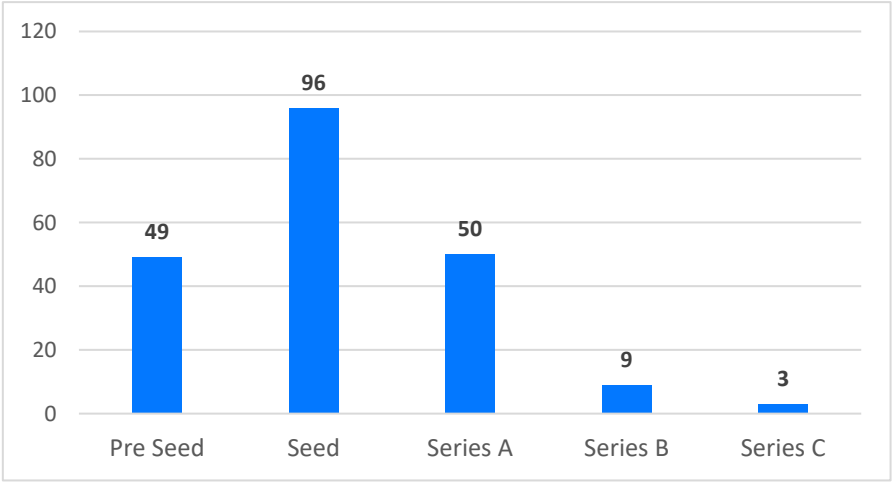
La variable category (SFO011) es destaca que la majoria d'empreses, el 63%, són del sector Digital Health, seguit de Medtech amb un 26% i la resta l'11% són de Biotech. Vegeu Taula 6.1.6 i Gràfic 6.1.4.

- SFO018 (Funding Stage)

Taula 6.1.7 Descriptiva de la variable SFO018

item	count	percent	cum_count	cum_percent
Seed	96	0.4637681	96	0.4637681
Series A	50	0.2415459	146	0.7053140
Pre Seed	49	0.2367150	195	0.9420290
Series B	9	0.0434783	204	0.9855072
Series C	3	0.0144928	207	1.0000000

Gràfic 6.1.5 Diagrama de barres de la variable SFO018



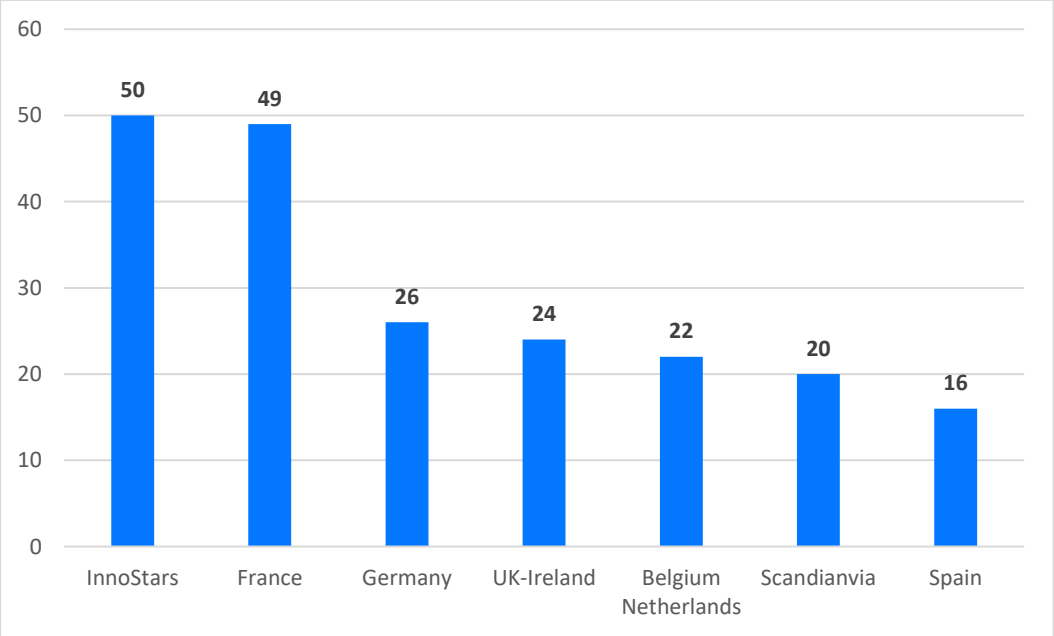
En la variable funding stage s'observa que la majoria d'empreses es troben en el nivell Seed de finançament, seguit dels nivells Series A i Pre Seed amb pràcticament la meitat d'empreses. Vegeu Taula 6.1.7 i Gràfic 6.1.5.

- SFO025 (CLC)

Taula 6.1.8 Descriptiva de la variable SFO025

item	count	percent	cum_count	cum_percent
InnoStars	50	0.2415459	50	0.2415459
France	49	0.2367150	99	0.4782609
Germany	26	0.1256039	125	0.6038647
UK-Ireland	24	0.1159420	149	0.7198068
Belgium-Netherlands	22	0.1062802	171	0.8260870
Scandinavia	20	0.0966184	191	0.9227053
Spain	16	0.0772947	207	1.0000000

Gràfic 6.1.6 Diagrama de barres de la variable SFO025



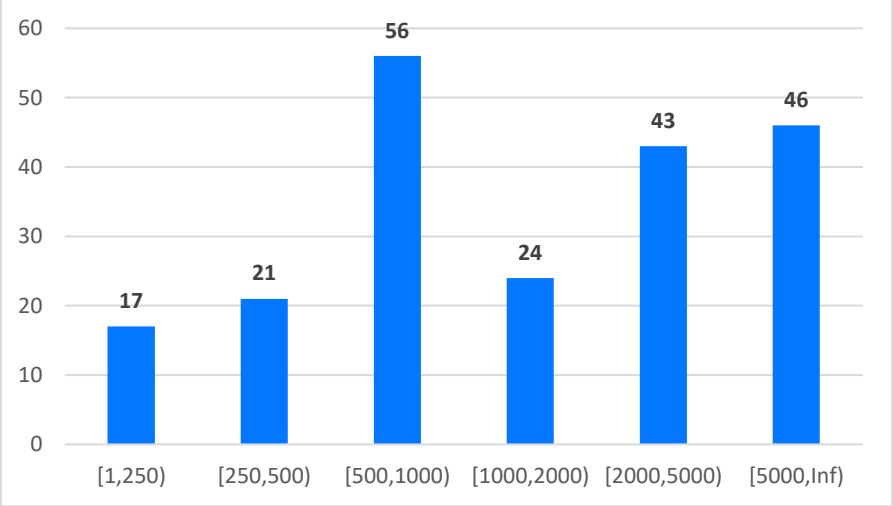
La variable CLC es distribueix de manera que entre el CLC de France i d'Innostars contenen pràcticament les mateixes empreses es troben la majoria d'empreses. Vegeu Taula 6.1.8 i Gràfic 6.1.6.

- SFO026 (Valuation)

Taula 6.1.9 Descriptiva de la variable SFO026

item	count	percent	cum_count	cum_percent
[500,1000)	56	0.2705314	56	0.2705314
[5000,Inf)	46	0.2222222	102	0.4927536
[2000,5000)	43	0.2077295	145	0.7004831
[1000,2000)	24	0.1159420	169	0.8164251
[250,500)	21	0.1014493	190	0.9178744
[1,250)	17	0.0821256	207	1.0000000

Gràfic 6.1.7 Diagrama de barres de la variable SFO026



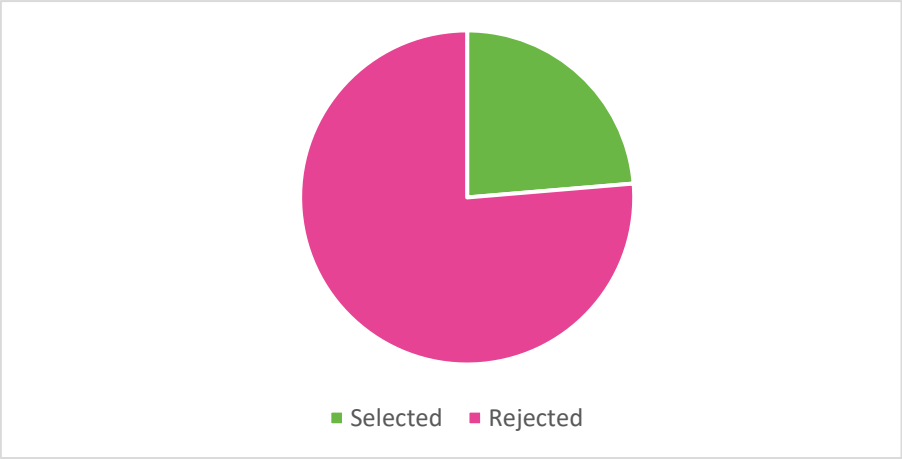
De la variable *valuation* destaca l'interval [500,100) k€ on es troben el 27% de les empreses. Vegeu Taula 6.1.9 i Gràfic 6.1.7.

- IFOSTAT (Status)

Taula 6.1.10 Descriptiva de la variable IFOSTAT

item	count	percent	cum_count	cum_percent
Rejected	158	0.763285	158	0.763285
Selected	49	0.236715	207	1.000000

Gràfic 6.1.8 Diagrama de sectors de la variable IFOSTAT



La variable IFOSTAT, què és la variable resposta del nostra model es distribueix amb el 76% de les empreses *rejected* i el 24% *selected*. Vegeu Taula 6.1.10 i Gràfic 6.1.8.

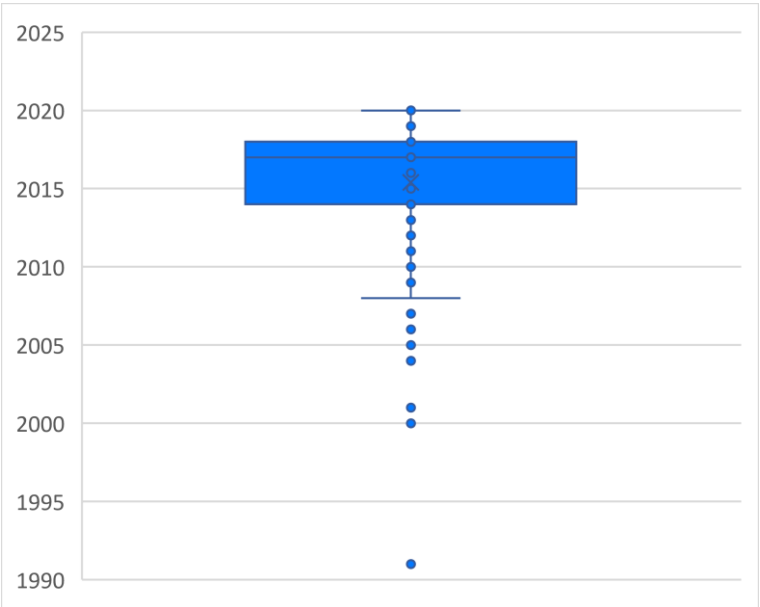
6.2. DESCRIPCIÓ UNIVARIANT VARIABLES QUANTITAVIES

- SFO007 (Incorporation year)

Taula 6.2.1 Descriptiva de la variable SFO007

	mean	sd	median	min	max	range	skew	kurtosis	se
X1	2015.401	4.401963	2017	1991	2020	29	-2.513492	8.923353	0.3059576

Gràfic 6.2.1 Diagrama de caixa de la variable SFO007



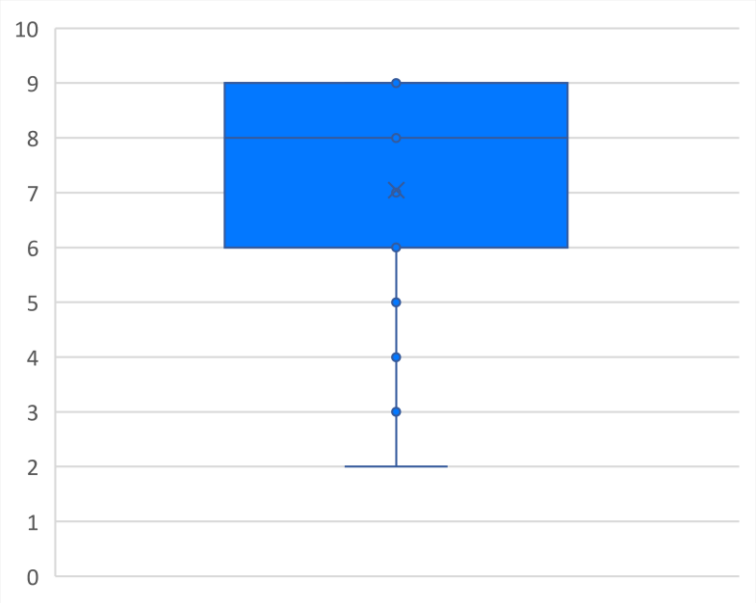
La distribució de la variable *Incorporation year* està centrada a l'any 2017 en mediana i a l'any 2015,4 en mitjana. Conté alguns *outliers* destacar sobretot el de l'any 1991. Vegeu Taula 6.2.1 i Gràfic 6.2.1.

- SFO012 (TRL)

Taula 6.2.2 Descriptiva de la variable SFO012

	mean	sd	median	min	max	range	skew	kurtosis	se
X1	7.048309	1.858485	8	2	9	7	-0.7158102	-0.7024231	0.1291736

Gràfic 6.2.2 Diagrama de caixa de la variable SFO012



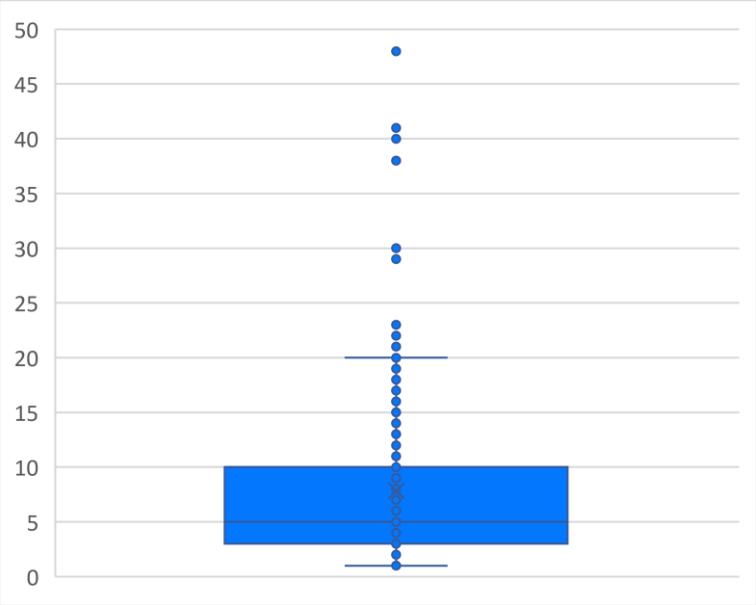
La distribució de la variable TRL està bastant centrada i no té *outliers*, és d'esperar ja que conté valors entre 1 i 9, en mitjana està centrada a 7 i en mediana a 8. Vegeu Taula 6.2.2 i Gràfic 6.2.2.

- SFO014 (FTE)

Taula 6.2.3 Descriptiva de la variable SFO014

	mean	sd	median	min	max	range	skew	kurtosis	se
X1	7.845411	7.742536	5	1	48	47	2.602179	7.839254	0.5381434

Gràfic 6.2.3 Diagrama de caixa de la variable SFO014



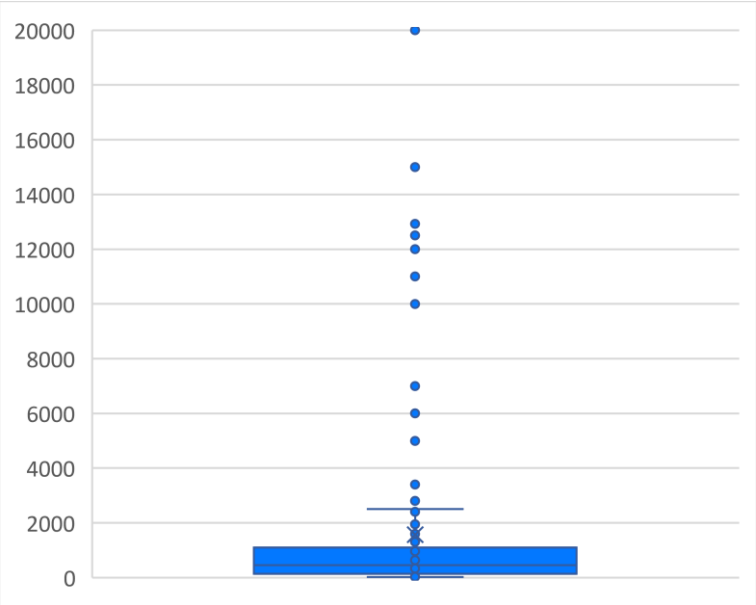
La distribució de la variable FTE conté alguns outliers, està centrada a 5 en mediana i a 7,8 en mitjana. Vegeu Taula 6.2.3 i Gràfic 6.2.3.

- SFO019 (Total funding)

Taula 6.2.4 Descriptiva de la variable SFO019

	mean	sd	median	min	max	range	skew	kurtosis	se
X1	1562.596	3256.001	450	25	20000	19975	3.482293	12.9294	226.3077

Gràfic 6.2.4 Diagrama de caixa de la variable SFO019



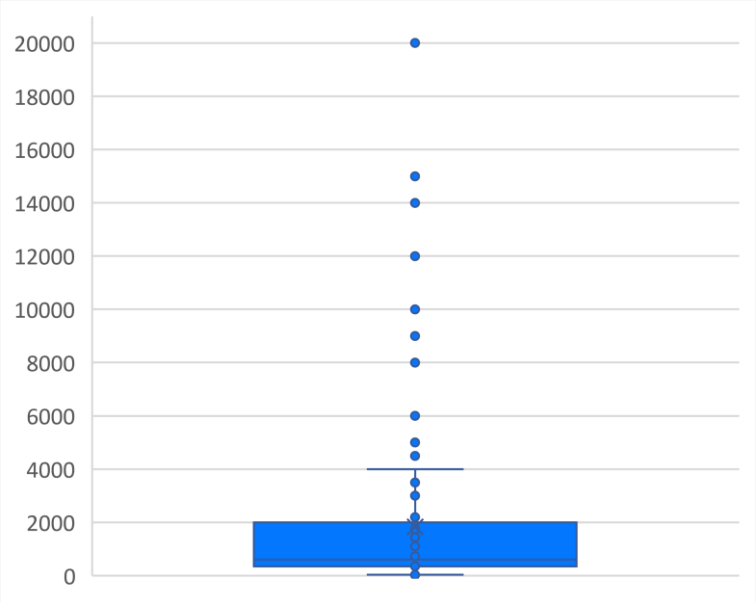
La variable total funding conté bastant valors extrems per sobre del límit superior del diagrama. Està centrat a 1562 en mitjana i a 450 en mediana. Vegeu Taula 6.2.4 i Gràfic 6.2.4.

- SFO021 (Equity sought)

Taula 6.2.5 Descriptiva de la variable SFO021

	mean	sd	median	min	max	range	skew	kurtosis	se
X1	1824.686	3028.841	600	30	20000	19970	2.984022	9.877905	210.519

Gràfic 6.2.5 Diagrama de caixa de la variable SFO021



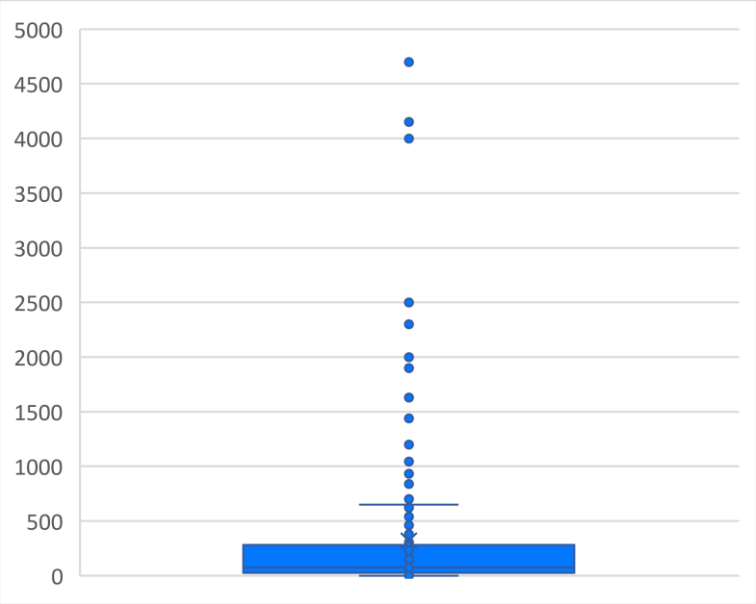
La variable *equity sought* està centrada a 600 en mediana i a 1824 en mitjana, té bastants *outliers* per sobre el límit superior. Vegeu Taula 6.2.5 i Gràfic 6.2.5.

- SFO022 (*Revenue*)

Taula 6.2.6 Descriptiva de la variable SFO022

	mean	sd	median	min	max	range	skew	kurtosis	se
X1	318.7729	664.5781	76	1	4700	4699	3.905447	18.0882	46.19137

Gràfic 6.2.6 Diagrama de caixa de la variable SFO022



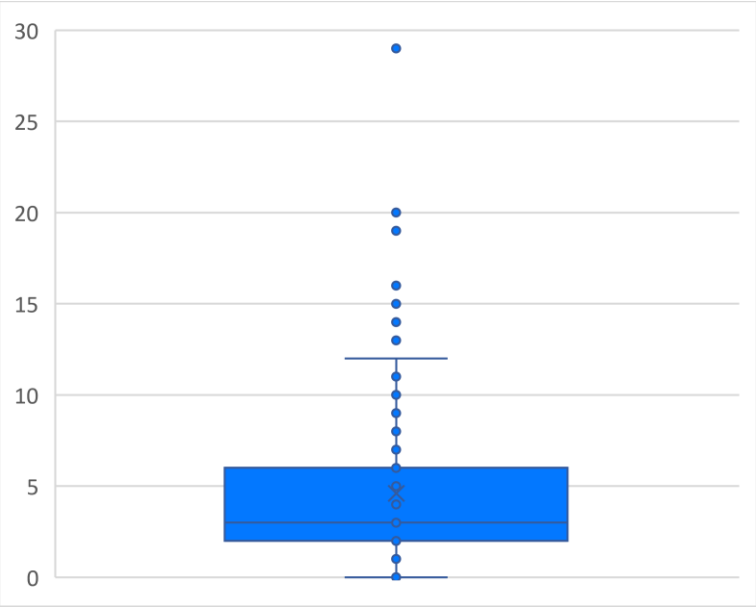
La variable *revenue* conté bastant valors extrems per sobre del límit superior del diagrama. Està centrat a 318,77 en mitjana i a 76 en mediana. Vegeu Taula 6.2.6 i Gràfic 6.2.6.

- Antic

Taula 6.2.7 Descriptiva de la variable antic

	mean	sd	median	min	max	range	skew	kurtosis	se
X1	4.599034	4.401963	3	0	29	29	2.513492	8.923353	0.3059576

Gràfic 6.2.7 Diagrama de caixa de la variable antic



La variable antic conté pocs valors extrems, destacar que el valor màxim és el 29. La variable està centrada a 3 en mediana i a 4,6 en mitjana. Vegeu Taula 6.2.7 i Gràfic 6.2.7.

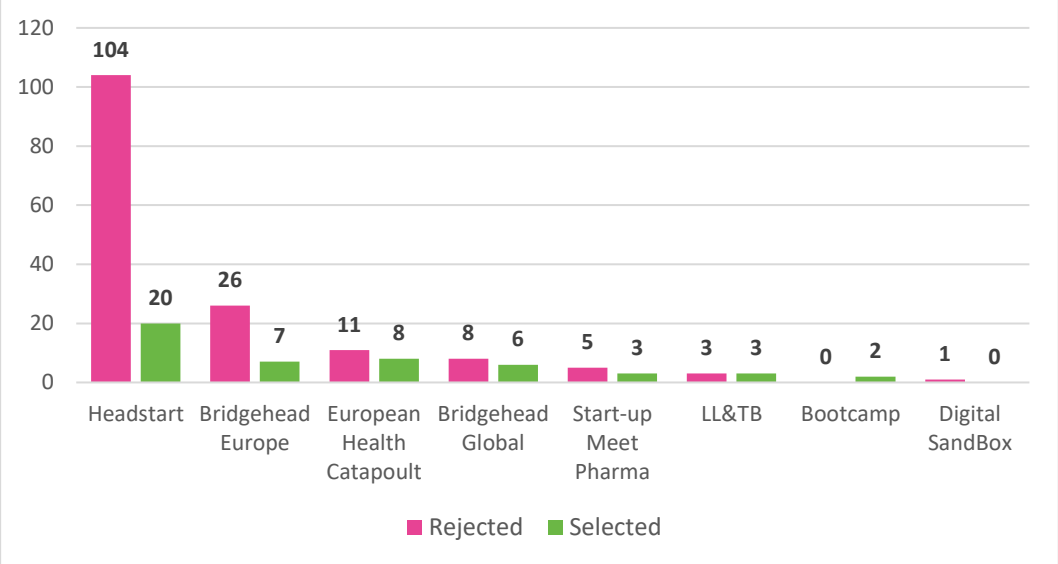
6.3. DESCRIPCIÓ BIVARIANT VARIABLES QUALITATIVES AMB VARIABLE RESPOSTA (STATUS)

- SFO000 (*EIT Health Programme*)

Taula 6.3.1 Descriptiva de les variables SFO000 i IFOSTAT

	Rejected	Selected
Bootcamp	0	2
Bridgehead Europe	26	7
Bridgehead Global	8	6
Digital SandBox	1	0
European Health Catapult	11	8
Headstart	104	20
LL&TB	3	3
Start-up Meet Pharma	5	3

Gràfic 6.3.1 Diagrama de barres de les variables SFO000 i IFOSTAT



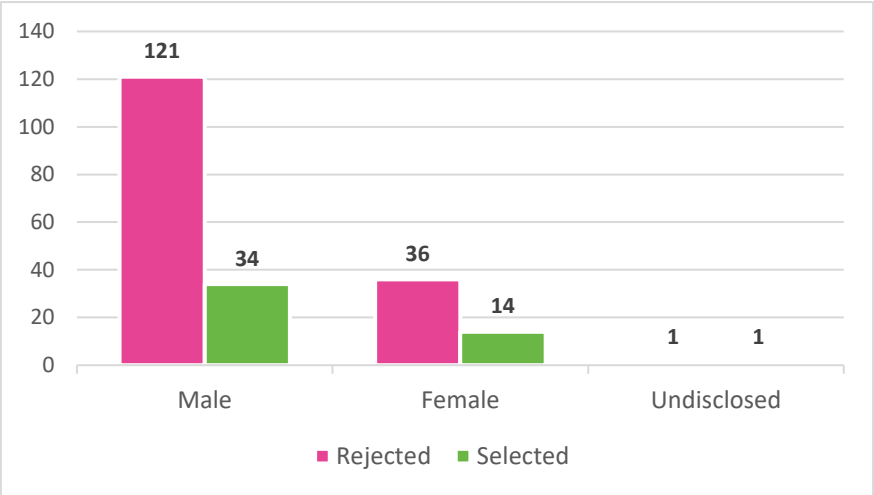
El programa *Headstart* és el que conté més *selected* ja que és el que conté més programes. Vegeu Taula 6.3.1 i Gràfic 6.3.1.

- SFO005 (*Gender*)

Taula 6.3.2 Descriptiva de les variables SFO005 i IFOSTAT

	Rejected	Selected
Male	121	34
Female	36	14
Undisclosed	1	1

Gràfic 6.3.2 Diagrama de barres de les variables SFO005 i IFOSTAT



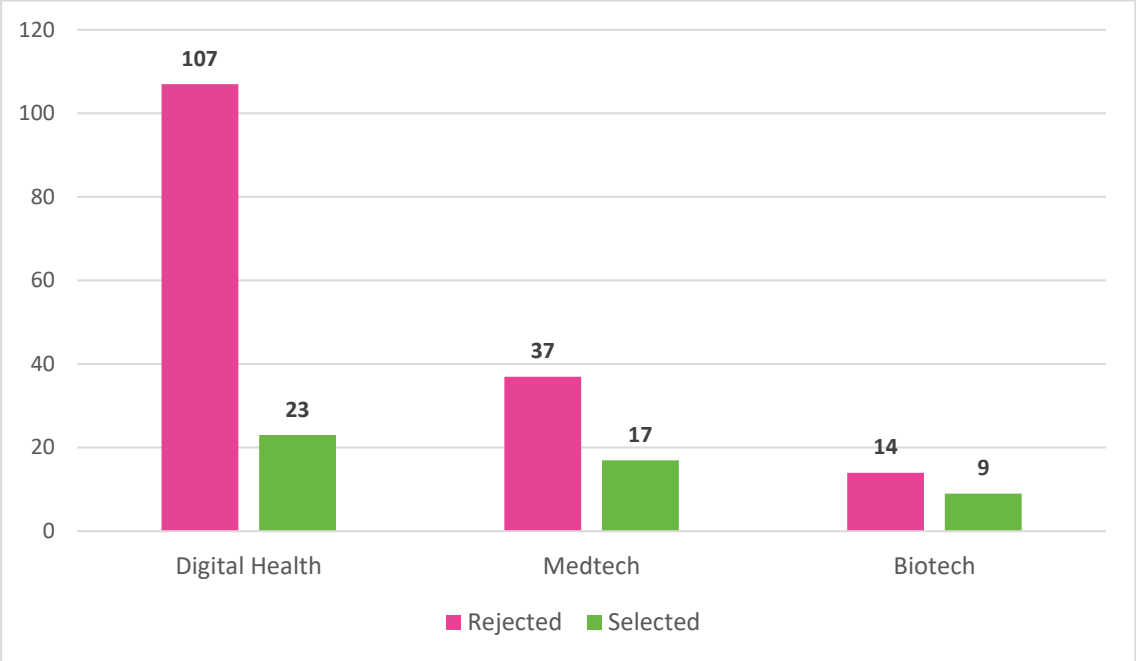
La distribució del *status* segons el sexe està bastant equilibrat. Vegeu Taula 6.3.2 i Gràfic 6.3.2.

- SFO011 (*Category*)

Taula 6.3.3 Descriptiva de les variables SFO011 i IFOSTAT

	Rejected	Selected
Biotech	14	9
Medtech	37	17
Digital Health	107	23

Gràfic 6.3.3 Diagrama de barres de les variables SFO011 i IFOSTAT



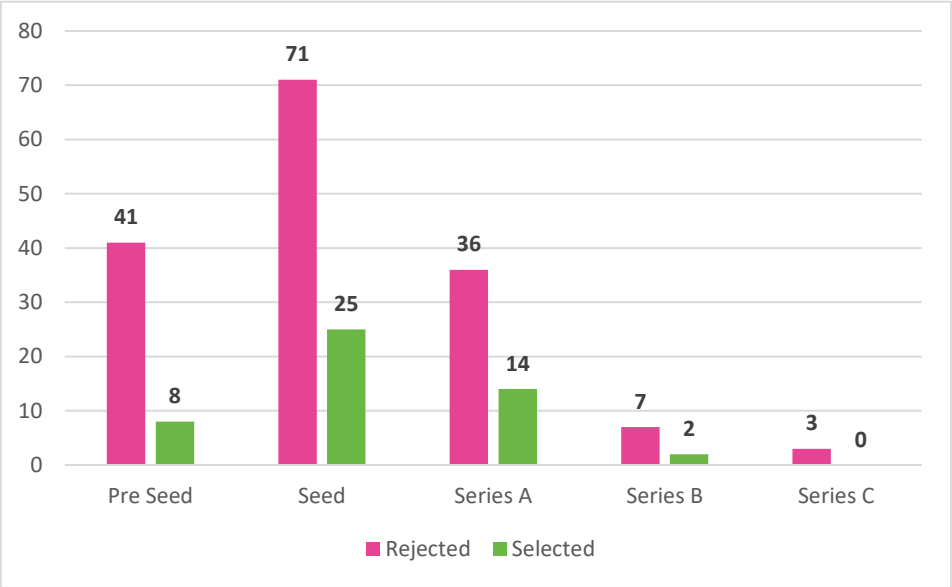
La distribució de la variable status segons la categoria es bastant proporcional, la categoria *digital Health* és la què té més empreses seleccionades. Vegeu Taula 6.3.3 i Gràfic 6.3.3.

- SFO018 (*Funding Stage*)

Taula 6.3.4 Descriptiva de les variables SFO018 i IFOSTAT

	Rejected	Selected
Pre Seed	41	8
Seed	71	25
Series A	36	14
Series B	7	2
Series C	3	0

Gràfic 6.3.4 Diagrama de barres de les variables SFO018 i IFOSTAT



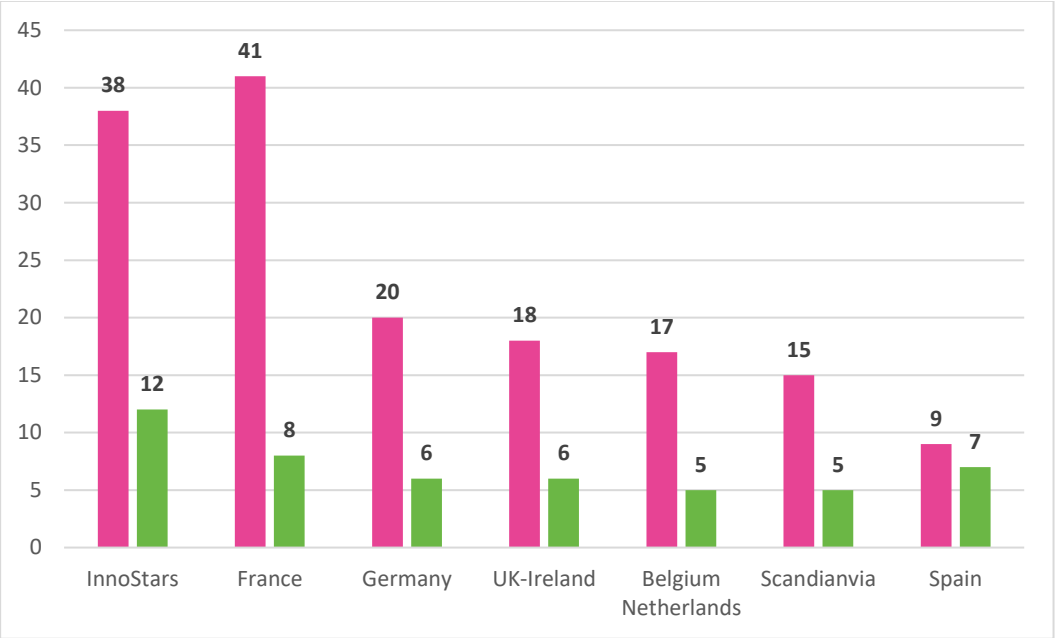
De la variable *Funding stage* destaca la categoria *Seed* com la que conté més empreses seleccionades i *series A* que conté menys empreses que *Pree Seed* però més empreses seleccionades. Vegeu Taula 6.3.4 i Gràfic 6.3.4.

- SFO025 (CLC)

Taula 6.3.5 Descriptiva de les variables SFO025 i IFOSTAT

	Rejected	Selected
Belgium-Netherlands	17	5
France	41	8
Germany	20	6
Scandinavia	15	5
Spain	9	7
UK-Ireland	18	6
InnoStars	38	12

Gràfic 6.3.5 Diagrama de barres de les variables SFO025 i IFOSTAT



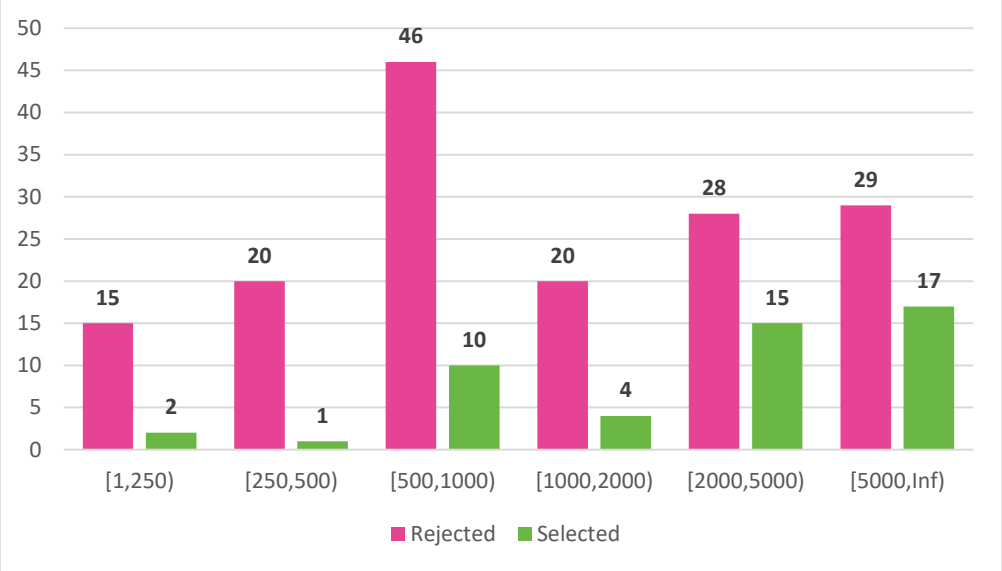
La distribució de la variable *status* segons el *CLC* està bastant equilibrada, destacar *InnoStars* com el *CLC* amb més empreses seleccionades. Vegeu Taula 6.3.5 i Gràfic 6.3.5.

- SFO026 (*Valuation*)

Taula 6.3.6 Descriptiva de les variables SFO026 i IFOSTAT

	Rejected	Selected
[1,250)	15	2
[250,500)	20	1
[500,1000)	46	10
[1000,2000)	20	4
[2000,5000)	28	15
[5000,Inf)	29	17

Gràfic 6.3.6 Diagrama de barres de les variables SFO026 i IFOSTAT

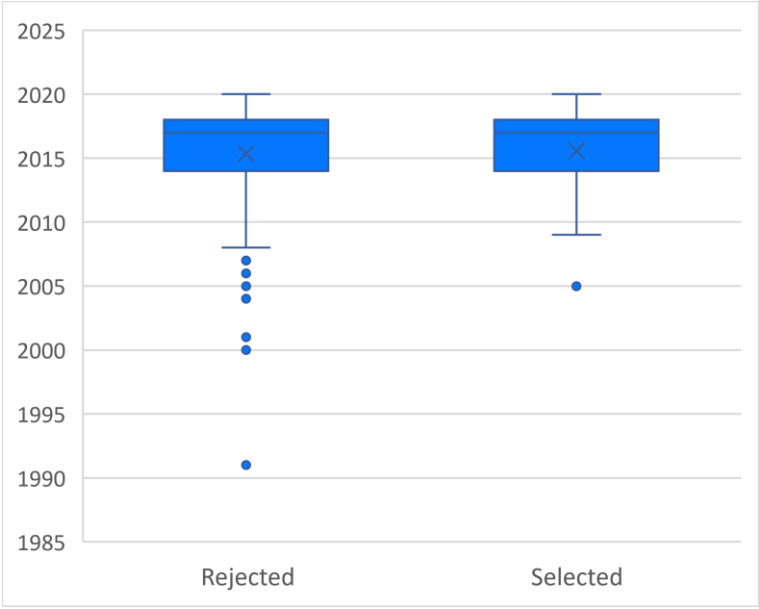


En la distribució de la variable *status* respecte la variable *valuation* observem que als intervals més alts el nombre d'empreses seleccionades és bastant més gran que a la resta. Vegeu Taula 6.3.6 i Gràfic 6.3.6.

6.4. DESCRIPCIÓ BIVARIANT VARIABLES QUANTITATIVES AMB VARIABLE RESPOSTA (STATUS)

- SFO007 *(Incorporation year*)

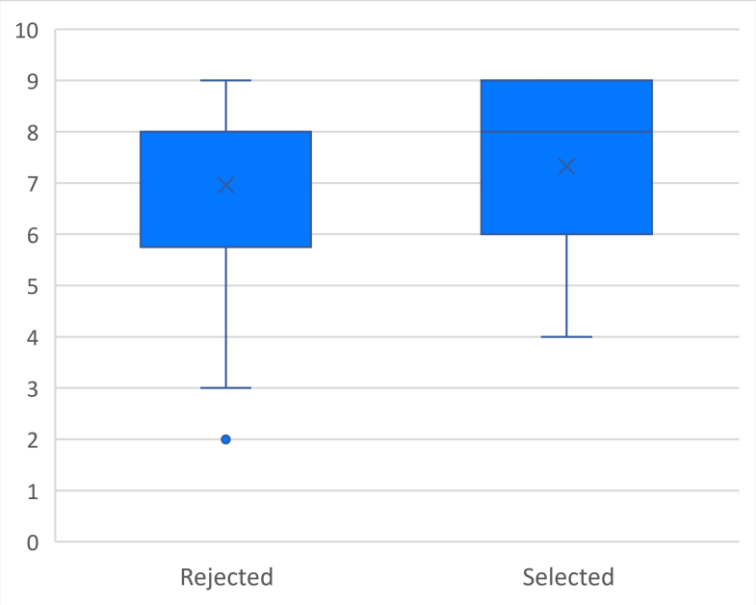
Gràfic 6.4.1 Diagrama de caixa de les variables SFO007 i IFOSTAT



La distribució de la variable *Incorporation year* segons l'*status* és bastant similar, lleugerament més alt al *selected*. Es destaca que en les *rejected* hi ha més *outliers*, per sota el límit inferior. Vegeu Gràfic 6.4.1.

- SFO012 (TRL)

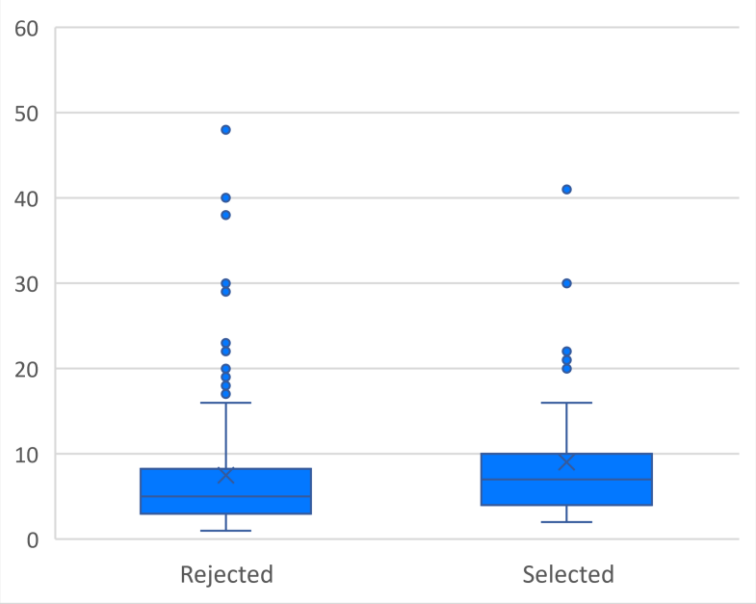
Gràfic 6.4.2 Diagrama de caixa de les variables SFO012 i IFOSTAT



En la distribució de la variable TRL segons l'*status*, es veu un TRL superior en les empreses *selected*. Vegeu Gràfic 6.4.2.

- SFO014 (FTE)

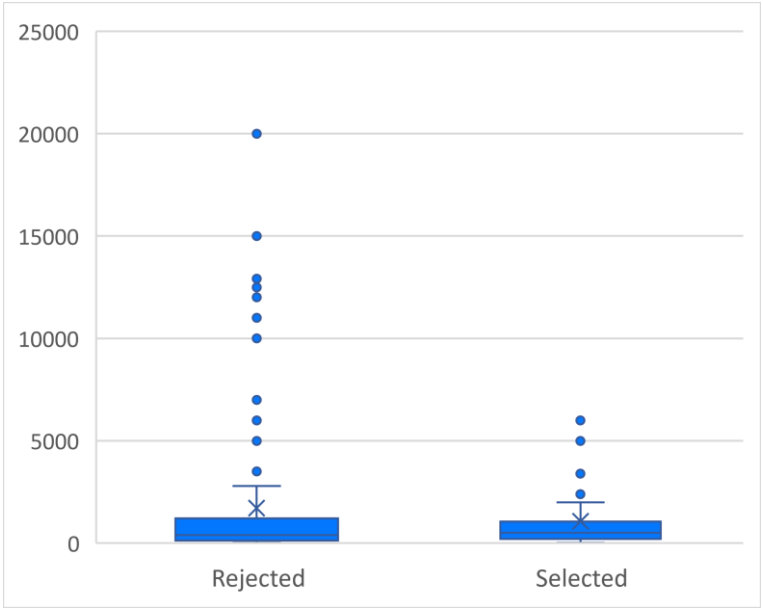
Gràfic 6.4.3 Diagrama de caixa de les variables SFO014 i IFOSTAT



La distribució de la variable FTE és superioren mediana en *l'status selected*. Vegeu Gràfic 6.4.3.

- SFO019 (*Total funding*)

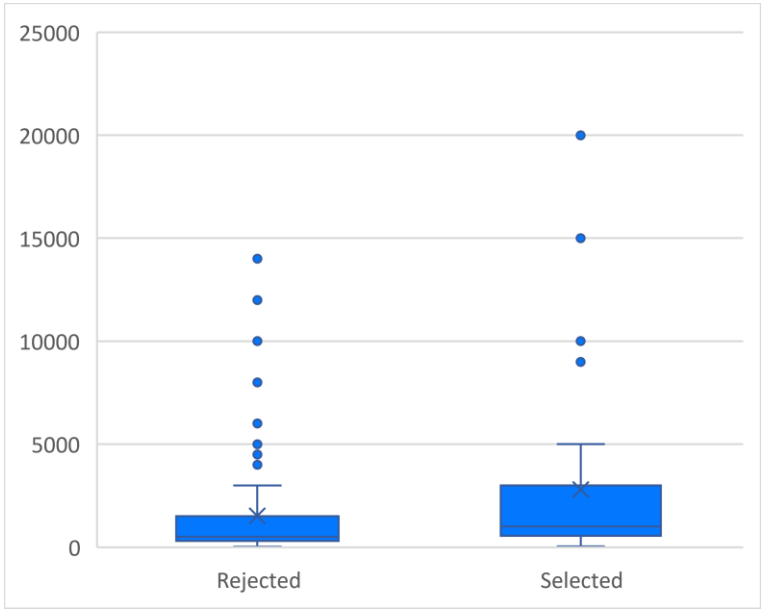
Gràfic 6.4.4 Diagrama de caixa de les variables SFO019 i IFOSTAT



La distribució de la variable *Total funding* està bastant equilibrada en els dos *status*. Vegeu Gràfic 6.4.4.

- SFO021 (*Equity sought*)

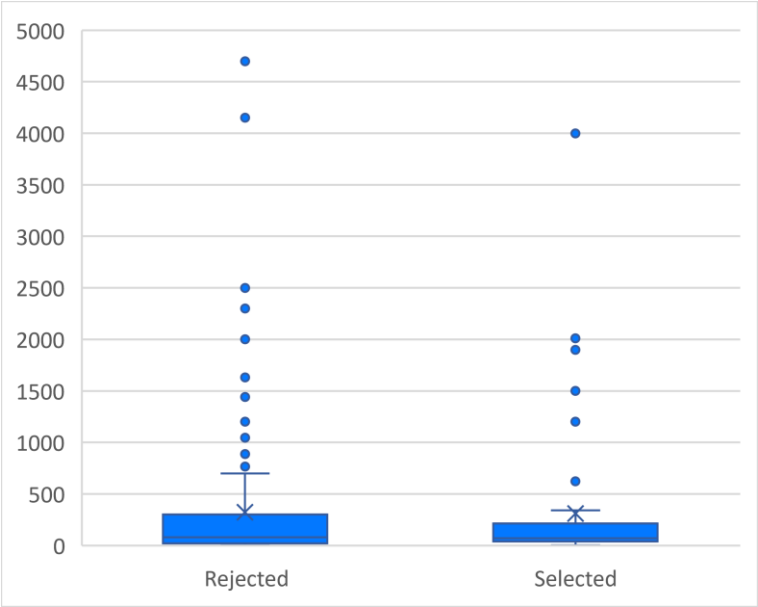
Gràfic 6.4.5 Diagrama de caixa de les variables SFO021 i IFOSTAT



La distribució de la variable *equity sought* segons *l'status* és superior en *l'status selected* en mediana, també té més variabilitat. Vegeu Gràfic 6.4.5.

- SFO022 (*Revenue*)

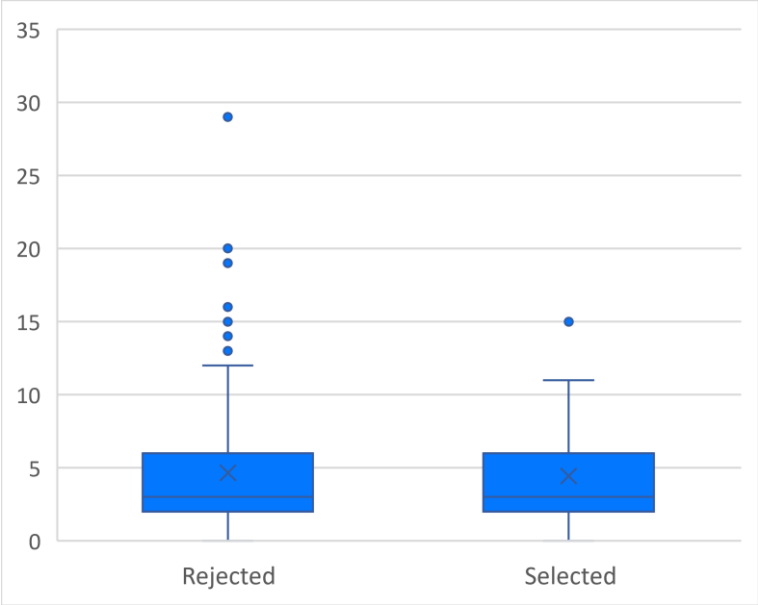
Gràfic 6.4.6 Diagrama de caixa de les variables SFO022 i IFOSTAT



La distribució de la variable *revenue* és molt similar pels dos *status*, hi ha una mica més de variabilitat en els *rejected*. Vegeu Gràfic 6.4.6.

- Antic

Gràfic 6.4.7 Diagrama de caixa de les variables Antic i IFOSTAT



La variable antic és lleugerament superior en mediana en *l'status rejected* sobre *l'status selected*. La distribució de *l'status rejected* té més valors extrems. Vegeu Gràfic 6.4.7.

7. MODELS LINEALS GENRALITZATS DE RESPOSTA BINARIA

7.1. AJUST I SELECCIÓ DELS MODELS

Tal i com s'ha explicat al apartat de metodologia, primerament es va dividir aleatòriament cada base de dades en dos:

- Mostral: Selecció aleatòria del 80 % de les dades
- Extramostral: Selecció aleatòria del 20 % de les dades

D'aquesta manera es s'ajustaran els models sobre la base de dades mostral i es faran les prediccions sobre les dades extramostrals.

L'ajust del models es farà amb la funció *glm* (*General Linear Models*) amb la família binomial ja que la variable resposta *status* (*IFOSTAT*) és binaria i el *link* és *logit* que és el *link* per defecte.

Finalment les variables seleccionades per l'ajust dels models són: SFO011, SFO012, SFO014, SFO018 i antic , com a "resta de camps" més les variables econòmiques depenent de la base de dades.

7.1.1. SELECCIÓ DELS MODELS I

Un cop ajustats els models additius amb les bases de dades mostrals compararem les sortides dels diferents models, les significacions dels paràmetres i l'ajust.

- **Model 1: Model additiu què conté totes les variables de la base de dades 1**

  IFOSTAT= β_0+ β_1**Medtech*+ β_2**Biotech*+ β_3**Other* + β_4*FTE+ β_5*TRL+ β_6**Seed* + β_7**Series A* + β_8**Series B* + β_9**Series C* + β_{10}*antic

Taula 7.1.2.1 *Summary* del model 1

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0,919	0,330	-2,779	0,005
SFO0112	-0,501	0,249	-2,008	0,045
SFO0113	-1,057	0,256	-4,133	0,000
SFO0115	-0,006	0,745	-0,008	0,994
SFO014	0,017	0,009	1,920	0,055
SFO012	0,067	0,049	1,378	0,168
SFO0182	0,217	0,215	1,009	0,313
SFO0183	0,386	0,278	1,390	0,165
SFO0184	0,504	0,567	0,888	0,375
SFO0185	-0,802	1,078	-0,744	0,457
antic	-0,013	0,022	-0,577	0,564

D'aquest primer model les categories *medtech* (SFO00112) i *biotech* (SFO00113) de la variable *category* són significatives al 5% i al 1%, respectivament. La variable FTE (SFO014) és significativa al 10% . Vegeu Taula 7.1.2.1.

Amb el test òmnibus, test anova del model nul front el model, s'obté un p-valor de 0,0001742,per tant, el model és globalment significatiu.

- **Model 2: Model additiu què conté totes les variables de la base de dades 2**

IFOSTAT= β_0+ β_1**Medtech*+ β_2**Biotech*+ β_3**Other* + β_4*FTE+ β_5*TRL+ β_6**Seed* + β_7**Series A* + β_8**Series B* + β_9**Series C* + β_{10}*antic+ β_{11}*SFO019

Taula 7.1.2.2 *Summary* del model 2

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0,595	0,442	-1,345	0,179
SFO0112	-0,400	0,338	-1,186	0,236
SFO0113	-1,354	0,349	-3,881	0,000
SFO0115	-13,975	882,744	-0,016	0,987
SFO014	0,052	0,020	2,538	0,011
SFO012	0,033	0,066	0,503	0,615
SFO0182	0,237	0,285	0,832	0,406
SFO0183	0,439	0,389	1,127	0,260
SFO0184	0,034	0,785	0,044	0,965
SFO0185	-12,478	882,743	-0,014	0,989
antic	-0,073	0,041	-1,765	0,078
SFO019	0,000	0,000	-2,299	0,021

Del model 2 la categoria *biotech* (SFO00113) de la variable *category* és significativa al 1%. La variable FTE i la variable (SFO019) són significatives al 5% i la variable antiguitat de l'empresa (antic) és significativa al 10% . Vegeu Taula 7.1.2.2.

Amb el test òmnibus, test *anova* del model nul front el model, s'obté un p-valor de 3.733e-05 ,per tant, el model és globalment significatiu.

- **Model 3: Mode additiu què conté totes les variables de la base de dades 3**

IFOSTAT= β_0+ β_1**Medtech*+ β_2**Biotech*+ β_3**Other* + β_4*FTE+ β_5*TRL+ β_6**Seed* + β_7**Series A* + β_8**Series B* + β_9**Series C* + β_{10}*antic+β_{11}*SFO021

Taula 7.1.2.3 Summary del model 3

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0,978	0,415	-2,359	0,018
SFO0112	-0,253	0,315	-0,802	0,422
SFO0113	-0,985	0,330	-2,982	0,003
SFO0115	-14,805	1455,398	-0,010	0,992
SFO014	0,000	0,022	0,016	0,987
SFO012	0,073	0,063	1,173	0,241
SFO0182	-0,006	0,258	-0,022	0,982
SFO0183	0,320	0,393	0,815	0,415
SFO0184	-1,629	1,181	-1,379	0,168
SFO0185	-14,141	809,316	-0,017	0,986
antic	-0,056	0,039	-1,414	0,157
SFO021	0,000	0,000	2,678	0,007

Del model 3 la categoria *biotech* (SFO00113) de la variable *category* i la variable (SFO021) són significatives al 1%. Vegeu Taula 7.1.2.3.

Amb el test òmnibus, test *anova* del model nul front el model, s'obté un p-valor de 8,241e-05 ,per tant, el model és globalment significatiu.

- **Model 4: Model què conté totes les variables de la base de dades 4**

  IFOSTAT= β_0+ β_1**Medtech*+ β_2**Biotech*+ β_3**Other* + β_4*FTE+ β_5*TRL+ β_6**Seed* + β_7**Series A* + β_8**Series B* + β_9**Series C* + β_{10}*antic+β_{11}*SFO022

Taula 7.1.2.4 *Summary* del model 4

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2,619	0,766	-3,420	0,001
SFO0112	-0,298	0,523	-0,570	0,569
SFO0113	-1,133	0,499	-2,272	0,023
SFO0115	-14,401	1498,857	-0,010	0,992
SFO014	0,025	0,025	1,006	0,315
SFO012	0,257	0,101	2,539	0,011
SFO0182	0,463	0,405	1,141	0,254
SFO0183	0,586	0,496	1,181	0,238
SFO0184	0,075	0,972	0,077	0,938
SFO0185	-15,099	1126,876	-0,013	0,989
antic	-0,042	0,045	-0,932	0,352
SFO022	0,000	0,000	-0,385	0,700

Del model 4 la categoria *biotech* (SFO00113) de la variable *category* i la variable són significatives al 5%. Vegeu Taula 7.1.2.4.

Amb el test òmnibus, test *anova* del model nul front el model, s'obté un p-valor de 0,02287 ,per tant, el model és globalment significatiu. Vegeu Taula 7.1.2.4.

- **Model 5: Model què conté totes les variables amb la base de dades 5**

  IFOSTAT= β_0+ β_1**Medtech*+ β_2**Biotech*+ β_3**Other* + β_4*FTE+ β_5*TRL+ β_6**Seed* + β_7**Series A* + β_8**Series B* + β_9**Series C* + β_{10}*antic+ β_{11}*SFO0262+ β_{12}*SFO0263+ β_{13}*SFO0264+ β_{14}*SFO0265 + β_{15}*SFO0266

Taula 7.1.2.5 Summary del model 5

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1,256	0,498	-2,523	0,012
SFO0112	-0,111	0,275	-0,405	0,686
SFO0113	-0,782	0,287	-2,729	0,006
SFO0115	1,502	1,556	0,965	0,334
SFO014	0,012	0,008	1,465	0,143
SFO012	0,012	0,055	0,216	0,829
SFO0182	0,007	0,255	0,029	0,977
SFO0183	-0,110	0,349	-0,314	0,753
SFO0184	0,161	0,604	0,266	0,790
SFO0185	-0,896	1,110	-0,807	0,419
antic	-0,021	0,028	-0,751	0,452
SFO0262	-0,385	0,536	-0,718	0,473
SFO0263	0,553	0,405	1,366	0,172
SFO0264	0,569	0,468	1,216	0,224
SFO0265	1,525	0,418	3,647	0,000
SFO0266	0,807	0,440	1,833	0,067

Del model 5 la categoria *biotech* (SFO00113) de la variable *category* i la categoria 5 de la variable *valuation* són significatives al 1%, la categoria 6 de la variable *valuation* és significativa al 10%. Vegeu Taula 7.1.2.5.

Amb el test òmnibus, test *anova* del model nul front el model, s'obté un p-valor de 1,595e-07, per tant, el model és globalment significatiu.

- **Model 6: Model què conté totes les variables de la base de dades 6**

IFOSTAT= β_0+ β_1*_Medtech_+ β_2*_Biotech_+ β_3*_Other_ + β_4*FTE+ β_5*TRL+ β_6*_Seed_ + β_7*_Series A_ + β_8*_Series B_ + β_9*_Series C_ + β_{10}*antic+ β_{11}*SFO0262+ β_{12}*SFO0263+ β_{13}*SFO0264+ β_{14}*SFO0265 + β_{15}*SFO0266 +β_{16}*SFO022+ β_{17}*SFO019+β_{18}*SFO021

Taula 7.1.2.6 *Summary* del model 6

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2,501	1,445	-1,731	0,084
SFO0112	-0,422	0,672	-0,628	0,530
SFO0113	-1,769	0,673	-2,630	0,009
SFO014	0,042	0,039	1,080	0,280
SFO012	0,130	0,131	0,996	0,319
SFO0182	0,896	0,623	1,439	0,150
SFO0183	0,037	0,775	0,048	0,961
SFO0184	-0,429	1,314	-0,327	0,744
SFO0185	-14,251	1001,834	-0,014	0,989
antic	-0,038	0,065	-0,590	0,555
SFO0262	0,024	1,559	0,015	0,988
SFO0263	0,725	1,216	0,596	0,551
SFO0264	0,369	1,319	0,280	0,780
SFO0265	1,276	1,256	1,016	0,310
SFO0266	1,744	1,304	1,337	0,181
SFO022	0,000	0,000	-0,943	0,346
SFO019	0,000	0,000	-1,054	0,292
SFO021	0,000	0,000	0,868	0,385

Del model 6 la categoria *biotech* (SFO00113) de la variable *category* és significativa al 1%. Vegeu Taula 7.1.2.6.

Amb el test òmnibus, test *anova* del model nul front el model, s'obté un p-valor de 0,004735 ,per tant, el model és globalment significatiu.

- **Model 7: Model què conté totes les variables de la base de dades 7**

  IFOSTAT= β_0+ β_1**Medtech*+ β_2**Biotech*+ β_3**Other* + β_4*FTE+ β_5*TRL+ β_6**Seed* + β_7**Series A* + β_8**Series B* + β_9**Series C* + β_{10}*antic+ β_{11}*SFO022+ β_{12}*SFO019

Taula 7.1.2.7 Summary del model 7

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1,399	0,870	-1,607	0,108
SFO0112	-0,619	0,568	-1,089	0,276
SFO0113	-1,683	0,538	-3,127	0,002
SFO014	0,044	0,027	1,613	0,107
SFO012	0,099	0,107	0,920	0,358
SFO0182	0,846	0,530	1,598	0,110
SFO0183	0,937	0,632	1,483	0,138
SFO0184	0,384	0,953	0,403	0,687
SFO0185	-13,638	1020,162	-0,013	0,989
antic	-0,044	0,050	-0,867	0,386
SFO022	0,000	0,000	-0,199	0,842
SFO019	0,000	0,000	-0,855	0,392

Del model 7 la categoria *biotech* (SFO00113) de la variable *category* és significativa al 1%. Vegeu Taula 7.1.2.7.

Amb el test òmnibus, test *anova* del model nul front el model, s'obté un p-valor de 0,01568 ,per tant, el model és globalment significatiu.

Taula 7.1.2.8 Resum dels models I

Model	n	g.ll.	AIC	BIC	Loglink	Variables significatives
m1	655	11	780,3	829,6	-379,2	SFO0112 SFO0113 SFO014
m2	412	12	451,8	500,1	-231,9	SFO0113 SFO014 SFO019 Antic
m3	461	12	522,9	572,6	-249,5	SFO0113 SFO021
m4	262	12	286,2	328,98	-131,1	SFO0113
m5	558	16	651,9	721,08	-309,9	SFO0113 SFO0265 SFO0266
m6	166	18	176,2	232,25	-70,12	SFO0113
m7	182	12	206,4	244,9	-91,2	SFO0113

A la Taula 7.1.2.8 tenim una comparació dels diferents models amb diferents paràmetres de bondat d'ajust: *AIC, BIC i Loglink*. Són models a partir de la mateixa base de dades on tenen la majoria de dades comunes. Per tant, es pot fer una comparació per l'elecció del millor model. El millor model segons els tres criteris és el 6, el qual intentarem millorar. Es seleccionen també el model 4 i 7 ja són els quela bondat d'ajust és pròxima al del model 6.

Respecte les variables significatives s'observen variables comunes en els diferents models: categoria *biotech* (SFO00113) de la variable *category* es comuna en tots els models i variable SFO014 (FTE) es comuna en dos models. La significació utilitzada és al 10%.

7.1.2. SELECCIÓ DELS MODELS II

Un cop seleccionats els millors models segons els paràmetres de bondat d'ajust s'intentarà millorar-los. Per fer-ho s'utilitza la funció *step* que a partir d'un model fa comparacions de models additius anant traient variables per tal d'aconseguir el millor model segons el criteri AIC.

- **Model 4.1 – Model 4 millorat amb la funció *step***

  IFOSTAT= β_0+ β_1**Medtech*+ β_2**Biotech*+ β_3**Other* + β_4*TRL

Taula 7.1.3.1 *Summary* del model 4.1

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2,387	0,716	-3,335	0,001
SFO0112	-0,412	0,512	-0,805	0,421
SFO0113	-1,147	0,490	-2,340	0,019
SFO0115	-15,297	1010,419	-0,015	0,988
SFO012	0,276	0,096	2,882	0,004

Després d'aplicar la funció *step* al model 4 el model additiu seleccionat per la funció *step* conté les variables *category* i TRL.

El model 4.1 té significatives la categoria *biotech* de la variable *category* i la variable TRL al 5% i al 1%, respectivament. Vegeu Taula 7.1.3.1.

El model és globalment significatiu amb un p-valor de 0,003748.

- **Model 6.1 – Model 6 millorat amb la funció step**

  IFOSTAT= β_0+ β_1**Medtech*+ β_2**Biotech*+ β_3*FTE+ β_4*TRL+ β_5*antic

Taula 7.1.3.2 *Summary* del model 6.1

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1,107	0,874	-1,266	0,205
SFO0112	-0,807	0,598	-1,348	0,178
SFO0113	-2,134	0,589	-3,625	0,000
SFO014	0,047	0,026	1,847	0,065
SFO012	0,180	0,117	1,536	0,125
antic	-0,094	0,056	-1,684	0,092

Un cop utilitzada la funció *step* al model 6 la funció *step* ens dona el model additiu amb les variables *category*, *TRL* i FTE com a millor model.
Del model 6.1 la categoria biotech és significativa amb un nivell de significació del 1% i les variables antic i FTE són significatives amb un nivell de significació del 10%. Vegeu Taula 7.1.3.2.
El model és globalment significatiu amb un p-valor de 0,0006606.

- **Model 7.1 – Model 7 millorat amb la funció *step***

  IFOSTAT= β_0+ β_1**Medtech*+ β_2**Biotech*+ β_3*FTE+ β_4*SFO019

Taula 7.1.3.3 *Summary* del model 7.1

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0,304	0,448	-0,678	0,498
SFO0112	-0,529	0,524	-1,010	0,312
SFO0113	-1,524	0,498	-3,059	0,002
SFO014	0,042	0,021	1,994	0,046
SFO019	0,000	0,000	-1,404	0,160

El model 7.1 és el resultat d'utilitzar la funció, el model additiu en qüestió conté les variables *category*, FTE i total *funding*.

El model 7.1 té significativa al 1% la categoria *biotech* de la variable *category* i la variable FTE és significativa la 5%. Vegeu Taula 7.1.3.3.

El model és globalment significatiu amb un p-valor de 0,002553.

Taula 7.1.3.4 Resum dels models II

Model	n	g.ll.	AIC	BIC	Loglink	Variables significatives
m4	262	12	278,8	296,7	-134,42	SFO0113 SFO012
m6	166	18	166,7	185,3	-77,33	SFO0113 SFO014 Antic
m7	182	12	199,4	215,42	-94,7	SFO0113 SFO014

A la Taula 7.1.3.4 es pot veure les diferents comparacions dels models millorats, el millor model segons els tres paràmetres de bondat d'ajust és el model 6 millorat segons amb la funció *step*.

Les variables significatives amb un nivell de significació del 10% del model seleccionat són la categoria *biotech* (SFO00113) de la variable *category*, la variable FTE (SFO014) i la variable antic. La categoria SFO0113 es comuna en els altres dos models i la variable SFO014 en un model.

7.1.3. SELECCIÓ DELS MODELS III

Després d'haver seleccionat el model 6 com a millor model es compararà el model amb diferents links:

Taula 7.1.3.1 Resum dels models III

Model	n	g.ll.	AIC	BIC	Loglink	Variables significatives
m6 link logit	166	18	166,17	185,14	-77,03	SFO0113 SFO014 Antic
m6 link probit	166	18	166,39	185,26	-77,19	SFO0113 SFO014 Antic
m6 link cloglog	166	18	166,76	185,43	-77,38	SFO0113 SFO014

El millor models segons els paràmetres de bondat d'ajust és el model 6 amb el *link lògit*. A partir d'ara serà anomenat com a model final. Vegeu Taula 7.1.3.1.

Respecte a la significació de les variables, les variables significatives al 10% del model final són la categoria *biotech* (SFO00113) de la variable *category*, la variable FTE (SFO014) i la variable antic. Vegeu Taula 7.1.3.2. El model és globalment significatiu.

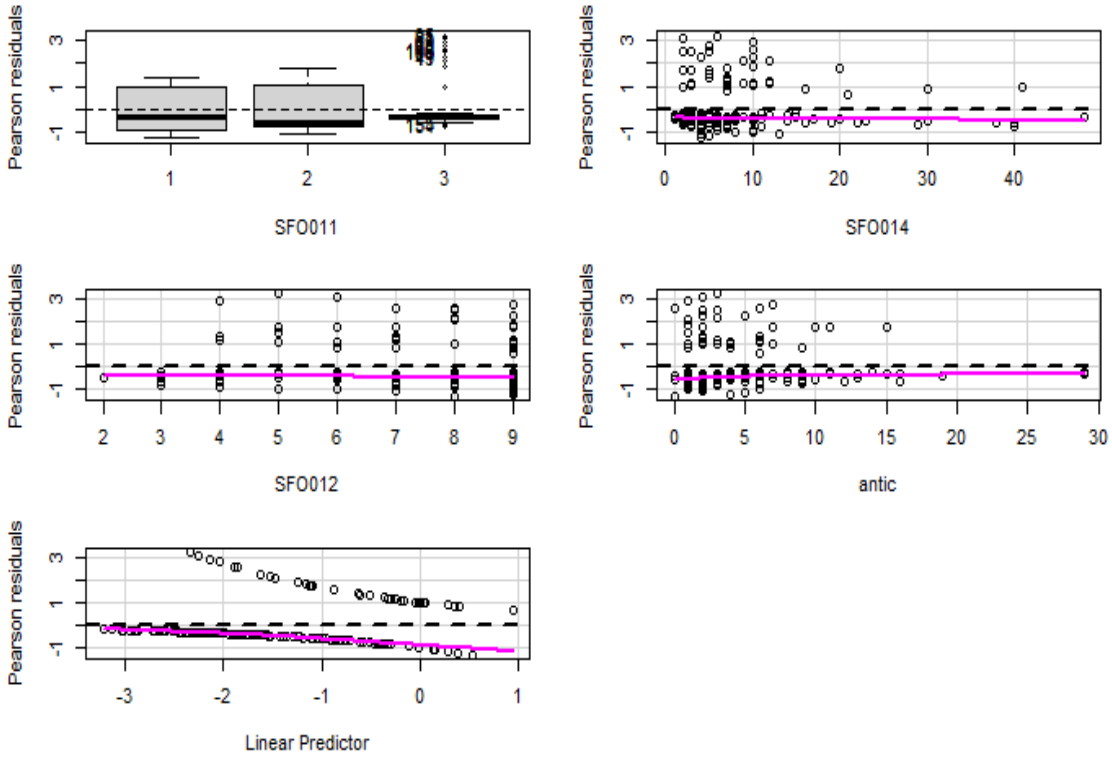
$$IFOSTAT = \beta_0 + \beta_1 * Medtech + \beta_2 * Biotech + \beta_3 * FTE + \beta_4 * TRL + \beta_5 * antic$$

Taula 7.1.3.2 Summary del model final

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1,107	0,874	-1,266	0,205
SFO0112	-0,807	0,598	-1,348	0,178
SFO0113	-2,134	0,589	-3,625	0,000
SFO014	0,047	0,026	1,847	0,065
SFO012	0,180	0,117	1,536	0,125
antic	-0,094	0,056	-1,684	0,092

7.2. VALIDACIÓ MODEL SELECCIONAT

Gràfic 7.2.1 Gràfics de residus del model final



En els gràfic 7.2.1 s'observa el residus per cada una de les variables, per validar-los han d'estar al voltant del 0. Es poden donar per bons ja que en cada una de les variables estan més o menys al voltant del 0.

Com s'ha vist el model es globalment significatiu, és el millor model que s'ha pogut ajustar i s'acaben de validar els residus ,per tant, ja podem considerar què és el millor model vàlid per realitzar prediccions i interpretar els coeficients.

7.3. PREDICCIONS MODEL SELECCIONAT

Taula 7.3.1 Taula prediccions total dades

	0	1
0	281	88
1	159	115

Taula 7.3.2 Taula indicadors total dades

Encert	Sensibilitat	Especifitat	Pred.positiu	Pred.negatiu
0.6158631	0.5665025	0.6386364	0.419708	0.7615176

Taula 7.3.3 Taula prediccions extramostrals

	0	1
0	17	8
1	12	4

Taula 7.3.4 Taula indicadors extramostrals

Encert6.1	Sensibilitat6.1	Especifitat6.1	Pred.positiu6.1	Pred.negatiu6.1
0.5121951	0.3333333	0.5862069	0.25	0.68

Ja que la base de dades no està balancejada i la proporció de 0 és el 75%, per fer les prediccions s'ha assignat una predicció 0 quan la probabilitat és superior a 0,75 en comptes de fer servir l'habitual 0,5.

A les taules 7.3.1 i 7.3.2 es troben les prediccions de la base de dades 1, ja que el model seleccionat conté variables de la resta de camps. De la taula de dalt les columnes són els valors reals i les files són les prediccions. Les prediccions són bastant normals. La especificitat que es refereix al percentatge de 0 predits com a 0 que realment són 0 és del 63,8% un resultat acceptable. Pel que fa, la sensibilitat, el percentatge d'1 predits com a 1 que realment són 1, és una mica més baix 56'6%, un resultat bastant regular.

Les següents dos taules 7.3.3 i 7.3.4 són sobre les dades extramostrals, el nombre de dades no és molt gran. La sensibilitat és 56,65% bastant regulat i la sensibilitat 33,3%.

Com a conclusió de les prediccions deixen bastant que desitjar, però no són dolentes del tot. Igualment no és pot afirmar que sigui un bon model predictiu. Per tant, les prediccions no les considerem com ha bones del tot.

7.4. RESULTATS DEL MODELS

Pel que fa al model final ha estat seleccionat com el model que ajusta millor les dades respecte als altres models. Tot i això, s'ha vist variables significatives comunes en tots els models. S'han realitzat els gràfics de residus i s'han donat com a vàlids. Finalment, s'han fet prediccions amb el model i s'ha descartat com un bon model predictiu ,sobretot, per que la sensibilitat és molt baixa.

Com ha últim pas, quedaria interpretar els coeficients de les variables del model, ja que ens donarà informació del signe i l'efecte que produeixen les variables explicatives sobre la variable resposta. Al ser un model de resposta binaria amb *link* lògit, calcularem *l'odds ratio* de cada coeficient per veure l'efecte que tenen sobre la variable resposta, per fer-ho s'elevarà cada un dels valors dels coeficients a e.

- SFO0112 (categoria *Medtech* variable *category*): β_1= -0.8066; OR=$e^{-0.8066}$=0,4463
  L'Odds Ratio de la categoria *medtech* indica que una empresa del sector *medtech* té un 65% menys de probabilitats de ser acceptada respecte la categoria de referència, en aquest cas la categoria *digital health*. En aquest cas el coeficient **no és significatiu** per tant, estadísticament es com si el coeficient fos 0, per tant, l'efecte de categoria *Medtech* no és significativament diferent al de la variable Digital Health.

- SFO0113 (categoria *Biotech* variable *category*): β_1= -2,1337; OR=$e^{-2,1337}$ =0,1183
  L'Odds Ratio de la categoria *Biotech* indica que una empresa del sector *medtech* té un 88,2% menys de probabilitats de ser acceptada respecte la categoria de referència, en aquest cas la categoria *digital health*. En aquest cas el coeficient **és significatiu**.

- SFO012 (variable TRL): β_1= 0,18; OR=$e^{0,18}$ =1,1975
  L'Odds Ratio de la variable TRL indica que per cada nivell que augmenta de TRL té un 19,75% més de probabilitats de ser acceptada. En aquest cas el coeficient **no és significatiu** per tant, estadísticament es com si el coeficient fos 0, per tant, l'efecte de la variable TRL sobre *l'Status* **no és significatiu**.

- SFO014 (variable FTE): β_1= 0,047; OR=$e^{0,047}$ =1,04842
  L'Odds Ratio de la variable FTE indica que per cada unitat de FTE que augmenta té un 4,84% més de probabilitats de ser acceptada. En aquest cas el coeficient **és significatiu**.

- Antic: β_1= -0,094; OR=$e^{-0,094}$ =0,9099
  L'Odds Ratio de la variable antic indica que per cada any d'antiguitat disminueixen un 9% les probabilitats de ser acceptada. En aquest cas el coeficient **és significatiu**.

El nivell de significació utilitzat ha estat del 10%

CONCLUSIONS

La majoria d'objectius s'han assolit. Primerament, l'especificació d'un model lineal generalitzat, que tingui en compte la naturalesa de la variable resposta de tipus qualitativa binària. S'han especificat uns quants models usant diferents bases de dades per perdre el mínim de variables possible i s'ha seleccionat el millor model segons els diferents criteris de bondat d'ajust. Pel que fa la validació del model és globalment significatiu i a partir dels gràfics de residus no s'ha vist cap anomalia, per tant, s'han donat com a vàlids.

Les prediccions del model, que és el següent objectiu, no s'han pogut donar com a bones del tot. El model és el millor que hem pogut ajustar però no té una bona capacitat predictiva. Les prediccions correctes ronden entre el 50 i el 60%, per tant no podem dir que siguin bones. Així doncs aquest objectiu no s'ha pogut assolir satisfactòriament.

El següent objectiu era conèixer les variables que tenen més efecte sobre la variable resposta status (si l'empresa a estat seleccionada o no) i quantificar aquest efecte, en termes estadístics, identificant els coeficients significatius i interpretar-los. Les variables rellevants són la categoria ***Biotech*** de la variable ***category***, la variable FTE (*Full Time Equivalent*) i la variable **antic** (antiguitat de l'empresa). Pel que fa la categoria ***Biotech*** el coeficient és negatiu el que ens indica que una empresa d'aquesta categoria és seleccionada respecte la categoria de referència (*digital Health*), concretament un 88,2% menys . El coeficient de la variable **FTE** és positiu i, per tant, ens indica que per cada treballador a jornada complerta que augmenti augmentaran les probabilitats de ser ***selected***, concretament, un 4,84%. L'últim coeficient significatiu és el de la variable **antic**, que és negatiu, és a dir, per cada any d'antiguitat de l'empresa disminueix la probabilitat de ser seleccionada, un 9%.

Per tant una empresa que no sigui ***biotech***, amb pocs anys d'antiguitat i bastants treballadors seria el prototip segons el model de tenir més probabilitats per ser seleccionada. L'explicació de que una empresa que sigui *biotech* es seleccioni menys ve donada en que són moltes menys nombroses que les dues altres categories en termes absoluts com s'ha pogut veure en els gràfics, però en termes relatius si que és superior a les altres categories, per tant, com que es presenten menys empreses *Biotech* que de la resta de categories se'n seleccionen menys en termes absoluts. El motiu de la significació del coeficient de la variable FTE, podria ser que hi ha programes pensats per empreses més madures que tinguin ja un equip de treball més gran; per tant, com més gran sigui l'equip de treball més coneixement reunit hi haurà i serà una empresa més "treballada". Amb les empreses noves passaria el mateix com més treballadors hi hagi més ben treballada estarà l'empresa i més probabilitats de ser seleccionada. Finalment, el coeficient de la variable antic el podem explicar ja que la majoria d'empreses que es presenten són de creació recent, atès que la majoria de programes estan pensats per aquest tipus d'empreses.

L'últim objectiu, el d'aplicar els coneixements apresos durant el grau, també s'ha assolit. S'han aplicat coneixements referents de les assignatures de programació i bases de dades per netejar i treballar sobre les dades. I sobretot s'han aplicat els coneixements tan pràctics com teòrics apresos en les assignatures de models lineals. A part d'aquest coneixement tan pràctic com teòric s'ha intentat mantenir el rigor i la exigència que ens han inculcat els professors des del primer dia.

Finalment, tots els coeficients tenen sentit i tenen concordança però el de antic i FTE per alguns casos poden ser contradictoris, per tant, després de veure el resultat dels coeficients, tindria sentit en un futur estudi fer un model agrupant programes similars o individualitzat per cada programa, ja que hi ha programes destinats a diferents tipus d'empreses. Ara mateix resulta inviable realitzar aquest estudi ja que la quantitat de dades es insuficient. Com que cada cop s'estan recollint les dades de millor forma i amb tots els camps sense tenir valors buits en un parell d'anys es podria realitzar aquest mateix estudi però com s'ha dit individualitzat o agrupat per programes similars.

BIBLIOGRAFIA

Cramer, J. S. (1999). Predictive performance of the binary logit model in unbalanced samples. *Journal of the Royal Statistical Society Series D: The Statistician*, *48*(1), 85–94. https://doi.org/10.1111/1467-9884.00173

Espinos Alegre, C. (2017). *El Sector MedTech*. https://upcommons.upc.edu/handle/2117/112092

King, G., & Zeng, L. (2003). Logistic regression in rare events data. *Journal of Statistical Software*, *8*, 137–163. https://doi.org/10.18637/jss.v008.i02

Moreno, J. de J., Garcia, A., & Pablo, F. (2003). *Análisis de la relación entre el crecimiento empresarial, la edad de la empresa y la estructura de propiedad*. *5*, 41. file:///G:/Análisis de la relación entre le crecimiento empresarial, la edad de la empresa y la estructura de propiedad.pdf

Salas-Eljatib, C., Fuentes-Ramirez, A., Gregoire, T. G., Altamirano, A., & Yaitul, V. (2018). A study on the effects of unbalanced data when fitting logistic regression models in ecology. *Ecological Indicators*, *85*(April 2017), 502–508. https://doi.org/10.1016/j.ecolind.2017.10.030

RDocumentation. *Step: Choose a model by AIC in a Stepwise Algorithm* https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/step

Wikipedia. (2021). *Modelo Lineal Generalizado* https://es.wikipedia.org/wiki/Modelo_lineal_generalizado

Wikipedia. (2020). *Criterio de información de Akaike* https://es.wikipedia.org/wiki/Criterio_de_informaci%C3%B3n_de_Akaike

Wikipedia. (2020). *Criterio de información bayesiano* https://es.wikipedia.org/wiki/Criterio_de_informaci%C3%B3n_bayesiano

EIT Health. https://eithealth.eu/what-we-do/

GREENE, W.H. (1999): Análisis econométrico. 3a Ed. Prentice Hall.

WOOLDRIDGE, J. (2009): Introducción a la Econometría. Un enfoque moderno. 4a Ed. Cengage Learning Eds.

Akaike H. (1974): *A new look at the statistical model identification. IEEE Transactions on Automatic Control*. AC-19:716-23

Clogg CC, Rubin DB, Schenker N, Schultz B & Weidman L.: *Multiple imputation of industry and occupation codes in census public-use samples using Bayesian logistic regression.*

Journal of the American Statistical Association 1991; 86:68 -78.

Firth D. *Bias reduction of maximum likelihood estimates*. Biometrika 1993; 80:27-38.

Heinze G, & Schemper M. (2002): *A solution to the problem of separation in logistic regression*. Statistics in Medicine 21:2409-19.

Hosmer D, Lemeshow S. *Goodness-of-fit tests for the multiple logistic regression model. Commun Stat Part A Theor Meth*. 1980;A10:1043-1069.

Hosmer, D. W., & Lemeshow, S. (1989): *Applied logistic regression*. New York: Wiley

McCullagh P. & Nelder JA. (1989). *Generalized Linear Models*. Champman & Hall: CRC

Fox J. *Applied Regression Analysis and Generalized Linear Models*. Sage Publications, 2nd Edition 2008.

Fox J., Weisberg, S., *An R Companion to Applied Regression*.Sage Publications, 2nd Edition 2011.

Dobson A., Barnett A. *An introduction to Generalized Linear Models*. Third Edition Chapman and Hall, 2008.

ANNEX

```
#Complimentació i depuració de la base dades
codis <- unique(dd$IFOCMP[duplicated(dd$IFOCMP)]) #vector amb codis de les
empreses repetides
length(codis)

drep<-dd[dd$IFOCMP%in% codis,] #base de dades amb les empreses repetides
nrow(drep)
cops2<-names(which(table(drep$IFOCMP)==2))
length(cops2)
d2cops<-dd[dd$IFOCMP %in% cops2,]
nrow(d2cops)

sum(is.na(drep[,7]))
sum(is.na(drep[,6]))
for (i in nrow(drep)-1:1) {
    if (is.na(drep[i,7])){
        if(!is.na(drep[i+1,7]) & drep$IFOCMP[i]==drep$IFOCMP[i+1]){
            drep[i,7]<-drep[i+1,7]
            drep[i,6]<-drep[i+1,6]
        }
        else if(!is.na(drep[i-1,7]) & drep$IFOCMP[i]==drep$IFOCMP[i-1]){
            drep[i,7]<-drep[i-1,7]
            drep[i,6]<-drep[i-1,6]
        }
    }
}

sum(is.na(drep[,7]))
sum(is.na(drep[,6]))

sum(is.na(drep[,10]))

for (i in (nrow(drep)-1):1) {
    if (is.na(drep[i,10])){
        if(!is.na(drep[i+1,10]) & drep$IFOCMP[i]==drep$IFOCMP[i+1]){
            drep[i,10]<-drep[i+1,10]
        }
        else if(!is.na(drep[i-1,10]) & drep$IFOCMP[i]==drep$IFOCMP[i-1]){
            drep[i,10]<-drep[i-1,10]

        }
    }
}

sum(is.na(d2cops[,10]))
head(drep)
!is.na(drep[,10]) & drep$IFOCMP[i]==drep$IFOCMP[i+1]
```

```
#Funciona

sum(is.na(d2cops[,7]))
sum(is.na(d2cops[,6]))
for (i in 1:nrow(d2cops)) {
    if (is.na(d2cops[i,7]) & d2cops$IFOCMP[i]==d2cops$IFOCMP[i+1]){
        d2cops[i,7]<-d2cops[i+1,7]
        d2cops[i,6]<-d2cops[i+1,6]
    }
}

sum(is.na(d2cops[,7]))
sum(is.na(d2cops[,6]))

sum(is.na(d2cops$SFO11))
for (i in 1:nrow(d2cops)) {
    if (is.na(d2cops$SFO11[i]) & d2cops$IFOCMP[i]==d2cops$IFOCMP[i+1]){
        d2cops$SFO11[i]<-d2cops$SFO11[i+1]
    }
}
sum(is.na(d2cops$SFO11))


nrow(d2cops)
nrow(dd)
ddn2<-dd[!(dd$IFOCMP %in% cops2),]
nrow(ddn2)
dd2<-rbind(ddn2,d2cops)
nrow(dd2)
dd2<-dd2[order(dd2$IFOSTAT,decreasing = T),]
dd2<-dd2[order(dd2$IFOCMP),]


#S'eliminen les empreses duplicades conservant les seleccionades

length(unique(dd2$IFOCMP))
table(duplicated(dd2$IFOCMP))
dd1<-dd2[!duplicated(dd2$IFOCMP),]
nrow(dd1)


bdades<-na.omit(dclean)

dcleantot<-dcleantot[!duplicated(dcleantot$IFOCMP),]


dclean8$IFOSTAT<-as.factor(dclean8$IFOSTAT)
```

```
dclean8$SFO018<-as.factor(dclean8$SFO018)
dclean8$SFO011<-as.factor(dclean8$SFO011)
dclean8$SFO006<-as.factor(dclean8$SFO006)
dclean8$SFO005<-as.factor(dclean8$SFO005)
dclean8$SFO014<-as.numeric(dclean8$SFO014)
dclean8$SFO007<-as.numeric(dclean8$SFO007)
dclean8<-dclean8[dclean8$SFO014<1000,]
dclean8$SFO012<-as.numeric(dclean8$SFO012)
nrow(dclean8)
dclean8$antic<-(dclean8$IFOYEAR-dclean8$SFO007)

str(dcleantot)
dcleantot$IFOSTAT<-as.factor(dcleantot$IFOSTAT)
dcleantot$SFO018<-as.factor(dcleantot$SFO018)
dcleantot$SFO011<-as.factor(dcleantot$SFO011)
dcleantot$SFO006<-as.factor(dcleantot$SFO006)
dcleantot$SFO005<-as.factor(dcleantot$SFO005)
dcleantot$SFO014<-as.numeric(dcleantot$SFO014)
dcleantot$SFO007<-as.numeric(dcleantot$SFO007)
dcleantot<-dcleantot[dcleantot$SFO014<1000,]
dcleantot$SFO012<-as.numeric(dcleantot$SFO012)
dcleantot$SFO019<-as.numeric(dcleantot$SFO019)
dcleantot$SFO021<-as.numeric(dcleantot$SFO021)
dcleantot$SFO022<-as.numeric(dcleantot$SFO022)
dcleantot$SFO026<-as.factor(dcleantot$SFO026)
nrow(dcleantot)
dcleantot$antic<-(dcleantot$IFOYEAR-dcleantot$SFO007)


dclean2<-na.omit(dcleantot)
nrow(dclean2)



#Base de dades amb el SFO019
dclean19<-dcleantot[,-(15:17)]
dclean19<-na.omit(dclean19)
nrow(dclean19)
table(dclean19$SFO019)

#Base de dades amb el SFO026
dclean26<-dcleantot[,-(14:16)]
dclean26<-na.omit(dclean26)
nrow(dclean26)
table(dclean26$SFO026)

#Base de dades amb el SFO021
dclean21<-dcleantot[,-14]
dclean21<-dclean21[,-(15:16)]
```

```
dclean21<-na.omit(dclean21)
nrow(dclean21)

#Base de dades amb el SFO022
dclean22<-dcleantot[,-17]
dclean22<-dclean22[,-(14:15)]
head(dclean22)
dclean22<-na.omit(dclean22)
nrow(dclean22)

#Base de dades amb el SFO019 i SFO022
dclean2219<-dcleantot[,-17]
dclean2219<-dclean2219[,-15]
dclean2219<-na.omit(dclean2219)
nrow(dclean2219)



#Creem les dos bases de dades mostrals i extramostrals pel model 1

set.seed(12)
total1<-nrow(dclean8)
n1<-round(total1*0.8)
tot1<-1:total1
muestra1 <- sample(1:total1, n1, replace= F)
dmostra1 <- as.data.frame(dclean8[muestra1,])
extra1<-as.vector(1:total1%in% muestra1)
estram1<-tot1[!extra1]
dextramostra1 <- as.data.frame(dclean8[estram1,])
nrow(dmostra1)
nrow(dextramostra1)
total1


#Creem el model m1
m1<-glm(IFOSTAT~SFO011+SFO014+SFO012+SFO018+antic,family = binomial,
dmostra1)



#Creem les dos bases de dades amb la mostra i les dades extramostrals pel
model 2 amb la variable SFO019

set.seed(13)
total2<-nrow(dclean19)
n2<-round(total2*0.8)
tot2<-1:total2
muestra2 <- sample(1:total2, n2, replace= F)
dmostra2 <- as.data.frame(dclean19[muestra2,])
```

```
extra2<-as.vector(1:total2%in% muestra2)
estram2<-tot2[!extra2]
dextramostra2 <- as.data.frame(dclean19[estram2,])
nrow(dmostra2)
nrow(dextramostra2)
total2


#Creem el model m2
m2<-glm(IFOSTAT~SFO011+SFO014+SFO012+SFO018+antic+SFO019,family = binomial,
dmostra2)
summary(m2)

#Creem les dos bases de dades amb la mostra i les dades extramostrals pel
model 3 amb la variable SFO021
set.seed(15)
total3<-nrow(dclean21)
n3<-round(total3*0.8)
tot3<-1:total3
muestra3 <- sample(1:total3, n3, replace= F)
dmostra3 <- as.data.frame(dclean21[muestra3,])
extra3<-as.vector(1:total3%in% muestra3)
estram3<-tot3[!extra3]
dextramostra3 <- as.data.frame(dclean21[estram3,])
nrow(dmostra3)
nrow(dextramostra3)
total3

#Creem el model m3
m3<-glm(IFOSTAT~SFO011+SFO014+SFO012+SFO018+antic+SFO021,family = binomial,
dmostra3)
summary(m3)

#Creem les dos bases de dades amb la mostra i les dades extramostrals pel
model 4 amb la variable SFO022
set.seed(17)
total4<-nrow(dclean22)
n4<-round(total4*0.8)
tot4<-1:total4
muestra4 <- sample(1:total4, n4, replace= F)
dmostra4 <- as.data.frame(dclean22[muestra4,])
extra4<-as.vector(1:total4%in% muestra4)
estram4<-tot4[!extra4]
dextramostra4 <- as.data.frame(dclean22[estram4,])
nrow(dmostra4)
nrow(dextramostra4)
total4

#Creem el model m4
```

```
m4<-glm(IFOSTAT~SFO011+SFO014+SFO012+SFO018+antic+SFO022,family = binomial,
dmostra4)
summary(m4)

#Creem les dos bases de dades amb la mostra i les dades extramostrals pel
model 5 amb la variable SFO026
set.seed(45)
total5<-nrow(dclean26)
n5<-round(total5*0.8)
tot5<-1:total5
muestra5 <- sample(1:total5, n5, replace= F)
dmostra5 <- as.data.frame(dclean26[muestra5,])
extra5<-as.vector(1:total5%in% muestra5)
estram5<-tot5[!extra5]
dextramostra5 <- as.data.frame(dclean26[estram5,])
nrow(dmostra5)
nrow(dextramostra5)
total5

#Creem el model 5

m5<-glm(IFOSTAT~SFO011+SFO014+SFO012+SFO018+antic+SFO026,family = binomial,
dmostra5) #camp antiguetat
summary(m5)


#Creem les dos bases de dades amb la mostra i les dades extramostrals pel
model 6 amb les variables econòmiques
set.seed(23)
total6<-nrow(dclean2)
n6<-round(total6*0.8)
tot6<-1:total6
muestra6 <- sample(1:total6, n6, replace= F)
dmostra6 <- as.data.frame(dclean2[muestra6,])
extra6<-as.vector(1:total6%in% muestra6)
estram6<-tot6[!extra6]
dextramostra6 <- as.data.frame(dclean2[estram6,])
nrow(dmostra6)
nrow(dextramostra6)
total6

#Creem el model m6
m6<-
glm(IFOSTAT~SFO011+SFO014+SFO012+SFO018+antic+SFO026+SFO022+SFO019+SFO021,f
amily = binomial, dmostra6)
summary(m6)
nrow(dmostra6)

set.seed(43)
```

```
total7<-nrow(dclean2219)
n7<-round(total7*0.8)
tot7<-1:total7
muestra7 <- sample(1:total7, n7, replace= F)
dmostra7 <- as.data.frame(dclean2219[muestra7,])
extra7<-as.vector(1:total7%in% muestra7)
estram7<-tot7[!extra7]
dextramostra7 <- as.data.frame(dclean2219[estram7,])
nrow(dmostra7)
nrow(dextramostra7)

m7<-glm(IFOSTAT~SFO011+SFO014+SFO012+SFO018+antic+SFO022+SFO019,family =
binomial, dmostra7)
summary(m7)

#Taules selecció I
data.frame(AIC(m1,m2,m3,m4,m5,m6,m7),BIC(m1,m2,m3,m4,m5,m6,m7),"loglik"=c(l
ogLik(m1),logLik(m2),logLik(m3),logLik(m4),logLik(m5),logLik(m6),logLik(m7)
))



#S'observa que els millors models són el m4, m6 i m7

#Selecció de models II

step(m4) #step model 4
#model 4.1
m4.1<-glm(formula = IFOSTAT ~ SFO011 + SFO012, family = binomial, data =
dmostra4)
summary(m4.1)

step(m6)#step model 6
#model 6.1
m6.1<-glm(formula = IFOSTAT ~ SFO011 + SFO014 + SFO012 + antic,
    family = binomial, data = dmostra6)
summary(m6.1)

step(m7)#step model 7
#model 7.1
m7.1<-glm(formula = IFOSTAT ~ SFO011 + SFO014 + SFO019, family = binomial,
data = dmostra7)
summary(m7.1)

#Taules selecció models II
data.frame(AIC(m4.1,m6.1,m7.1),BIC(m4.1,m6.1,m7.1),"loglik"=c(logLik(m4.1),
logLik(m6.1),logLik(m7.1)))
```

```
#Selecció de models III
m6.1l<-glm(formula = IFOSTAT ~ SFO011 + SFO014 + SFO012 + antic,
    family = binomial(link = "probit"), data = dmostra6)
summary(m6.1l)
m6.1cl<-glm(formula = IFOSTAT ~ SFO011 + SFO014 + SFO012 + antic,
    family = binomial(link = "cloglog"), data = dmostra6)
summary(m6.1cl)
summary(m6)

data.frame(AIC(m6.1,m6.1l,m6.1cl),BIC(m6.1,m6.1l,m6.1cl),"loglik"=c(logLik(
m6.1),logLik(m6.1l),logLik(m6.1cl)))

mfinal<-m6.1
summary(mfinal)
residualPlots(mfinal)
mfinal0<-glm(formula = IFOSTAT ~ 1, family = binomial(link = "logit"),
    data = dmostra6)
anova(mfinal0,mfinal,test = "Chisq")
Anova(mfinal,test.statistic="LR")


#Significacions globals


m10<-glm(formula = IFOSTAT ~ 1, family = binomial(link = "probit"),
    data = dmostra1)
anova(m10,m1,test = "Chisq")

m20<-glm(formula = IFOSTAT ~ 1, family = binomial(link = "probit"),
    data = dmostra2)
anova(m20,m2,test = "Chisq")

m30<-glm(formula = IFOSTAT ~ 1, family = binomial(link = "probit"),
    data = dmostra3)
anova(m30,m3,test = "Chisq")

m40<-glm(formula = IFOSTAT ~ 1, family = binomial(link = "probit"),
    data = dmostra4)
anova(m40,m4.1,test = "Chisq")

m50<-glm(formula = IFOSTAT ~ 1, family = binomial(link = "probit"),
    data = dmostra5)
anova(m50,m5,test = "Chisq")

m60<-glm(formula = IFOSTAT ~ 1, family = binomial(link = "probit"),
    data = dmostra6)
anova(m60,m6.1,test = "Chisq")

m70<-glm(formula = IFOSTAT ~ 1, family = binomial(link = "probit"),
```

```
    data = dmostra7)
anova(m70,m7.1,test = "Chisq")

#Capacitat predictora model final

d811<-dclean8[!(dclean8$SFO011==5),]

prob.vot <- predict(mfinal,d811,ty="response")
pres.est<- ifelse(prob.vot<0.25,0,1)
t <- table(pres.est,d811$IFOSTAT)
t

Encert <- sum(diag(t))/sum(t)
Sensibilitat <- t[2,2]/sum(t[,2])
Especifitat <- t[1,1]/sum(t[,1])
Pred.positiu <- t[2,2]/sum(t[2,])
Pred.negatiu <- t[1,1]/sum(t[1,])

Indicadors <-
data.frame(Encert,Sensibilitat,Especifitat,Pred.positiu,Pred.negatiu)
Indicadors

prob.vot6.1 <- predict(mfinal,dextramostra6,ty="response")
pres.est6.1 <- ifelse(prob.vot6.1<0.25,0,1)
t6.1 <- table(pres.est6.1,dextramostra6$IFOSTAT)
t6.1

Encert6.1 <- sum(diag(t6.1))/sum(t6.1)
Sensibilitat6.1 <- t6.1[2,2]/sum(t6.1[,2])
Especifitat6.1 <- t6.1[1,1]/sum(t6.1[,1])
Pred.positiu6.1 <- t6.1[2,2]/sum(t6.1[2,])
Pred.negatiu6.1 <- t6.1[1,1]/sum(t6.1[1,])

Indicadors6.1 <-
data.frame(Encert6.1,Sensibilitat6.1,Especifitat6.1,Pred.positiu6.1,Pred.ne
gatiu6.1)
Indicadors6.1

#coeficients model final
exp(coef(mfinal))
```