Week1: Survival Analysis

November 11, 2020

Contents

A. Consider the TRACE data of the timereg package.

We consider the time to death given by time and the event status defined from $status! = 0$. One important prognostic factor for death is VF (ventricular fibrillation). Some observations are censored and some subjects die, and $status! = 0$ describes this. Assume that we have independent censoring given the covariates.

1. To see if VF is important for death I did the following analyses

```
library(mets)
data(TRACE)
TRACE$death <- (TRACE$status!=0)*1

gg <- glm(death~vf,TRACE,family=binomial())
summary(gg)
```

What can we conclude based on this analysis. How do we interpret the parameters for the specific data. Make a statement of what we have learned. Are the parameters useful, is the test of interest ? The parameters estimated can be expressed by computing $P(T < C|X)$. To get this you need to mix out over for example $C = c$ given X.

2. Alternatively to see if people die early or late we can also compute the average time depending on vf

```
library(mets)
data(TRACE)
TRACE$death <- (TRACE$status!=0)*1
gg <- lm(time~vf,TRACE)
summary(gg)
```

What can we conclude based on this analysis. How do we interpret the parameters for the specific data. Make a statement of what we have learned. You may start by finding the survival distribution of $\min(T, C)$ given X. Then the mean can be gotten directly from this survival distribution.

B. This exercise is about generating data that follows a specific hazard.

Let T have hazard $\alpha(t)$ (that is nice and smooth). Define $A(t) = \int_0^t \alpha(s)ds$, $S(t) = \exp(-A(t))$ and let $A^{-1}(t)$ denote the inverse of $A(t)$. You may assume all regularity that you need.

1. Show that if $E \sim Exp(1)$ so E is exponential with rate 1, then the survival distribution of $A^{-1}(E)$ is $S(t)$.

2. What is the distribution of $A^{-1}(A(x) + E)$ with x $\in]0, \infty]$ when $E \sim Exp(1)$.

3. Use this to generate data from a $\alpha(t) \equiv 0.1$. Generate also a censoring time and return the right censored survival data. Estimate the cumulative hazard of the censored sample (in R).

4. Show that if $T_1 \sim \alpha_1(t)$ and $T_2 \sim \alpha_2(t)$ and T_1 and T_2 are independent then $\min(T_1, T_2) \sim \alpha_1(t) + \alpha_2(t)$. Hint: write up survival function.

5. We now wish to generate data from a piece-wise constant hazards model. A piecewise constant hazards model with $\lambda_1 = 0.1$ in $[0, 10[$ and $\lambda_2 = 0.2$ in $]10, 100]$ can be generated based on 1. We compute $V = A^{-1}(E)$ when $E < A(100)$ (and status=1) and let $V = 100$ when

$E > A(100)$ and return this a censored observation with status=0. What is $P(V = 100, \text{status} = 0)$ the probability of getting such an observation. What is for $t < 100$ $P(V > t)$. We can also construct a realization from this model based on two exponentials with rate λ_1 and and λ_2, respectively. Describe, how this is can be done, and check that this generates data from the same distribution.

6. Now given covariates X we assume that T has hazards $\alpha(t)\exp(X^T\beta)$. How can we generate survival data that follows this model.

C. The Weibull model

Let T^* have hazard given by

$$\lambda\gamma(\lambda t)^{\gamma-1} exp(X^T\beta) \quad (1)$$

so that the cumulative hazard is $(\lambda t)^\gamma exp(X^T\beta)$.

1. With $Y = \log(T^*)$, show that

$$Y = \alpha + \tilde{\beta}^T X + \sigma W,$$

where $\alpha = -\log(\lambda)$, $\sigma = \gamma^{-1}$, $\tilde{\beta} = -\sigma\beta$, and W has the extreme value distribution: $P(W > w) = \exp(-\exp(w))$.

2. Do a Weibull regression with VF as a covariate (for the TRACE data), then estimate the survivor function $P(T^* > t)$ and construct the associated 95\ pointwise confidence intervals (using the delta theorem). Do it for the two groups of VF.

3. Make the plots of the estimated survivor functions and their confidence intervals.

4. Can we say anything about how useful these survival predictions are ? Does the model fit ? Here you may compare to the non-parametric estimates based on the Nelson-Aalen estimator of the cumulative hazards, or estimating the survival function using the Kaplan-Meier:

```
ss <- survfit(Surv(time,event)~vf,data=TRACE); kmplot(ss)
```

D. Consider the TRACE data of the timereg package.

We consider the time to death given by time and the event status defined from $status! = 0$. One important prognostic factor for death is VF (ventricular fibrillation). The Nelson-Aalen estimator will estimate the cumulative hazard under independent censoring, and we will study it further later in the course.

1. Estimate the hazard in the constant hazard model (exponential model) for two groups of the data depending on VF (VF=0, VF=1).
2. Compute the Nelson-Aalen estimator for these two groups. What do this say about 1). You can get these from the survfit function with the plot option fun="cumhaz" and this is a non-parametric estimator of the cumulative hazard that we shall meet later, or via the phreg function of mets (see below).
3. Approximate the cumulative hazard with a piecewise constant hazards model (using the estimates from the Nelson-Aalen). Simulate data that looks like the two VF groups based on the piecewise constant approximation. hint: use the rchaz where you can give the cumulative hazar function of the timereg package.
4. In all of the above we choose to censor all observations at 7 (for example).
5. Check that it works by fitting the model also to the simulated data and compare.

Setting up things for R

```
library(mets)
data(TRACE)
TRACE <- trafsform(TRACE,event=(status!=0))

dsum(TRACE,event+time~vf)
ll <- dsum(TRACE,event+time~vf)
## rates
ll$event/ll$time

```

```
## use also poisson regression to estimate

## Nelson-Aalen giving cumulative hazard
ss <- survfit(Surv(time,event)~vf,data=TRACE)
kmplot(ss,fun="cumhaz")

## Nelson-Aalen giving cumulative hazard
ss <- phreg(Surv(time,event)~+strata(vf),data=TRACE)
bplot(ss)

## weibull regression below
weifit <- survreg(Surv(time,event)~sex,data=TRACE)
summary(weifit)
```