Models for binary response and survival data

Ana Vázquez
Anna Espinal
Olga Julià

January 20th, 2015

Contents

1. Motivation
2. Basic concepts: discrete time
3. Extended dataset
   - Original dataset
   - Extended dataset
4. Models for binary response
   - Link logit
   - Link cloglog
5. Models for binary response and discrete survival data
6. Dataset of Singer and Willet
7. Conclusions

Motivation

- Usually time to event is measured on a continuous scale
- Sometimes can be measured on a discrete scale giving a discrete response variable.
- Discrete observed times usually implies tied data.
- Need for specific methods of analyzing discrete survival data.

Basic concepts: discrete time

- Let T be a discrete random variable defined as the time until the event of interest.
- $t_1 < t_2 < \cdots < t_j < \ldots$ with probability mass function $p_j = P(T = t_j)$, $j = 1, 2, \ldots$ and $\sum_j p_j = 1$

Concept	Formula
Survival function	$S(t_j) = P(T > t_j) = \sum_{i:t_i > t_j} p_i$
Hazard function	$h(t_j) = P(T = t_j \mid T \geq t_j) = \frac{p_j}{S(t_{j-1})}, j = 1, 2, \ldots$
	$h(t_j) = 1 - \frac{S(t_j)}{S(t_{j-1})}$

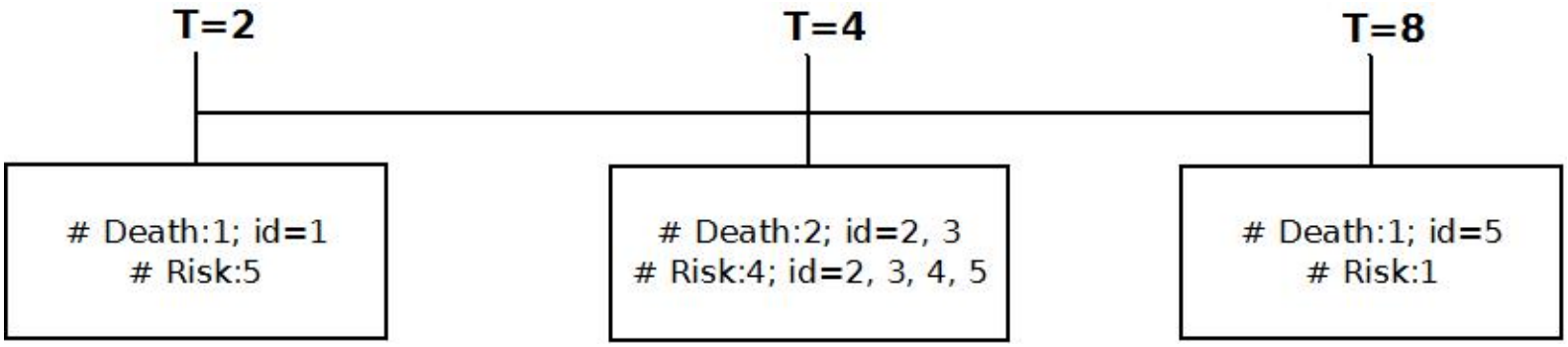
Original dataset

Original dataset:

id	T	δ	Z
1	2	1	0
2	4	1	0
3	4	1	1
4	4	0	1
5	8	1	1

- Risk set: all individuals with $T \geq t_j$

Original dataset



Empirical estimates of hazard:

$$\begin{array}{|l|}\hline h(t_j) = P(T = t_j | T \geq t_j) \\ \hline h(t_1) = \frac{1}{5} = h(t = 2) \\ \hline h(t_2) = \frac{2}{4} = h(t = 4) \\ \hline h(t_3) = 1 = h(t = 8) \\ \hline \end{array}$$

Extended dataset

Extended dataset:

id	T^{order}	T	δ	D_1	D_2	D_3	Y	Z
1	1	2	1	1	0	0	1	0
2	2	4	1	1	0	0	0	0
3	2	4	1	1	0	0	0	1
4	.	4	0	1	0	0	0	1
5	3	8	1	1	0	0	0	1

Extended dataset

Extended dataset:

id	T^{order}	T	δ	D_1	D_2	D_3	Y	Z
1	1	2	1	1	0	0	1	0
2	2	4	1	1	0	0	0	0
2	**2**	**4**	**1**	**0**	**1**	**0**	**1**	**0**
3	2	4	1	1	0	0	0	1
3	**2**	**4**	**1**	**0**	**1**	**0**	**1**	**1**
4	.	4	0	1	0	0	0	1
4	**.**	**4**	**0**	**0**	**1**	**0**	**0**	**1**
5	3	8	1	1	0	0	0	1
5	**3**	**8**	**1**	**0**	**1**	**0**	**0**	**1**

Extended dataset

Extended dataset:

id	T^{order}	T	δ	D_1	D_2	D_3	Y	Z
1	1	2	1	1	0	0	1	0
2	2	4	1	1	0	0	0	0
2	2	4	1	0	1	0	1	0
3	2	4	1	1	0	0	0	1
3	2	4	1	0	1	0	1	1
4	.	4	0	1	0	0	0	1
4	.	4	0	0	1	0	0	1
5	3	8	1	1	0	0	0	1
5	3	8	1	0	1	0	0	1
5	**3**	**8**	**1**	**0**	**0**	**1**	**1**	**1**

Extended dataset

Empirical estimates of hazard:

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

Extended dataset

Empirical estimates of hazard:

Original dataset		Extended dataset
$h(t_j) = P(T = t_j \mid T \geq t_j)$	$=$	$h(t_j^{order}) = P(Y = 1 \mid D_j = 1)$
$h(t_1) = \frac{1}{5} = h(t = 2)$	$=$	$h(t_1^{order}) = \frac{1}{5}$
$h(t_2) = \frac{2}{4} = h(t = 4)$	$=$	$h(t_2^{order}) = \frac{2}{4}$
$h(t_3) = 1 = h(t = 8)$	$=$	$h(t_3^{order}) = 1$

Extended dataset

- k different uncersored times.
- $D_1, D_2, \ldots, D_k$ time dummy variables.
- T^{order} indicates the number of risk sets that each individual belongs (number of rows for each individual).
- Y a **binary variable** taking value equal 1 for the last row of individuals with uncensored time.

T^{order}	T	δ	D_1	D_2	.	D_k	Z_1	.	Z_p	Y
1	t_1	δ_1	1	0	.	0	z_{11}	.	z_{1p}	y_{11}
2	t_2	δ_2	1	0	0	0	z_{21}	.	z_{2p}	y_{12}
2	t_2	δ_2	0	1	0	0	z_{21}	.	z_{2p}	y_{22}
.	.	.	.	.	.	.	.	.	.	
.	.	.	.	.	.	.	.	.	.	
.	.	.	.	.	.	.	.	.	.	
n	t_n	δ_n	1	0	0	0	z_{n1}	.	z_{np}	y_{1n}
n	t_n	δ_n	0	1	0	0	z_{n1}	.	z_{np}	y_{2n}
n	t_n	δ_n	0	0	1	0	z_{n1}	.	z_{np}	y_{3n}
.	.	.	.	0	.	.	.	.	.	
n	t_n	δ_n	0	0	0	1	z_{n1}	.	z_{np}	y_{nn}

Extended dataset

In each of the datasets can be defined the following functions:

- Original dataset: $h(t_j) = P(T = t_j | T \geq t_j)$
- Extended dataset: $h(t_j^{order}) = P(Y = 1 | D_j = 1)$

Extended dataset

In each of the datasets can be defined the following functions:

- Original dataset: $h(t_j) = P(T = t_j | T \geq t_j)$
- Extended dataset: $h(t_j^{order}) = P(Y = 1 | D_j = 1)$

then can be proved that

$$h(t_j) = h(t_j^{order})$$

Models for binary response

Let Y be the outcome of interest:

$$Y = \begin{cases} 1 & \text{presence} \\ 0 & \text{absence} \end{cases}$$

- $Y_1, \ldots, Y_n$ i.i.d
- $Z_i = (Z_{1i}, Z_{2i}, \ldots, Z_{pi})$ vector of covariates

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

Models for binary response: link logit

Two link functions:

- Link *logit*:

$$logit(p) = ln(\frac{p}{1-p}) = \alpha + \beta_1 Z_1 + \ldots + \beta_p Z_p \Leftrightarrow$$

$$p = \frac{e^{(\alpha+\beta_1 Z_1+\cdots+\beta_p Z_p)}}{1 + e^{(\alpha+\beta_1 Z_1+\cdots+\beta_p Z_p)}}$$

The likelihood function:

$$\begin{aligned} L(\alpha, \beta | Y, \mathbf{Z}) &= \prod_{i=1}^{n} P(Y_i = 1|Z_i)^{y_i} P(Y_i = 0|Z_i)^{1-y_i} \\ &= \prod_{i=1}^{n} \left(\frac{e^{\alpha+\beta' Z_i}}{1 + e^{\alpha+\beta' Z_i}} \right)^{y_i} \left(\frac{1}{1 + e^{\alpha+\beta' Z_i}} \right)^{1-y_i} \end{aligned}$$

Models for binary response: link cloglog

- Link *cloglog*:

$$cloglog(p) = ln(-ln(1-p)) = \eta + \gamma_1 Z_1 + \ldots + \gamma_p Z_p \Leftrightarrow$$

$$p = 1 - exp\left(-e^{\eta+\gamma_1 Z_1+\ldots+\gamma_p Z_p}\right)$$

The likelihood function:

$$\begin{aligned} L(\eta, \gamma | Y, \mathbf{Z}) &= \prod_{i=1}^{n} P(Y_i = 1|Z_i)^{y_i} P(Y_i = 0|Z_i)^{1-y_i} \\ &= \prod_{i=1}^{k} \left(1 - exp\left(-e^{\eta+\gamma' Z_i}\right)\right)^{y_i} \left(exp\left(-e^{\eta+\gamma' Z_i}\right)\right)^{1-y_i} \end{aligned}$$

Models for binary response and discrete survival data

From $h(t_j^{order}) = P(Y = 1|D_j = 1)$, possible models in Extended dataset:

$$
\begin{aligned}
ln\left(\frac{h(t^{order}|Z)}{1 - h(t^{order}|Z)}\right) &= \alpha_1 D_1 + \alpha_2 D_2 + \cdots + \alpha_k D_k + \beta_l Z \\
ln(-ln(1 - h(t^{order}|Z))) &= \eta_1 D_1 + \eta_2 D_2 + \cdots + \eta_k D_k + \beta_{cl} Z
\end{aligned}
$$

OBSERVATIONS:

- There are more than one intercept, in particular it has one for every moment of time, so it is like a time-dependent intercept. For example, if $D_1 = 1$ the others D_i are 0, evaluating the first time in the both models.

Models for binary response and discrete survival data

OBSERVATIONS:

- There are more than one intercept, in particular it has one for every moment of time, so it is like a time-dependent intercept. For example, if $D_1 = 1$ the others D_i are 0, evaluating the first time in the both models.
- The values of Y comes from Bernoulli variables which are not identically distributed

Software

Models with binary response may be estimated using the standard functions/procedures:

Model	R	SAS
Logit	`glm(,family=binomial)`	`proc logistic`
Cloglog	`glm(,family=binomial(link= ‘‘ cloglog’’))`	`proc logistic,link=cloglog`

Real example

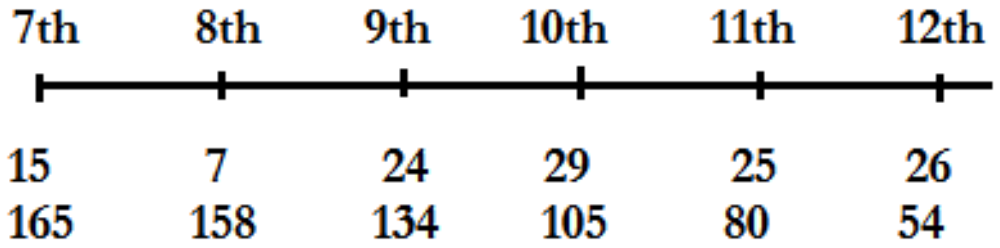
Capaldi, Crosby and Stoolmiller (1996) study: Predicting the timing of first sexual intercourse for at-risk adolescent males. This data was analized in Singer and Willet (2003)

- 180 school boys were tracked from the 7th through the 12th grade
- Outcome: when they had sex for the first time
- At the end of the follow-up in 12th grade, 54 boys (30%) were still virgins. These observations are censored.
- Several characteristics were available in the dataset but we will focus on the Parental transition variable (PT).

Example: Descriptive Results

$n = 180$

% cens$=\frac{54}{180} = 30\%$



Time	PT=0	PT=1	Total
7	2 13.33%	13 86.67%	15
8	2 28.57%	5 71.43%	7
9	8 33.33%	16 66.67%	24
10	8 27.59%	21 72.41%	29
11	10 40%	15 60%	25
12	42 52.5%	38 47.5%	80
Total	72	108	180

Example: Extended dataset

id	T^{order}	Time	Cens	PT	D_1	D_2	D_3	D_4	D_5	D_6	Y
1	1	7	1	1	1	0	0	0	0	0	1
2	1	7	1	0	1	0	0	0	0	0	1
...	...	...	...	...	...	...	...	...	...		
16	2	8	1	1	1	0	0	0	0	0	0
16	2	8	1	1	0	1	0	0	0	0	1
17	2	8	1	1	1	0	0	0	0	0	0
17	2	8	1	1	0	1	0	0	0	0	1
...	...	...	...	...	...	...	...	...	...	...	...

Example: Models

With the exentended dataset models for binary response with logit link and link cloglog that will apply:

- Link logit:

$$ln(\frac{h}{1-h}) = \alpha_1 D_1 + \alpha_2 D_2 + \alpha_3 D_3 + \alpha_4 D_4 + \alpha_5 D_5 + \alpha_6 D_6 + \beta_l PT$$

- Link cloglog:

$$ln(-ln(1-h)) = \eta_1 D_1 + \eta_2 D_2 + \eta_3 D_3 + \eta_4 D_4 + \eta_5 D_5 + \eta_6 D_6 + \beta_{cl} PT$$

Example: Results

Method	$\hat{\beta}$	LCI_β	UCI_β
Breslow	0.695	0.314	1.077
Efron	0.778	0.395	1.16
EPL	0.867	0.442	1.29
AL	0.782	0.399	1.166
Logit	0.874	0.455	1.309
Cloglog	0.785	0.41	1.180

Real example: PH vs Models for binary response

Assuming a Proportional Hazard models (Cox, 1972) and using the standard approaches for tied data, there is a relationship between the effect of covariate among models:

Cox model	Models for binary response
$\hat{\beta}_{EPL}$	$\hat{\beta}_{I}$
$\hat{\beta}_{AL}$	$\hat{\beta}_{cI}$

Real example: PH vs Models for binary response

Assuming a Proportional Hazard models (Cox, 1972) and using the standard approaches for tied data, there is a relationship between the effect of covariate among models:

Cox model	Models for binary response
$\hat{\beta}_{EPL}$	$\hat{\beta}_{I}$
$\hat{\beta}_{AL}$	$\hat{\beta}_{cI}$

- EPL: real discrete time variable
- AL: grouped continuous time variable

Real example: PH vs Models for binary response

Assuming a Proportional Hazard models (Cox, 1972) and using the standard approaches for tied data, there is a relationship between the effect of covariate among models:

Cox model	Models for binary response
$\hat{\beta}_{EPL}$	$\hat{\beta}_{I}$
$\hat{\beta}_{AL}$	$\hat{\beta}_{cI}$

- EPL: real discrete time variable
- AL: grouped continuous time variable

NOTE: Cox model does not assume that the hazard function is a probability

Real example: PH vs Models for binary response

Method	$\hat{\beta}$	LCI_β	UCI_β
Breslow	0.695	0.314	1.077
Efron	0.778	0.395	1.16
EPL	**0.867**	**0.442**	**1.29**
AL	0.782	0.399	1.166
Logit	**0.874**	**0.455**	**1.309**
Cloglog	0.785	0.41	1.180

Real example: PH vs Models for binary response

Method	$\hat{\beta}$	LCI_β	UCI_β
Breslow	0.695	0.314	1.077
Efron	0.778	0.395	1.16
EPL	0.867	0.442	1.29
AL	**0.782**	**0.399**	**1.166**
Logit	0.874	0.455	1.309
Cloglog	**0.785**	**0.41**	**1.180**

Conclusions

Advantages of using models for binary response instead of Cox models with ties:

- Computationally faster
- Same estimations than the ones coming assuming a PH model
- These models takes into account that the risk is a probability
- Easy implementation in the usual software

It is needed having the extended dataset from the original data

Bibliography

- Klein, J.P and Moeschberger M.L, *Survival Analysis: Techniques for Censored and Truncated Data* (2nd Edition). Springer (2003).
- D.R Cox and D.Oakes,*Analysis of Survival Data*.Chapman and Hall.
- Hosmer, D. and S. Lemeshow . *Applied Logistic Regression*(2nd Edition). New York: John Wiley & Sons (2000)
- P. McCullagh and Nelder, J. A. Nelder FRS, *Generalized Linear Models* (2nd Edition). Chapman and Hall.
- Therneau, T.M and Grambsch, P.M , *Modeling Survival Data. Extending the Cox Model*. Springer (2000)
- Singer,J.D and Willett, J.B , *Applied Longitudinal Data Analysis. Modeling change and event occurrence*. OXFORD University Express (2003). Springer (2000)

Questions

THANK YOU FOR YOUR ATTENTION!