Conditional Models with Time-Dependent Coefficients under Censoring and Truncation

BIANCA TEODORESCU

Institut de Statistique, Université catholique de Louvain, BELGIUM

Séminaire Jeunes Chercheurs, 1 February 2008

Outline of the talk

- the model
- LS estimation
- goodness-of-fit test
- bootstrap approximation
- real-data analysis

The Model

For each individual i, $i = 1, \ldots, n$, observations consist of $(Z_i, \delta_i, T_i, X_i)$ for $Z_i \geq T_i$, where

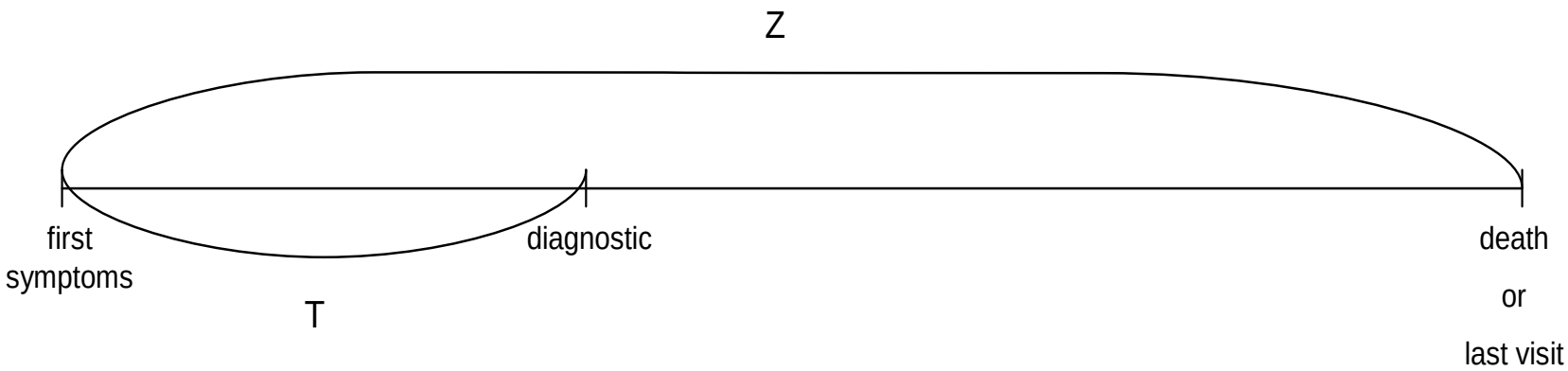
- $Z_i = \min\{Y_i, C_i\} \overset{i.i.d.}{\sim} H$
  with • Y_i = the life-time, $Y \overset{i.i.d.}{\sim} F$
  • C_i = the censoring-time, $C \overset{i.i.d.}{\sim} G$
- $\delta_i = I_{\{Y_i \leq C_i\}}$
- T_i = the truncation-time, $T \overset{i.i.d.}{\sim} L$
- X_i = the covariates
- (Z_i, T_i) is only observed if $Z_i \geq T_i$

Assumption: Y independent of (T, C), given X

Example

Description of the data:

- retrospective study (using the archives of the hospital)
- 945 patients diagnosed with gastric adenocarcinoma between january 1975 and december 1993 in Hospital Xeral-Calde of Lugo, Spain
- patiens were followed up until 1st of august 1996
- TNM classification
  (extension of tumor: initial, advanced or not known)
- we are interested in the time between the first symptoms and death of the patient



For each individual i ($i = 1, \ldots, 945$) we observed:

- $Z_i =$ time from first symptoms to death or last control
- $\delta_i =$ death indicator $(0 =$ alive, $1 =$ dead$)$
- $T_i =$ time between first symptoms and diagnosis (truncation time)
- TNM classification ($1 =$initial, $2 =$advanced, $3 =$not known)
- age at diagnosis

The Model

Interested in the relation between $S(z|\boldsymbol{X}) = P(Y > z|\boldsymbol{X})$ and $\boldsymbol{X}$

Use Generalized Linear Models:

$$\phi(S(z|\boldsymbol{X})) = \boldsymbol{\beta}'(z)\boldsymbol{X}$$

where: - ϕ = **known** link function (monotonic & differentiable in $[0, 1]$)

- $\boldsymbol{\beta}(z)$ = a $(p+1) \times 1$ - vector of **unknown** time-dependent regression parameters

- $S(z|\boldsymbol{X})$ = the **unknown** survival function that will be estimated **nonparametricaly**

The Model

$$\phi(S(z|\boldsymbol{X})) = \boldsymbol{\beta}'(z)\boldsymbol{X}$$

Examples of link functions:

1. **Proportional hazards:** $\phi(u) = \log(-\log(u))$

$$\phi(S(z|\boldsymbol{X})) = \log(H(z|\boldsymbol{X})) = \log(H_0(z)) + \beta_1(z)X + \ldots + \beta_p(z)X^p$$

2. **Log-logistic (proportional odds model):** $\phi(u) = \log\left(\frac{u}{1-u}\right)$

$$\phi(S(z|\boldsymbol{X})) = \log\left(\frac{S(z|\boldsymbol{X})}{1 - S(z|\boldsymbol{X})}\right) = \log(H_0(z)) + \beta_1(z)X + \ldots + \beta_p(z)X^p$$

3. **Additive hazards:** $\phi(u) = -log(u)$

$$\phi(S(z|\boldsymbol{X})) = H(z|\boldsymbol{X}) = \beta_0(z) + \beta_1(z)X + \ldots + \beta_p(z)X^p$$

LS estimation

If $\boldsymbol{X} = X$ (one-dimensional continuous covariate)

$$\phi(S(z|X)) = \beta_0(z) + \beta_1(z)X + \ldots + \beta_p(z)X^p$$

Estimation procedure for $\beta_j(z)$ ($j = 0, \ldots, p$) for a fixed z:

Step 1: Use the estimator of Iglesias-Peréz and González-Manteiga (1999) for the conditional distribution (involves smoothing) to obtain:

$$\hat{S}_n(z|X_1), \ldots, \hat{S}_n(z|X_n)$$

where $\hat{S}_n(z|x) = 1 - \hat{F}_n(z|x) = \prod_{i=1}^{n} \left(1 - \frac{I_{\{Z_i \leq z, \delta_i = 1\}} B_{ni}(x)}{\sum_{j=1}^{n} I_{\{T_j \leq u \leq Z_j\}} B_{nj}(x)}\right)$, $B_{ni}(x)$ are Nadaraya-Watson weights, K is a known kernel that depends on a bandwidth sequence $h = h_n \to 0$.

LS estimation

Step 2: Use classical LS estimation (replacing $S(z|X)$ by $\hat{S}_n(z|X)$) to obtain the estimators of $\beta_j(z)$ ($j = 0, \ldots, p$), denoted by $\hat{\boldsymbol{\beta}}(z) = (\hat{\beta}_0(z), \ldots, \hat{\beta}_p(z))^{'}$, where

$$\hat{\boldsymbol{\beta}}_j(z) = (\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}\phi(\hat{S}(z|X_j))$$

Repeat the same procedure for all possible z (in practice only uncensored points Z_i are needed).

Asymptotic properties of $\hat{\boldsymbol{\beta}}(z)$

Theorem: *Under certain assumptions, the estimator* $\hat{\boldsymbol{\beta}}(z)$ *is consistent and* $n^{1/2}(\hat{\boldsymbol{\beta}}(z) - \boldsymbol{\beta}(z))$ *is asymptotically normal with zero mean and a certain variance-covariance matrix* $\boldsymbol{A}$.

Remarks:

- a similar LS estimation procedure can be used for a model with only discrete covariates or with a combination of discrete covariables and a one-dimensional continuous one
- dealing with a one-dimensional continuous covariate, involves choosing a smoothing parameter (h) for the estimation of $S(z|X)$ (*a solution:* use bootstrap)
- a ML estimation procedure for the model was proposed by Jung(1996) and Subramanian (2001, 2004)

Simulation studies

LS estimator when we have a one-dimensional continuous covariate, censoring and truncation

Our model: $\phi(S(z|X)) = \beta_0(z) + \beta_1(z)X$

- $X \sim U[0,1]$
- $Y|_{X=x} \sim \exp(4x)$
- $C|_{X=x} \sim \exp(dx)$, with d that gives different censoring %
- $T|_{X=x} \sim \exp(rx)$, with r that gives different truncation %
- $\phi(u) = \log(u)$ (additive risk model) $\Rightarrow$ true model: $\phi(S(z|X)) = -4zX$
- $\phi(S(z|X=0)) = \beta_0(z)$ and $\phi(S(z|X=1)) = \beta_0(z) + \beta_1(z)$
- sample size $n = 100$
- $M = 1000$ Monte Carlo simulations
- h is chosen by bootstrap among $\{0.2, 0.25, 0.3, 0.35, 0.4\}$

Simulation studies for a fixed $z = 0.3$

- **only censoring**

Cens	$\beta_0(z) = 0$		$\beta_1(z) = -1.2$		$\phi(S(z\|x_2)) = -1.2$	
perc	Bias	MSE	Bias	MSE	Bias	MSE
20	-0.0974	0.0234	0.2799	0.1595	0.1826	0.0758
40	-0.1225	0.0281	0.4068	0.2343	0.2843	0.1176

- **censoring and truncation**

Cens	Trunc	$\beta_0(z) = 0$		$\beta_1(z) = -1.2$		$\phi(S(z\|X = 1)) = -1.2$	
perc	perc	Bias	MSE	Bias	MSE	Bias	MSE
20	10	-0.1391	0.0488	0.2708	0.2028	0.1317	0.0755
	20	-0.1524	0.0583	0.2923	0.2283	0.1399	0.0851
40	10	-0.1506	0.0494	0.3072	0.2178	0.1566	0.0859
	20	-0.1692	0.0662	0.3126	0.2370	0.1434	0.0853

Goodness-of-Fit Tests

Hypothesis: suppose ϕ and p known

$$H_0 : \exists \boldsymbol{\beta}(z) \in \mathbb{R}^{p+1} \text{ s.t. } \phi(S(z|X)) = \beta_0(z) + \beta_1(z)X + \ldots + \beta_p(z)X^p$$

$$H_1 : \text{The model does not hold for any } \boldsymbol{\beta}(z) \in \mathbb{R}^{p+1}$$

Test statistic:

$$\boxed{T_n = n\sqrt{h} \int_0^\tau \hat{\Phi}_n(\hat{\boldsymbol{\beta}}(z))dz}$$

where

$$\hat{\Phi}_n(\hat{\boldsymbol{\beta}}(z)) = \frac{1}{n} \sum_{r=1}^{n} \left(\phi(\hat{S}_n(z|X_r)) - (\hat{\beta}_0(z) + \hat{\beta}_1(z)X_r + \ldots + \hat{\beta}_p(z)X_r^p) \right)^2$$

Goodness-of-Fit Tests

Theorem: *Under H_0 and some regularity conditions,*

$$\boxed{T_n - b_{0h} \xrightarrow{d} N(0, V)}$$

with $$b_{0h} = h^{-1/2} K^{(2)}(0) \int_0^\tau \int_x \sigma^2(z, x) dx dz$$

$$V = K^{(4)}(0) \int_0^\tau \int_x \sigma^4(z, x) dx dz$$

Remark: In practice obtaining the critical values is not very easy

solution: use bootstrap

Bootstrap Approximation

Under H_0 we have

$$\phi(S(z|X)) = \boldsymbol{\beta}'(z)\boldsymbol{X}$$

How to bootstrap?

1. choose a "pilot" bandwidth g to estimate

- $S(z|X_j) \Rightarrow \hat{S}_g(z|X_j)$
- $G(z|X_j) \Rightarrow \hat{G}_g(z|X_j)$
- $L(z|X_j) \Rightarrow \hat{L}_g(z|X_j)$

2. using $\hat{S}_g(z|X_j)$ estimate $\boldsymbol{\beta}(z)$ by the LS procedure $\Rightarrow \hat{\boldsymbol{\beta}}_g(z)$

then re-estimate $S(z|X_j)$ by $\tilde{S}_g(z|X_j) = \phi^{-1}(\hat{\boldsymbol{\beta}}_g'(z)\boldsymbol{X})$

Bootstrap Approximation

3. For fixed B and for $b = 1, \ldots, B$,

a) for each X_j, $j = 1, \ldots, n$, draw samples from

- $\tilde{S}_g(z|X_j) \Rightarrow Y^*_{j,b}$
- $\hat{G}_g(z|X_j) \Rightarrow C^*_{j,b}$
- $\hat{L}_g(z|X_j) \Rightarrow T^*_{j,b}$

$$\Rightarrow (Z^*_{1,b}, \delta^*_{1,b}, T^*_{1,b}, X_1), \ldots, (Z^*_{n,b}, \delta^*_{n,b}, T^*_{n,b}, X_n)$$

b) use bandwidth h to estimate $\hat{S}^*_{h,b}(z|X) \Rightarrow \hat{\boldsymbol{\beta}}^*_{h,b}(z)$ by the LS procedure

c) compute

$$T^*_n = n\sqrt{h} \int_0^\tau \hat{\Phi}_n(\hat{\boldsymbol{\beta}}^*_{h,b}(z))dz \simeq n\sqrt{h} \sum_{i=1}^{n-1} \hat{\Phi}_n(\hat{\boldsymbol{\beta}}^*_{h,b}(Z_i))(Z_{i+1} - Z_i)$$

4. $T^*_{n,[(1-\alpha)B]}$ approximates the $(1-\alpha)$ quantile of the distribution of T_n under H_0.

Simulation Studies

$$H_0 : \log(S(z|X)) = -4zX$$
$$H_1 : \log(S(z|X)) = -4zX + aa \cdot z \sin\left(\frac{\pi X}{2}\right)$$

- $X \sim U[4, 10]$
- $Y|_{X=x} \sim \exp(-4x + aa \cdot \sin\left(\frac{\pi x}{2}\right))$, with aa that indicates different departures from H_0
- $C|_{X=x} \sim \exp(dx)$, with d that gives different censoring %
- $T|_{X=x} \sim \exp(rx)$, with r that gives different censoring %
- $\phi(u) = \log(u)$ (additive risk model)
- sample size $n = 100$
- Monte Carlo simulations $M = 1000$
- bootstrap resamples $B = 500$
- $h \in \{1.2, 1.5, 1.8, 2.1, 2.4\}$ and $g = 2h$

Simulation Studies

Cens		H_0	H_1				
perc	h	aa=0	aa=4	aa=8	aa=12	aa=16	aa=20
20	1.2	0.031	0.077	0.200	0.445	0.752	0.931
	1.5	0.029	0.114	0.261	0.531	0.788	0.945
	1.8	0.039	0.113	0.305	0.534	0.820	0.945
	2.1	0.075	0.122	0.315	0.525	0.801	0.943
	2.4	0.080	0.140	0.275	0.422	0.650	0.870
40	1.2	0.024	0.051	0.107	0.225	0.420	0.660
	1.5	0.021	0.050	0.151	0.294	0.519	0.762
	1.8	0.038	0.065	0.180	0.325	0.535	0.757
	2.1	0.051	0.082	0.157	0.296	0.506	0.701
	2.4	0.068	0.106	0.134	0.231	0.376	0.562

Simulation Studies

Cens perc	Trunc perc	h	H_0 aa=0	H_1 aa=4	aa=8	aa=12	aa=16	aa=20
20	10	1.2	0.051	0.067	0.191	0.385	0.662	0.893
		1.5	0.056	0.104	0.238	0.432	0.733	0.939
		1.8	0.057	0.107	0.227	0.445	0.768	0.960
		2.1	0.064	0.116	0.244	0.441	0.804	0.923
		2.4	0.068	0.139	0.269	0.395	0.693	0.857
	20	1.2	0.061	0.076	0.167	0.373	0.651	0.872
		1.5	0.026	0.092	0.199	0.421	0.702	0.931
		1.8	0.037	0.071	0.233	0.433	0.794	0.901
		2.1	0.036	0.111	0.269	0.415	0.692	0.912
		2.4	0.038	0.091	0.270	0.372	0.833	0.842

Simulation Studies

Cens perc	Trunc perc	h	H_0 aa=0	H_1 aa=4	aa=8	aa=12	aa=16	aa=20
40	10	1.2	0.035	0.059	0.081	0.213	0.398	0.676
		1.5	0.047	0.063	0.141	0.286	0.461	0.691
		1.8	0.023	0.054	0.157	0.316	0.536	0.734
		2.1	0.024	0.073	0.156	0.278	0.497	0.706
		2.4	0.025	0.095	0.143	0.234	0.366	0.535
	20	1.2	0.041	0.060	0.090	0.204	0.412	0.645
		1.5	0.046	0.059	0.135	0.273	0.458	0.692
		1.8	0.031	0.053	0.141	0.312	0.556	0.725
		2.1	0.028	0.071	0.122	0.226	0.492	0.699
		2.4	0.032	0.092	0.127	0.330	0.359	0.522

Real-data Analysis

Description of the data:

- 945 patients diagnosed with gastric adenocarcinoma between 1975 and 1993 in Hospital Xeral-Calde of Lugo, Spain
- 2 TNM classifications (extension of tumor)

For each individual i ($i = 1, \ldots, 945$) we observed:

- $Z_i =$ time from first symptoms to death or last control
- $\delta_i =$ death indicator $(0 =$ alive, $1 =$ dead$)$
- $T_i =$ time between first symptoms and diagnosis (truncation time)
- TNM classification ($1 =$initial, $2 =$advanced, $3 =$not known)
- age at diagnosis

Real-data Analysis

$H_0 : \exists \boldsymbol{\beta}(z) \in \mathbb{R}^4$ s.t.

$$\phi(S(z|\boldsymbol{X})) = \beta_0(z) + \beta_1(z)X_1 + \beta_2(z)X_2 + \beta_3(z)X_3$$

H_a : The model does not hold for any $\boldsymbol{\beta}(z) \in \mathbb{R}^4$

where ϕ is one of the following:

- $\phi_1(u) = -\log(u)$ (additive hazards model)
- $\phi_2(u) = \log(-\log(u))$ (proportional hazards)
- $\phi_3(u) = \log(\frac{u}{1-u})$ (proportional odds)

$$X_1 = \begin{cases} 1 & , \text{ TNM=2} \\ 0 & , \text{ otherwise} \end{cases} \qquad X_2 = \begin{cases} 1 & , \text{ TNM=3} \\ 0 & , \text{ otherwise} \end{cases}$$

$X_3 =$ Age at diagnosis $- 67.71606$

$h = 35$ and $g = 2h$ (min age $= 22$, max age $= 98$)

Real-data Analysis

- $B = 500$ bootstrap simulations
- $\alpha = 0.05$

- for ϕ_1 (additive hazards) p-value=$0.02 \Rightarrow$ reject H_0
- for ϕ_2 (proportional hazards) p-value=$0.54 \Rightarrow$ do not reject H_0
- for ϕ_3 (proportional odds) p-value=$0.006 \Rightarrow$ reject H_0

$\Rightarrow$ Keep the proportional hazards model.

Proportional hazards model

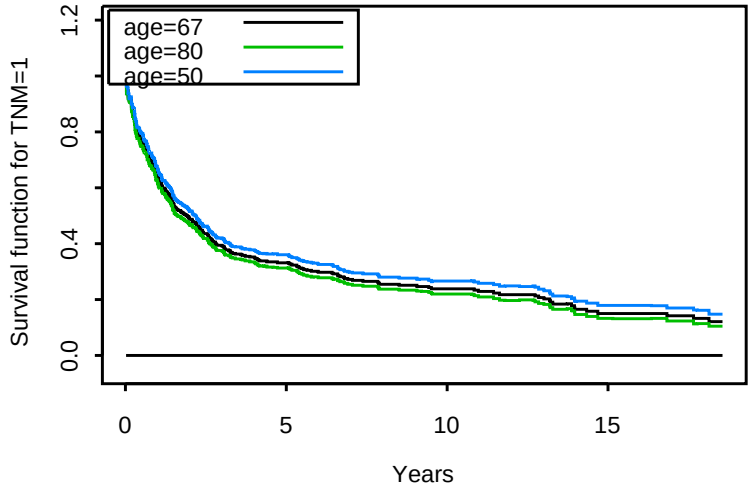
$$\log(H(z|\boldsymbol{X})) = \beta_0(z) + \beta_1(z)X_1 + \beta_2(z)X_2 + \beta_3(z)X_3$$

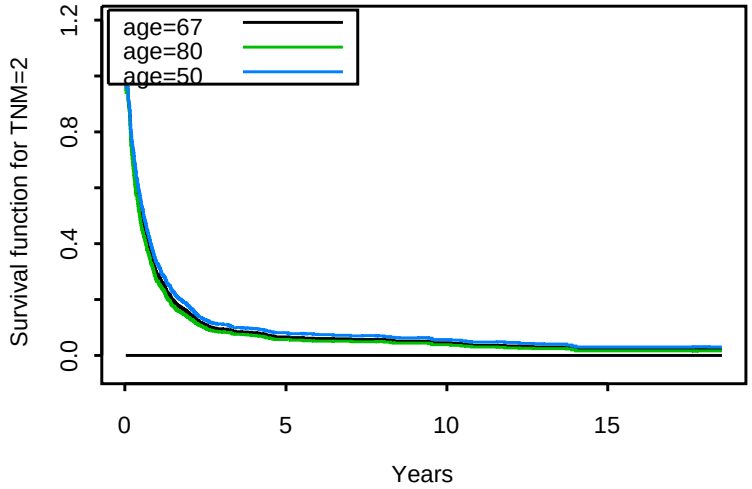
or

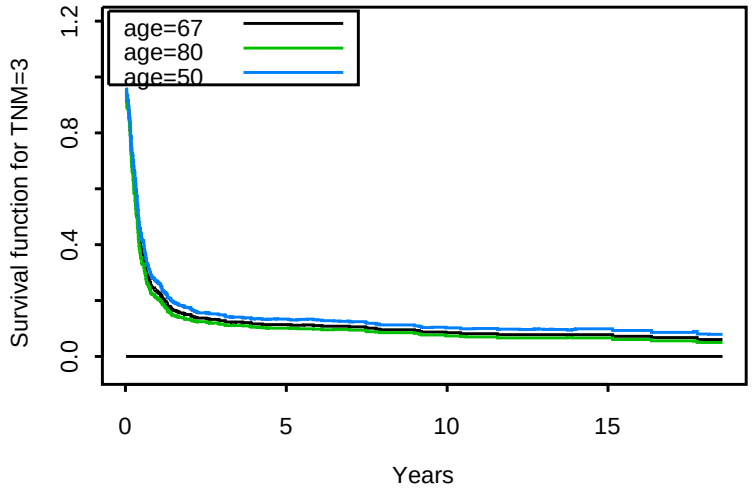
$$H(z|\boldsymbol{X}) = H_0(z)\exp(\beta_0(z) + \beta_1(z)X_1 + \beta_2(z)X_2 + \beta_3(z)X_3)$$

- $H_1(Z|X_3) = H_0(z)\exp(\beta_3(z)X_3)$ for $TNM = 1$
- $H_2(Z|X_3) = H_0(z)\exp(\beta_1(z) + \beta_3(z)X_3)$ for $TNM = 2$
- $H_3(Z|X_3) = H_0(z)\exp(\beta_2(z) + \beta_3(z)X_3)$ for $TNM = 3$

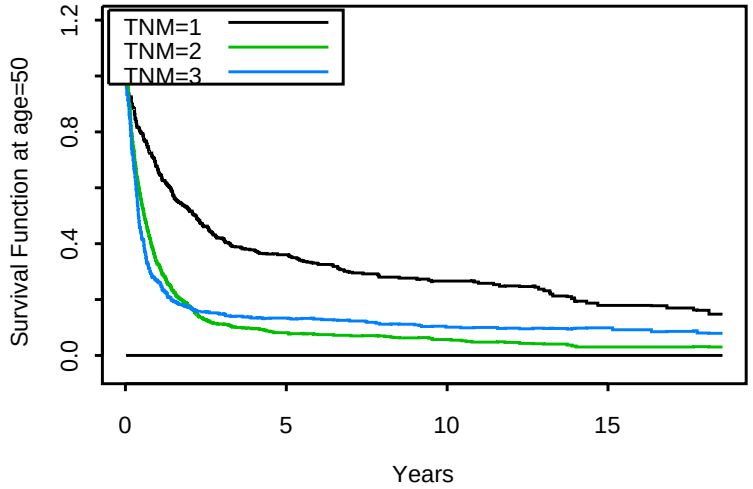
Comparison of survival curves

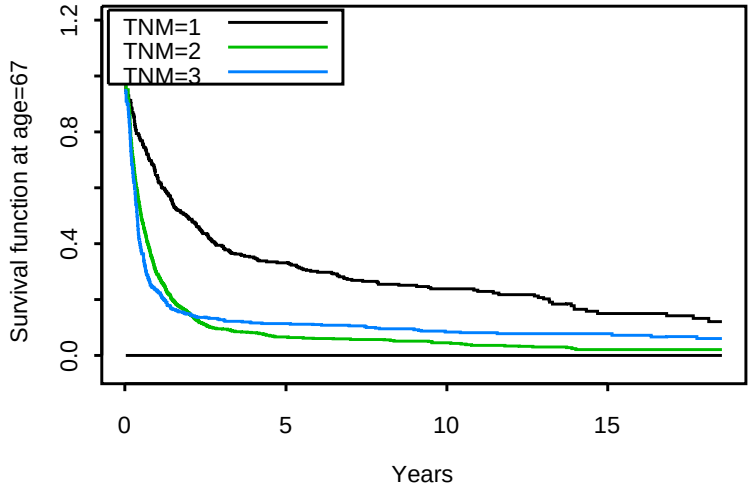


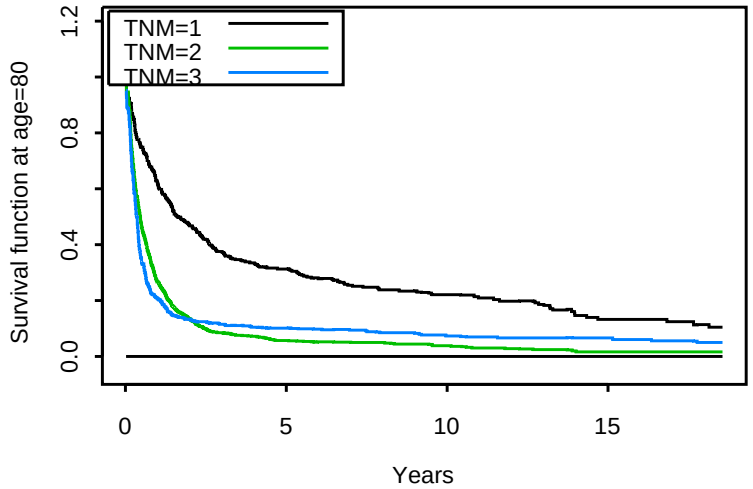




Comparison of survival curves







References

- CASARIEGO VALES, E., PITA FERNÁNDEZ, S., and all (2001) - Supervivencia en 2.334 pacientes con cáncer gástrico y factores que modifican el prognóstico, *Med. Clin. (Barc)*, 117, 361-365
- CAO, R. and GONZÁLEZ-MANTEIGA, W. (2002) - Goodness-of-fit tests for conditional models under censoring and truncation. *Journal of Econometrics*, 143, 166-190.
- IGLESIAS-PÉREZ, C. and GONZÁLEZ-MANTEIGA, W. (1999) - Strong Representation of a Generalized Product-Limit Estimator for Truncated and Censored Data with Some Applications, *Nonparametric Statistics*, 10, 213-244
- TEODORESCU, B., VAN KEILEGOM, I. and CAO, R. (2005) - Generalized Conditional Linear Models under Left Truncation and Right Censoring, (submitted)