TREBALL DE FI DE MÀSTER

# Màster Interuniversitari en Estadística i Investigació Operativa UPC-UB

**Títol: Models per a temps de supervivència discrets**

**Autor: Ana Vázquez Fariñas**

**Directora: Olga Julià de Ferran**
**Departament de probabilitat, lògica i estadística, UB**

**Co-directora: Anna Espinal Berenguer**
**Servei d'Estadística Aplicada, UAB**

**Universitat: Universitat Politècnica de Catalunya – Universitat de Barcelona**

**Convocatòria**: Juny 2015



UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

UNIVERSITAT DE BARCELONA

Universitat Politècnica de Catalunya
Facultat de Matemàtiques i Estadística

Treball de fi de Màster

# Models per a temps de supervivència discrets

Ana Vázquez Fariñas

Directora: Olga Julià de Ferran i Anna Espinal Berenguer

Departament de probabilitat, lògica i estadística, UB
Servei d'Estadística Aplicada, UAB

A Sergio, als meus pares Alfonso i Ana, i també a Jose i Antonia.

# Agraïments

En primer lloc, vull agrair a les meves directores, la Dra. Olga Julià i la Dra. Anna Espinal, per tot el que m'han aportat tant a nivell acadèmic com a nivell personal. Agrair també la seva dedicació a aquest treball, tot el suport i ajuda facilitada així com tots els suggeriments fets al llarg d'aquest treball. Moltes gràcies per la vostra amabilitat.

Vull agrair també a la resta dels meus companys del Servei d'Estadística Aplicada (Llorenç, Anabel, Oliver i Ester) per les aportacions realitzades a aquest treball.

Finalment, agrair a la meva familia i amics tot el suport donat durant aquest període així com a en Sergio per ajudar-me i animar-me a continuar amb el meu aprenentatge.

# Resum

**Paraules clau:** Models de resposta binària, Model de Riscos Proporcionals, Versemblança Parcial

**MSC2000:** 62J12, 62N01, 62N02, 62N03

En l'anàlisi de la supervivència l'objectiu és analitzar el temps fins un esdeveniment d'interès. Normalment aquest temps s'assumeix que prové d'una variable aleatòria contínua. No obstant, pot ser que, per arrodoniment o falta de precisió en les mesures, tinguem un temps mesurat en una escala discreta. El fet que el temps sigui una variable discreta fa que es puguin produir empats, és a dir, que hi hagi més d'un individu amb el mateix temps. Aquest fet té conseqüències a l'hora d'utilitzar la Versemblança Parcial (PL) del model de Cox, ja que les observacions empatades en principi no se sap com ordenar-les.
En aquest treball es comparen diferents models per estimar l'efecte de les covariants quan el temps és discret: El model de Cox amb diferents metodologies per tractar els empats (les aproximacions Breslow i Efron, i les metodologies Discrete i Exact) i els models de resposta binària amb link *lògit* o *clog-log*.
Les diferents aproximacions i metodologies per la PL del model de Cox, defineixen de manera diferent com tractar l'ordre dels individus empatats, i a més, tenen unes suposicions diferents: la metodologia Discrete té la suposició que el temps realment és discret, mentre que la metodologia Exact té la suposició que la variable temps prové de la discretització d'una variable contínua.
D'altre banda, donat que el temps es considera una variable aleatòria discreta la funció de risc és una probabilitat. És per això que també es poden utilitzar models de resposta binària. Per poder-los aplicar però, s'ha de construir la base de dades expandida, que és una redefinició de la base de dades original, on cada individu es desplega amb tants registres com conjunts de risc participa. A més en aquesta base de dades expandida hi ha un indicador per cada temps on es produeixi alguna mort, i una variable binària relacionada amb el temps i l'indicador de censura. Els avantatges d'utilitzar aquests models és que ofereixen la possibilitat de fer servir tota la metodologia i software pels models de resposta binària, com és el cas de la regressió logística.
S'ha realitzat una comparació d'aquestes metodologies així com una justificació de les equivalències entre elles. A més s'han dut a terme estimacions de l'efecte de les covariants d'aquests models mitjançant dades simulades i mitjançant dades de l'àmbit veterinari.
S'ha obtingut que la metodologia Discrete i el model de resposta binària amb link *logit* donen estimacions concordants així com la metodologia Exact i el model de resposta binària amb link *clog-log*. Una de les aportacions més importants, és que en el cas de la metodologia Discrete, el terme $e^{\beta}$ no es pot interpretar com un HR, que seria l'habitual en cas que la variable temps fos contínua, sinó que correspon a l'OR. Per poder calcular el HR, seria adequat utilitzar els models de resposta binària, ja que permeten obtenir estimacions del HR en cada instant de temps.

# Abstract

**Keywords:** Models for binary response, Proportional Hazards model, Partial Likelihood

**MSC2000:** 62J12, 62N01, 62N02, 62N03

In survival analysis the main goal is to analyze the time until an event of interest. Usually this time is assumed that comes from a continuous random variable. However, you may have rounded values or lack of precision in the measurements, giving a sample of discrete times. Because this time is measured in a discrete scale, ties could be present. So one individual may have the same value than other. This has implications when using the Partial Likelihood (PL) for fitting a Cox model because you do not know how to sort tied observations.
This study compares different models to estimate the effect of covariates when time is discrete: the Cox model with different methods to deal with ties (Breslow and Efron approaches and methodologies Exact and Discrete) and models for a binary response with link *logit* or *clog-log*.
The different approaches and methodologies used for dealing with tied data in the PL are based on several sorting criteria. Even though for a real discrete times is not possible sorting the tied values. Each of the methodology has different assumptions: Discrete methodology assumes that time is really a random discrete variable; Exact methodology assumes that time variable comes from the discretization of a continuous variable; Breslow and Efron approaches are corrections of the PL for no ties.
When time is a discrete random variable then the risk function becomes a probability. This is the reason because you can use also models for a binary response. However to apply them, you need an extended dataset, coming from the original data. In this extended dataset each individual has as many records as risk sets is involved. There is also a set of indicators for each time in which a death occurs. A binary variable related to the time and the censoring indicator is also defined. A conditional probability that this binary variable equals 1 is equivalent to the risk function of the original time.
The advantages of using these models is that standard software usually allows fitting models for a binary response, as is the case of logistic regression.
These study includes methodological hints as a proof of the equivalence between some methodologies for deal with ties in a Cox model and the models for a binary response with

links *logit* and *clog-log*. A simulation study for showing that depending on the model, the magnitude $e^{\beta}$ can not be interpreted as a HR. In fact for the Discrete methodology and Logit model it is an $OR$. To calculate the HR would be easier using models for a binary response, because you only have to inverse the link. Finally, these methodologies has been applied in a veterinary study, where the goal is analyzing the time to PMWS. This time was really measured in a discrete scale (weeks). Moreover in the model has been also introduced time varying covariates.

# Índex general

# Capítol 1
# Conceptes bàsics de l'anàlisi de supervivència

El temps fins a un esdeveniment habitualment està mesurat en una escala contínua. Per diverses raons però ens podem trobar temps mesurats en escala discreta, ja sigui perquè la variable és de naturalesa discreta (nombre de vegades fins que...) o bé perquè és el resultat d'un arrodoniment o agrupació en intervals del temps. Aquesta peculiaritat implica l'ús de metodologies específiques de l'anàlisi de la supervivència per a temps definits per variables aleatòries discretes.

## 1.1. Introducció

Sigui $T$ una variable aleatòria no negativa definida com el temps fins a l'esdeveniment d'interès $\xi$. Aquest succés $\xi$ pot ser la mort, l'aparició d'un tumor, el desenvolupament d'una malaltia...
Una dificultat de l'anàlisi de la supervivència és que sovint la informació sobre la supervivència dels individus és incompleta (el temps exacte fins que es produeix l'esdeveniment d'interès no s'observa), ja sigui perquè $\xi$ succeeix abans que l'individu entri en l'estudi, o bé perquè quan finalitza l'estudi, $\xi$ encara no ha succeït o perquè es coneix que $\xi$ ha succeït en un interval de temps. Aquesta característica particular de l'anàlisi de la supervivència s'anomena **censura**. Segons quin cas s'hagi produït, dels que s'han comentat, la censura és diferent:

- Si per un individu l'esdeveniment d'interès encara no ha succeït al final del període de l'estudi, l'observació es diu que està **censurada per la dreta**.
- Si per un individu l'esdeveniment d'interès ha succeït abans d'entrar a l'estudi, l'observació es diu que està **censurada per l'esquerra**.
- Si per un individu l'esdeveniment d'interès no es pot observar exactament i només es coneix que ha succeït en un cert interval de temps, l'observació es diu que està **censurada en un interval**.

Aquest treball només considerarà censura per la dreta i no informativa, és a dir, que el coneixement del temps de censura d'un individu no proporciona més informació

sobre la supervivència futura de l'individu que la que s'obtindria si aquest hagués continuat en l'estudi.

Les funcions rellevants en l'anàlisi de la supervivència són:

- La funció de supervivència, $S(t)$
- La funció de distribució, $F(t)$
- La funció de densitat, $f(t)$
- La funció de risc, $h(t)$
- La funció de risc acumulat, $H(t)$

Aquestes funcions serveixen per il·lustrar diferents aspectes de la v.a $T$ i si en coneixem una, les altres queden determinades de manera unívoca.

En aquest capítol es presenten les definicions d'aquestes funcions en els dos casos: quan $T$ és una variable aleatòria absolutament contínua i quan és discreta. Per comoditat, en algunes interpretacions, suposarem que l'esdeveniment d'interès $\xi$ és la mort.

## 1.2. Conceptes bàsics per temps absolutament continus

Suposem que la v.a $T$ és absolutament contínua i no negativa amb densitat $f(t)$.

#### Funció de supervivència

La funció bàsica per poder descriure els fenòmens de temps fins a l'esdeveniment d'interès $\xi$ és la **funció de supervivència,** $S(t)$, que es defineix com la probabilitat que un individu sobrevisqui més enllà d'un temps $t$, és a dir, experimentar l'esdeveniment després de $t$. Es defineix com

$$S(t) = P(T > t) \tag{1.1}$$

$S(t)$ pot adoptar diferents formes, però sempre $S(0) = 1$, decreix monòtonament i convergeix a 0 quan $t \to \infty$.

La funció de supervivència es pot definir a partir de la funció de distribució o a partir de la funció de densitat:

- A partir de la funció de distribució, $F(t)$:

$$S(t) = 1 - P(T \leq t) = 1 - F(t)$$

- A partir de la funció de densitat, $f(t)$:

$$S(t) = P(T > t) = \int_t^\infty f(t)dt \longleftrightarrow f(t) = -\frac{d}{dt}S(t)$$

**Funció de risc**

Una altra manera d'estudiar la distribució de $T$ és mitjançant la **funció de risc,** $h(t)$**,** que descriu com es comporta la taxa de morir en un interval petit donat que la persona està viva a l'inici d'aquest interval.
Es defineix com:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t | T \geq t)}{\Delta t} \tag{1.2}$$

$h(t)$ és una funció no negativa, que expressa com el risc de patir $\xi$ va canviant amb el temps, i ens aporta la mateixa informació que la funció de supervivència, però ho fa en termes de velocitat o taxa.

De (1.2), es pot interpretar que $h(t)\Delta t$ és la probabilitat que un individu amb un temps com a mínim de $t$ experimenti l'esdeveniment $\xi$ en l'interval següent $(t, t+\Delta t]$.

La funció de risc es pot definir a partir de la funció de supervivència com:

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt}\ln(S(t))$$

La funció de risc pot prendre diversos patrons. Algun d'ells són:

- Risc creixent: poblacions que envelleixen amb l'edat.
- Risc decreixent: poblacions que s'enforteixen amb el temps.
- Risc constant: poblacions que no envelleixen ni s'enforteixen.
- Risc amb forma de banyera: poblacions amb funció de risc a l'inici decreixent, després constant durant un període de temps i finalment creixent.
- Risc amb forma de gepa: poblacions amb funció de risc creixent al principi i decreixent després.

Sovint també es pot utilitzar la **funció de risc acumulat**, $H(t)$, que es defineix com:

$$H(t) = \int_0^t h(s)ds = -\ln(S(t)) \tag{1.3}$$

Tot i que aquesta funció és molt útil teòricament i gràficament, no té una interpretació intuïtiva directa.

A partir de (1.3) es pot definir també la següent relació entre la funció de risc acumulat i la supervivència:

$$S(t) = e^{-H(t)} = \exp\left(-\int_0^t h(u)du\right)$$

## 1.3. Conceptes bàsics per a temps discrets

Suposem que la v.a $T$ és discreta, que pot prendre els valors $t_1 < t_2 < \ldots < t_k < \ldots$ amb funció de massa de probabilitat $p(t_j) = P(T = t_j)$, $\forall j \; j = 1, \ldots, k, \ldots$ on $p(t_j) > 0$ i $\sum_j p(t_j) = 1$.

### Funció de supervivència

Es defineix la **funció de supervivència** com:

$$S(t) = P(T > t) = \sum_{t_j > t} p(t_j) \tag{1.4}$$

Essent una funció no creixent esglaonada, amb $S(0) = 1$ i $\lim_{t\to\infty} S(t) = 0$.

La funció de supervivència es pot escriure com el producte de les probabilitats de supervivència condicionades en cada moment. És a dir:

$$S(t) = \prod_{t_j \leq t} P(T > t_j | T > t_{j-1}) = \prod_{t_j \leq t} \frac{S(t_j)}{S(t_{j-1})} \tag{1.5}$$

### Funció de risc

La **funció de risc** es defineix com la probabilitat que a un individu li passi l'esdeveniment $\xi$ a $t_j$ condicionada a que en $t_{j-1}$ encara no l'hi ha passat:

$$h(t_j) = P(T = t_j | T \geq t_j) = P(T = t_j | T > t_{j-1}) = \frac{p(t_j)}{S(t_{j-1})}, j = 1, 2, \ldots \tag{1.6}$$

entenent $S(t_0) = 1$.
Donat que $p(t_j) = S(t_{j-1}) - S(t_j)$ i utilitzant (1.6), es compleix que:

$$h(t_j) = \frac{p(t_j)}{S(t_{j-1})} = \frac{S(t_{j-1}) - S(t_j)}{S(t_{j-1})} = 1 - \frac{S(t_j)}{S(t_{j-1})}$$

A partir de (1.5) s'estableix la relació entre la funció de supervivència i la funció de risc:

$$S(t) = \prod_{t_j \leq t} (1 - h(t_j)) \tag{1.7}$$

La **funció de risc acumulat** es defineix com:

$$H(t) = \sum_{t_j \leq t} h(t_j) \tag{1.8}$$

S'observa que a partir d'aquesta definició la relació que hi ha entre la funció de supervivència i el risc acumulat en el cas de temps continus, $S(t) = e^{-H(t)}$, ara no es compleix. Alguns autors (Cox & Oakes, 1984), proposen una definició alternativa per $H(t)$ amb l'objectiu de mantenir aquesta relació:

$$H(t) = -\sum_{t_j \leq t} \ln(1 - h(t_j)), \tag{1.9}$$

on ara, sí que es pot provar $S(t) = e^{-H(t)}$.

Utilitzant que $\ln(1 - x) \simeq x$ per $x$ pròxims a zero, s'observa que si $h(t_j)$ és petit (proper a 0) aleshores (1.8) i (1.9) donen valors molt semblants.

## 1.4. Similituds i diferències: temps continus vs temps discrets

A partir de les definicions de $S(t)$, $h(t)$ i $H(t)$, es poden trobar certes similituds a nivell d'interpretació d'aquestes funcions pel cas de temps absolutament continus i discrets.

- La funció de supervivència no té cap peculiaritat quan el temps és discret.
- Per la funció de risc: a diferència del que s'ha obtingut per temps absolutament continus, en cas que el temps sigui una v.a discreta, $h(t)$ és una probabilitat. Aquest fet és molt important, ja que permetrà abordar les anàlisis fent servir models per a probabilitats (models de resposta binària amb transformació *logit* i *clog-log*), tal i com s'explicarà en capítols posteriors.
- Per la funció de risc acumulat: com s'ha comentat, per mantenir la similitud entre els dos casos, és convenient definir $H(t)$, per temps discrets, com (1.9).

# Capítol 2
# Model de Riscos Proporcionals (Cox, 1972)

Normalment en l'anàlisi de supervivència s'està interessat en la comparació de dos o més grups. Quan els grups són similars excepte, per exemple, pel tractament, es podrien utilitzar proves no paramètriques. Però és més habitual que els individus dels grups tinguin característiques addicionals que puguin afectar al seu resultat, com per exemple, l'edat, el gènere, el nivell socioeconòmic, etc. Aquestes variables es poden utilitzar com a covariables per explicar la variable resposta i poder identificar possibles factors de risc.

En aquest capítol es presenta el model de Riscos Proporcionals, formulat per Cox (1972) per temps continus, així com les diferents aproximacions i metodologies relacionades amb aquest model quan hi ha empats en les dades, és a dir, que hi hagi més d'un individu amb el mateix temps.

## 2.1. Notació i model de Cox

Sigui $T$ una v.a absolutament contínua no negativa que representa el temps fins a l'esdeveniment d'interès $\xi$. Es considera una mostra de grandària $n$ donada per:

- $t_i$ temps observat corresponent a l'individu $i$ amb $i = 1, \ldots, n$,
- $\delta_i$ l'indicador de censura corresponent a l'individu $i$ amb $i = 1, \ldots, n$,

$$\delta_i = \begin{cases} 1 & \text{temps complet} \\ 0 & \text{temps censurat} \end{cases}$$

- $Z_i = (Z_{1i}, Z_{2i}, \ldots, Z_{pi})$ vector de covariables o factors corresponent a l'individu $i$ amb $i = 1, \ldots, n$.

Sigui $h(t|Z)$ la funció de risc en el temps $t$ per un individu amb vector de covariables $Z$. El model de Riscos Proporcionals (Cox, 1972) estableix que:

$$h(t|Z) = h_0(t)e^{\beta' Z} = h_0(t)\exp\left(\sum_{k=1}^{p}\beta_k Z_k\right) \tag{2.1}$$

on $h_0(t)$ és la funció de risc basal, és a dir, funció de l'individu de referència i $\beta = (\beta_1, \beta_2, \ldots, \beta_p)'$ correspon al vector de paràmetres de regressió.

El model de Cox (2.1) es considera un model semiparamètric donat que inclou una part paramètrica i una altra part no paramètrica:

(1) La part paramètrica correspon a $\exp(\sum_{k=1}^{p}\beta_k Z_k)$.
(2) La part no paramètrica és la funció de risc basal $h_0(t)$. Aquesta correspon a una funció no especificada que s'estima un cop s'ha realitzat l'estimació dels paràmetres $\beta$.

El model de Cox té com a objectiu estimar l'efecte de les covariants $\beta = (\beta_1, \beta_2, \ldots, \beta_p)'$ mitjançant la maximització de la **funció de versemblança parcial**, que es presentarà en les següents seccions. El model de Cox és un **model de Riscos Proporcionals**, ja que si agafem dos individus amb covariables $Z_i$ i $Z_j$, la relació entre les seves funcions de risc és:

$$\frac{h(t|Z_{ki})}{h(t|Z_{kj})} = \frac{h_0(t)\exp(\sum_{k=1}^{p}\beta_k Z_{ki})}{h_0(t)\exp(\sum_{k=1}^{p}\beta_k Z_{kj})} = \exp\left(\sum_{k=1}^{p}\beta_k(Z_{ki} - Z_{kj})\right) \tag{2.2}$$

on s'observa que (2.2), anomenat **Hazard Ratio**, no depèn del temps sinó només dels predictors i de $\beta$.

## 2.2. Funció de versemblança parcial

En el model de Cox els paràmetres s'estimen maximitzant el logaritme de la **funció de versemblança parcial (Partial likelihood, PL)**. Per definir-la cal suposar que:

- Hi ha $r$ temps de mort diferents i que no hi ha empats. Per tant hi ha $n - r$ observacions censurades.
- $t_{(1)}, t_{(2)}, \ldots, t_{(r)}$ són els temps de mort ordenats.
- $R_j = R_{t_{(j)}}$ correspon al conjunt d'individus a risc en el temps $t_{(j)}$, és a dir, tots els individus que tenen un temps més gran o igual a $t_{(j)}$.

La probabilitat que l'individu $m$ amb covariants $Z_m$ mori a $t_m$, donat que hi ha un individu de $R_m$ que mor en aquest instant de temps, s'escriu com:

$$P(\text{l'individu } m \text{ mori a } t_m | \text{ una mort a } t_m, \text{ el conjunt de risc és } R_m) =$$

$$= \frac{P(\text{l'individu m mori a } t_m | \text{ sobreviscut fins a } t_m)}{P(\text{una mort a } t_m | \text{el conjunt de risc és } R_m)} = \frac{h(t_m|Z_{(m)})}{\sum_{l\in R(t_m)} h(t_m|Z_l)} =$$

$$= \frac{h_0(t_m)\exp(\beta' Z_{(m)})}{\sum_{l\in R(t_m)} h_0(t_m)\exp(\beta' Z_l)} = \frac{\exp(\beta' Z_{(m)})}{\sum_{l\in R(t_m)} \exp(\beta' Z_l)}$$

La funció de versemblança parcial s'obté multiplicant aquestes probabilitats condicionades per cada $m$, pel que s'acaba obtenint:

$$PL(\beta_1,\ldots,\beta_p) = \prod_{m=1}^{r} L_m = \prod_{m=1}^{r} \frac{\exp(\sum_{k=1}^{p}\beta_k Z_{(m)k})}{\sum_{l\in R(t_{(m)})}\exp(\sum_{k=1}^{p}\beta_k Z_{lk})} \tag{2.3}$$

on s'observa que el numerador només depèn de la informació dels individus que s'han mort, en canvi, en el denominador s'utilitza la informació de tots els individus a risc (incloent individus que després seran censurats).

La PL es tracta com una funció de versemblança, de manera que per fer inferència s'utilitza de la manera habitual, per tant, per estimar l'efecte de les covariants s'ha de calcular el logaritme de la PL:

$$\ln(PL(\beta)) = \sum_{m=1}^{r}\sum_{k=1}^{p}\exp(\beta_k Z_m k) - \sum_{m=1}^{r}\ln\left[\sum_{l\in R(t_{(m)})}\exp\left(\sum_{k=1}^{p}\beta_k Z_{lk}\right)\right] \tag{2.4}$$

Les estimacions dels paràmetres s'obtindrien maximitzant (2.3) o equivalentment (2.4).

## 2.3. Estimació dels paràmetres i interpretació en el model de Cox

En aquest model s'hi poden incloure covariables qualitatives (sexe, estadi de l'enfermetat, etc.) o quantitatives (pressió arterial, edat, etc.).
Les variables independents que són conegudes a l'inici de l'estudi s'anomenen **covariables basals fixes**, en canvi si les covariables poden canviar després de l'inici de l'estudi s'anomenen **covariables canviants en el temps**.

Donat el model de Cox definit a (2.1) i les estimacions dels paràmetres maximitzant (2.4), en el cas de dos individus $i$ i $j$ que només es diferencien en la k-èssima covariable, que suposem qualitativa i pren el valor 0 per $i$ i 1 per $j$, aleshores el $HR = e^{\beta_k}$ per qualsevol temps $t$; representa la raó de riscos entre un grup i l'altre

de la variable $Z_k$. Si la variable qualitativa pot prendre $s$ valors, aleshores es necessiten $s-1$ variables *dummies*, que indiquen a quin grup pertany i són mútuament excloents.

Utilitzant la normalitat assimptòtica dels estimadors de màxima versemblança, l'interval de confiança per l'$HR = e^{\beta_k}$ es calcula com:

$$\exp\left(\hat{\beta}_k \pm z_{\frac{\alpha}{2}}\sqrt{\widehat{var}(\hat{\beta}_k)}\right)$$

Si $Z_k$ és una covariable quantitativa el $\widehat{HR} = e^{\hat{\beta}_k}$, representa la raó mitjana de riscos al augmentar en una unitat la covariable $Z_k$. També pot ser d'interès estimar la raó de riscos al incrementar $Z_k$ en $c$ unitats. D'aquesta manera es tindrà $\widehat{HR} = e^{c\hat{\beta}_k}$. En aquest últim cas l'interval de confiança de l'$HR$ és:

$$\exp\left(c\hat{\beta}_k \pm z_{\frac{\alpha}{2}}\sqrt{\widehat{var}(c\hat{\beta}_k)}\right)$$

## 2.4. Model de Cox quan hi ha empats en el temps

La funció de versemblança parcial presentada en l'apartat anterior, està pensada per temps absolutament continus, però a la pràctica ens podem trobar amb valors del temps agrupats o bé discrets, el que podria donar lloc a valors empatats. Aquest fet fa que (2.3) no sigui apropiada, ja que la PL es basa en l'ordenació dels temps de mort, i en cas d'empats s'ha de buscar una manera de poder "ordernar-los".
En aquesta secció es presenten diferents aproximacions i metodologies per estimar els paràmetres del model de Cox amb temps empatats. Per fer-ho, es necessita la notació següent:

- $t_1 < t_2 < \ldots < t_r$ els diferents temps de mort,
- $Z_i = (Z_{1i}, \ldots, Z_{pi})$ el vector de covariables per l'individu $i$ amb $i = 1, \ldots n$,
- $D_m$ conjunt d'individus amb temps no censurat igual a $t_m$,
- $d_m$ nombre d'individus amb temps no censurats igual a $t_m$, és a dir, el cardinal de $D_m$,
- $s_m = \sum_{i:t_i=t_m} Z_i$ és la suma de les covariables de tots els individus que moren a $t_m$,
- $n_j$ nombre d'individus a risc a $t_j$.

Quan hi ha empats en el temps, la PL es construeix com el producte de les següents probabilitats, per cada temps complet:

$$\begin{aligned} &P(\text{morin els } D_m \text{ individus a } t_m | \text{ hi ha } d_m \text{ morts en } t_m, \text{ el conjunt de risc és } R_m) = \\ &= \frac{P(\text{morin els } D_m \text{ individus a } t_m | \text{ els } D_m \text{ individus estan en risc a } t_m)}{P(d_m \text{ morts a } t_m | \text{el conjunt de risc és } R_m)} \end{aligned} \tag{2.5}$$

Les diferents alternatives per definir aquestes probabilitats es presenten a continuació.

### 2.4.1. Breslow

Aquesta aproximació, va ser proposada per Breslow (1974), on a partir de les probabilitats (2.5) la PL es defineix com:

$$\prod_{m=1}^{r} \frac{\exp(\beta' s_m)}{[\sum_{l \in R(t_{(m)})} \exp(\beta' Z_l)]^{d_m}} \tag{2.6}$$

Cal notar que:

- Tots els individus que moren a $t_m$ tenen el mateix denominador.
- Si per cada temps complet el nombre de morts $d_j$ és petit i/o el nombre d'individus a risc $n_j$ és gran (i per tant $\frac{d_j}{n_j}$ és petit), aleshores aquesta és una bona aproximació (Zhang (2005)).
- No obstant, si aquestes condicions no es compleixen, aquesta aproximació pot ser poc fiable. Per aquest motiu, Efron (1977) va suggerir una altra aproximació.

### 2.4.2. Efron

Aquesta aproximació va ser proposada per Efron (1977). En la construcció de la PL en cada temps el denominador va disminuint proporcionalment, de manera que la PL queda definida com:

$$\prod_{m=1}^{r} \frac{\exp(\beta' s_m)}{\prod_{k=1}^{d_m} [\sum_{l \in R(t_m)} \exp(\beta' Z_l) - \frac{k-1}{d_m} \sum_{l \in D(t_m)} \exp(\beta' Z_l)]} \tag{2.7}$$

on $D(t_m) = \{j : t_j = t_m\}$

- S'observa que els $d_m$ individus que moren a $t_m$ contribueixen amb diferents pesos.
- Tant aquest tractament d'empats com el de Breslow no presenten cap problema computacionalment.

### 2.4.3. Discrete

Aquest mètode va ser proposat per Cox (1972). No assumeix que hi pugui haver un ordenament subjacent dels temps de supervivència empatats, sinó que els empats que s'observen són empats "verdaders", és a dir, les morts sí succeeixen en el mateix temps i no es pot dir que una passi abans que l'altra.

Tenint en compte aquest fet, en el denominador de les probabilitats amb les que es construeix la PL,

$$P(\text{morin els } D_m \text{ individus a } t_m| \text{ hi ha } d_m \text{ morts en } t_m, \text{ el conjunt de risc és } R_m)$$

es té en compte tots els possibles subconjunts (sense reemplaçament) de $d_m$ individus que es poden fer dins del conjunt de risc.
Per a les probabilitats $\pi_{it} = P(\text{individu i es mori a t}|\text{ sobreviscut fins a t})$ s'assumeix un model de resposta binària amb link *logit* (amb intercepts canviats en el temps):

$$\ln\left(\frac{\pi_{it}}{1-\pi_{it}}\right) = \alpha_t + \beta Z_i \Longrightarrow \pi_{it} = \frac{e^{\alpha_t+\beta Z_i}}{1+e^{\alpha_t+\beta Z_i}} \text{ i } 1-\pi_{it} = \frac{1}{1+e^{\alpha_t+\beta Z_i}}$$

Cal notar que $\pi_{it}$ i $(1-\pi_{it})$ tenen el mateix denominador.

Substituïnt $\pi_{it_m}$ i $(1-\pi_{it_m})$ a les probabilitats de la PL s'obté:

$$\frac{\frac{\exp(\beta^{'} s_m)}{\prod_{j\in R(t_m)}(1+e^{\alpha_{t_m}+\beta Z_j})}}{\sum_{l\in R_{d_m}(t_m)} \frac{\exp(\beta' s_l)}{\prod_{j\in R(t_m)}(1+e^{\alpha_{t_m}+\beta Z_j})}} = \frac{\exp(\beta^{'} s_m)}{\sum_{l\in R_{d_m}(t_m)} \exp(\beta' s_l)}$$

on $R_{d_m}(t_m)$ és el conjunt de tots els subconjunts de $d_m$ individus escollits sense reemplaçament del conjunt de risc $R(t_m)$.

Per tant, la PL amb aquesta metodologia es defineix com:

$$\prod_{m=1}^{r} \frac{\exp(\beta^{'} s_m)}{\sum_{l\in R_{d_m}(t_m)} \exp(\beta' s_l)} \tag{2.8}$$

S'observa que:

- En el producte (2.8) el terme m-èssim representa la probabilitat condicionada d'observar les morts dels individus del conjunt $D_m$ donat que hi ha $d_m$ morts en $t_m$ i el conjunt de risc és $R(t_m)$.
- El nombre de termes del denominador és $\binom{n_m}{d_m}$, que serà gran si $d_m$ i $n_m$ són grans.
- Donat que en el denominador hi ha tots els possibles subconjunts de $d_m$ individus dins $R_m$, es pot donar una interpretació de $\frac{\text{casos favorables}}{\text{casos possibles}}$.
- A nivell computacional, aquesta aproximació pot trigar molt de temps si en algun instant de temps hi ha un nombre elevat d'empats.

**OBSERVACIÓ:** el nom d'aquest mètode és el que utilitza el software SAS en el `proc phreg`. Altres autors (Therneau & Grambsch, 2000) anomenen aquest tractament d'empats com Exact Partial Likelihood (EPL).

### 2.4.4. Exact

Aquest mètode va ser proposat per Kalbfleisch & Prentice (1980). Assumeix que el temps de supervivència té una distribució contínua, i per tant, els temps empatats, en realitat són diferents i tan sols és degut a que la mesura del temps no té la suficient precisió o per arrodoniment.
Al no tenir coneixement de l'ordre dels temps dels individus empatats de $D_m$, es considera la suma per totes les possibles permutacions d'aquests individus, i per tant, en cada terme, el conjunt de risc varia segons la permutació.

Amb això, la funció de versemblança parcial amb aquest mètode s'expressa de la forma:

$$\prod_{j=1}^{r} \sum_{P \in Q_j} \frac{\exp(\beta' s_j)}{\prod_{r=1}^{d_j} \left[ \sum_{l \in R(t_{(j)}, p_r)} \exp(\beta' Z_l) \right]}$$

on $Q_j$ és el conjunt de permutacions de $i_1, i_2, \ldots, i_{d_j}$, $P = (p_1, p_2, \ldots, p_{d_j})$ és un element de $Q_j$, i $R(t_{(j)}, p_r)$ és el conjunt de risc corresponent a la permutació $Q_j$ quan els individus $p_1, p_2, \ldots, p_r$ ja no hi són.

- Els resultats obtinguts amb aquest mètode són similars als que s'obtenen amb Efron, tal i com es veurà més endavant.
- A nivell computacional, quan hi ha molts empats, donat que la versemblança parcial creix de manera molt rápida en el nombre de termes, pot requerir molt temps de computació.

**OBSERVACIÓ:** el nom d'aquest mètode és el que utilitza el software SAS en el `proc phreg`. Altres autors (Therneau & Grambsch, 2000) anomenen aquest tractament d'empats com Average likelihood (AL).

## 2.5. Software

Els tractaments d'empats per l'estimació dels paràmetres d'un model de Cox es poden realitzar des de diferents softwares. A continuació es presenten una relació de les funcions o procediments per implementar aquests mètodes amb SAS i R:

| Tractament d'empats | Software | Funció/procediment | Opció |
|---|---|---|---|
| Breslow | R | `coxph` | `ties`="breslow" |
| | SAS | `proc phreg` | `ties`=Breslow |
| Efron | R | `coxph` | `ties`="efron" |
| | SAS | `proc phreg` | `ties`=Efron |
| Discrete | R | `coxph` | `ties`="exact" |
| | SAS | `proc phreg` | `ties`=discrete |
| Exact | R | | |
| | SAS | `proc phreg` | `ties`=exact |

TAULA 2.1. Software

Tal i com es mostra a la taula 2.1, en R la funció bàsica per estimar els paràmetres en el model de Cox per a temps amb empats, és la funció `coxph` indicant a l'opció `ties` quin mètode d'empats es desitja realitzar. El mètode que realitza per defecte és Efron i el mètode Exact no està implementat. Pel que fa al SAS, tots els mètodes d'empats es realitzen amb el procediment `proc phreg`, on en l'opció `ties` s'indica el mètode d'empats que es vol realitzar.

## 2.6. Exemple

Com a il·lustració de tot el que s'ha explicat en aquest capítol es calcularà la corresponent funció de versemblança parcial, segons el cas. Per simplificar considerarem una única covariant $Z$.

Siguin les dades donades per:

| id | $T$ | $\delta$ | Z |
|---|---|---|---|
| 1 | $t_1$ | 1 | $z_1$ |
| 2 | $t_2$ | 1 | $z_2$ |
| 3 | $t_3$ | 0 | $z_3$ |
| 4 | $t_4$ | 1 | $z_4$ |
| 5 | $t_5$ | 1 | $z_5$ |

TAULA 2.2. Dades

on suposem que $t_1 = t_2 < t_3 < t_4 < t_5$. Per tant, els dos primers individus tenen temps empatats.
Donat que hi ha 3 temps complets diferents $(t_1, t_4, t_5)$ implica que la PL tindrà 3 termes:

$$L(\beta) = L_1(\beta)L_4(\beta)L_5(\beta)$$

on $L_m(\beta)$ és la component de la PL corresponent al $m$-èssim temps de supervivència complet. Com que els temps $t_4$ i $t_5$ corresponen a temps complets diferents, els termes $L_4(\beta)$ i $L_5(\beta)$ es construirien com en el cas sense empats, i serà sempre igual per totes les propostes d'empats. Així és:

$$L_4(\beta) = \frac{\exp(z_4\beta)}{\exp(z_4\beta) + \exp(z_5\beta)} \text{ i que } L_5 = 1$$

Així ens centrarem en $L_1(\beta)$ per a cadascuna de les alternatives de la PL per temps empatats.

**Breslow:**
Tal i com s'ha definit anteriorment, la PL amb aquesta proposta es pot escriure com:

$$L(\beta) = \prod_{m=1}^{r} L_m(\beta) = \prod_{m=1}^{r} \frac{\exp(\beta' s_m)}{\sum_{l \in R_{d_m}(t_{(m)})} \exp(\beta' s_l)}$$

Per tant, el terme $L_1(\beta)$ es calcularia com:

$$L_1(\beta) = \frac{e^{z_1\beta}}{e^{z_1\beta} + e^{z_2\beta} + e^{z_3\beta} + e^{z_4\beta} + e^{z_5\beta}} \times \frac{e^{z_2\beta}}{e^{z_1\beta} + e^{z_2\beta} + e^{z_3\beta} + e^{z_4\beta} + e^{z_5\beta}} = \frac{e^{(z_1+z_2)\beta}}{\left[\sum_{l=1}^{5} e^{z_l\beta}\right]^2}$$

**Efron:**
La PL amb aquest tractament d'empats té la forma:

$$L(\beta) = \prod_{m=1}^{r} L_m(\beta) = \prod_{m=1}^{r} \frac{\exp(\beta' s_m)}{\prod_{k=1}^{d_m} [\sum_{l \in R(t_m)} \exp(\beta' Z_l) - \frac{k-1}{d_m} \sum_{l \in D(t_m)} \exp(\beta' Z_l)]}$$

on $D(t_m) = \{j : t_j = t_m\}$

Per tant, el terme $L_1(\beta)$ es calcularia com:

$$L_1(\beta) = \frac{e^{z_1\beta}}{e^{z_1\beta} + e^{z_2\beta} + e^{z_3\beta} + e^{z_4\beta} + e^{z_5\beta}} \times \frac{e^{z_2\beta}}{e^{z_1\beta} + e^{z_2\beta} + e^{z_3\beta} + e^{z_4\beta} + e^{z_5\beta} - \frac{1}{2}\left(e^{z_1\beta} + e^{z_2\beta}\right)}$$

**Discrete:**
La PL amb aquest tractament d'empats té la forma:

$$L(\beta) = \prod_{m=1}^{r} L_m(\beta) = \prod_{m=1}^{r} \frac{\exp(\beta' s_m)}{\sum_{l \in R_{d_m}(t_m)} \exp(\beta' s_l)}$$

on $R_{d_m}(t_m)$ és el conjunt de tots els subconjunts de $d_m$ individus escollits sense reemplaçament del conjunt de risc $R(t_m)$.
Com s'ha dit anteriorment aquestes $L_m$ representen la probabilitat condicionada d'observar les morts del conjunt $D_m$ donat que hi ha $d_m$ morts en $t_m$ i el conjunt de risc és $R(t_m)$. Això aplicat a l'exemple:

$$L_1(\beta) = P(\text{els individus 1 i 2 morin}| \text{ hi ha 2 morts entre els 5 individus})$$

Es pot demostrar que aquesta probabilitat és:

$$L_1(\beta) = \frac{e^{\beta(z_1)}e^{\beta(z_2)}}{e^{\beta(z_1+z_2)}+e^{\beta(z_1+z_3)}+e^{\beta(z_1+z_4)}+e^{\beta(z_1+z_5)}+e^{\beta(z_2+z_3)}+e^{\beta(z_2+z_4)}+e^{\beta(z_2+z_5)}+e^{\beta(z_3+z_4)}+e^{\beta(z_3+z_5)}+e^{\beta(z_4+z_5)}}$$

$$L_1(\beta) = \frac{e^{\beta(z_1+z_2)}}{\sum_{l \in R_{d_m}(t_{(m)})} e^{\beta s_l}}$$

**Exact:**
La PL amb aquesta metodologia té la forma:

$$\prod_{j=1}^{r} \sum_{P \in Q_j} \frac{\exp(\beta' s_j)}{\prod_{r=1}^{d_j} \left[ \sum_{l \in R(t_{(j)}, p_r)} \exp(\beta' Z_l) \right]}$$

on $Q_j$ és el conjunt de permutacions de $i_1, i_2, \ldots, i_{d_j}$, $P = (p_1, p_2, \ldots, p_{d_j})$ és un element de $Q_j$ i $R(t_{(j)}, p_r)$ és el conjunt de risc corresponent a la permutació $Q_j$ quan els individus $p_1, p_2, \ldots, p_r$ ja no hi són.

Es pot veure que $L_1(\beta)$ és:

$$L_1(\beta) = \frac{e^{z_1\beta}}{e^{z_1\beta} + e^{z_2\beta} + e^{z_3\beta} + e^{z_4\beta} + e^{z_5\beta}} \times \frac{e^{z_2\beta}}{e^{z_2\beta} + e^{z_3\beta} + e^{z_4\beta} + e^{z_5\beta}} + \\ + \frac{e^{z_2\beta}}{e^{z_2\beta} + e^{z_1\beta} + e^{z_3\beta} + e^{z_4\beta} + e^{z_5\beta}} \times \frac{e^{z_1\beta}}{e^{z_1\beta} + e^{z_3\beta} + e^{z_4\beta} + e^{z_5\beta}}$$

Les dues primeres components de $L_1$ estan suposant que primer es produeix la mort de l'individu 1 i després la del 2. Per contra, la tercera i quarta component estan suposant que primer s'ha produït la mort de l'individu 2 i després la del 1.

# Capítol 3
# Models lineals generalitzats

Els models lineals, ANOVA, ANCOVA... es basen en les següents suposicions:

- Els errors segueixen una distribució Normal i independents.
- Homocedastacitat: variància constant.
- La variable resposta es relaciona linealment amb les variables independents.

Però en moltes ocasions, pot passar que algunes d'aquestes suposicions no es compleixin, com per exemple la no normalitat de la variable resposta.
Alguns d'aquests problemes es poden solucionar mitjançant transformacions de la variable resposta (com per exemple prenent logaritmes), però no es garanteix que amb aquestes transformacions es corregeixi l'heterocedasticitat o la falta de normalitat. Altres transformacions més complexes com l'exponencial o les potències fan que sigui complicat interpretar els resultats obtinguts del model.

Una alternativa a realitzar transformacions de la variable resposta i quan no hi ha normalitat, és utilitzar els models lineals generalitzats (GLM, Nelder & Wedderburn (1972)). Aquests models són una extensió dels models lineals, que permeten utilitzar distribucions no normals dels errors i variàncies no constants.
Per exemple es pot establir un GLM quan la variable resposta és un variable binària o un recompte.

Un model lineal generalitzat té les següents components:

- **Component sistemàtica:** Especifica quines són les variables explicatives, que entren en forma d'efectes fixes en un model lineal. És a dir, les variables $Z_j$ es relacionen mitjançant $\alpha + \beta_1 Z_1 + \beta_2 Z_2 + \ldots + \beta_p Z_p$ amb la variable resposta.
- **Funció link:** Considerant $\mu = E(Y)$, essent $Y$ la variable resposta, aleshores la funció *link* s'especifica com una funció $g(\cdot)$ que relaciona la component sistemàtica amb $\mu$ de la següent manera:

$$g(\mu) = \alpha + \beta_1 Z_1 + \beta_2 Z_2 + \ldots + \beta_p Z_p$$

## 3.1. Models lineals generalitzats per dades binàries

En moltes situacions la variable resposta $Y$ té només dues categories:

$$Y = \begin{cases} 1 & \text{la característica està present,} \\ 0 & \text{la característica no està present.} \end{cases}$$

Com que $Y$ és una variable binària, un model lineal

$$Y = \alpha + \beta_1 Z_1 + \beta_2 Z_2 + \ldots + \beta_p Z_p + \epsilon$$

no es pot establir perquè no compleix les suposicions abans comentades.
En lloc d'això, es modelitza la probabilitat:

$$p = P(Y = 1|Z) = P(Y = 1|Z_1, \ldots, Z_p)$$

A partir d'aquí, es pot establir una relació lineal amb les covariants fent servir diferents transformacions.

## 3.2. Models per resposta binària amb link logit

Com que $p \in [0, 1]$, la funció logit de $p$, és a dir, $\ln\left(\frac{p}{1-p}\right) \in \mathbb{R}$, es pot posar com a combinació lineal de les variables explicatives $Z_1, \ldots, Z_p$:

$$logit(p_z) = \ln\left(\frac{p_z}{1-p_z}\right) = \beta_0 + \beta_1 Z_1 + \ldots + \beta_p Z_p$$

el que s'anomena model de regressió logística. Aquesta expressió és equivalent a:

$$p_z = \frac{e^{(\beta_0+\beta_1 Z_1+\ldots+\beta_p Z_p)}}{1 + e^{(\beta_0+\beta_1 Z_1+\ldots+\beta_p Z_p)}} \tag{3.1}$$

De (3.1) es pot extreure que $Z_i$ és un factor de risc (factor protector) de $Y$ si $\beta_i > 0$ ($\beta_i < 0$), i la variable resposta no depèn de $Z_i$ si $\beta_i = 0$.

### 3.2.1. Estimació per màxima versemblança dels paràmetres

Els paràmetres del model es poden estimar mitjançant el mètode de màxima versemblança: assumint una mostra d'observacions independents, $(Y_i, Z_j)$ $i = 1, \ldots, n$ i $j = 1, \ldots, p$.

L'expressió de la funció de versemblança és:

$$\begin{aligned} L(\beta') &= \prod_{i=1}^{n} P(Y_i|Z_i)P(Z_i) \propto \prod_{i=1}^{n} P(Y_i|Z_i) \\ &= \prod_{i=1}^{n} P(Y_i=1|Z_i)^{Y_i} P(Y_i=0|Z_i)^{1-Y_i} \\ &= \prod_{i=1}^{n} \left(\frac{e^{\beta_0+\beta' Z_i}}{1+e^{\beta_0+\beta' Z_i}}\right)^{Y_i} \left(\frac{1}{1+e^{\beta_0+\beta' Z_i}}\right)^{1-Y_i} \end{aligned}$$

on $\beta' = (\beta_1, \ldots, \beta_p)$.

A partir d'aquesta, es calcula la funció de log-versemblança:

$$\begin{aligned} l(\beta') = \ln(L(\beta')) &= \ln\left[\prod_{i=1}^{n} \left(\frac{e^{\beta_0+\beta' Z_i}}{1+e^{\beta_0+\beta' Z_i}}\right)^{Y_i} \left(\frac{1}{e^{\beta_0+\beta' Z_i}}\right)^{1-Y_i}\right] \\ &= \sum_{i=1}^{n} \left[Y_i \ln\left(\frac{e^{\beta_0+\beta' Z_i}}{1+e^{\beta_0+\beta' Z_i}}\right) + (1-Y_i)\ln\left(1-\frac{e^{\beta_0+\beta' Z_i}}{1+e^{\beta_0+\beta' Z_i}}\right)\right] \end{aligned}$$

Per trobar el valor de $\beta_0, \beta_1, \ldots, \beta_p$ que maximitza $L(\beta')$ s'han de resoldre les equacions de versemblança, que són:

$$\sum_{i=1}^{n} \left[Y_i - \frac{e^{\beta_0+\beta' Z_i}}{1+e^{\beta_0+\beta' Z_i}}\right] = 0 \tag{3.2}$$

$$\sum_{i=1}^{n} Z_i \left[Y_i - \frac{e^{\beta_0+\beta' Z_i}}{1+e^{\beta_0+\beta' Z_i}}\right] = 0 \tag{3.3}$$

Les expressions (3.2) i (3.3) no són lineals respecte $\beta_0, \beta_1, \ldots, \beta_p$, i per tant, es requereixen mètodes numèrics per trobar la seva solució.

### 3.2.2. Interpretació dels paràmetres

Una de les raons de la popularitat de la regressió logística és pel fet que no únicament tenen interpretació els paràmetres estimats $\hat{\beta}$ sinó també el terme $e^{\hat{\beta}}$.

Sigui $Z_k$ una covariable dicotòmica del model de regressió logística. L'odds ratio ($OR$) associat a $Z_k = 1$ i ajustat per la resta de covariants s'expressa com:

$$OR_{Z_k} = \frac{odds(p|Z_1, \ldots, Z_k = 1, \ldots, Z_p)}{odds(p|Z_1, \ldots, Z_k = 0, \ldots, Z_p)} = e^{\beta_k}$$

on $odds(p) = \frac{p}{1-p}$.

Aquest $OR$ representa la raó dels odds de grup $Z_k = 1$ respecte el grup $Z_k = 0$. L'estimador de l'$OR_{Z_k}$ i el seu interval de confiança és:

$$\widehat{OR}_{Z_k} = e^{\hat{\beta}_k}$$

$$IC(OR_{Z_k}) = \exp\left(\hat{\beta}_k \pm z_{\frac{\alpha}{2}}\sqrt{\widehat{var}(\hat{\beta}_k)}\right) = \widehat{OR}_{Z_k}\exp\left(\pm z_{\frac{\alpha}{2}}\sqrt{\widehat{var}(\hat{\beta}_k)}\right)$$

En cas que la variable $Z_k$ té $s$ categories, es necessiten $s-1$ variables *dummies* i s'aplicaria el mateix argument anterior.

Si $Z_k$ és una covariable quantitativa, l'$OR = e^{\beta}$ representa la raó dels odds a l'augmentar en una unitat $Z_k$ ($OR = e^{c\beta}$ correspon a la raó de l'odds a l'augmentar en $c$ unitats la covariable $Z_k$).

## 3.3. Models per resposta binària amb link clog-log

També es poden utilitzar altres tipus de transformacions, per modelar aquesta $p$ mitjançant un model lineal. L'altra transformació en la que estem interessats és la transformació *clog-log* (McCullagh & Nelder, 1991); expressa el $\ln(-\ln(1-p))$ com a combinació lineal de les variables explicatives $Z_1, \ldots, Z_p$:

$$\textit{clog-log}(p_z) = \ln\left(\ln\left(\frac{1}{1-p_z}\right)\right) = \ln(-\ln(1-p_z)) = \gamma_0 + \gamma_1 Z_1 + \ldots + \gamma_p Z_p$$

el que anomenarem model Clog-log. Aquesta expressió és equivalent a:

$$p_z = 1 - \exp(-e^{\gamma_0+\gamma_1 Z_1+\ldots+\gamma_p Z_p})$$

### 3.3.1. Estimació per màxima versemblança dels paràmetres

Els paràmetres del model poden ser estimats, igual que abans, amb el mètode de màxima versemblança: assumint una mostra d'observacions independents, $(Y_i, Z_j)$ $i = 1, \ldots, n.$ i $j = 1, \ldots, p.$
L'expressió de la funció de versemblança és:

$$\begin{aligned} L(\gamma') &= \prod_{i=1}^{n} P(Y_i|Z_i)P(Z_i) \propto \prod_{i=1}^{n} P(Y_i|Z_i) \\ &= \prod_{i=1}^{n} P(Y_i = 1|Z_i)^{Y_i} P(Y_i = 0|Z_i)^{1-Y_i} \\ &= \prod_{i=1}^{n} \left(1 - \exp(-e^{\gamma_0+\gamma' Z_i})\right)^{Y_i} \left(\exp(-e^{\gamma_0+\gamma' Z_i})\right)^{1-Y_i} \end{aligned}$$

on $\gamma^{'} = (\gamma_1, \ldots, \gamma_p)$.

A partir d'aquesta, trobem la funció de log-versemblança:

$$\begin{aligned} l(\gamma^{'}) = \ln(L(\gamma^{'})) &= \ln\left[\prod_{i=1}^{n}\left(1-\exp\left(-e^{\gamma_0+\gamma^{'}Z_i}\right)\right)^{Y_i}\left(\exp\left(-e^{\gamma_0+\gamma^{'}Z_i}\right)\right)^{1-Y_i}\right] \\ &= \sum_{i=1}^{n}\left[Y_i\ln\left(1-\exp\left(-e^{\gamma_0+\gamma^{'}Z_i}\right)\right)+(1-Y_i)\left(-e^{\gamma_0+\gamma^{'}Z_i}\right)\right] \end{aligned}$$

Per trobar el valor de $\gamma_0, \gamma_1, \ldots, \gamma_p$ que maximitza $L(\gamma^{'})$ s'han de resoldre les equacions de versemblança, que són:

$$\sum_{i=1}^{n} e^{\gamma_0+\gamma^{'}Z_i}\left(\frac{Y_i\exp(-e^{\gamma_0+\gamma^{'}Z_i})}{1-\exp(-e^{\gamma_0+\gamma^{'}Z_i})} - Y_i e^{\gamma_0+\gamma^{'}Z_i}\right) = 0 \tag{3.4}$$

$$\sum_{i=1}^{n} Z_i e^{\gamma_0+\gamma^{'}Z_i}\left(\frac{Y_i\exp(-e^{\gamma_0+\gamma^{'}Z_i})}{1-\exp(-e^{\gamma_0+\gamma^{'}Z_i})} - Y_i Z_i e^{\gamma_0+\gamma^{'}Z_i}\right) = 0 \tag{3.5}$$

Les expressions (3.4) i (3.5) no són lineals respecte $\gamma_0, \gamma_1, \ldots, \gamma_p$ i per tant, es requereixen de mètodes numèrics per trobar la seva solució.

## 3.4. Relació entre els paràmetres dels models de resposta binària

A partir dels models per resposta binària amb link *logit* i link *clog-log* es pot establir una relació entre els pàrametres d'aquests models.

Suposem que tenim una única variable explicativa $Z$, aleshores:

- Model amb link *logit*:

$$logit(p_z) = \ln\left(\frac{p_z}{1-p_z}\right) = e^{\beta_0+\beta_1 Z} \Leftrightarrow p_z = \frac{e^{\beta_0+\beta_1 Z}}{1+e^{\beta_0+\beta_1 Z}}$$

- Model amb link *clog-log*:

$$clog\text{-}log(p_z) = \ln(-\ln(1-p_z)) = \gamma_0+\gamma_1 Z \Leftrightarrow p_z = 1-\exp(-e^{\gamma_0+\gamma_1 Z})$$

Tenint en compte que en tots dos models es modelitza la mateixa $p_z$, es poden igualar les dues expressions i trobar la relació entre els paràmetres.

Per trobar la relació que hi ha entre la $\gamma_0$ del model amb link *clog-log* i els paràmetres amb link *logit*, considerem $Z = 0$ i igualem les dues expressions de $p_z$:

$$\begin{aligned}
\frac{e^{\beta_0}}{1+e^{\beta_0}} = 1-\exp(-e^{\gamma_0}) &\Leftrightarrow \\
e^{\beta_0} = 1-\exp(-e^{\gamma_0})+e^{\beta_0}-e^{\beta_0}\exp(-e^{\gamma_0}) &\Leftrightarrow \\
\exp(-e^{\gamma_0})+e^{\beta_0}\exp(-e^{\gamma_0}) = 1 &\Leftrightarrow \\
\exp(-e^{\gamma_0})[1+e^{\beta_0}] = 1 &\Leftrightarrow \\
-e^{\gamma_0}+\ln(1+e^{\beta_0}) = 0 &\Leftrightarrow \\
\gamma_0 = \ln(\ln(1+e^{\beta_0}))
\end{aligned}$$

De manera anàloga s'obtindria la següent relació entre el $\gamma_1$ del model *clog-log* i els paràmetres del model *logit*:

$$\gamma_1 = \ln\left(\ln(1+e^{\beta_0+\beta_1}\right) - \gamma_0$$

## 3.5. Software

Els models de respota binària es poden estimar des de qualsevol software. A continuació es presenten les funcions o procediments per estimar aquests models mitjançant el software SAS i R.

| Models per resposta binària | Software | Funció/procediment | Opció |
|---|---|---|---|
| Link logit | R | `glm` | `family=‘‘binomial’’` |
| | SAS | `proc logistic` | |
| Link clog-log | R | `glm` | `family=‘‘binomial’’,link=‘‘cloglog’’` |
| | SAS | `proc logistic` | `link=cloglog` |

TAULA 3.1. Software

## 3.6. Versemblança dels models de supervivència discrets i la seva relació amb els models de resposta binària

Suposem que la mostra prové d'una variable temps amb distribució discreta:

$$(t_1,\delta_1,Z_1),(t_2,\delta_2,Z_2),\ldots,(t_n,\delta_n,Z_n)$$

on com s'ha vist fins ara: les $t_i$ corresponen als temps observats, $\delta_i$ als indicadors de censura (1 si és un temps complet i 0 si és un temps censurat) i $Z_i$ les covariants de l'individu $i$.
La versemblança és proporcional a:

$$L \approx P(T_i = t_i \text{ si } \delta_i = 1 \cup T_i > t_i \text{ si } \delta_i = 0, i = 1\ldots,n | Z_i, i = 1\ldots,n) \qquad (3.6)$$

Utilitzant la independència de les observacions, la versemblança es pot escriure de forma tradicional com:

$$L \approx \prod_{i=1}^{n} \left[ P(T = t_i|Z_i)^{\delta_i} P(T > t_i|Z_i)^{(1-\delta_i)} \right] \tag{3.7}$$

Donada la definició de la funció de risc: $h(t_j|Z) = P(T = t_j|T \geq t_j, Z)$, es pot relacionar la funció de risc amb les probabilitats anteriors com:

$$P(T = t_i|Z_i) = h(t_i|Z_i) \prod_{m;t_m<t_i} (1 - h(t_m|Z_i))$$

$$P(T > t_i|Z_i) = \prod_{m;t_m\leq t_i} (1 - h(t_m|Z_i))$$

Subtituïnt a (3.7) les igualtats anteriors s'obté la versemblança en termes de la funció de risc:

$$L \approx \prod_{i=1}^{n} \left[ h(t_i|Z_i) \prod_{m;t_m<t_i} (1 - h(t_m|Z_i)) \right]^{\delta_i} \left[ \prod_{m;t_m\leq t_i} (1 - h(t_m|Z_i)) \right]^{(1-\delta_i)} \tag{3.8}$$

Es defineix $r_{im} = \delta_i \mathbb{1}(t_i = t_m)$. Subtituïnt a la funció anterior:

$$L \approx \prod_{i=1}^{n} \prod_{m;t_m\leq t_i} \left( \frac{h(t_m|Z_i)}{1 - h(t_m|Z_i)} \right)^{r_{im}} (1-h(t_m|Z_i)) = \prod_{i=1}^{n} \prod_{m;t_m\leq t_i} h(t_m|Z_i)^{r_{im}} (1-h(t_m|Z_i))^{1-r_{im}} \tag{3.9}$$

Que és la mateixa versemblança que la d'un model de regressió amb resposta binària, ja que si $r_{im} = 1$ s'obté $h(t_m|Z_i)$ i si $r_{im} = 0$ s'obté $1 - h(t_m|Z_i)$.

S'observa que:

- El nombre de termes d'aquesta funció de versemblança és $\sum_{i=1}^{n} k_i$ on $k_i = \#\{j; t_j \leq t_i\}$.
- Cada individu contribueix amb tants factors definits en funció del seu temps. L'individu amb temps més petit contribueix amb un sol factor i l'individu amb temps més gran amb el nombre màxim.
- Quan en un temps no hi ha cap mort, el risc de morir s'estimaria per 0, i per tant, el corresponent factor de la versemblança maximitza en 1.
- Aquesta funció de versemblança correspon a la que obtindriem a partir d'una mostra de variables Bernoulli independents però no identicament distribuides amb paràmetre $h_m$.
- Cal notar que $h_m = P(r_{im} = 1)$. A més $r_{im}$ es pot redefinir en funció del temps i la censura.

Aquesta versemblança s'obté de manera natural a partir de la següent base de dades:

| $T^{order}$ | T | $\delta$ | $D_1$ | $D_2$ | . | $D_r$ | $Z_1$ | . | $Z_p$ | Y |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | $t_1$ | $\delta_1$ | 1 | 0 | . | 0 | $z_{11}$ | . | $z_{1p}$ | $y_{11}$ |
| 2 | $t_2$ | $\delta_2$ | 1 | 0 | 0 | 0 | $z_{21}$ | . | $z_{2p}$ | $y_{12}$ |
| 2 | $t_2$ | $\delta_2$ | 0 | 1 | 0 | 0 | $z_{21}$ | . | $z_{2p}$ | $y_{22}$ |
| . | . | . | . | . | . | . | . | . | . | |
| . | . | . | . | . | . | . | . | . | . | |
| . | . | . | . | . | . | . | . | . | . | |
| n | $t_n$ | $\delta_n$ | 1 | 0 | 0 | 0 | $z_{n1}$ | . | $z_{np}$ | $y_{1n}$ |
| n | $t_n$ | $\delta_n$ | 0 | 1 | 0 | 0 | $z_{n1}$ | . | $z_{np}$ | $y_{2n}$ |
| n | $t_n$ | $\delta_n$ | 0 | 0 | 1 | 0 | $z_{n1}$ | . | $z_{np}$ | $y_{3n}$ |
| . | . | . | . | 0 | . | . | . | . | . | |
| n | $t_n$ | $\delta_n$ | 0 | 0 | 0 | 1 | $z_{n1}$ | . | $z_{np}$ | $y_{nn}$ |

TAULA 3.2. Base de dades

on:

- $T^{order}$ pren tants valors com temps observats diferents hi hagi. En cas de temps censurats, aquesta variable té un valor faltant (*missing*).
- $T$ i $\delta$ són, respectivament, les variables temps i indicador de censura de la base de dades original.
- $D_j = \{D_j, j = 1, \ldots, r\}$: n'hi haura tantes com temps complets diferents hi hagi. Són variables *dummies* que indiquen si l'individu $i$ està en el conjunt de risc del temps $t_j$.
- $Z_1, \ldots, Z_p$ corresponen a les covariants.
- La variable $Y$, és un v.a binària que indica si l'indiviu és viu ($Y = 0$) o mort ($Y = 1$) en cada moment, mentre l'individu estigui a risc.
- En relació amb $r_{im}$, s'observa que $r_{im} = Y$.

Quan es treballa amb aquesta base de dades, tal i com es veurà en les pròximes seccions, la variable resposta serà la v.a $Y$.

Com s'ha comentat, la versemblança obtinguda (3.9) és la d'una variable aleatòria $Y \sim Bernoulli(h)$, per tant, es poden aplicar els models de resposta binària explicats en els apartats anteriors.

- Si a la $h$ s'aplica el link *logit*, el què es modelitzarà és $\ln\left(\frac{h}{1-h}\right)$.
- Si a la $h$ s'aplica el link *clog-log*, el què es modelitzarà és $\ln(-\ln(1-h))$.

## 3.7. Aplicació a les dades de supervivència

A la PL definida a (2.3), s'observa que el numerador només depèn de la informació dels individus que han mort, en canvi en el denominador s'utilitza la informació de tots els individus que encara no han mort en aquell temps (incloent individus que després seran censurats). A partir dels conjunts de risc de cada factor de la PL es podrien classificar els individus en funció de quants conjunts de risc participen, és a dir:

- Els individus amb temps complet o censurat a $t_{(1)}$ participen en 1 conjunt de risc: $R(t_{(1)})$.
- Els individus amb temps complet o censurat a $t_{(2)}$ participen en 2 conjunts de risc: $R(t_{(1)})$ i $R(t_{(2)})$.
- ...
- Els individus amb temps complet o censurat a $t_{(r)}$ participen en r conjunts de risc: $R(t_{(1)}), R(t_{(2)}), \ldots, R(t_r)$.

Aleshores, a partir d'aquí, en les següents seccions d'aquest capítol, presentarem com redefinir la base de dades original, a partir de la informació de quants conjunts de risc participa cada individu, per tal de poder utilizar els models de resposta binària per la modelització del temps fins a l'esdeveniment.

### 3.7.1. Base de dades expandida - exemple

Per explicar com construir aquesta nova base de dades, partirem d'un exemple per després fer l'extensió al cas general.

Suposem aquesta base de dades per analitzar:

| id | T | $\delta$ | Z |
|---|---|---|---|
| 1 | 2 | 1 | 0 |
| 2 | 4 | 1 | 0 |
| 3 | 4 | 1 | 1 |
| 4 | 4 | 0 | 1 |
| 5 | 8 | 1 | 1 |

TAULA 3.3. Exemple: Base de dades

S'observa que:

- Hi ha 3 temps complets diferents (2, 4, 8) i un temps censurat.
- El primer individu participa en un únic conjunt de risc: $R(t_1) = R(t = 2)$.
- Els individus 2, 3 i 4 participen en 2 conjunts de risc: $R(t = 2)$ i $R(t = 4)$.
- L'individu 5 participa en 3 conjunts de risc: $R(t = 2)$, $R(t = 4)$ i $R(t = 8)$.

Donat que el temps és discret es poden calcular les funcions de risc empíriques a partir de les definicions del Capítol 1 com $h(t_j) = P(T = t_j | T \geq t_j)$. Amb això s'obté:

| $h(t_j) = P(T = t_j \mid T \geq t_j)$ |
|---|
| $\tilde{h}(t_1) = \frac{1}{5} = \tilde{h}(t = 2)$ |
| $\tilde{h}(t_2) = \frac{2}{4} = \tilde{h}(t = 4)$ |
| $\tilde{h}(t_3) = 1 = \tilde{h}(t = 8)$ |

TAULA 3.4. Funció de risc empírica ($\tilde{h}$)

Com s'ha comentat al principi, tenint en compte en quants conjunts de risc participa cada individu, es pot reexpressar la informació de la taula 3.3, com:

| id | $T^{order}$ | T | $\delta$ | $D_1$ | $D_2$ | $D_3$ | Z | Y |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 2 | 1 | 1 | 0 | 0 | 0 | 1 |
| 2 | 2 | 4 | 1 | 1 | 0 | 0 | 0 | 0 |
| 2 | 2 | 4 | 1 | 0 | 1 | 0 | 0 | 1 |
| 3 | 2 | 4 | 1 | 1 | 0 | 0 | 1 | 0 |
| 3 | 2 | 4 | 1 | 0 | 1 | 0 | 1 | 1 |
| 4 | . | 4 | 0 | 1 | 0 | 0 | 1 | 0 |
| 4 | . | 4 | 0 | 0 | 1 | 0 | 1 | 0 |
| 5 | 3 | 8 | 1 | 1 | 0 | 0 | 1 | 0 |
| 5 | 3 | 8 | 1 | 0 | 1 | 0 | 1 | 0 |
| 5 | 3 | 8 | 1 | 0 | 0 | 1 | 1 | 1 |

TAULA 3.5. Base de dades expandida

S'observa que en les files de color vermell hi ha la informació dels individus que participen en el primer conjunt de risc, en les de color blau, hi ha la informació dels individus que participen en el segon conjunt de risc i per últim, la informació dels individus que participen el tercer conjunt de risc es troba en les files de color verd.

Per la construcció i interpretació d'aquesta base de dades cal tenir en compte:

- A la nova base de dades, cada individu contribueix amb tants registres com conjunts de risc participa.
- La variable $T^{order}$ pren tants valors com temps observats diferents hi hagi. Els individus censurats tenen *missing* aquesta variable, donat que no participen amb un temps no censurat, però s'han de tenir en compte perquè sí participen en els conjunts de risc.
- Les variables $T$ i $\delta$ són les variables corresponents, respectivament, al temps i l'indicador de censura de la base de dades original.
- De $D_j$ n'hi haurà tantes com temps complets diferents hi hagin. Són variables *dummies* que indiquen si el temps d'aquell individu és més gran o igual que $t_j$. Per exemple l'individu 2, que correspon al segon temps complet, té en el primer registre $D_1 = 1$ donat que el seu temps és més gran o igual que el primer temps complet, i en el segon registre $D_2 = 1$ el seu temps és més gran o igual que el segon temps complet.

- La variable $Y$, és una v.a binària, que indica si l'indiviu es mor en aquell temps mentre està a risc. Per l'individu 2, que li correspon al segon temps complet, per tant, en el primer registre, la informació que tenim és que $D_1 = 1$ però no es mor en aquell temps, per tant li correspon $Y = 0$. En canvi, en el segon registre, la informació és que $D_2 = 1$ i que es mor en aquell temps ($\delta = 1$ i $T^{order} = 2$) per tant li correspon $Y = 1$. Així, en el cas d'individus censurats aquesta variable sempre prendrà el valor 0.

Quan es treballa amb la base de dades expandida, la variable resposta és la v.a $Y$ i donat que és una variable binària es podran aplicar els models de resposta binària.

Amb la base de dades original s'ha estimat la funció de risc en cada instant de temps utilitzant la seva definició: $h(t_j) = P(T = t_j | T \geq t_j)$. Amb la base de dades expandida la funció que permetria obtenir el risc de la base de dades original és $h(t_j^{order}) = P(Y = 1 | D_j = 1)$. Si s'aplica a la base de dades expandida de l'exemple:

| $h(t_j^{order}) = P(Y = 1 \mid D_j = 1)$ |
|---|
| $\tilde{h}(t_1^{order}) = P(Y = 1 \mid D_1 = 1) = \frac{1}{5}$ |
| $\tilde{h}(t_2^{order}) = P(Y = 1 \mid D_2 = 1) = \frac{2}{4}$ |
| $\tilde{h}(t_3^{order}) = P(Y = 1 \mid D_3 = 1) = \frac{1}{1}$ |

TAULA 3.6. Funció de risc empírica amb la base de dades expandida ($\tilde{h}$)

d'on s'obtenen les mateixes funcions de risc empíriques de la taula 3.4.

### 3.7.2. Base de dades expandida - generalització

A continuació es generalitza el procediment vist a la secció anterior. Sigui:

- $t_1 < t_2 < \ldots < t_r$ els diferents temps no censurats,
- $\delta_i$ l'indicador de censura per l'individu $i = 1, \ldots n$,
- $Z_i = (Z_{1i}, \ldots, Z_{pi})$ el vector de covariables per l'individu $i = 1, \ldots n$,
- $d_m$ nombre d'individus amb temps no censurats iguals a $t_m$.

Partint de la base de dades original:

| ID | T | $\delta$ | $Z_1$ | . | $Z_p$ |
|---|---|---|---|---|---|
| 1 | $t_1$ | $\delta_1$ | $z_{11}$ | . | $z_{1p}$ |
| 2 | $t_2$ | $\delta_2$ | $z_{21}$ | . | $z_{2p}$ |
| . | . | . | . | . | . |
| n | $t_n$ | $\delta_n$ | $z_{n1}$ | . | $z_{np}$ |

TAULA 3.7. Base de dades original

I seguint el raonament explicat a la secció anterior, a partir de la taula 3.7, es contrueix la base de dades expandida com:

| $T^{order}$ | T | $\delta$ | $D_1$ | $D_2$ | . | $D_r$ | $Z_1$ | . | $Z_p$ | Y |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | $t_1$ | $\delta_1$ | 1 | 0 | . | 0 | $z_{11}$ | . | $z_{1p}$ | $y_{11}$ |
| 2 | $t_2$ | $\delta_2$ | 1 | 0 | 0 | 0 | $z_{21}$ | . | $z_{2p}$ | $y_{12}$ |
| 2 | $t_2$ | $\delta_2$ | 0 | 1 | 0 | 0 | $z_{21}$ | . | $z_{2p}$ | $y_{22}$ |
| . | . | . | . | . | . | . | . | . | . | |
| . | . | . | . | . | . | . | . | . | . | |
| . | . | . | . | . | . | . | . | . | . | |
| n | $t_n$ | $\delta_n$ | 1 | 0 | 0 | 0 | $z_{n1}$ | . | $z_{np}$ | $y_{1n}$ |
| n | $t_n$ | $\delta_n$ | 0 | 1 | 0 | 0 | $z_{n1}$ | . | $z_{np}$ | $y_{2n}$ |
| n | $t_n$ | $\delta_n$ | 0 | 0 | 1 | 0 | $z_{n1}$ | . | $z_{np}$ | $y_{3n}$ |
| . | . | . | . | 0 | . | . | . | . | . | |
| n | $t_n$ | $\delta_n$ | 0 | 0 | 0 | 1 | $z_{n1}$ | . | $z_{np}$ | $y_{nn}$ |

TAULA 3.8. Base de dades expandida

Pel que s'ha vist, en cada base de dades es pot definir el risc de morir a cada moment com:

- Base de dades original: $h(t_j) = P(T = t_j | T \geq t_j)$.
- Base de dades expandida: $h(t_j^{order}) = P(Y = 1 | D_j = 1)$.

A continuació es demostra la igualtat entre les dues funcions.

Partim del risc calculat amb la base de dades expandida:

$$P(Y = 1|D_j = 1) = \frac{P(Y = 1 \cap D_j = 1)}{P(D_j = 1)} = \frac{P(T^{order} = j|D_j = 1)}{P(T^{order} \geq j|D_j = 1)} = P(T = t_j|T \geq t_j) = h(t_j)$$

i veiem que s'obté la funció de risc de la base de dades original. Això pel cas en que no es considera cap covariant.

En cas que el risc depengui d'alguna covariant, la igualtat també es manté:

$$P(Y = 1|D_j = 1, Z) = \frac{P(Y = 1 \cap D_j = 1|Z)}{P(D_j = 1|Z)} = \frac{P(T^{order} = j|Z, D_j = 1)}{P(T^{order} \geq j|Z, D_j = 1)} = \\ = P(T = t_j|T \geq t_j, Z) = h(t_j|Z)$$

### 3.7.3. Models per resposta binària - base de dades expandida

Donat que la variable resposta a la base de dades expandida és la v.a $Y \sim Bernoulli$, els models que es poden utilitzar són els models de resposta binària. En concret s'aplicaran els models presentats a l'inici d'aquest capítol, on el que es modelitzarà és: $h(t^{order}|Z) = h(t|Z)$ amb $h(t^{order}|Z) = P(Y = 1|D_j = 1, Z)$.

- Model de resposta binària amb link *logit*:

$$\ln\left(\frac{h(t^{order}|Z)}{1-h(t^{order}|Z)}\right)=\ln\left(\frac{h(t|Z)}{1-h(t|Z)}\right)=\alpha_1 D_1+\alpha_2 D_1+\cdots+\alpha_r D_r+\beta_l Z$$

- Model de resposta binària amb link *clog-log*:

$$\ln(-\ln(1-h(t^{order}|Z)))=\ln(-\ln(1-h(t|Z)))=\eta_1 D_1+\eta_2 D_1+\cdots+\eta_r D_r+\beta_{cl} Z$$

S'observa que en aquests models de resposta binària:

- Hi ha més d'un terme independent, en concret n'hi ha un per cada instant de temps, ja que, per exemple quan $D_1=1$ la resta de les $D_j$ són 0, i el que s'esta fent és avaluar els models en el primer instant de temps, i així successivament.
- Cal notar que la variable resposta $Y$ es construeix a partir de variables $D_1,\ldots,D_r$ que són variables independents però no identicament distribuides i de l'indicador de censura ($\delta$).
- Es poden estimar els riscos en cada instant de temps:
  - Pel model de resposta binària amb link *logit*, l'estimació de la funció de risc, en cada instant de temps és:

$$h(t_j^{order}|Z)=h(t_j|Z)=\frac{e^{\alpha_j+\beta_l Z}}{1+e^{\alpha_j+\beta_l Z}}$$

    per tant, la funció de risc basal (quan totes les covariants són 0), en cada instant de temps s'estimaria com:

$$h(t_j^{order}|Z=0)=h(t_j|Z=0)=\frac{e^{\alpha_j}}{1+e^{\alpha_j}}$$

  - Pel model de resposta binària amb link *clog-log*, l'estimació de la funció de risc, en cada instant de temps és:

$$h(t_j^{order}|Z)=h(t_j|Z)=1-\exp(-e^{\eta_j+\beta_{cl}Z})$$

    per tant, la funció de risc basal (quan totes les covariants són 0), en cada instant de temps s'estimaria com:

$$h(t_j^{order}|Z=0)=h(t_j|Z=0)=1-\exp(-e^{\eta_j})$$

# Capítol 4
# Relació entre el model de Cox i els models lineals generalitzats

En els capítols anteriors, s'ha vist que quan es vol analitzar dades de temps de supervivència on hi ha valors empatats, es poden analitzar via:

- El model de Riscos Proporcionals amb les diferents metodologies per tractar dades amb empats.
- El model de resposta binària amb link *logit* o *clog-log*, però prèviament s'ha de construir la base de dades expandida.

El fet d'utlitzar un model o un altre, es pot fer donat que existeix una relació entre el risc obtingut a partir de la base de dades original i l'expandida: $h(t|Z) = h(t^{order}|Z)$. En aquest capítol es mostraran algunes indicacions sobre la relació entre la PL (Versemblança Parcial) del model de Cox i la versemblança dels models amb resposta discreta.

Suposem que la mostra de temps prové d'una distribució discreta:

$$(t_1, \delta_1, Z_1), (t_2, \delta_2, Z_2), \ldots, (t_n, \delta_n, Z_n)$$

on igual que al llarg d'aquest treball: les $t_i$ corresponen als temps observats, $\delta_i$ als indicadors de censura (0 si és un temps censurat i 1 si no ho és) i $Z_i$ les covariants de l'individu $i$.
La funció de versemblança per estimar un model general pel temps de supervivència, amb vector de paràmetres $(\alpha', \beta')$ es pot escriure com:

$$L(\alpha', \beta') = \prod_{i=1}^{n} h(t_i|\alpha', \beta', Z)^{\delta_i} S(t_i|\alpha', \beta', Z)$$

Si s'assumeix un model de Riscos Proporcionals, aleshores ($h(t_i|\alpha', \beta, Z) = h_0(t_i|\alpha')e^{\beta' Z}$) i s'obté:

$$\begin{aligned} L(\alpha', \beta') &= \prod_{i=1}^{n} h(t_i|\alpha', \beta', Z)^{\delta_i} S(t_i|\alpha', \beta', Z) = \\ &= \prod_{i=1}^{n} h_0(t_i|\alpha')^{\delta_i} \exp(\beta' Z)^{\delta_i} S_0(t_i|\alpha')^{\exp(\beta' Z)} \end{aligned}$$

amb log-versemblança:

$$l(\alpha', \beta') = \ln(L(\alpha', \beta')) = \sum_{i=1}^{n} \delta_i \ln(h_0(t_i|\alpha')) + \delta_i \beta' Z + \exp(\beta' Z) \ln(S_0(t_i|\alpha'))$$

on s'observa que cal especificar la forma de $h_0(t_i|\alpha')$ per poder maximitzar aquesta funció.

Normalment quan s'assumeix un model de Riscos Proporcionals en comptes d'aquesta versemblança, es considera la funció PL definida en el Capítol 2. Tal i com s'ha comentat en el capítols anteriors, per definir-la s'utilitzen les següents probabilitats:

- Sense empats: $P$(l'individu $m$ mori a $t_m$| una mort a $t_m$, el conjunt de risc és $R_m$)
- Amb empats: $P$(morin els $D_m$ individus a $t_m$| hi han $d_m$ morts en $t_m$, el conjunt de risc és $R_m$)

Cal notar que aquestes probabilitats només depenen de l'ordre dels temps amb què als subjectes els succeix l'esdeveniment d'interès. De fet, com que el model de Cox estableix una relació entre la funció de risc i les covariants, amb la funció PL s'obté una eina que permet estimar l'efecte de les covariants tenint en compte només l'ordre en que succeixen els events i no el valor concret del temps en que succeixen. Així doncs, no és necessari especificar la forma de la funció de risc basal $h_0(t_i|\alpha')$.

Quan hi ha empats, les diferents metodologies proposades al Capítol 2 (Breslow, Efron, Discrete o Exact), utilitzen el mateix argument, de manera que les probabilitats de la funció PL només depenen de l'ordenació dels subjectes i no del valor del temps. No obstant, notem però que aquestes ordenacions es plantejen de maneres diferents segons la metodologia.
Per tant, més en general, quan l'estimació dels paràmetres de les covariants $\beta_1$,$\beta_2$,..., $\beta_p$, només depengui de l'ordre del temps, tindrà sentit plantejar un argument de PL i no de versemblança.

Aquest raonament es pot plantejar perquè el model de Cox s'expressa a l'escala del risc on, com s'ha vist, només importa l'ordre del temps i no el seu valor. No obstant, si estem a l'escala del temps aquest argument no és vàlid. Per exemple en els models de vida accelerada, on el que es modelitza és el temps, aquest argument no es podria fer servir.

## 4.1. Il·lustració

En aquest apartat es mostra un exemple per posar de manifest aquesta relació entre la versemblança i la PL quan s'assumeix un model de Cox. El que es preten és avaluar si la versemblança i la PL són iguals pel que fa a la forma o només tenen en comú el punt on maximitzen les dues funcions.

Per il·lustrar-ho s'ha generat una variable aleatòria temps amb distribució $T \sim Exp(\lambda)$, s'ha calculat la L (Versemblança) i la PL per aquesta distribució i numèricament s'han comparat les dues funcions.

**Pàrametres de la il·lustració:**

- L'efecte del grup tractament s'ha fixat en $\beta = 1$.
- Covariant binària $Z \sim Bin(n, p = 0,3)$.
- S'han considerat totes les dades completes ($\delta = 1$).
- Les grandàries mostrals utilitzades són 100, 500 i 1000.

La programació de la L, PL, gràfics i estimacions del paràmetre, que es mostren en aquesta secció, s'han realitzat mitjançant el software Maple 18.

**Versemblança d'un model amb distribució Exponencial:**

En el cas que la variable segueixi una distribució $Exp(\lambda_z)$:

$$L(\lambda) = \prod_{i=1}^{n} f(t_i|\lambda_z)^{\delta_i} S(t_i|\lambda_z)^{1-\delta_i} = \prod_{i=1}^{n} (e^{-z_i\beta} \exp(-e^{-z_i\beta} t_i))^{\delta_i} (\exp(-e^{-z_i\beta} t_i))^{1-\delta_i}$$

$$l(\beta) = \ln(L(\beta)) = \sum_{i=1}^{n} -\delta_i z_i \beta - e^{-z_i\beta} t_i$$

**Versemblança Parcial:**

Tal i com s'ha definit (2.3) la PL del model de Cox és:

$$PL(\beta_1, \ldots, \beta_p) = \prod_{m=1}^{r} L_m = \prod_{m=1}^{r} \frac{\exp(\sum_{k=1}^{p} \beta_k Z_{(m)k})}{\sum_{l \in R(t_{(m)})} \exp(\sum_{k=1}^{p} \beta_k Z_{lk})}$$

on $r$ és el nombre de dades no censurades, i en aquesta il·lustració $n = r$ per com s'han generat les dades.

**Resultats:**
A continuació es mostren les estimacions de l'efecte de la covariant, $\beta$, per les simulacions de 100, 500 i 1000, així com el valor de la L i PL en $\beta$.

| n | $\hat{\beta}_L$ | $\ln(L(\hat{\beta}))$ | $\hat{\beta}_{PL}$ | $\ln(PL(\hat{\beta}))$ |
|---|---|---|---|---|
| 100 | 1,138 | -139,54 | 1,149 | -351,26 |
| 500 | 0,920 | -681,54 | 0,853 | -2574,82 |
| 1000 | 1,05 | -1308,30 | 1,03 | -5812,96 |

TAULA 4.1. Resultats L i PL

on s'observa que les estimacions del paràmetre $\beta$, per les dues metodologies són concordants, i per tant, les dues funcions maximitzen en punts propers.

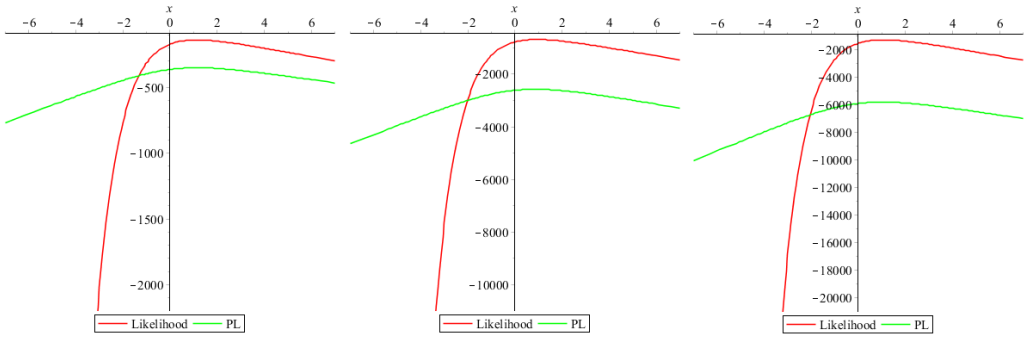


FIGURA 4.1. Esquerre a dreta: gràfics L i PL per les mides mostral 100, 500 i 1000

En els 3 gràfics anteriors, hi ha representades cadascuna de les dues funcions segons la grandària mostral. Es pot observar que són funcions amb formes diferents, però totes dues maximitzen en punts propers. Per tant, per estimar l'efecte d'una covariant, tant és utilitzar la L com la PL.

En les següents seccions d'aquest capítol, es donen algunes indicacions per justificar els resultats concordants entre els estimadors dels paràmetres pels models de resposta binària i les metodologies per a definir la PL d'un model de Cox per tractar dades amb empats.

## 4.2. Relacions per temps discrets

A continuació es mostren les diferents relacions per temps discrets.

Al capítol 3 s'ha vist que si la mostra de temps prové d'una distribució discreta:

$$(t_1, \delta_1, Z_1), (t_2, \delta_2, Z_2), \ldots, (t_n, \delta_n, Z_n)$$

on les $t_i$ corresponen als temps observats, $\delta_i$ als indicadors de censura (0 si és un temps censurat i 1 en cas contrari) i $Z_i$ les covariants de l'individu $i$, aleshores la versemblança és proporcional a:

$$L \approx \prod_{i=1}^{n} \prod_{m=1}^{t_i} \left( \frac{h_m(Z)}{1-h_m(Z)} \right)^{r_{im}} (1-h_m(Z)) = \prod_{i=1}^{n} \prod_{m=1}^{t_i} h_m(Z)^{r_{im}} (1-h_m(Z))^{1-r_{im}} \tag{4.1}$$

i donat que el risc és una probabilitat, definida per $h(t_j|Z) = P(T = t_j|T \geq t_j, Z)$, prenent una parametrització mitjançant el link *logit*, tal que

$$h(t_j|Z) = \frac{e^{\alpha_j + \beta_l Z}}{1 + e^{\alpha_j + \beta_l Z}}$$

aleshores, la versemblança que s'obté a (4.1) és la versemblança que s'obtindria a partir de la base de dades expandida, on el que es modelitza és $h(t^{order}|Z) = h(t|Z)$ i està parametritzada amb el link *logit*.

Així doncs obtenim una igualtat entre la versemblança per a una mostra d'observacions censurades o no d'una v.a discreta i la versemblança d'un model de resposta binària amb link *logit*.

### 4.2.1. Relació entre logit i la PL amb la metodologia Discrete

En el capítol 2 s'han presentat les diferents alternatives de la PL en el cas d'empats, però donat que s'està suposant que el temps és discret la més adequada és la metodologia Discrete ja que és l'única que té en compte la naturalesa discreta de les dades.

Tal i com s'ha definit, en les probabilitats que formen els factors de la PL, $\pi_{it} = P$(individu i es mori a t| sobreviscut fins a t), si s'assumeix un model amb link *logit* com s'ha vist a la Secció 2.4.3, s'obté la PL de la metodologia Discrete:

$$\prod_{m=1}^{r} \frac{\exp(\beta' s_m)}{\sum_{l \in R_{d_m}(t_m)} \exp(\beta' s_l)} \tag{4.2}$$

on $R_{d_m}(t_m)$ és el conjunt de tots els subconjunts de $d_m$ individus escollits sense reemplaçament del conjunt de risc $R(t_m)$.

Donat que la versemblança (4.1) està expressada a l'escala del risc, on només es té en compte l'ordre del temps, es pot fer un argument de PL. Si a la L i a la PL es parametritza el risc, de la mateixa manera, s'arriba a les mateixes estimacions. Aquesta és la raó per la qual no hi ha discrepàncies entre les estimacions pel model de resposta binària amb link *logit* i les obtingudes per la metodologia Discrete per estimar els paràmetres del model de Cox.

És important remarcar que en el model de resposta binària amb link *logit*, a banda de poder donar interpretació al vector de paràmetres $\beta'$, també s'interpreta el terme $e^{\beta} = OR$. Per aquest motiu, si les estimacions dels paràmetres mitjançant el model *Logit* i la metodologia Discrete són concordants, implica que el terme $e^{\beta}$ també ho serà, i per tant, no es podrà donar una interpretació d'$HR$ a la magnitud $e^{\beta}$ de la metodologia Discrete, sinó que s'haurà de fer en termes d'$OR$.

## 4.3. Relacions per temps continus agrupats

Seguint el mateix argument que a la secció 4.1.1, si a la funció de versemblança definida en (4.1), el risc es parametritza mitjançant el link *clog-log*, s'obté la mateixa versemblança que s'obtindria amb la base de dades expandida amb $h(t^{order}|Z)$ parametritzat amb el link *clog-log*.
De fet, Prentice & Gloeckler (1978), presenten una versió equivalent al model de Riscos Proporcionals, quan el temps és discret, suposant que les dades han sigut agrupades, i mostra com el link *clog-log* és el lligam natural entre la funció de risc i les covariants.

A continuació es demostra que aquest link *clog-log* surt de manera natural en aquest cas.
Sigui $U$ la variable temps contínua i $T$ la variable discreta, corresponent a l'agrupació de $U$. La variable discreta $T = t_j$ si $U$ pren algun valor en l'interval $(t_{j-1}, t_j]$ amb $(t_{j-1}, t_j]$ intervals disjunts, amb $0 = t_0 < t_1 < \ldots < t_j$.
Assumint un model de Cox per la variable temps contínua, $U$, aleshores:

$$\begin{aligned}
P(T = t_i|Z) &= P(t_{j-1} \leq U < t_j|Z) = S_U(t_{j-1}) - S_U(t_j) \\
&= \exp\left(-\int_0^{t_{j-1}} \lambda(t|Z)dt\right) - \exp\left(-\int_0^{t_j} \lambda(t|Z)dt\right) \\
&= \exp\left(-\int_0^{t_{j-1}} \lambda_0(t)e^{\beta' Z}dt\right) - \exp\left(-\int_0^{t_j} \lambda_0(t)e^{\beta' Z}dt\right) \\
&= \exp\left(-\int_0^{t_{j-1}} \lambda_0(t)e^{\beta' Z}dt\right)\left[1 - \exp(-\int_{t_{j-1}}^{t_j} \lambda_0(t)e^{\beta' Z}dt)\right] \\
&= S_U(t_{j-1}|Z)\left[1 - \exp\left(-e^{\beta' Z}\int_{t_{j-1}}^{t_j} \lambda_0(t)dt\right)\right]
\end{aligned}$$

on $S_U(t)$ és la funció de supervivència de $U$.

A partir d'aquesta expressió, la funció de risc de $T$ es pot escriure com:

$$\begin{aligned} h(t_j|Z) &= P(T=t_j|T\geq t_j,Z)=\frac{P(T=t_j|Z)}{P(T\geq t_j|Z)} \\ &= \frac{S_U(t_{j-1}|Z)\left(1-\exp\left(-e^{\beta' Z}\int_{t_{j-1}}^{t_j}\lambda_0(t)dt\right)\right)}{S_U(t_{j-1}|Z)} \\ &= 1-\exp\left(-e^{\beta' Z}\int_{t_{j-1}}^{t_j}\lambda_0(t)dt\right) \\ &= 1-\exp\left(-e^{\beta' Z+\eta_j}\right) \end{aligned} \tag{4.3}$$

amb $\eta_j=\ln\left(\int_{t_{j-1}}^{t_j}\lambda_0(t)dt\right)$

Així doncs aplicant una transformació *clog-log* s'obté:

$$\ln(-\ln(1-h(t_j|Z)))=\beta' Z+\eta_j$$

on els paràmetres $\beta'$ són exactament els paràmetres de les covariants que intervenen en la funció de risc del model de Riscos Proporcionals assumit.

### 4.3.1. Relació entre clog-log i la PL amb la metodologia Exact

En el Capítol 2 es van presentar les diferents alternatives de la PL en cas d'empats, però donat que ara s'està suposant que el temps prové de la discretització d'una variable contínua, la més adequada és la metodologia Exact. Amb aquesta metodologia la PL que s'obté és:

$$\prod_{j=1}^{k}\sum_{P\in Q_j}\frac{\exp(\beta' s_j)}{\prod_{r=1}^{d_j}\left[\sum_{l\in R(t_{(j)},p_r)}\exp(\beta' Z_l)\right]}$$

on $Q_j$ és el conjunt de permutacions de $i_1,i_2,\ldots,i_{d_j}$, $P=(p_1,p_2,\ldots,p_{d_j})$ és un element de $Q_j$ i $R(t_{(j)},p_r)$ és el conjunt de risc corresponent a la permutació $Q_j$ quan els individus $p_1,p_2,\ldots,p_r$ ja no hi són.

Igual que en el cas de la metodologia Discrete, a la versemblança (4.1) només es té en compte l'ordre dels temps i per tant es pot fer un argument de PL. Com que la metodologia Exact té la suposició al darrere que la variable temps és en realitat contínua per a la qual s'ha assumit la hipòtesi de proporcionalitat de Cox, és per això que té sentit considerar la contribució a la PL de totes les permutacions possibles dels individus empatats en cada temps (totes les possibles maneres d'ordenar els individus que moren al mateix moment), el que dóna lloc a la metodologia Exact que s'acaba d'esmentar.

Per altra banda, tal i com s'ha vist a (4.3) el link natural, per establir un model pel risc discret on apareguin els paràmetres del model de Cox és el *clog-log*. Donat que

$h(t|Z) = h(t^{order}|Z)$, el model de resposta binària que s'hauria d'aplicar, a la base de dades expandida, és el que parametritza el risc amb el link *clog-log* i en aquest cas la L i la PL són funcions del mateix paràmetre $\beta'$ i per tant les estimacions que s'obtenen són equivalents.

## 4.4. Conclusions

En els capítols anteriors, s'han explicat les diferents metologies que es poden utilitzar a l'hora d'analitzar dades amb temps de supervivència discrets; les diferents propostes del model de Cox amb tractament d'empats (Breslow, Efron, Discrete i Exact), i els models de resposta binària amb link *logit* i *clog-log*, un cop s'expandeix la base de dades original.

Donat que algunes d'aquestes metodologies tenen unes suposicions diferents sobre la naturalesa de la variable temps, en el sentit que unes assumeixen que el temps és realment discret i altres que el temps correspon a la discretització d'una variable contínua, en aquest capítol s'han vist les relacions dels diferents models en funció d'aquesta suposició. En concret s'han donat indicacions de les següents relacions:

$$
\begin{array}{ccc}
L & \xrightarrow{\text{v. discretes}} & logit \\
\| & & \| \\
PL & \xrightarrow{\text{v. discretes}} & Discrete \\
L & \xrightarrow{\text{v. contínues}} & clog-log \\
\| & & \| \\
PL & \xrightarrow{\text{v. contínues}} & Exact
\end{array}
$$

# Capítol 5
# Simulacions

Per posar de manifest les relacions explicades en el capítol anterior, entre les diferents metodologies per a temps discrets a l'hora d'estimar els paràmetres de la covariant en un model de Cox, així com pels models de resposta binària amb link *logit* i *clog-log*, es presenta un estudi de simulació. Un exemple amb dades reals, és al següent capítol.

## 5.1. Metodologia

Es defineixen dos escenaris de simulació:

- El temps realment és una v.a discreta.
- El temps prové d'una v.a contínua, però el què s'observa són els valors agrupats.

En aquest estudi un cop s'han generat les dades, s'ha establert el model de Cox per al temps. Per a l'estimació dels paràmetres s'han aplicat els diferents mètodes per a temps amb empats (Breslow, Efron, Discrete i Exact definits al Capítol 2). A més, a partir de la base de dades, s'ha generat la base de dades expandida a partir de la qual s'han establert els models de resposta binària amb link *logit* i *clog-log*. Per comoditat en aquest capítol els anomenarem *Logit* i *Clog-log*. Com a recordatori, s'observa que:

- **Base de dades original:** Model de Cox amb empats $h(t|Z) = h_0(t)e^{\beta' Z}$
- **Base de dades expandida:** Models de resposta binària

$$\begin{aligned} \text{Logit: } \ln\left(\frac{h(t|Z)}{1-h(t|Z)}\right) &= \alpha_1 D_1 + \alpha_2 D_2 + \cdots + \alpha_r D_r + \beta_l' Z' \\ \text{Clog-log: } \ln(-\ln(1-h(t|Z))) &= \eta_1 D_1 + \eta_2 D_2 + \cdots + \eta_r D_r + \beta_{cl}' Z' \end{aligned}$$

on $h(t_j|Z) = h(t_j^{order}|Z) = P(Y = 1|D_j = 1, Z')$, $r$ és el nombre de temps no censurats i les $D_j$ són variables indicadores de si l'individu té un temps més gran o igual que $t_j$.

Del fet que en els models de resposta binària s'obtinguin tant estimacions de l'efecte de la covariant $\beta_l'$ o $\beta_{cl}'$ com dels coeficients corresponents al temps $(\alpha_1, \alpha_2, \ldots, \alpha_r)$ o $(\gamma_1, \gamma_2, \ldots, \gamma_r)$ fa que es pugui calcular el risc per cada temps i per tant, també el HR en cada instant de temps, definits com:

- Pel model *Logit*:

$$HR_j = \frac{\frac{e^{\alpha_j+\beta_l}}{1+e^{\alpha_j+\beta_l}}}{\frac{e^{\alpha_j}}{1+e^{\alpha_j}}} = e^{\beta_l}\frac{1+e^{\alpha_j}}{1+e^{\alpha_j+\beta_l}} = e^{\beta_l}\frac{1-p_{tract}}{1-p_{control}} \tag{5.1}$$

  on $p_{tract} = P(Y=1|D_j=1, Z=1)$ i $p_{control} = P(Y=1|D_j=1, Z=0)$
- Pel model *Clog-log*:

$$HR_j = \frac{1-\exp(e^{\eta_j+\beta_{cl}})}{1-\exp(e^{\eta_j})} \tag{5.2}$$

Les distribucions que s'han utilitzat per la variable temps són distribucions amb risc constant:

- Pel cas discret s'ha assumit la distribució geomètrica amb paràmetre $p_z$:

$$h(T_z = t) = P(T_z = t|T_z \geq t) = \frac{p_z(1-p_z)^t}{1-\sum_{j=0}^{t-1} p_z(1-p_z)^j} = \frac{p_z(1-p_z)^t}{1-p_z[\frac{1-(1-p_z)^t}{1-(1-p_z)}]} = p_z$$

- Pel cas de dades agrupades, s'ha assumit la distribució exponencial amb paràmetre $\lambda_z$, agrupant els valors en intervals de la mateixa longitud, excepte l'últim, de forma que el risc de la variable agrupada és constant, excepte en l'últim instant de temps. En efecte, si $T$ prové de la discretització, en intervals $[t_{j-1}, t_j)$, d'una variable $U \sim Exp(\lambda)$, aleshores $h(t_j)$ és

$$h(t_j) = 1 - e^{-\lambda(t_j - t_{j-1})} \tag{5.3}$$

La generació de les dades s'ha realitzat mitjançant el software R v3.1.3 mentre que la creació de la base dades expandida i l'anàlisi s'ha realitzat mitjançant el software SAS v9.3, SAS Institute Inc., Cary, NC, USA.

## 5.2. Simulació del temps amb distribució geomètrica

Els paràmetres utilitzats en la simulació són:

- Covariant binària: $Z \sim Bern(p = 0,3)$.
- Es fixa $HR = 2$: $HR = \frac{0,6}{0,3}$.
- Temps geomètric de paràmetre $p_c$ pel grup control: $T_{control} \sim Geom(0,3)$.
- Temps geomètric de paràmetre $p_t$ pel grup tractament: $T_{tract} \sim Geom(0,6)$. Si la $T_{control} \sim Geom(0,3)$ i $T_{tract} \sim Geom(0,6)$, aleshores l'$OR = 3,5$.
- Variable temps: $T_z = T_{control}(1-Z) + T_{tract}Z$, on $T_z \sim Geom(p_z)$. A l'utilitzar la función `rgeom` (implementada en `R`) per $T_{control}$ i $T_{tract}$ cal sempre sumar 1 ja que `rgeom` dóna la distribució geomètrica que comença en el zero.
- S'ha agrupat $T_z$ en 5 categories, de manera que es tindrà 5 temps diferents.
- S'ha fixat la censura a l'últim instant de temps per simplificar el problema.

- S'han generat 100 mostres de grandàries mostrals: 100, 150, 250, 500, 1500, 2500 i 5000.

### 5.2.1. Resultats

A la taula 5.1 per la grandària mostral del 5000, es mostren els següents resultats:

- L'estimació de l'efecte de la covariant: $\hat{\beta}$.
- L'estimació de l'error estàndard de l'efecte de la covariant: $Se(\hat{\beta})$.
- El terme $e^{\hat{\beta}}$.
- L'estimació de l'error estàndard de $e^{\hat{\beta}}$: $Se(e^{\hat{\beta}})$.

Per la resta de grandàries mostrals, els resultats en format taula, es troben a l'annex.

| Mètode | $\hat{\beta}$ | $Se(\hat{\beta})$ | $e^{\hat{\beta}}$ | $Se(e^{\hat{\beta}})$ |
|---|---|---|---|---|
| Breslow | 0,693 | 0,026 | 2,000 | 0,063 |
| Efron | 0,911 | 0,035 | 2,489 | 0,088 |
| Discrete | 1,253 | 0,051 | 3,507 | 0,184 |
| Logit | 1,254 | 0,052 | 3,509 | 0,184 |
| Exact | 0,944 | 0,037 | 2,571 | 0,096 |
| Clog-log | 0,944 | 0,037 | 2,572 | 0,096 |

TAULA 5.1. Resultats estimacions geomètrica, n=5000

Com es pot observar en la Figura 5.1 (per totes les grandàries mostrals), s'observa que les metodologies Exact, Efron i el model *Clog-log* donen estimacions concordants de l'efecte de la covariant amb valors molt propers; mentre que la metodologia Discrete i el model *Logit* donen estimacions del paràmetre semblants entre ells, però diferents de l'altre bloc.

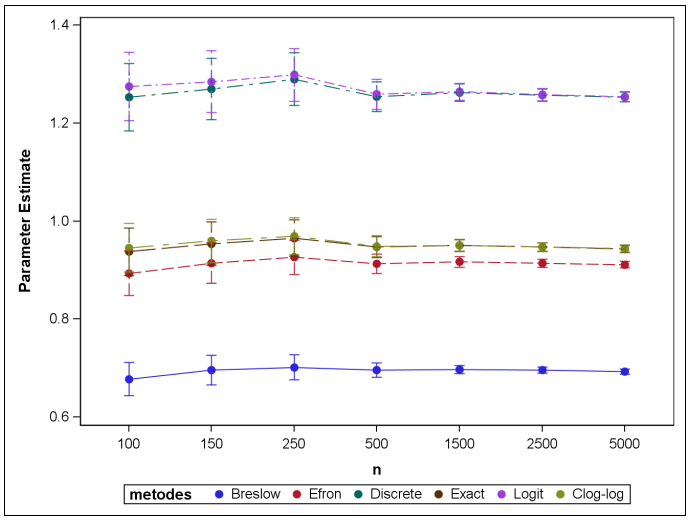


FIGURA 5.1. Resultats estimacions $\beta$ geomètrica

S'ha de tenir en compte que la variable temps segueix una distribució geomètrica de paràmetre $p_z$, que és una probabilitat, de manera que depenent de com es relacioni aquesta $p_z$ amb les covariants, donarà peu de manera natural a quin link cal escollir pels models de resposta binària.
Si es relaciona la $p_z$ amb les covariants mitjançant:

- el link *logit*, com $\ln(\frac{p_z}{1-p_z}) = \alpha_1 D_1 + \alpha_2 D_2 + \cdots + \alpha_5 D_5 + \beta_l Z$ aleshores el model de resposta binària, *Logit* i per tant, la metodologia Discrete del model de Cox per a temps amb empats, són les que millor estimarien el paràmetre.
- el link *clog-log*, com $\ln(-\ln(1-p_z)) = \eta_1 D_1 + \eta_2 D_2 + \cdots + \eta_5 D_5 + \beta_{cl} Z$ aleshores el model de resposta binària *Clog-log* i la metodologia Exact del model de Cox per a temps amb empats, són els que millor s'aproximarien al paràmetre.

Per tant, no és correcte fixar-se només en l'estimació de l'efecte de la covariant, ja que no es pot donar resposta a la pregunta de quina metodologia és més adient, donat que depenent de com es relacioni la $p_z$ amb les covariant aniran millor unes o altres. Per aquest motiu, per comparar els resultats de les 6 metodologies s'ha fet mitjançant alguna magnitud que sigui independent del link com és el $HR$, que s'ha fixat en 2.

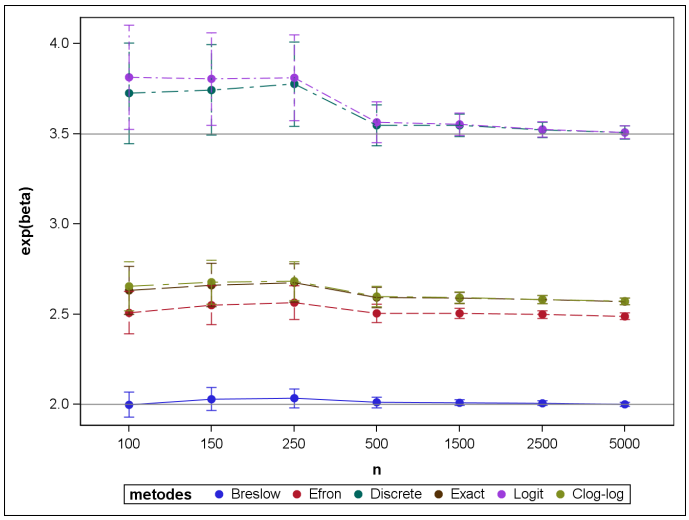


FIGURA 5.2. Resultats estimacions $e^{\beta}$

A la figura anterior igual que en les estimacions dels paràmetres (Figura 5.1) hi ha 3 blocs de resultats per a les estimacions de la magnitud $e^{\beta}$; *Logit* i Discrete per una banda, i *Clog-log*, Exact i Efron per l'altra. És important notar que l'única metodologia on realment el terme $e^{\beta}$ correspon al HR és Breslow, i es veu en el gràfic que és l'única que estima el $HR = 2$ fixat en la simulació.
Per altra banda, com s'ha comentat en el capítol anterior, pels models de resposta binària *Logit*, i per tant, per la metodologia Discrete del model de Cox el terme $e^{\beta}$ no fa referència al $HR$ sinó a l'$(OR)$. Aquest fet es pot observar en el gràfic anterior, on es veu com en el model *Logit* i Discrete, el terme $e^{\beta}$ s'aproxima, a mesura que la mida mostral augmenta, a l'$OR$ fixat en 3,5.
Pel que fa al model de resposta binària *Clog-log*, el terme $e^{\beta}$, igual que en el cas anterior, no correspon al HR, sinó que:

$$e^{\beta} = \frac{\ln(1 - p_{tract})}{\ln(1 - p_{contol})}$$

per tant, la magnitud $e^{\beta}$ no es pot interpretar com un HR en el model *Clog-log* ni en la metodologia Exact.

Aplicant les definicions (5.1, 5.2) a les dades simulades s'obtenen les següents estimacions del HR, en cada instant de temps, pels models de resposta binària *Logit* i *Clog-log* (a la taula 5.2 només es mostren els resultats per la grandària mostral del 5000, la resta es troben a l'annex).
Pels models de resposta binària, a part d'obtenir estimacions de l'efecte de la covariant, també es disposa de les estimacions del coeficients corresponents al temps. Això permet poder calcular el risc en cada instant de temps, segons cada model, tal i com s'ha comentat a l'inici del capítol.

| T | Logit | | Clog-log | |
|---|---|---|---|---|
| | $\widehat{HR}$ | $Se(\widehat{HR})$ | $\widehat{HR}$ | $Se(\widehat{HR})$ |
| 1 | 2,001 | 0,058 | 2,001 | 0,055 |
| 2 | 2,000 | 0,053 | 2,000 | 0,051 |
| 3 | 2,002 | 0,052 | 2,001 | 0,050 |
| 4 | 2,001 | 0,062 | 2,000 | 0,055 |
| 5 | 2,001 | 0,060 | 2,000 | 0,053 |

TAULA 5.2. HR estimats, n=5000

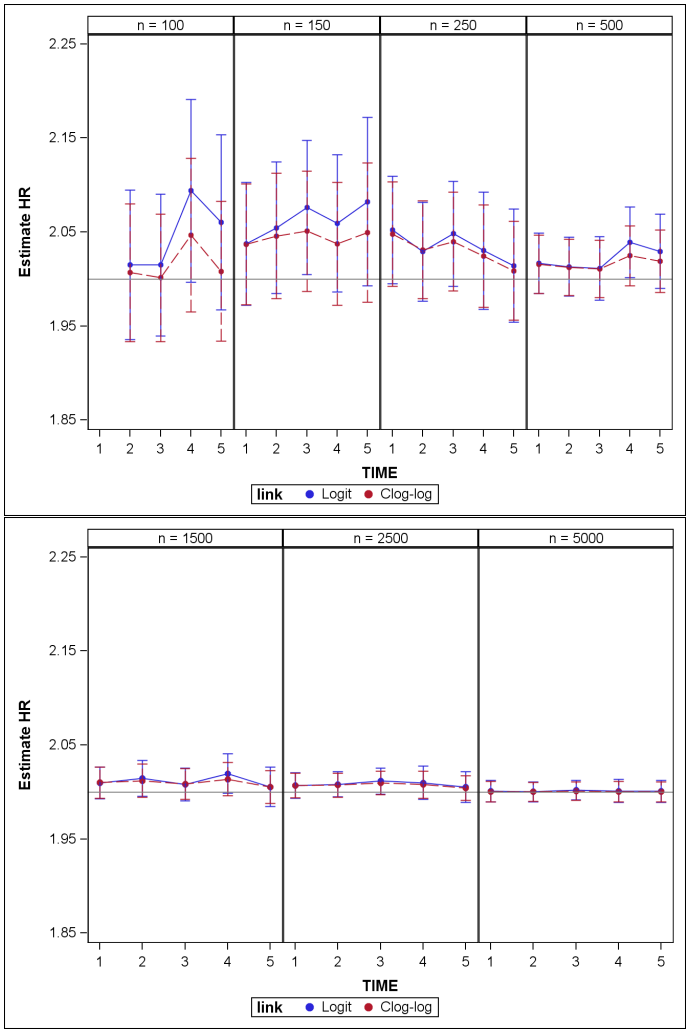


FIGURA 5.3. HR estimats pels models de resposta binària

On s'observa que pel HR, les estimacions en tots els models són molt semblants entre elles i estimen correctament el valor final d'aquesta magnitud. De fet cal recordar que com s'ha vist en el Capítol 3, a partir dels paràmetres del model *Logit* es pot

arribar als paràmetres del model *Clog-log* i per tant també s'obtenen HR semblants.

Com a conclusió, en el cas que el temps sigui realment un variable discreta:

- Els dos models de resposta binària estimen correctament els HR en cada instant de temps.
- Breslow, és l'única metodologia del model de Cox per a temps amb empats on el terme $e^{\hat{\beta}} = \widehat{HR}$.
- Amb la metodologia Discrete del model de Cox per a temps amb empats el terme $e^{\hat{\beta}}$ estima l'OR associat a la covariant.
- Amb la metodologia Exact del model de Cox per a temps amb empats, el terme $e^{\hat{\beta}}$ estima la magnitud $\frac{\ln(1-p_{tract})}{\ln(1-p_{contol})}$.

## 5.3. Simulació del temps amb distribució exponencial

Els paràmetres utilitzats en la simulació són:

- Covariant binària: $Z \sim Bern(p = 0,3)$.
- Es fixa HR=2: $HR = \frac{0,6}{0,3}$.
- Temps exponencial de paràmetre $\lambda_c$ pel grup control: $T_{control} \sim Exp(0,3)$.
- Temps exponencial de paràmetre $\lambda_t$ pel grup tractament: $T_{tract} \sim Exp(0,6)$.
- Variable temps: $T_z = T_{control}(1-Z) + T_{tract}Z$, on $T_z \sim Exp(\lambda_z)$.
- S'ha agrupat $T_z$ en 5 categories, de manera que es tindrà 5 temps diferents. La manera d'agrupar-los ha estat en intervals de longitud 0,75 excepte l'últim.
- S'ha fixat la censura en l'últim instant de temps per simplifacar el problema.
- S'ha generat 100 mostres de les següents mides mostrals: 100, 150, 250, 500, 1500, 2500 i 5000.

Tal i com s'ha calculat a (5.3) la funció de risc i el $HR$ per la variable discretitzada, amb els paràmetres de la generació s'obté:

$$h(t^{'}_{control}) = 1 - e^{-0,3(0,75)} \ , \ h(t^{'}_{exp}) = 1 - e^{-0,6(0,75)} \Longrightarrow HR = \frac{h(t^{'}_{control})}{h(t^{'}_{exp})} = 1,799$$

### 5.3.1. Resultats

A la taula 5.3 per la grandaria mostral de 5000, es mostren els següents resultats:

- L'estimació de l'efecte de la covariant: $\hat{\beta}$.
- L'estimació de l'error estàndard de l'efecte de la covariant: $Se(\hat{\beta})$.
- El terme $e^{\hat{\beta}}$.
- L'estimació de l'error estàndard de $e^{\hat{\beta}}$: $Se(e^{\hat{\beta}})$.

| Mètode | $\hat{\beta}$ | $Se(\hat{\beta})$ | $e^{\hat{\beta}}$ | $Se(e^{\hat{\beta}})$ |
|---|---|---|---|---|
| Breslow | 0,565 | 0,028 | 1,760 | 0,049 |
| Efron | 0,681 | 0,034 | 1,976 | 0,067 |
| Discrete | 0,831 | 0,043 | 2,298 | 0,098 |
| Logit | 0,831 | 0,043 | 2,299 | 0,098 |
| Exact | 0,694 | 0,035 | 2,003 | 0,070 |
| Clog-log | 0,694 | 0,035 | 2,003 | 0,070 |

TAULA 5.3. Resultats estimacions exponencial, n=5000

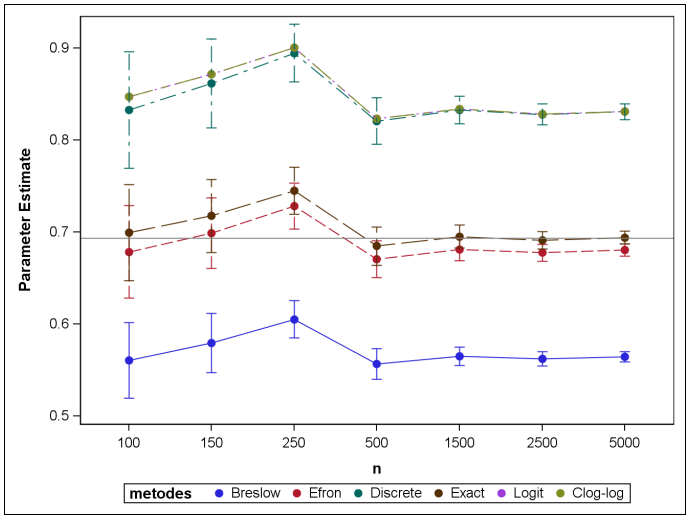


FIGURA 5.4. Resultats estimacions $\beta$ exponencial

En aquest cas, $\lambda_c = e^{\beta_0}$ i $\lambda_t = e^{\beta_0+\beta_1}$, i donat que $\lambda_c = 0,3$ i $\lambda_t = 0,6$, s'obté que l'efecte de la covariant és $\beta_1 = 0,693$. Així doncs, en aquest cas, és fàcil poder comparar el paràmetre amb les estimacions obtingudes en els 6 casos.

A la Figura 5.4, es pot veure com les estimacions del model *Logit* i la metodologia Discrete del model de Cox per temps amb empats donen estimacions concordants entre ells però diferents de les metodologies Exact, Efron i *Clog-log* que també són concordants entre ells. En aquest cas però, sí que es pot dir que són aquests 3 últims models els que estimen correctament el paràmetre corresponent a l'efecte de la covariant.

Pel que fa a les estimacions de $e^{\beta}$, que és el que hauria de ser el $HR$, s'observa, en la Figura 5.5, que en aquest cas, el terme $e^{\hat{\beta}}$ a partir del model *Clog-log* i de la metodologia Exact sí que són bones estimacions del $HR$ amb valor 2, que s'obté a partir de la variable contínua exponencial. A l'hora de trobar el $HR$ de la variable exponencial discretitzada, com que $e^{\beta} \neq HR$ s'han d'estimar els $HR$ en cada instant de temps a partir dels models de resposta binària, tal i com s'ha fet en la simulació anterior.

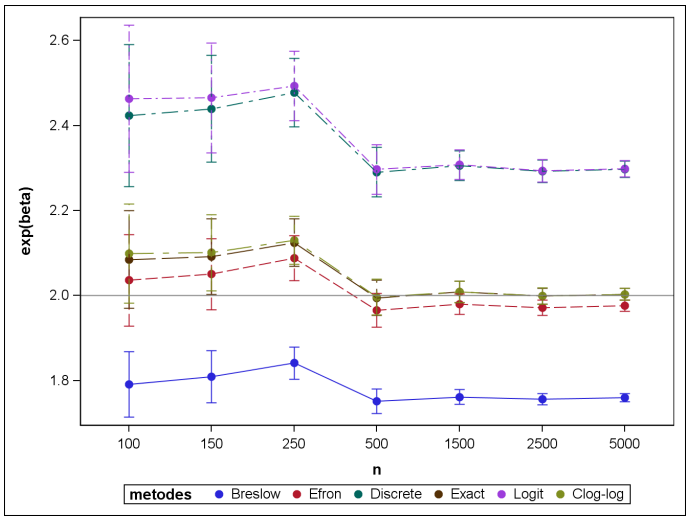


FIGURA 5.5. Resultats estimacions $e^{\beta}$

| T | Logit | | Clog-log | |
|---|---|---|---|---|
| | $\hat{HR}$ | $Se(\widehat{HR})$ | $\widehat{HR}$ | $Se(\widehat{HR})$ |
| 1 | 1,826 | 1,442 | 1,802 | 0,050 |
| 2 | 1,823 | 0,055 | 1,801 | 0,050 |
| 3 | 1,823 | 0,056 | 1,800 | 0,050 |
| 4 | 1,823 | 0,055 | 1,801 | 0,050 |
| 5 | 2,001 | 0,027 | 1,551 | 0,032 |

TAULA 5.4. HR estimats, n=5000

A la Figura 5.6, s'hi pot veure que els HR estimats pels dos models de resposta binària són bones estimacions (sense biaix i poca variabilitat) del HR per a la variable que prové de la discretització d'un temps continu. Cal observar que l'estimació en l'últim instant de temps, no és correcte, donat que en aquest valor la llargada de l'interval és diferent a la resta, ja que s'han acumulat molts més valors. Per tant, el pes d'aquest valor no és el mateix que en els anteriors, amb la qual cosa, les corresponents estimacions en aquest temps contenen un biaix pel que fa el HR.

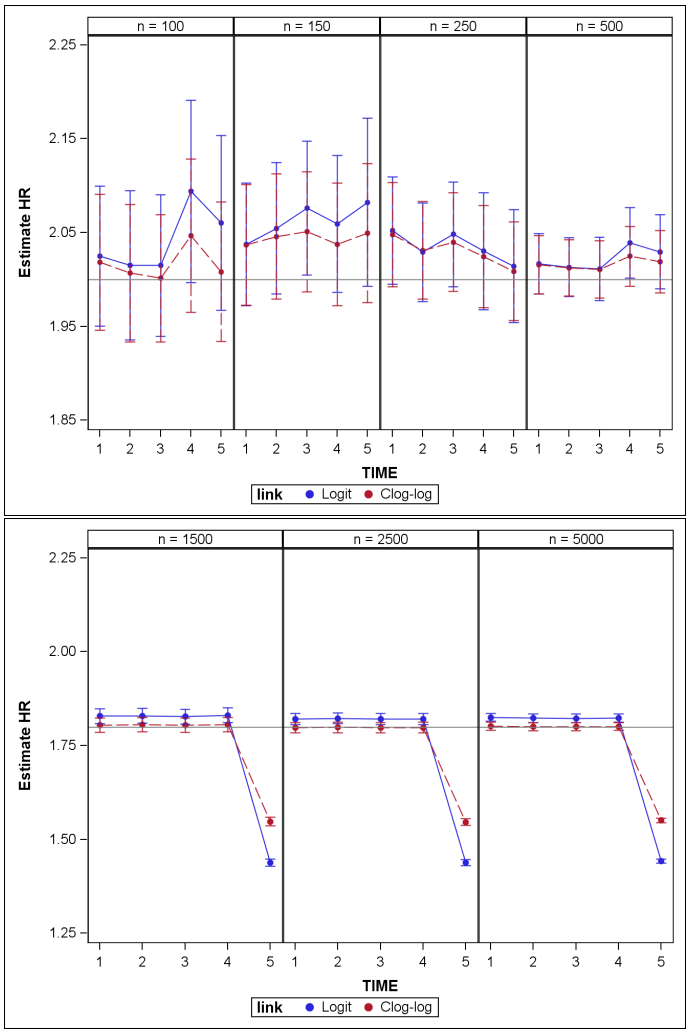


FIGURA 5.6. HR estimats pels models de resposta binària

Com a conclusió, en el cas de tenir temps discrets que provenen d'agrupar els valors d'una variable contínua:

- Els paràmetres obtinguts per les metodologies Exact i Efron del model de Cox per a temps amb empats i el model *Clog-log* donen bones estimacions de l'efecte de la covariant.
- El terme $e^{\hat{\beta}}$ obtingut amb la metodologia Exact del model de Cox per a temps amb empats i amb el model *Clog-log* corresponen a estimacions del HR de la variable contínua no de la variable discretitzada.
- Els dos models de resposta binària estimen correctament els HR en cada instant de temps.

# Capítol 6
# Aplicació a l'anàlisi del temps fins a malaltia

En aquest capítol es presenten els resultats d'aplicar les metodologies estudiades al llarg d'aquest treball.

Les dades analitzades han estat cedides pel Centre de Recerca en Sanitat Animal (CReSA) de la Universitat Autònoma de Barcelona. L'objectiu és comparar els resultats d'una anàlisi prèvia feta amb metodologia estàndard de l'anàlisi de la supervivència (Grau-Roma et al., 2012), amb els resultats que s'obtindrien si s'aplica el model adient per a temps discret.

## 6.1. Descripció de les dades i disseny de l'estudi

La Circovirosi porciona (CP) és una malaltia molt comú que afecta porcs de cria (Grau-Roma et al., 2011; Madec et al., 2008) i es considera que té un greu impacte econòmic en la producció porcina (Armstrong i Bishop, 2004).
El principal signe clínic de CP és la pèrdua de pes, però també pot incloure pal·lidesa de la pell, icterícia (coloració groguenca de la pell), o diarrea entre altres (Harding, 1998).
Els criteris per al diagnòstic de CP inclouen la presència de signes clínics compatibles, una depleció limfocitària amb infiltració granulomatosa moderada a greu, i una moderada a elevada quantitat de circovirus porcí tipus 2 (PCV2) a les lesions limfoides. El PCV2 és considerat l'agent infecciós essencial pel desenvolupament de CP. No obstant això hi ha altres factors que poden desencadenar el desenvolupament de la CP en porcs infectats amb PCV2.

Per a dur a terme l'estudi de CP es va dissenyar un protocol en el qual es van recollir dades en 6 instants de temps establerts (1, 3, 7, 11 i 15 setmanes i l'última mesura del temps correspon al moment de la necròpsia). Alguns porcs tenen un total de 5 mesures donat que l'observació de la setmana 15 ja correspon a la del moment de la necròpsia. Cal posar de manifest que quan un animal tenia signes de malaltia es sacrificava, i amb l'objectiu de caracteritzar la CP també es sacrificava

algun animal prèviament sa.

Com que l'objectiu era estudiar els animals que desenvolupaven CP, aquests es van monitoritzar fins a la detecció i confirmació diagnòstica de la malaltia a la granja. Així doncs, es va definir el temps fins a CP com a variable d'interès, inicialment aquestes dades es van analitzar utilitzant tècniques estàndard de l'anàlisi de la supervivència, on el desenvolupament de CP es va considerar com l'esdeveniment d'interès. Es va utilitzar l'edat de l'animal com la variable de temps, començant al dia 0 i acabant amb l'edat a la necròpsia. Aquesta edat es va tractar com a variable temps contínua.
Donat que només hi havia 6 moments del temps, i el temps estava agrupat en setmanes, seria més adient tractar-la com a variable discreta, i aplicar les metodologies específiques per aquest cas.

Al llarg d'aquests 6 instants de temps, es van realitzar anàlisis serològiques, per detectar factors que podien influir en el desenvolupament de CP. Els més rellevants i utilitzats a l'article són:

- Edat de la necròpsia: edat del porc a la necròpsia.
- Classificació: variable binària que indica si el porc en el moment de la necròpsia té CP o alguna altra cosa.
- Circovirus porcí tipus 2 (PCV2): les determinacions de PCV2 es van expressar com títols d'anticossos, els quals es van transformar a l'escala logarítmica.
- Parvovirus porcí (PPV): anticossos de PPV mesurats en unitats de denistat òptica (OD).
- Virus de la grip porcina (SIV): anticossos de SIV mesurats en unitats de denistat òptica (OD). També hi ha aquesta variable com a binària que pot prendre els valors positiu o negatiu, per indicar si el nivells d'anticossos supera un cert llindar.
- Salmonella (SAL): variable binària que indica si en cada instant de temps el porc està infectat o no de Salmonella.
- Síndrome respiratòria i reproductiva porcina (PRRSV): variable binària que indica si en cada instant de temps el porc té nivells superiors a un cert llindar de OD, d'aquest virus.

Els porcs en el moment de la necròpsia es van classificar en tres categories: porcs amb la malaltia d'interès (CP), porcs sans que no han desenvolupat la malaltia (Sans) i porcs que tenien signes de tenir CP però al final no es van diagnosticar com a CP en base a les anàlisis laboratorials (porcs prims, no CP). A Grau-Roma et al., 2012 només van considerar els porcs que van desenvolupar CP i els Sans, la resta es van excloure de l'estudi. Donat que un dels objectius és comparar resultats amb les anàlisis fetes, també s'aplicarà aquest criteri.

La grandària mostral era de 108 porcs, però s'analitzen els 65, que satisfan el criteri abans comentat.

En aquest estudi hi havia dues hipòtesis:

- El desenvolupament de CP pot estar influenciat per la seroconversió dels patògens descrits. La seroconversió es defineix com un increment sostingut en el temps, a partir de la setmana 11. La variable resultant acaba siguent una variable binària que indica si el porc ha seroconvertit per aquell patogen o no.
- El desenvolupament de CP pot estar influenciat pel nivell d'immunitat maternal. Aquest nivell es defineix com la mesura dels anticossos enfront els diferents patògens a les 3 setmanes de vida del porc.

Per poder donar resposta a les hipòtesis de l'estudi, a partir de les variables explicades anteriorment es defineixen les següents variables:

(1) Variables de seroconversió:
   - Seroconversió PCV2: variable indicadora de si el porc ha seroconvertit de PCV2.
   - Seroconversió PPV: variable indicadora de si el porc ha seroconvertit de PPV.
   - Seroconversió SIV: variable indicadora de si el porc ha seroconvertit de SIV.
   - Seroconversió SAL: variable indicadora de si el porc ha seroconvertit de SAL.
   - Seroconversió PRRSV: variable indicadora de si el porc ha seroconvertit de PRRSV.

(2) Variables d'immunitat maternal: correspon als valors de les següents variables a les tres setmanes:
   - PCV2: valors serològics enfront de PCV2 (en títols).
   - PPV: valors serològics enfront de PPV (en OD).
   - SIV: valors serològics enfront de SIV (en OD).
   - SAL: variable indicadora de si a les 3 setmanes tenia anticossos enfront Salmonella.
   - PRRSV: variable indicadora de si a les 3 setmanes tenia un valor serològic positiu o negatiu d'aquest.

Els objectius que es plantegen en aquest capítol són:

- Analitzar el temps fins el desenvolupament de CP utilitzant les metodologies per a dades de supervivència discretes.
- Incorporar millores i proposar futures línies d'investigació: covariables canviants en el temps o *competing risks*.

Totes les anàlisis es realitzen amb el software SAS v9.3, SAS Institute Inc., Cary, NC, USA.

## 6.2. Anàlisi descriptiva

En aquest apartat s'explora la base de dades i es descriuen totes les variables.

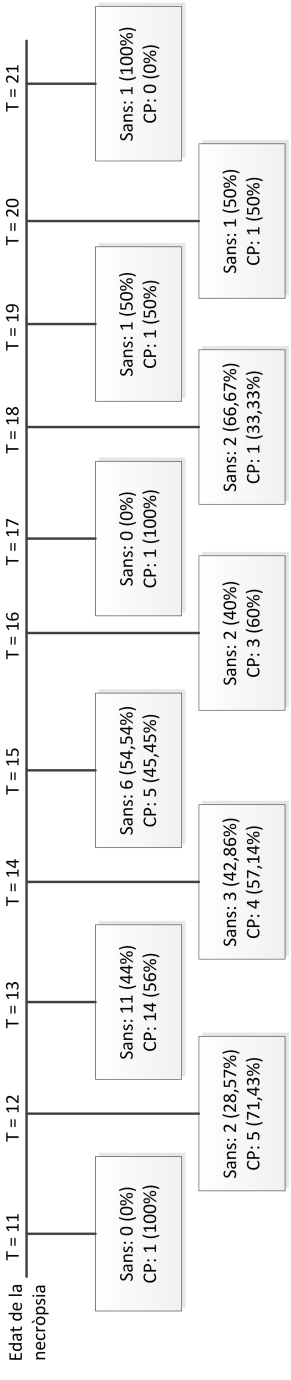


FIGURA 6.1. Distribució dels animals en funció del temps de mort

A la Figura 6.1, es mostra la distribució dels porcs sans i malalts, en cada instant de temps. Es pot veure que hi ha un total de 36 porcs que van desenvolupar la malaltia (correspon a un 55.38%) de la mostra. Pel que fa als diferents temps, la majoria d'animals es van sacrificar a les setmanes 13 i 15 (un total de 25 i 11 porcs respectivament). La mort a la setmana 11 correspon a un porc malalt, cosa que podria dir que aquest porc tenia signes molts clars de patir aquesta malaltia. Per altra banda, el porc sacrificat a l'última setmana correspon a un porc sa que es va matar perquè el període d'estudi finalitzava.

| Variable | *N* | % |
|---|---|---|
| Variable de seroconversió | | |
| Seroconversió PCV2 (Sí) | 55 | 84,62% |
| Seroconversió PPV (Sí) | 4 | 6,15% |
| Seroconversió SIV (Sí) | 36 | 55,38% |
| Seroconversió SAL (Sí) | 19 | 29,23% |
| Seroconversió PRRSV (Sí) | 1 | 1,54% |
| Variable d'immunitat maternal | | |
| PRRSV (+) | 24 | 36,92% |
| SAL (+) | 9 | 13,85% |

Taula 6.1. Descriptiva covariables qualitatives

| Variable | *N* | Mitjana | Mediana | Desv. Est. | Mínim | Màxim |
|---|---|---|---|---|---|---|
| Variables d'immunitat maternal | | | | | | |
| PCV2 | 65 | 2,814 | 2,500 | 0,712 | 1,300 | 4,300 |
| SIV | 65 | 0,793 | 0,817 | 0,260 | 0,261 | 1,231 |
| PPV | 65 | 2,757 | 2,903 | 0,826 | 0 | 3,204 |

Taula 6.2. Descriptiva covariables quantitatives

A les taules 6.1 i 6.2 hi ha l'anàlisis descriptiva de les covariables d'interès. Pel que fa a les seroconversions, un 84,62% van seroconvertir de PCV2 en algun moment del temps, un 55,38% de SIV i un 29,23% de Salmonela. Cal notar, que la seroconversió de PPV i PRRSV és menys habitual. Per a la seroconversió de SIV, és normal que hi hagi un percentatge baix de casos, donat que la serocoversió d'aquest virus sol produir-se en períodes de temps que van més enllà de les 21 setmanes (que és la setmana d'estudi). Per la seroconversió de PRRSV, no és sorprenent que hi hagi un únic cas de seroconversió atès que la majoria de les granges que van participar en l'estudi són estables enfront aquest virus.
Pel que fa a les covariants d'immunitat maternal, s'observa que un 36,92% i 13,85% tenien valors positius de PRRSV i SAL.
Per les variables quantitatives d'immunitat maternal, pel PCV2 i PPV les mitjanes d'anticossos són semblants, mentre que per la variable SIV la mitjana és inferior amb un valor de 0,793.

### 6.2.1. Anàlisi descriptiva bivariant respecte l'indicador de malaltia

Pel que fa l'anàlisis descriptiva bivariada de les diferents covariants respecte l'indicador de malaltia, destaca a la taula 6.3 que respecte les variables de seroconversió s'observa que tant si van seroconvertir com si no, el percentatge d'animals amb CP és més elevat excepte per la seroconversió de PPV, on dels 4 que van seroconvertir la meitat va desenvolupar CP.
D'altra banda, de les variables qualitatives d'immunitat maternal, la distribució d'aquestes variables està equilibrada entre sans i malalts. I per les variables d'immunitat maternal de la taula 6.4, s'observa que tant la mitjana d'anticossos enfront PCV2 com a PPV, són lleugerament superiors en el grup de sans. Això és degut a que els anticossos que li passa la mare al porc són més elevats en els porcs sans que en els malalts, és a dir, tenen més anticossos a les 3 setmanes.

| | Indicador de malaltia | |
|---|---|---|
| | Sans | CP |
| Variables de seroconversió | | |
| Seroconversió PCV2 | | |
| No | 3 (30,00%) | 7 (70,00%) |
| Sí | 26 (47,27%) | 29 (52,73%) |
| Seroconversió PPV | | |
| No | 27 (44,26%) | 34 (55,74%) |
| Sí | 2 (50,00%) | 2 (50,00%) |
| Seroconversió SIV | | |
| No | 12 (41,38%) | 17 (58,62%) |
| Sí | 17 (47,22%) | 19 (52,78%) |
| Seroconversió SAL | | |
| No | 20 (43,48%) | 26 (56,52%) |
| Sí | 9 (47,37%) | 10 (52,63%) |
| Seroconversió PRRSV | | |
| No | 29 (45,31%) | 35 (54,69%) |
| Sí | 0 (0%) | 1 (100%) |
| Variables d'immunitat maternal | | |
| PRRSV | | |
| + | 17 (41,46%) | 24 (58,54%) |
| - | 12 (50,00%) | 12 (50,00%) |
| SAL | | |
| + | 1 (11,11%) | 8 (88,89%) |
| - | 28 (50,00%) | 28 (50,00%) |

TAULA 6.3. Covariables qualitatives vs Indicador de malaltia

| | Censura | |
|---|---|---|
| | Sans | CP |
| Variables d'inmunitat maternal PCV2 | | |
| Mitjana | 2,893 | 2,750 |
| Mediana | 3,100 | 2,500 |
| Desv. Estàndard | 0,704 | 0,722 |
| Mínim | 1,300 | 1,300 |
| Màxim | 4,300 | 4,300 |
| SIV | | |
| Mitjana | 0,778 | 0,805 |
| Mediana | 0,817 | 0,826 |
| Desv. Estàndard | 0,294 | 0,232 |
| Mínim | 0,261 | 0,450 |
| Màxim | 1,206 | 1,231 |
| PPV | | |
| Mitjana | 2,794 | 2,727 |
| Mediana | 2,903 | 2,903 |
| Desv. Estàndard | 0,794 | 0,861 |
| Mínim | 0,000 | 0,000 |
| Màxim | 3,204 | 3,204 |

TAULA 6.4. Covariables quantitatives vs Indicador de malaltia

Donat que per la seroconversió de PPV i PRRSV només hi ha 4 i 1 cas respectivament, aquestes dues variables no es tindran en compte en les posteriors anàlisis.

## 6.3. Models de Cox tractant el temps com a continu

Aquests resultats són els que es van presentar en el treball Grau-Roma et al.,2012 i tan sols es descriuen perquè serveixin com a referència per comparar amb els resultats dels models obtinguts amb les metodologies proposades en aquest treball. Es presenta primer el model de Riscos Proporcionals bivariant, és a dir, analitzant l'efecte de cadascuna de les variables per separat, i després el model de Cox incloent totes les covariables conjuntament.

| Variables de seroconversió | $\hat{\beta}$ | $Se(\hat{\beta})$ | $\widehat{HR}$ | $IC(HR)$ | p.valor |
|---|---|---|---|---|---|
| Seroconversió PCV2 (Sí vs No) | -0,391 | 0,423 | 0,677 | (0,295; 1,553) | 0,357 |
| Seroconversion SIV (Sí vs No) | -0,576 | 0,342 | 0,562 | (0,288; 1,098) | 0,092 |
| Seroconversió SAL (Sí vs No) | 0,618 | 0,416 | 1,855 | (0,820; 4,196) | 0,138 |

TAULA 6.5. Resultats dels models bivariants

| Variables d'immunitat maternal | $\hat{\beta}$ | $Se(\hat{\beta})$ | $\widehat{HR}$ | $IC(HR)$ | p.valor |
|---|---|---|---|---|---|
| PCV2 | -0,701 | 0,254 | 0,496 | (0,302; 0,816) | 0,006 |
| SIV | -0,072 | 0,676 | 0,931 | (0,248; 3,500) | 0,916 |
| PPV | 0,044 | 0,182 | 1,045 | (0,731; 1,493) | 0,807 |
| PRRSV (+ vs -) | -0,754 | 0,390 | 0,470 | (0,219; 1,012) | 0,053 |
| SAL (+ vs -) | 0,752 | 0,410 | 2,121 | (0,949; 4,742) | 0,067 |

TAULA 6.6. Resultats dels models bivariants

A les taules 6.5 i 6.6, es troben els resultats dels models bivariants. De les variables corresponents a la seroconversió, només és estadísticament significativa (amb un nivell de significació del 10%) la seroconversió de SIV, amb un $\widehat{HR} = 0,562$, el que implica que el risc de patir CP és gairebé el doble en els porcs que no han seroconvertit respecte els que sí . Pel que fa a les variables d'immunitat maternal, les variables estadísticament significatives són:

- PCV2: obtenim un $\widehat{HR} = 0,496$, el que implica que a l'augmentar en mitjana una unitat els anticossos enfront PCV2 el risc de patir CP disminueix. Això vol dir que tenir més anticossos enfront PCV2 (més defenses) a les poques setmanes de néixer és un factor protector (ajuda a no desenvolupar CP).
- PRRSV (+ vs -): obtenim un $\widehat{HR} = 0,470$, el que implica que el risc de patir CP és el doble en els que tenen aquesta variable negativa respecte els que la tenen positiva.
- SAL (+ vs -): obtenim un $\widehat{HR} = 2,121$, el que implica que el risc de patir CP és el doble en els que tenen aquesta variable positiva respecte els que la tenen negativa.

| Variable | $\hat{\beta}$ | $Se(\hat{\beta})$ | $\widehat{HR}$ | $IC(HR)$ | p.valor |
|---|---|---|---|---|---|
| Variables de seroconversió | | | | | |
| Seroconversió PCV2 (Sí vs No) | -0,685 | 0,455 | 0,504 | (0,207; 1,230) | 0,132 |
| Seroconversion SIV (Sí vs No) | -0,733 | 0,422 | 0,481 | (0,210; 1,099) | 0,082 |
| Seroconversió SAL (Sí vs No) | 0,276 | 0,501 | 1,318 | (0,494; 3,517) | 0,581 |
| Variables d'immunitat maternal | | | | | |
| PCV2 | -0,691 | 0,321 | 0,501 | (0,267; 0,940) | 0,031 |
| SIV | -0,736 | 1,167 | 0,479 | (0,049; 4,713) | 0,528 |
| PPV | -0,092 | 0.247 | 0,912 | (0,562; 1,481) | 0,794 |
| PRRSV (+ vs -) | -0,242 | 0,516 | 0,785 | (0,285; 2,159) | 0,638 |
| SAL (+ vs -) | 0,531 | 0,445 | 1,701 | (0,712; 4,068) | 0,232 |

TAULA 6.7. Resultats dels model amb totes les covariants

En el model de Cox, incloent totes les variables d'interès, s'obtenen els següents resultats:

- Per les variables de seroconversió només és estadísticament significativa la Seroconversió SIV (amb un nivell de significació del 10%) on s'obté un $\widehat{HR} = 0,481$, és a dir, el risc de desenvolupar CP és el doble pels que no han seroconvertit en relació als que sí .
- Per les variables d'immunitat maternal l'única variable estadísticament significativa és els anticossos enfront PCV2, igual que en el cas bivariant. Així doncs, tenir nivells alts d'anticossos de PCV2 a les 3 setmanes fa disminuir el risc de desenvolupar CP.

Aquestes taules contenen, a més a més, la quantificació de l'efecte de les covariants mitjançant el $\widehat{HR}$ i el seu interval de confiança, resultats no inclosos a l'article on només es presenten en termes de factor protector, de risc o no significatiu. D'aquesta manera, ara a més de poder dir si és un factor protector o de risc, es pot quantificar la relació entre els riscos.

## 6.4. Models de Cox tractant el temps com a discret

En aquest apartat, s'apliquen els models més adients per temps de supervivència discrets presentats en aquest treball. Com s'ha comentat a l'inici d'aquest capítol, la variable d'interès "temps fins al desenvolupament de CP"està agrupada en setmanes, és per tant adequat utilitzar la metodologia Exact per estimar un model de Cox ja que té la suposició que el temps discret prové de l'agrupació d'una variable contínua, com seria aquest cas. Per altra banda, dels models de resposta binària presentats, seria adient utilitzar el model de resposta binària amb link *clog-log*, ja que com s'ha dit en el Capítol 4, les estimacions que s'obtenen són equivalents amb els resultats de la metodologia Exact.

Concretament, els models que s'apliquen són:

- El model de Cox amb la metodologia Exact.
- El model de resposta binària amb link *clog-log*. Pel que fa a aquest model, prèviament s'ha d'expandir la base de dades. Cal tenir en compte que en total hi ha 10 temps diferents de mort (donat que per la setmana 21, l'única dada és censurada i per aquest motiu el porc sacrificat en aquesta setmana, participarà a la base de dades expandida amb 10 registres, donat que participa en 10 conjunts de risc). De manera que la base de dades expandida tindrà 10 variables *dummies* $D_j$, i si $Z'$ correspon al vector de covariants, el model que s'aplicarà és:

$$\ln(-\ln(1-h(t|Z))) \quad = \quad \sum_{j=1}^{10} \eta_j D_j + \beta'_{cl} Z'$$

  on $Z'$ és un vector amb les mateixes covariants utilitzades en l'apartat anterior.

Els models establerts, des del punt de vista de temps discrets, inclouen les mateixes covariants que les de l'apartat anterior, de manera que hi haurà els resultats pels models bivariants i els resultats del model amb totes les covariants.

Donat que s'ha vist que la metodologia Exact i el model de resposta binària amb link *clog-log* donen estimacions concordants, els resultats presentats corresponen només a aquest últim model.

**Models bivariants:**

| Variables de seroconversió | $\hat{\beta}$ | $Se(\hat{\beta})$ | $e^{\hat{\beta}}$ | $IC(e^{\beta})$ | p.valor |
|---|---|---|---|---|---|
| Seroconversió PCV2 (Sí vs No) | -0,418 | 0,428 | 0,658 | (0,304; 1,640) | 0,329 |
| Seroconversió SIV (Sí vs No) | -0,622 | 0,344 | 0,537 | (0,274; 1,059) | 0,070 |
| Seroconversió SAL (Sí vs No) | 0,713 | 0,417 | 2,041 | (0,871; 4,617) | 0,087 |

TAULA 6.8. Resultats dels models bivariants

| Variables d'immunitat maternal | $\hat{\beta}$ | $Se(\hat{\beta})$ | $e^{\hat{\beta}}$ | $IC(e^{\beta})$ | p.valor |
|---|---|---|---|---|---|
| PCV2 | -0,768 | 0,258 | 0,464 | (0,278; 0,758) | 0,003 |
| SIV | -0.096 | 0,678 | 0,908 | (0,246; 3,431) | 0,887 |
| PPV | 0,048 | 0,182 | 1,049 | (0,766; 1,575) | 0,791 |
| PRRSV (+ vs -) | -0,830 | 0,391 | 0,436 | (0,193; 0,911) | 0,034 |
| SAL (+ vs -) | 0,869 | 0,414 | 2,384 | (1,027; 5,253) | 0,044 |

TAULA 6.9. Resultats dels models bivariants

A les taules 6.8 i 6.9, hi ha els resultats dels models bivariants. De les variables de seroconversió, s'han trobat les següents variables estadísticament significatives (amb un nivell de significació del 10%):

- Seroconversió SIV (Sí vs No): obtenim un $\widehat{HR} = 0,537$, i per tant, el risc de patir CP és gairebé el doble en els que no han seroconvertit respecte els que sí.
- Seroconversió SAL (Sí vs No): obtenim un $\widehat{HR} = 2,041$, que vol dir que el risc de patir CP és el doble en els que sí han seroconvertit respecte els que no.

Pel que fa a les variables d'immunitat maternal, les variables estadísticament significatives són:

- PCV2: obtenim un $\widehat{HR} = 0,464$, el que implica que a l'augmentar en mitjana un títol els anticossos de PCV2 el risc de patir CP disminueix.
- PRRSV (+ vs -): obtenim un $\widehat{HR} = 0,436$, el que implica que el risc de patir CP només és del 40% pels positius en relació al que tenen els negatius.
- SAL (+ vs -): obtenim un $\widehat{HR} = 2,384$, el que implica que el risc de patir CP és més del doble en els que tenen aquesta variable positiva respecte els negatius.

Una de les diferències respecte dels models bivariats, tractant el temps com a variable contínua, és que ara la Seroconversió de Salmonella sí que és estadísticament significativa. Per la resta de variables hi ha una concordança amb els resultats presentats a Grau-Roma et al., 2012.

Una de les aportacions més importants en aquest estudi, és que al tractar el temps com a discret ens permet utilitzar els models de resposta binària (en aquest cas amb link *clog-log*), de manera que a més de poder obtenir les estimacions dels efectes de les covariants, també es poden estimar els riscos en cada instant de temps complet.

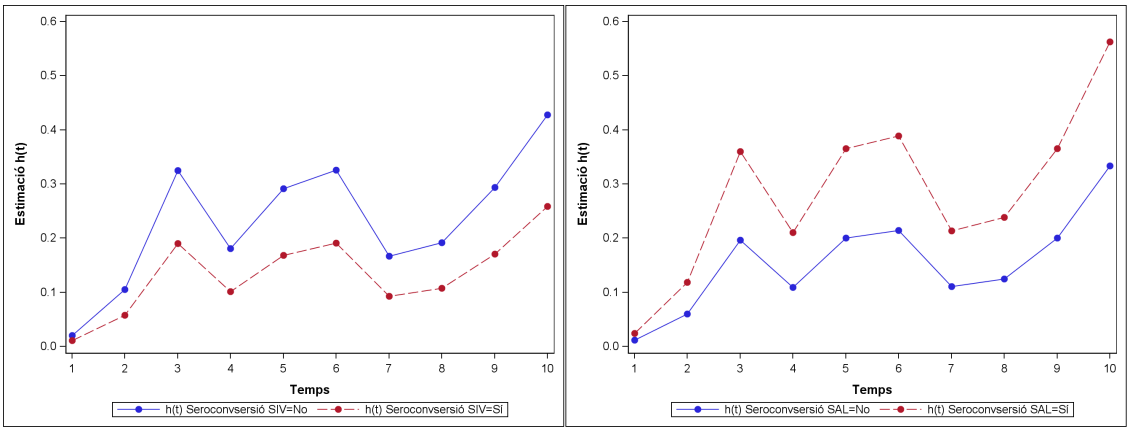


FIGURA 6.2. Funció de risc estimada per la variable Seroconversió SIV i Seroconversion SAL

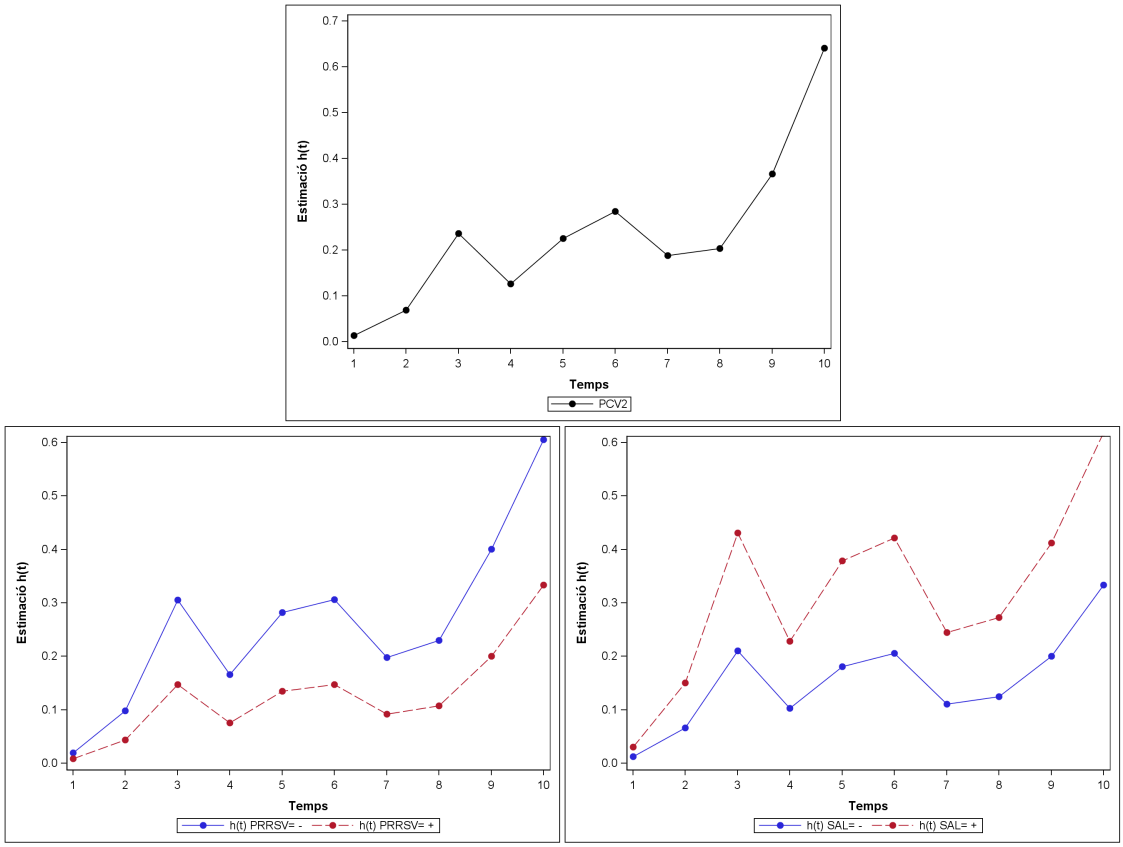


FIGURA 6.3. Funcions de risc estimades per les variables d'immunitat maternal PCV2, PRRSV i SAL

Als gràfics de la Figura 6.2 i 6.3, es representen les funcions de risc estimades, amb el model *Clog-log*, de cadascuna de les covariants estadísticament significatives. En aquests gràfics s'observa:

- En tots, l'eix del temps va de 1 a 10, per indicar que hi ha 10 temps complets. Però l'1 representa 11 setmanes, fins al 10 que representa 20 setmanes.
- En els gràfics de les variables qualitatives, hi ha representat el risc per cadascuna de les categories de la covariant. Per exemple, per la variable Seroconversió SIV, hi ha representat el risc dels que no han seroconvertit (blau) i dels que sí (vermell). I el mateix per la resta de covariants.
- De l'única covariant quantitativa estadísticament significativa (anticossos enfront PCV2) es representa el risc mitjà per cadascun del temps.

**Model multivariant:** en aquest cas es presenta el model parsimoniós amb només les variables estadísticament significatives.

| Variable | $\hat{\beta}$ | $Se(\hat{\beta})$ | $e^{\hat{\beta}}$ | $IC(e^{\beta})$ | p.valor |
|---|---|---|---|---|---|
| Variables de seroconversió | | | | | |
| Seroconversió SIV (Sí vs No) | -0,563 | 0,349 | 0,569 | (0,287; 1,134) | 0,107 |
| Variables d'immunitat maternal | | | | | |
| PCV2 | -0,761 | 0,264 | 0,467 | (0,276; 0,773) | 0,004 |

Taula 6.10. Resultats dels model amb totes les covariants

- Seroconversio SIV (Sí vs No): s'obté un $\widehat{HR} = 0,569$, el que indica que el risc de patir CP és pels que han seroconvertit la meitat dels que no. Aquesta variable és significativa amb un nivell de significació del 10%.
- PCV2 (immunitat maternal): s'obté un $\widehat{HR} = 0,467$, el que vol dir que a l'augmentar en mitjana els anticossos de PCV2 en un títol redueix el risc de desenvolupar CP.

El motiu pel qual no s'obtenen gaires discrepàncies entre els resultats corresponents a temps continu i discret és perquè hi ha molts temps complets i hi ha pocs empats en cada cas, cosa que fa que els resultats amb aquestes metodologies siguin molt semblants al model de Riscos Proporcionals amb temps continu.

## 6.5. Propostes de millora

A continuació es presenten les diferents propostes de millora que es s'ha realitzat en aquestes anàlisis.

### 6.5.1. Millora I - Covariants canviants en el temps

En els apartats anteriors d'aquest capítol, s'ha analitzat la dependència de la funció de risc en funció de diverses covariants, però aquestes covariants s'han tractat com a variables fixes.

Per les característiques d'aquest estudi, els porcs han estat mesurats al llarg del temps, de manera que es tenen els valors dels 6 instants de temps de les variables PCV2, PPV, SIV, PRRSV o SAL.

En aquesta secció es planteja un model de Cox amb covariants mesurades al llarg de l'estudi, de manera que s'aplicaran les tècniques de l'anàlisi de la supervivència amb covariants canviants en el temps.

El model de Riscos Proporcionals, explicat en el Capítol 2, estableix la següent relació entre la funció de risc en el temps $t$ per un individu amb vector de covariants Z:

$$h(t|Z) = h_0(t)e^{\beta' Z} = h_0(t)\exp\left(\sum_{k=1}^{p}\beta_k Z_k\right)$$

Generalitzant aquest model a la situació en la que algunes covariants són canviants en el temps; sigui $Z_j(t)$ el valor de la covariant $j$ al temps $t$. Aleshores el model de Cox es pot escriure com:

$$h(t|Z(t)) = h_0(t)\exp(\beta' Z(t)) = h_0(t)\exp\left[\sum_{k=1}^{p}\beta_k Z_k(t)\right]$$

Per la construcció de la PL en aquest cas, suposem:

- $t_i$ temps observat corresponent a l'individu $i$ amb $i = 1, \ldots, n$,
- Suposem $r$ temps observats diferents,
- $\delta_i$ l'indicador de censura corresponent a l'individu $i$ amb $i = 1, \ldots, n$,
- $Z_i(t) = [Z_{i1}(t), \ldots, Z_{ip}(t)]$, el vector de covariants de l'individu $j$.

Amb el mateix argument emprat al Capítol 2, la PL queda definida com:

$$PL(\beta) = \prod_{m=1}^{r}\frac{\exp\left[\sum_{h=1}^{p}\beta_h Z_{(m)h}(t_m)\right]}{\sum_{l\in R(t_m)}\exp\left[\sum_{h=1}^{p}\beta_h Z_{lh}(t_m)\right]}$$

El procés d'estimacions del vector de paràmetres $\beta^{'}$ es fa seguint el mateix argument explicat en el Capítol 2.
Pel que fa el cas d'empats, Kalbfeisch & Prentice (1980) apunten a que es pot fer el mateix argument de generalització.
Per aplicar-ho a aquestes dades, igual que en el cas de les covariants fixes, es adient utilitzar la metodologia Exact, ja que és la que suposa que la variable temps prové de la discretització d'una variable contínua, que seria aquest cas. I seguint les relacions indicades en el Capítol 4, aquesta metodologia donaria resultats concordants amb el model de resposta binària amb link *clog-log*, de manera que només es presentaran els resultats utilitzant aquest model.

**Model Clog-log:**

Per aplicar el model de resposta binària amb link *clog-log*, com s'ha explicat en el capítols anteriors, s'ha de construir la base de dades expandida. El mecanisme explicat per construir-la es pot aplicar també en aquest cas. L'únic que ara, la covariant no prendrà el mateix valor en cada registre. El valor que prendrà la covariant serà el valor que tenia originàriament en el moment de temps que marca la $D_j$ corresponent. Per exemple, el porc amb necròpsia a les 12 setmanes (que és el segon temps complet) tindrà dos registres, el valor de la covariant pel primer registre serà el que tenia a les 11 setmanes i per el segon registre serà el valor del moment de la necròpsia, i així successivament.

**Resultats:**

En el model, només s'ha inclòs les covariants SIV i SAL donat que són les úniques que van prenent valors diferents durant les 6 mesures (les altres covariants, tenen pocs canvis de positiu a negatiu al llarg de les 6 mesures, i normalment s'observa el canvi a la mesura corresponent a la necròpsia). La variable SAL ja correspon a una variable binària, però per aquesta part de covariants canviants en el temps s'utilitzarà la variable SIV binària donat que és més fàcil trobar els intervals on es produeixen canvis de valors.

**Models bivariants:**

| Variable | $\hat{\beta}$ | $Se(\hat{\beta})$ | $e^{\hat{\beta}}$ | $IC(e^{\beta})$ | p.valor |
|---|---|---|---|---|---|
| SIV (+ vs -) | -0,390 | 0,348 | 0,677 | (0,337; 1,355) | 0,263 |
| SAL (+ vs -) | 0,622 | 0,426 | 1,863 | (0,774; 4,241) | 0,144 |

TAULA 6.11. Resultats models bivariants

d'on s'observa que cap de les dues covariants canviants en el temps és estadísticament significativa.

**Model multivariant:** s'ha decidit fer un model multivariant incloent aquestes dues variables canviants en el temps i els anticossos enfront PCV2 a les tres setmanes. La variable seroconversió SIV no s'ha inclòs en aquest model (tot i que en el model multivariant discret és estadísticament significativa) perquè la covariant canviant en el temps, ja té en compte la informació de com varia aquesta variable al llarg del temps.

| Variable | $\hat{\beta}$ | $Se(\hat{\beta})$ | $e^{\hat{\beta}}$ | $IC(e^{\beta})$ | p.valor |
|---|---|---|---|---|---|
| SIV (+ vs -) | -0,573 | 0,361 | 0,564 | (0,273; 1,142) | 0,112 |
| SAL (+ vs -) | 0.520 | 0,440 | 1,682 | (0,680; 3,957) | 0,237 |
| PCV2 | -0,757 | 0,267 | 0,469 | (0,276; 0,780) | 0,004 |

TAULA 6.12. Resultats models bivariants

Per aquest model l'única variable estadísticament significativa és el nivell d'anticossos enfront PCV2 a les 3 setmanes. Igual que en les anàlisis anteriors, més nivells d'anticossos enfront PCV2 (més defenses), indica un menor risc de desenvolupar CP.

### 6.5.2. Millora II - Competing risk

Com s'ha comentat a la secció 6.1 d'aquest capítol, els porcs es podien classificar en tres categories: amb CP, Sans o Porcs prims no CP. Donat que a l'article només van considerar els casos de CP i Sans, l'altre grup no es va considerar. No obstant, el fet que un porc poguès desenvolupar CP o que es classifiquès com porc prim, no CP, es pot veure com a Competing Risks, ja que quan s'observa un dels dos events es modifica la capacitat d'observar l'altre.
Donat que no forma part dels objectius inicials d'aquest estudi, no s'ha pogut desenvolupar aquesta anàlisi, tot i que ja s'hi està treballant actualment.
Val a dir que per l'equip veterinari treballar un model amb els riscos en competència i poder obtenir resultats de la funció d'incidència ho troben interessant i els pot servir en els seus estudis.

# Capítol 7
# Discussió i conclusions

L'objectiu d'aquest treball, contextualitzat dins de l'anàlisi de supervivència:

(1) Una revisió de les metodologies i aproximacions per tractar valors empatats quan volem estimar un model de Cox.
(2) Una proposta d'utilitzar els models de resposta binària a partir de definir la base de dades expandida.

A partir d'aquesta revisió, s'ha pogut comparar els dos enfocs, a partir dels quals s'ha obtingut:

- Per temps de naturalesa discreta:
  (1) La metodologia Discrete per estimar un model de Cox és la més adient, donat que té la suposició al darrrere que el temps és una variable discreta, i dóna estimacions concordants als models de resposta binària amb link *logit*, realitzats a la base de dades expandida.
  (2) Quan el temps és continu, el $HR$ és igual a $e^{\beta}$. Però en aquest cas no es compleix, i a més en la metodologia Discrete, el terme $e^{\beta}$ correspon al OR. L'única aproximació en que $e^{\beta}$ correspon al HR és Breslow.
  (3) Per poder estimar el HR, es necessari utilitzar els models de resposta binària, ja que a partir de les estimacions dels paràmetres corresponents a les *dummies* $D_j$ conjuntament amb els paràmetres associats a l'efecte de les covariants, es pot obtenir una estimació dels riscos en cada instant de temps complet, i per tant, el HR per cada temps es pot calcular com el quocient d'aquests riscos.
- Per temps continus discretitzats:
  (1) La metodologia Exact del model de Cox és la més adient, donat que té la suposició al darrrere que el temps prové de la discretització d'una variable continua, i dona estimacions concordants als models de resposta binària amb link *clog-log*, realitzats a la base de dades expandida.
  (2) Pel que fa al terme $e^{\beta}$, amb la metodologia Exact, aquest correspon al HR, però de la variable contínua, no de la discretitzada.
  (3) Per obtenir el HR de la variable discretitzada, igual que abans, és adient fer-ho a través dels models de resposta binària, on es poden obtenir

estimacions de la funció de risc (i per tant del HR) en cada instant de temps.

Per altra banda hi ha avantatges i inconvenients per utilitzar models de resposta binària enfront al model de Cox amb les diferents metodologies per tractar els empats:

- Avantatges:
  (1) A l'utilitzar els models de respota binària (amb link *logit* o *clog-log* en funció del cas), el temps computacional és molt menor, ja que quan hi ha un nombre d'empats molt elevat, les metodologies Exact o Discrete poden trigar molt en obtenir els resultats.
  (2) Les estimacions són semblant a les obtingudes amb el model de Cox.
  (3) Un dels fets més important, a l'hora de treballar amb temps discrets, és que el risc és una probabilitat, i aquests models de resposta binària ho tenen en compte.
  (4) Pel que fa a la implementació és molt més senzilla i fàcil, ja que tots els softwares tenen les funcions necessàries per realitzar aquests models. L'únic que es necessita és construir la base de dades expandida, però es fa mitjançant una funció fàcil d'implementar.
  (5) Per la interpretació de les magnituds estimades, cal remarcar que $e^{\beta}$ no sempre correspon a un $HR$ sinó que a vegades estima un $OR$.
- Inconvenients: el principal és que es necessita que hi hagi alguna observació censurada igual al valor de l'últim temps de mort. Aquest requeriment només cal perquè les metodologies estàndards funcionin.

# Capítol 8
# Línies de futur

En aquest capítol es presenten línies futures de recerca sorgides a partir d'aquest treball.

(1) En primer lloc tal i com s'ha comentat en el capítol 2, la metodologia Exact del model de Cox no està implementada en `R`, de manera que una de les línies futures de treball serà crear un paquet addicional de `R`, on s'implementi aquesta metologia o ampliar la funció `coxph` del paquet `Survival`, on a l'opció `ties` sí que es permeti seleccionar aquesta metologia.

(2) En el capítol 4, s'han donat les indicacions per explicar el perquè les metodologies Exact i el model de resposta binària *clog-log* donen estimacions concordants així com la metodologia Discrete i el model *Logit*. Però cal acabar d'obtenir una demostració més formal d'aquestes relacions. De fet, cal observar que "passar"de la versemblança a la versemblança parcial es fa de forma habitual quan s'estableix un model de Cox on l'objectiu és la funció de risc.

(3) Per altra banda, per exemplificar les relacions entre aquests models, s'ha realitzat un estudi de simulació on les distribucions que s'han escollit són la geomètrica i l'exponencial que tenen risc constant i a més, s'ha fixat la censura en l'últim instant de temps. Per això també serà interessant realizar un estudi de simulació més general on s'utilitzin altres distribucions i també incloure algun tipus de censura aleatòria.

(4) Com s'ha comentat, un dels inconvenients d'utilitzar els models de resposta binària es que necessita que hi hagi alguna observació censurada igual al valor de l'últim temps de mort. Una de les línies de treball futur, és estudiar com adapatar les metodologies per aquests models quan totes les observacions corresponents a l'últim instant de temps són completes o censurades.

(5) Arrel de les dades del temps fins desenvulpar CP, també és interessant veure com es comporten aquests models discrets en el cas de tenir competing risk, i com hauria de ser la base de dades expandida i els models de resposta binària que s'haurien d'aplicar per aquest cas.

# Capítol 9
# Annex

| Mètode | $n$ | $\hat{\beta}$ | Std | $e^{\hat{\beta}}$ | $Std(e^{\hat{\beta}})$ |
|---|---|---|---|---|---|
| Breslow | 100 | 0,678 | 0,171 | 1,998 | 0,349 |
| Efron | 100 | 0,893 | 0,229 | 2,509 | 0,593 |
| Discrete | 100 | 1,253 | 0,348 | 3,725 | 1,403 |
| Logit | 100 | 1,275 | 0,354 | 3,814 | 1,455 |
| Exact | 100 | 0,938 | 0,246 | 2,633 | 0,673 |
| Cloglog | 100 | 0,956 | 0,247 | 2,656 | 0,683 |
| Breslow | 150 | 0,696 | 0,153 | 2,030 | 0,315 |
| Efron | 150 | 0,914 | 0,208 | 2,550 | 0,542 |
| Discrete | 150 | 1,270 | 0,314 | 3,744 | 1,266 |
| Logit | 150 | 1,285 | 0,317 | 3,804 | 1,298 |
| Exact | 150 | 0,954 | 0,222 | 2,661 | 0,607 |
| Cloglog | 150 | 0,960 | 0,223 | 2,677 | 0,614 |
| Breslow | 250 | 0,701 | 0,128 | 2,033 | 0,267 |
| Efron | 250 | 0,926 | 0,176 | 2,565 | 0,474 |
| Discrete | 250 | 1,290 | 0,270 | 3,776 | 1,180 |
| Logit | 250 | 1,299 | 0,271 | 3,811 | 1,197 |
| Exact | 250 | 0,966 | 0,188 | 2,674 | 0,536 |
| Cloglog | 250 | 0,969 | 0,188 | 2,684 | 0,539 |
| Breslow | 500 | 0,696 | 0,074 | 2,011 | 0,150 |
| Efron | 500 | 0,913 | 0,101 | 2,504 | 0,257 |
| Discrete | 500 | 1,250 | 0,153 | 3,547 | 0,567 |
| Logit | 500 | 1,260 | 0,153 | 3,564 | 0,571 |
| Exact | 500 | 0,947 | 0,107 | 2,593 | 0,284 |
| Cloglog | 500 | 0,949 | 0,107 | 2,597 | 0,285 |
| Breslow | 1500 | 0,697 | 0,041 | 2,010 | 0,082 |
| Efron | 1500 | 0,917 | 0,057 | 2,505 | 0,142 |
| Discrete | 1500 | 1,262 | 0,088 | 3,547 | 0,315 |
| Logit | 1500 | 1,264 | 0,088 | 3,553 | 0,315 |
| Exact | 1500 | 0,950 | 0,061 | 2,591 | 0,157 |
| Cloglog | 1500 | 0,951 | 0,061 | 2,592 | 0,157 |
| Breslow | 2500 | 0,696 | 0,032 | 2,007 | 0,063 |
| Efron | 2500 | 0,914 | 0,042 | 2,498 | 0,104 |
| Discrete | 2500 | 1,257 | 0,061 | 3,522 | 0,214 |
| Logit | 2500 | 1,258 | 0,061 | 3,526 | 0,215 |
| Exact | 2500 | 0,947 | 0,044 | 2,581 | 0,113 |
| Cloglog | 2500 | 0,947 | 0,044 | 2,582 | 0,113 |
| Breslow | 5000 | 0,693 | 0,026 | 2,000 | 0,063 |
| Efron | 5000 | 0,911 | 0,035 | 2,489 | 0,088 |
| Discrete | 5000 | 1,253 | 0,051 | 3,507 | 0,184 |
| Logit | 5000 | 1,254 | 0,052 | 3,509 | 0,184 |
| Exact | 5000 | 0,944 | 0,037 | 2,571 | 0,096 |
| Cloglog | 5000 | 0,944 | 0,037 | 2,572 | 0,096 |

TAULA 9.1. Distribució geomètrica: Resultats estimacions geomètrica

| N | T | Logit | | Clog-log | |
|---|---|---|---|---|---|
| | | $\widehat{HR}$ | $Se(\widehat{HR})$ | $\widehat{HR}$ | $Se(\widehat{HR})$ |
| 100 | 1 | 2,025 | 0,376 | 2,018 | 0,365 |
| | 2 | 2,015 | 0,401 | 2,007 | 0,368 |
| | 3 | 2,015 | 0,381 | 2,001 | 0,341 |
| | 4 | 2,094 | 0,490 | 2,047 | 0,413 |
| | 5 | 2,060 | 0,469 | 2,008 | 0,375 |
| 150 | 1 | 2,037 | 0,330 | 2,037 | 0,323 |
| | 2 | 2,055 | 0,354 | 2,046 | 0,336 |
| | 3 | 2,076 | 0,359 | 2,051 | 0,323 |
| | 4 | 2,059 | 0,368 | 2,037 | 0,329 |
| | 5 | 2,082 | 0,452 | 2,049 | 0,373 |
| 250 | 1 | 2,052 | 0,288 | 2,048 | 0,281 |
| | 2 | 2,029 | 0,264 | 2,031 | 0,262 |
| | 3 | 2,048 | 0,281 | 2,040 | 0,265 |
| | 4 | 2,030 | 0,316 | 2,024 | 0,275 |
| | 5 | 2,014 | 0,304 | 2,009 | 0,264 |
| 500 | 1 | 2,017 | 0,162 | 2,016 | 0,157 |
| | 2 | 2,013 | 0,158 | 2,012 | 0,150 |
| | 3 | 2,011 | 0,170 | 2,011 | 0,154 |
| | 4 | 2,039 | 0,189 | 2,025 | 0,161 |
| | 5 | 2,029 | 0,199 | 2,019 | 0,167 |
| 1500 | 1 | 2,010 | 0,085 | 2,010 | 0,083 |
| | 2 | 2,015 | 0,095 | 2,012 | 0,088 |
| | 3 | 2,008 | 0,088 | 2,009 | 0,082 |
| | 4 | 2,020 | 0,107 | 2,014 | 0,089 |
| | 5 | 2,005 | 0,105 | 2,005 | 0,088 |
| 2500 | 1 | 2,007 | 0,069 | 2,007 | 0,066 |
| | 2 | 2,008 | 0,068 | 2,007 | 0,065 |
| | 3 | 2,012 | 0,070 | 2,010 | 0,063 |
| | 4 | 2,010 | 0,088 | 2,008 | 0,073 |
| | 5 | 2,005 | 0,081 | 2,004 | 0,066 |
| 5000 | 1 | 2,001 | 0,058 | 2,001 | 0,055 |
| | 2 | 2,000 | 0,053 | 2,000 | 0,051 |
| | 3 | 2,002 | 0,052 | 2,001 | 0,050 |
| | 4 | 2,001 | 0,062 | 2,000 | 0,055 |
| | 5 | 2,001 | 0,060 | 2,000 | 0,053 |

TAULA 9.2. Distribució geomètrica: HR estimats pels models de resposta binària

| Mètode | $n$ | $\hat{\beta}$ | Std | $e^{\hat{\beta}}$ | $Std(e^{\hat{\beta}})$ |
|---|---|---|---|---|---|
| Breslow | 100 | 0,561 | 0,207 | 1,791 | 0,387 |
| Efron | 100 | 0,679 | 0,253 | 2,036 | 0,540 |
| Discrete | 100 | 0,833 | 0,320 | 2,424 | 0,841 |
| Logit | 100 | 0,847 | 0,325 | 2,464 | 0,872 |
| Exact | 100 | 0,699 | 0,263 | 2,084 | 0,579 |
| Cloglog | 100 | 0,706 | 0,266 | 2,099 | 0,588 |
| Breslow | 150 | 0,580 | 0,162 | 1,809 | 0,308 |
| Efron | 150 | 0,699 | 0,193 | 2,050 | 0,422 |
| Discrete | 150 | 0,862 | 0,243 | 2,440 | 0,636 |
| Logit | 150 | 0,872 | 0,246 | 2,466 | 0,651 |
| Exact | 150 | 0,718 | 0,200 | 2,092 | 0,447 |
| Cloglog | 150 | 0,722 | 0,201 | 2,101 | 0,452 |
| Breslow | 250 | 0,605 | 0,102 | 1,841 | 0,191 |
| Efron | 250 | 0,728 | 0,125 | 2,088 | 0,267 |
| Discrete | 250 | 0,895 | 0,158 | 2,478 | 0,407 |
| Logit | 250 | 0,901 | 0,159 | 2,493 | 0,412 |
| Exact | 250 | 0,745 | 0,129 | 2,125 | 0,282 |
| Cloglog | 250 | 0,748 | 0,130 | 2,130 | 0,284 |
| Breslow | 500 | 0,557 | 0,084 | 1,752 | 0,147 |
| Efron | 500 | 0,671 | 0,102 | 1,965 | 0,200 |
| Discrete | 500 | 0,821 | 0,128 | 2,291 | 0,293 |
| Logit | 500 | 0,824 | 0,128 | 2,297 | 0,295 |
| Exact | 500 | 0,685 | 0,105 | 1,994 | 0,209 |
| Cloglog | 500 | 0,686 | 0,105 | 1,997 | 0,210 |
| Breslow | 1500 | 0,565 | 0,050 | 1,761 | 0,089 |
| Efron | 1500 | 0,681 | 0,061 | 1,980 | 0,122 |
| Discrete | 1500 | 0,833 | 0,075 | 2,306 | 0,176 |
| Logit | 1500 | 0,834 | 0,075 | 2,308 | 0,176 |
| Exact | 1500 | 0,695 | 0,063 | 2,008 | 0,127 |
| Cloglog | 1500 | 0,696 | 0,063 | 2,009 | 0,127 |
| Breslow | 2500 | 0,562 | 0,038 | 1,756 | 0,067 |
| Efron | 2500 | 0,678 | 0,047 | 1,971 | 0,092 |
| Discrete | 2500 | 0,828 | 0,058 | 2,293 | 0,134 |
| Logit | 2500 | 0,829 | 0,058 | 2,294 | 0,134 |
| Exact | 2500 | 0,691 | 0,048 | 1,999 | 0,096 |
| Cloglog | 2500 | 0,692 | 0,048 | 1,999 | 0,096 |
| Breslow | 5000 | 0,565 | 0,028 | 1,760 | 0,049 |
| Efron | 5000 | 0,681 | 0,034 | 1,976 | 0,067 |
| Discrete | 5000 | 0,831 | 0,043 | 2,298 | 0,098 |
| Logit | 5000 | 0,831 | 0,043 | 2,299 | 0,098 |
| Exact | 5000 | 0,694 | 0,035 | 2,003 | 0,070 |
| Cloglog | 5000 | 0,694 | 0,035 | 2,003 | 0,070 |

TAULA 9.3. Distribució exponencial: Resultats estimacions

| N | T | Logit | | Clog-log | |
|---|---|---|---|---|---|
| | | $\widehat{HR}$ | $Se(\widehat{HR})$ | $\widehat{HR}$ | $Se(\widehat{HR})$ |
| 100 | 1 | 1,884 | 0,441 | 1,859 | 0,423 |
| | 2 | 1,884 | 0,469 | 1,858 | 0,435 |
| | 3 | 1,864 | 0,424 | 1,847 | 0,411 |
| | 4 | 1,877 | 0,445 | 1,854 | 0,424 |
| | 5 | 1,431 | 0,186 | 1,538 | 0,225 |
| 150 | 1 | 1,896 | 0,346 | 1,865 | 0,333 |
| | 2 | 1,895 | 0,341 | 1,864 | 0,329 |
| | 3 | 1,895 | 0,353 | 1,863 | 0,333 |
| | 4 | 1,894 | 0,337 | 1,863 | 0,325 |
| | 5 | 1,454 | 0,185 | 1,565 | 0,210 |
| 250 | 1 | 1,922 | 0,227 | 1,891 | 0,213 |
| | 2 | 1,916 | 0,207 | 1,888 | 0,200 |
| | 3 | 1,914 | 0,204 | 1,887 | 0,199 |
| | 4 | 1,918 | 0,218 | 1,889 | 0,205 |
| | 5 | 1,488 | 0,115 | 1,602 | 0,130 |
| 500 | 1 | 1,820 | 0,170 | 1,795 | 0,160 |
| | 2 | 1,818 | 0,160 | 1,794 | 0,154 |
| | 3 | 1,821 | 0,166 | 1,795 | 0,158 |
| | 4 | 1,824 | 0,171 | 1,797 | 0,159 |
| | 5 | 1,433 | 0,080 | 1,539 | 0,095 |
| 1500 | 1 | 1,828 | 0,099 | 1,805 | 0,096 |
| | 2 | 1,830 | 0,100 | 1,805 | 0,096 |
| | 3 | 1,827 | 0,099 | 1,804 | 0,096 |
| | 4 | 1,831 | 0,099 | 1,806 | 0,095 |
| | 5 | 1,438 | 0,049 | 1,548 | 0,059 |
| 2500 | 1 | 1,821 | 0,075 | 1,798 | 0,072 |
| | 2 | 1,823 | 0,074 | 1,799 | 0,071 |
| | 3 | 1,821 | 0,073 | 1,798 | 0,071 |
| | 4 | 1,821 | 0,075 | 1,798 | 0,072 |
| | 5 | 1,438 | 0,039 | 1,546 | 0,046 |
| 5000 | 1 | 1,826 | 0,054 | 1,802 | 0,050 |
| | 2 | 1,823 | 0,055 | 1,801 | 0,050 |
| | 3 | 1,823 | 0,056 | 1,800 | 0,050 |
| | 4 | 1,823 | 0,055 | 1,801 | 0,050 |
| | 5 | 1,442 | 0,027 | 1,551 | 0,032 |

TAULA 9.4. Distribució exponencial: HR estimats pels models de resposta binària

# Bibliografia


[1] Cox, D.R. (1972) Regression Models and Life-Tables (with discussion). *Journal of the Royal Stastistical Society, B,* **74**, 187-220.

[2] Cox, D.R. (1975) Partial likelihood. *Biometrika* **62**, 269-276.

[3] Klein, J.P. and Moeschberger, M.L. (2003) *Survival Analysis: Techniques for Censored Data and Truncated Data,* (2nd Edition). Springer.

[4] Cox, D.R. and Oakes, D. (1984) *Analysis of Survival Data* (1st Edition). Chapman and Hall, London.

[5] Hosmer, D. and Lemeshow, S. (2000) *Applied Logistic Regression* (2nd Edition). John Wiley & Sons, New York.

[6] McCullagh, P. and Nelder, J.A. (1991) *Generalized Linear Models* (2nd Edition). Chapman and Hall, London.

[7] Therneau, T.M. and Grambsch, P.M. (2000) *Modeling Survival Data. Extending the Cox model* (2nd Edition). Springer.

[8] Singer, J.E. and Willet, J.B. (2000) *Applied Longitudinal Data Analysis. Modeling change and event ocurrence*. Springer.

[9] Breslow, N.E. (1974) Covariance analysis of censored survival data. *Biometrics*, **30**, 89-100.

[10] Efron, B. (1977) The efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association*, **72**, 557-565.

[11] Kalbfleish, J.D. and Prentice, R.L. (1972) Contribution to the discussion of a paper by D.R Cox. *Journal of the Royal Statistical Society, B*, **34**, 215-216.

[12] Prentice, R.L. and Gloekler, L.A (1978) Regression analysis of grouped survival data with application to breast cancer data. *Biometrics,* **34**, 57-67.

[13] Kalbfleish, J.D. and Prentice, R.L. (1980) *The Statistical Analysis of Failure Time Data.* John Wiley & Sons, New York.

[14] Nelder, J.A. and Wedderburn, R.W.M. (1972) Generalized Linear Models. *Journal of the Royal Statistical Society*, **3**, 370-384.

[15] Collet, D. (2003) *Modelling Survival Data in Medical Research* (2nd Edition). Chapman and Hall, London.

[16] Zhang, D., *Lecture notes: Chapter 7* (2005).

[17] Grau-Roma, L., Stockmarr, A., Kristensen, C.S., Enoe, C., Lopez-Soria, S., Nofrarias, M., Bille-Hansen, V., Hjulsager, C.K., Sibila, M., Jorsal, S.E., Fraile, L., Baekbo, P., Vigre, H., Segales and Larsen, L.E. (2012) Infectious risk factors for individual postweaning multisystemic wasting syndrome (PMWS) development in pigs from affected farms in Spain and Denmark. *Elsevier Journal*, **93**, 1231-1240.

[18] Grau-Roma, L., Fraile, L., Segalés, J. (2011) Recent advances in the epidemiology, diagnosis and control of diseases caused by porcine circovirus type 2. *The Veterinary Journal* **187**, 23-32.

[19] Madec, F., Rose, N., Grasland, B., Cariolet, R., Jestin, A., (2008) Post-weaning multisystemic wasting syndrome and other PCV2-related problems in pigs: a 12-year experience. *Transboundary and Emerging Diseases* **55**, 273-283.

[20] Armstrong, D., Bishop, S.C. (2004) Does genetics or litter effect influence mortatily in PMWS. *In: Armstrong, D. (Ed.), Proceeding of International Pig Veterinary Society Congress*, 62.

[21] Harding, J. (1998) Postweaning multisystemic wasting syndrome: epidemiology and clinical presentation. *Journal of Swine Health & Production* **6**, 249-254.