Partial Likelihood and Models for binary response

Ana Vázquez[1], *Anna Espinal*[2], *Olga Julià*[3]

[1]ana.vazquez@uab.cat, Servei d'Estadística Aplicada, Universitat Autònoma de Barcelona
[2]anna.espinal@uab.cat, Servei d'Estadística Aplicada, Universitat Autònoma de Barcelona
[3]olgajulia@ub.edu, Departament de probabilitat, lògica i estadística, Universitat de Barcelona

Abstract

Usually time until an event is measured in continuous scale, but for various reasons we can find time measured in discrete scale. In this paper compares different models for analyzing discrete time from a set of covariates: the Proportional Hazards model (Cox, 1972) with different methodologies for tied data and models for binary response with link logit and cloglog.

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

AMS: AMS classification. 62J12, 62N01, 62N02, 62N03

1. Partial likelihood for tied event times

Let $((t_1, \delta_1), (t_2, \delta_2), \ldots, (t_n, \delta_n))$ be a sample of survival times where t_i are the observed times and δ_i the censorship indicator. Suppose that there are r different and uncensored times and $Z_i = (Z_{1i}, \ldots, Z_{pi})$ be the vector of covariates for individual $i = 1, \ldots, n$. To obtain estimates of the covariate effects, Cox (1972) proposed a semiparametric method based on the Partial Likelihood (PL) given by:

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

where $R(t_{(m)})$ is the risk set in $t_{(m)}$.
The PL factors correspond to the probabilities: $P($ individual dies at $t_m|$ one death at $t_m)$.

The PL is treated as a usual likelihood function and inferences are carried by usual way: the estimation of parameters is obtained by maximizing $\ln(PL(\beta))$.

Sometimes, due to the way that times is measured, there is tied values for the observed time. That is, more than one individual have the same observed time. Suppose that d_m individuals failling at t_m.
There are some alternatives taking into account for ties into the PL:

- Breslow (1974): all individuals failing in t_m have the same denominator in PL.
- Efron (1977): individuals failing in t_m contributes with different weights, due to the denominator decrease proportionally.

- Discrete (1972): this method assumes discrete times, therefore no underlying ordering of ties is considered. The denominator takes into account for all possible subsets (without replacement) of d_i individuals that we can take from the risk set.
- Exact (1980): this method assumes that the survival time comes from a continuous random variable and we observed tied values because data are grouped. The PL takes into account all possible orders of tied individuals. This method is often very close to the Efron approximation.

2. Likelihood for discrete times

Suppose our sample of times to event comes from a discrete random variable: $((t_1, \delta_1), (t_2, \delta_2), \ldots, (t_n, \delta_n))$ where t_i are the observed times and δ_i the censorship indicator.

The likelihood function is proportional to: $\prod_{i=1}^{n} \left[P(t = t_i)^{\delta_i} P(t > t_i)^{(1-\delta_i)} \right]$
We can relate the risk function, $h(\cdot)$, with previous probabilities as:

$$P(T = t_i) = h_{t_i} \prod_{m=1}^{t_i - 1} (1 - h_m) \text{ and } P(T > t_i) = \prod_{m=1}^{t_i} (1 - h_m)$$

where $h_j = P(T = j | T \geq j)$

Substituting the above equalities is obtained:

$$L \approx \prod_{i=1}^{n} \left[h_{t_i} \prod_{m=1}^{t_i - 1} (1 - h_m) \right]^{\delta_i} \left[\prod_{m=1}^{t_i} (1 - h_m) \right]^{(1-\delta_i)}$$

Taking $r_{im} = \delta_i \mathbb{1}(t_i = m)$, we obtained that:

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

which is the same likelihood function coming from a model for a binary response, where if $r_{im} = 1$ we have h_m and if $r_{im} = 0$ we have $(1 - h_m)$.

This is the likelihood for a response variable, $Bernoulli(h_m)$ that is we could also establish the usual models for a binary response:

- Model for binary response with link logit:

$$\ln \left(\frac{h(t_k|Z)}{1 - h(t_k|Z)} \right) = \alpha_k + \beta_l Z \Longleftrightarrow h(t_k|Z) = \frac{e^{\alpha_k + \beta_l Z}}{1 + e^{\alpha_k + \beta_l Z}}$$

- Model for binary response with link cloglog:

$$\ln \left(- \ln(1 - h(t_k|Z)) \right) = \eta_k + \beta_{cl} Z \Longleftrightarrow h(t_k|Z) = \exp(-e^{\eta_k + \beta_{cl} Z})$$

3. Relationships between Likelihood, Partial likelihood and Models for binary response

Regardless of the nature of time variable, according with Cox argument, to estimate the effect of covariates we can work directly using the PL without going through the likelihood.
following this argument, we can justify the next relationships depending on the nature of time:

- Time as a discrete variable:
  1. Likelihood *vs* Logit: If h_m, from (1) are parameterized using the link logit, so that, $h(t_m|Z) = \frac{e^{\alpha_m+\beta Z_l}}{1+e^{\alpha_m+\beta_l Z}}$, we just get the same likelihood corresponding to a model for binary response with link logit.
  2. Partial likelihood *vs* Discrete: If the probabilities used in the PL, are parameterized by the link logit, we get the PL of the Discrete method for dealing with tied data in a Cox model.
- Time as a continuous variable with grouped values:
  1. Likelihood *vs* Cloglog: If h_m, from (1) are parameterized using the link cloglog, so that, $h(t_m|Z) = \exp(-e^{\eta_m+\beta_{cl}Z})$, we just get the same likelihood corresponding to a model for binary response with link cloglog.
  2. Partial likelihood *vs* Cloglog: Prentice and Gloeckler (1978), presents an equivalent version of continuous Proportional Hazards model, when time is a discrete variable. If T comes from grouping a continuous variable U, and assuming a Cox model for the continuous time variable U, then the risk function at time t_j can be expressed as:

$$h(t_j|Z) \quad = \quad 1 - \exp\left(-e^{\beta' Z+\eta_j}\right)$$

     where $\eta_j = \ln\left(\int_{t_{j-1}}^{t_j} \lambda_0(t)dt\right)$. Note that this is exactly a model for a binary response with a link cloglog.
  3. Partial likelihood *vs* Exact: because the probabilities used in the PL and in the Exact method are based in the all possible orders of tied individuals.

4. Simulations

The relationships argued in the previous section have been analyzed by simulation studies for the two cases: time as a discrete variable and time as a continuous grouped variable. Geometric and exponential distributions have been used, respectively, because they have a constant hazard function.

In both simulations we show that Discrete and Logit give concordant estimates as well as Exact and Cloglog.
It is important to note that the magnitude e^{β} not always corresponds to Hazard Ratio (HR).

- Time as a discrete variable: the term e^{β} corresponds to HR only with Breslow approximation. For obtaining estimates of the HR, it must use models for binary response.
- Time as a continuous grouped variable: the term e^{β} corresponds to HR of original continuous variable with Exact and Cloglog. To obtain estimates of the HR for the grouped variablemust be used models for binary response.

As an illustration, Figure 1 displays a plot for the geometric and exponential distribution results. Different sample sizes are presented. HR=2 in both simulations. We can observe:

- In both cases, there are three different estimations depending on the used method: Breslow, Discrete and Logit, Exact and Cloglog.
- For the geometric distribution, $e^{\hat{\beta}}$ only is a good estimate for HR when Breslow is used. For the other methods $e^{\hat{\beta}}$ do not estimate HR but odds ratio (OR) in some cases.
- For the exponential distribution, $e^{\hat{\beta}}$ is a good estimate for HR when Efron, Exact and Cloglog are used.

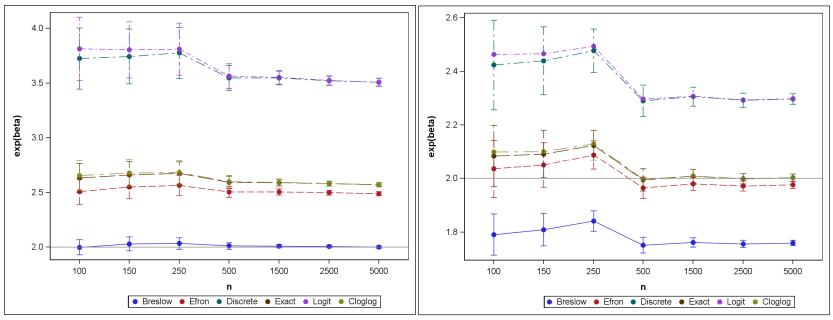

Figure 1: Geometric distribution (Left) Exponential distribution (Right)

5. Acknowledgments

MTM2012-38067-C02-01, 2014 SGR 464 and MTM2012-31118.

6. Bibliography

[1] Cox, D.R (1972), *Regression Models and Life-Tables (with discussion)*. Journal of the Royal Statistical Society, Vol.34, No.2, 187-220.

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

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