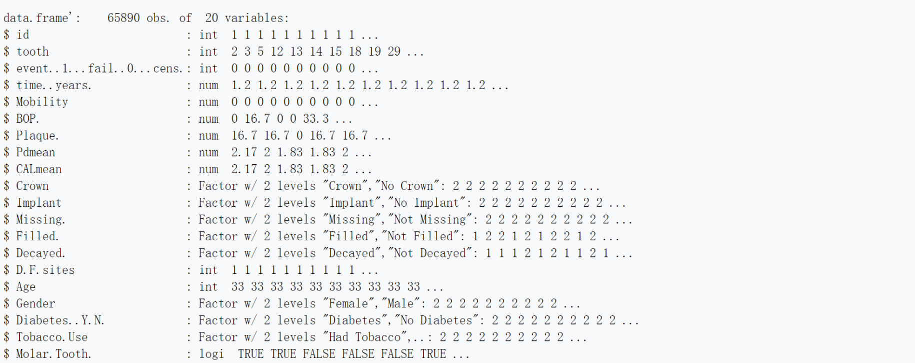
The Cure Model for Teeth Data

Fengyu Zhang

Institute of Statistics & Big Data
Renmin University of China

Data Overview

The data is about a kind of dental disease. The dataset contains 65890 observations of 20 variables. Here is a overview of the dataset:

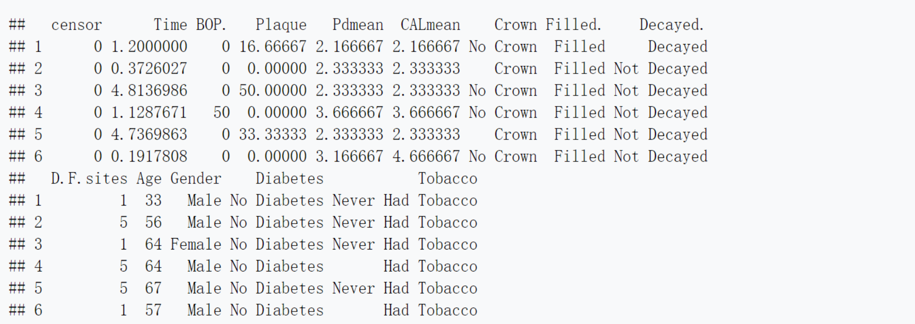


```
data.frame' :    65890 obs. of  20 variables:
$ id                          : int  1 1 1 1 1 1 1 1 1 1 ...
$ tooth                       : int  2 3 5 12 13 14 15 18 19 29 ...
$ event..1...fail..0...cens.  : int  0 0 0 0 0 0 0 0 0 0 ...
$ time..years.                : num  1.2 1.2 1.2 1.2 1.2 1.2 1.2 1.2 1.2 1.2 ...
$ Mobility                    : num  0 0 0 0 0 0 0 0 0 0 ...
$ BOP.                        : num  0 16.7 0 0 33.3 ...
$ Plaque.                     : num  16.7 16.7 0 16.7 16.7 ...
$ Pdmean                      : num  2.17 2 1.83 1.83 2 ...
$ CALmean                     : num  2.17 2 1.83 1.83 2 ...
$ Crown                       : Factor w/ 2 levels "Crown","No Crown": 2 2 2 2 2 2 2 2 2 2 ...
$ Implant                     : Factor w/ 2 levels "Implant","No Implant": 2 2 2 2 2 2 2 2 2 2 ...
$ Missing.                    : Factor w/ 2 levels "Missing","Not Missing": 2 2 2 2 2 2 2 2 2 2 ...
$ Filled.                     : Factor w/ 2 levels "Filled","Not Filled": 1 2 2 1 2 1 2 2 1 2 ...
$ Decayed.                    : Factor w/ 2 levels "Decayed","Not Decayed": 1 1 1 2 1 2 1 1 2 1 ...
$ D.F.sites                   : int  1 1 1 1 1 1 1 1 1 1 ...
$ Age                         : int  33 33 33 33 33 33 33 33 33 33 ...
$ Gender                      : Factor w/ 2 levels "Female","Male": 2 2 2 2 2 2 2 2 2 2 ...
$ Diabetes..Y.N.              : Factor w/ 2 levels "Diabetes","No Diabetes": 2 2 2 2 2 2 2 2 2 2 ...
$ Tobacco.Use                 : Factor w/ 2 levels "Had Tobacco",..: 2 2 2 2 2 2 2 2 2 2 ...
$ Molar.Tooth.                : logi  TRUE TRUE FALSE FALSE FALSE TRUE ...
```

Specific details are given in "Data_Description.txt".

Data Overview

In order to be operable, we select the observations with “tooth=2” which is a molar and rename some of the columns for convenience. Also, we remove some irrelevant variables. Here is the overview of the dataset that has been preprocessed:

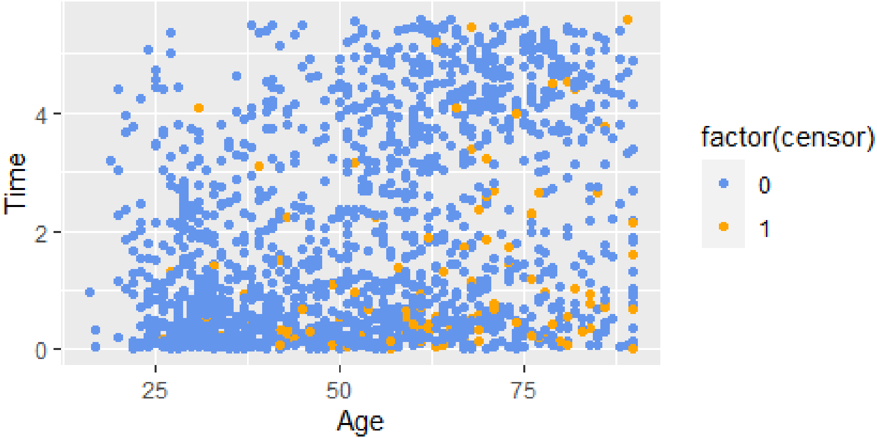


```
##   censor      Time BOP.   Plaque   Pdmean  CALmean    Crown Filled.    Decayed.
## 1      0 1.2000000    0 16.66667 2.166667 2.166667 No Crown  Filled     Decayed
## 2      0 0.3726027    0  0.00000 2.333333 2.333333    Crown  Filled Not Decayed
## 3      0 4.8136986    0 50.00000 2.333333 2.333333 No Crown  Filled Not Decayed
## 4      0 1.1287671   50  0.00000 3.666667 3.666667 No Crown  Filled Not Decayed
## 5      0 4.7369863    0 33.33333 2.333333 2.333333    Crown  Filled Not Decayed
## 6      0 0.1917808    0  0.00000 3.166667 4.666667 No Crown  Filled Not Decayed
##   D.F.sites Age Gender    Diabetes           Tobacco
## 1         1  33   Male No Diabetes Never Had Tobacco
## 2         5  56   Male No Diabetes Never Had Tobacco
## 3         1  64 Female No Diabetes Never Had Tobacco
## 4         5  64   Male No Diabetes       Had Tobacco
## 5         5  67   Male No Diabetes Never Had Tobacco
## 6         1  57   Male No Diabetes       Had Tobacco
```

*In practice, we need to transform the binary variables to 0-1 variables.

Data Overview

A first look at the censor rate.



Each point represents an observation and the 92% of the observations are censored. We can see that the censor rate is high and a cure model seems therefore appropriate for these data.

Cure model

- In survival analysis, one usually assumes that all subjects under study will eventually experience the event of interest.
- When the event of interest is the time until a patient progresses or relapses from a certain disease, then patients who are cured from the disease will never experience the event.
- Their survival time will be set to infinity.
- Cure models are survival models that have been developed to take this feature into account.

Cure model

Boag (1949) and Farewell (1982) originally proposed a mixture cure model which assumes that the survival function has the following form:

$$S(t|x,z) = P(T > t|x,z) = 1 - p(x) + p(x)S_u(t|z), \quad t \geqslant 0, \tag{1}$$

where

- $p(x) = \mathbb{P}(B = 1|X = x)$ is the conditional probability of being uncured (often referred to as the 'incidence') with $B = I(T < \infty)$ the latent uncured status.
- $S_u(t|z) = \mathbb{P}(T > t|B = 1, Z = z)$ is the conditional survival function for the uncured subjects (often referred to as the 'latency')

Here, the covariate vectors X and Z can contain (partially) the same covariates, but they can also be completely different.

Cure model

For the part of latency ($S_u(t|z)$), we consider a Cox propotional hazards (PH) model (Cox 1972) with the following form

$$S_u(t|z) = S_0(t)^{\exp(\beta^{\mathrm{T}} z)} \tag{2}$$

where $S_0(t) = \mathbb{P}(T > t|B = 1)$ is the baseline conditional survival function. The conditional hazard function is given by

$$\lambda_u(t|z) = \lambda_0(t) \exp(\beta^{\mathrm{T}} z),$$

where $\lambda_0(t)$ is the baseline hazard function.

Cure model

For the part of incidence (uncured rate $p(x)$), two models are considered.

1. Logistic model (common assumption):

$$p(x) = \frac{\exp(\gamma_0 + \gamma^{\mathrm{T}} x)}{1 + \exp(\gamma_0 + \gamma^{\mathrm{T}} x)}$$

   for some parameter vector γ and an intercept γ_0. The logistic model is easy to interpret and estimate

2. Single-index model:

$$p(x) = g(\gamma^{\mathrm{T}} x)$$

   for any smooth link function g with values between 0 and 1. The single-index model has nonparametric link function and therefore much more flexibel than the logistic model. Besides, it does not suffer from the curse-of-dimensionality problems.

Estimation

In survival analysis, we usually observe the couple (Y, δ) instead of the survival time T, where $Y = \min(T, C), \delta = I(T \leqslant C)$, and C is the censoring time. As often, we assume T and C are independent given the covariates X, Z. Denote $(Y_i, \delta_i, X_i, Z_i), i = 1, \ldots, n$ be i.i.d. realizations of (Y, δ, X, Z), the likelihood function takes the form

$$L = \prod_{i=1}^{n} \{p(X_i) f_u(Y_i|Z_i)\}^{\delta_i} \cdot [\{1 - p(X_i)\} + p(X_i) S_u(Y_i|Z_i)]^{1-\delta_i} . \qquad (3)$$

where $f_u(t|z) = -(d/dt) S_u(t|z)$ is the conditional density function. The likelihood has two types of contributions: from censored and from the uncensored observations.

Estimation

We use EM algorithm to handle the fact that the cure status B_i is unobserved. The complete-data likelihood is given by

$$
\begin{aligned}
L_c = & \prod_{i=1}^{n} \{p(X_i)\lambda_u(Y_i|Z_i)S_u(Y_i|Z_i)\}^{B_i\delta_i} \times \\
& \left[\{1 - p(X_i)\}^{1-B_i} + \{p(X_i)S_u(Y_i|Z_i)\}^{B_i}\right]^{1-\delta_i}
\end{aligned}
\tag{4}
$$

Then we need to calculate the conditional expectation of the log-likelihood given the observed data and the current parameter values. As the log-likelihood is linear in B, it is the same as computing

$$
\mathbb{E}(B_i|\mathcal{O}, \Theta^{(m-1)}) := W_i^{(m)},
$$

where $\mathcal{O} = \{(Y_i, \delta_i, X_i, Z_i), i = 1, \ldots, n\}$ are observed data and $\Theta = (\gamma, \beta, S_0)$ for logistic model and $\Theta = (\gamma, \beta, S_0, g)$ for single-index model.

Estimation

In M-step, we maximize the expected log-likelihood which is obtained by replacing B_i by $W_i^{(m)}$ in the equation (4):

$$
\begin{aligned}
\tilde{L}_c = & \prod_{i=1}^{n} \{p(X_i)\lambda_u(Y_i|Z_i)S_u(Y_i|Z_i)\}^{W_i^{(m)}\delta_i} \times \\
& \left[\{1 - p(X_i)\}^{1-W_i^{(m)}} + \{p(X_i)S_u(Y_i|Z_i)\}^{W_i^{(m)}}\right]^{1-\delta_i}.
\end{aligned}
\tag{5}
$$

After some algebra, $\tilde{L}_c$ can be written as the product of two parts:

$$
\begin{aligned}
\tilde{L}_c & = \prod_{i=1}^{n} \left[p(X_i)^{W_i^{(m)}}\{1 - p(X_i)\}^{1-W_i^{(m)}}\right] \times \prod_{i=1}^{n}\{\lambda_u(Y_i|Z_i)^{\delta_i} S_u(Y_i|Z_i)\}^{W_i^{(m)}} \\
& = \tilde{L}_1 \times \tilde{L}_2.
\end{aligned}
\tag{6}
$$

It can be maximized separately for the two parts of the model.

Estimation

Although the framework of the EM algotirhm is constructed, one problem is that how to estimate the parameters when we use the single-index model in the incidence part. Ichimura (1993) proposed a leave-one-out kernel estimator of $g(\gamma^{\mathrm{T}} X_i)$:

$$\sum_{j \neq i}^{n} \frac{K\left(\frac{\gamma^t X_i - \gamma^t X_j}{h}\right)}{\sum_{l \neq i}^{n} K\left(\frac{\gamma^t X_i - \gamma^t X_l}{h}\right)} B_j .$$

We need to replace B_j by $W_j^{(m)}$ obtained in the E-step and then the estimator becomes

$$\tilde{g}_{-i}^{(m)}\left(\gamma^t X_i\right) = \sum_{j \neq i}^{n} \frac{K\left(\frac{\gamma^t X_i - \gamma^t X_j}{h}\right)}{\sum_{l \neq i}^{n} K\left(\frac{\gamma^t X_i - \gamma^t X_l}{h}\right)} W_j^{(m)} \qquad (7)$$

The kernel estimator (7) is substitued in $\tilde{L}_1$, and γ is estimated by maximizing the likelihood.

Estimation

Another problem is that how to estimate the latency ($\tilde{L}_2$). Note that
$\tilde{L}_2 = \prod_{i=1}^n \left[\{\lambda_0 (Y_i) \exp (\boldsymbol{\beta}^t \boldsymbol{Z}_i)\}^{\delta_i} \exp \{-\Lambda_0 (Y_i) \exp (\boldsymbol{\beta}^t \boldsymbol{Z}_i)\} \right]^{W_i^{(m)}}$.
Sy and Taylor (2000) propose a profile likelihood approach to estimate β. First, given a fixed β, Λ_0 is estimated nonparametrically by

$$\sum_{j: Y_{(j)} \leq t} \frac{D_j}{\sum_{k \in R_j} W_k^{(m)} \exp (\boldsymbol{\beta}^t Z_k)}, \tag{8}$$

where $Y_{(j)}$ are order statitics, D_j is the number of events at time $Y_{(j)}$ and R_j is the risk set before $Y_{(j)}$. Second, we plug (8) in $\tilde{L}_2$, obtaining the partial likelihood

$$\breve{\mathrm{L}}_2 = \prod_{i=1}^n \left\{ \frac{\exp (\beta^t Z_i)}{\sum_{k \in R_i} W_k^{(m)} \exp (\beta^t Z_k)} \right\}^{\Delta_i} \tag{9}$$

The MLE of β denoted by $\hat{\beta}^{(m)}$ is obtained by maximizing (9). Then we plug $\hat{\beta}^{(m)}$ in (8) to obtain $\hat{\Lambda}_0^{(m)}(t)$. We do alternative iterations until convergence.

Data Application

Table: Parameter Estimations, Std.error and Wald's test

	SIC cure model			LC cure model		
Incidence	Estimate	Std.error	p-value	Estimate	Std.error	p-value
(intercept)	-	-	-	-4.54406	1.573907	0.003888
Age	0.56649	0.180257	0.0016741	0.031461	0.016599	0.058042
Gender	-0.05871	0.346062	0.8652774	-0.01342	0.55367	0.980664
BOP	0.6242	0.294932	0.0343093	1.692404	0.834437	0.04254
Plaque	-0.4325	0.269344	0.1083279	-0.00462	0.815525	0.995481
Pdmean	-0.08126	0.244503	0.7396307	0.152681	0.281289	0.587275
CALmean	0.303903	0.198951	0.12663	0.201554	0.165835	0.224218
latency	Estimate	Std.error	p-value	Estimate	Std.error	p-value
Age	-0.02421	0.010842	0.0255254	-0.02177	0.016181	0.178525
Gender	0.148134	0.220341	0.5013959	0.193319	0.572714	0.735703
BOP	0.815028	0.426396	0.055949	-0.29098	0.593112	0.623712
Plaque	-0.77723	0.392359	0.0476001	-0.68691	0.794442	0.387237
Pdmean	0.258177	0.160801	0.1083693	-0.02163	0.224441	0.923235
CALmean	0.343105	0.119735	0.0041631	0.359877	0.162896	0.027157

Data Application

- According to the table, for the latency part, the effects for age, gender, Plaque and CALmean have the same direction and the estimates are very close. Only CALmean affects significantly the survivial time of uncured subjects in both of the two models.
- For the incidence part, we compare the predicted error of the incidence. First we divided the dataset into a training and test subsut, following 2/3-1/3 recommendations of Hastie and Friedman (2009). We use the training set to estimate the parameters and calculate the prediction error which is given by

$$PE = -\sum_{j=1}^{n_{test}} \log\left[\hat{p}\left(x_j^{\text{test}}\right)^{\hat{w}_j}\left\{1-\hat{p}\left(x_j^{\text{test}}\right)\right\}^{1-\hat{w}_j}\right] \tag{10}$$

  After computing, the prediction error for the SIC model equals to 57.65, while it is equal to 70.93 for the LC model, which means that the SIC model performs better in predicting the uncured status.

Reference

Boag, J. W. "Maximum Likelihood Estimates of the Proportion of Patients Cured by Cancer Therapy." *Journal of the Royal Statistical Society. Series B (Methodological)*, 11(1):15–53 (1949).

Cox, D. R. "Regression Models and Life-Tables." *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220 (1972).

Farewell, V. T. "The Use of Mixture Models for the Analysis of Survival Data with Long-Term Survivors." *Biometrics*, 38(4):1041–1046 (1982).

Hastie, T. R., T. and Friedman, R. "The Elements of Statistical Learning: Data Mining, Inference, and Prediction." *New York: Springer* (2009).

Ichimura, H. "Semiparametric least squares (SLS) and weighted SLS estimation of single-index models." *Journal of Econometrics*, 58(1):71 – 120 (1993).

Sy, J. P. and Taylor, J. M. G. "Estimation in a Cox Proportional Hazards Cure Model." *Biometrics*, 56(1):227–236 (2000).