MATH-449 - Biostatistics
EPFL, Spring 2024
Problem Set 7

1. Let T be a survival time and A, Z be binary random variables. Suppose that T satisfies a proportional hazard model given A and Z, so that the hazard takes the form

$$\lim_{h\to 0^+} \frac{1}{h} P(t \le T < t+h | t \le T, A, Z) = \alpha_0(t) e^{0.1A+Z}.$$

   a) Write down the hazard ratio

$$\lim_{h\to 0^+} \frac{P(t \le T < t+h | t \le T, A=1, Z)}{P(t \le T < t+h | t \le T, A=0, Z)}.$$

   b) Show that the hazard of T given A at time t is given by

$$\frac{\alpha_0(t)}{P(t \le T|A)} \Big(e^{0.1A} e^{-\int_0^t \alpha_0(s)ds e^{0.1A}} P(Z=0|A) + e^{0.1A+1} e^{-\int_0^t \alpha_0(s)ds e^{0.1A+1}} P(Z=1|A) \Big) \tag{1}$$

   c) Calculate the hazard of T given Z

   d) Suppose $P(A=0|Z=0) = 2/3, P(A=0|Z=1) = 1/3$, and $P(Z=0) = 1/2$. Use the result from b) to write down the hazard ratio [1]

$$\lim_{h\to 0^+} \frac{P(t \le T < t+h | t \le T, A=1)}{P(t \le T < t+h | t \le T, A=0)}.$$

   Compare with the answer you got in a).

2. *(Simple linear regression on censored data - be careful!)*

   We revisit the simple linear regression model you have seen in introductory courses. For this, we assume that we have an i.i.d. sample $\{(\tilde{T}_i, A_i, D_i)\}_{i=1}^n$.[2] Suppose for the moment there is no censoring, i.e. $D_i = 1$ and $\tilde{T}_i = T_i$ for $i = 1, \cdots, n$. Assume a regression model on the form

$$T_i = \alpha + \beta A_i + \epsilon_i, \tag{2}$$

   where we assume that $E[\epsilon_i | A_i] = 0$. It is well known that the least squares estimators [3]

$$\hat{\beta} = \frac{cov_n(A,T)}{var_n(A)}, \qquad \hat{\alpha} = E_n[T] - \hat{\beta} E_n[A] \tag{3}$$

   consistently estimate the true regression parameters

$$\beta = \frac{cov(A,T)}{var(A)}, \qquad \alpha = E[T] - \beta E[A]. \tag{4}$$

   a) Suppose now that we have censoring, so that we don't have access to all T_i, $i \in \{1, \cdots, n\}$. Consider the subset of observations $\mathcal{D} = \{(T_i, A_i, D_i) : D_i = 1\}$ (i.e. omitting all censored individuals), and let $n_D = \sum_{i=1}^n I(D_i = 1)$. Argue that the estimators

$$\hat{\beta} = \frac{cov_{n_D}(A,T)}{var_{n_D}(A)}, \qquad \hat{\alpha} = E_{n_D}[T] - \hat{\beta} E_{n_D}[A]$$

[1] You should not attempt to write out the terms $P(t \le T | A=0)$ and $P(t \le T | A=1)$.

[2] Recall that $\tilde{T}_i = T_i$ if $D_i = 1$ (i.e. if individual i dies), and $\tilde{T}_i = T_i^*$ if $D_i = 0$ (i.e. if individual i is censored). A_i is a continuous covariate taking values in a neighbourhood of 0.

[3] We have used the subscript n to indicate finite sample versions, i.e. $E_n[A] = \frac{1}{n}\sum_{i=1}^n A_i$, and $var_n(A) = cov_n(A, A)$, where $cov_n(A, Z) = \frac{1}{n}\sum_{i=1}^n \big((A_i - E_n[A])(Z_i - E_n[Z])\big)$.

applied to the data in $\mathcal{D}$ will approach

$$\beta^{\mathcal{D}} = \frac{cov(A, T|D=1)}{var(A|D=1)}, \qquad \alpha^{\mathcal{D}} = E[T|D=1] - \beta^{\mathcal{D}} E[A|D=1]$$

in large samples.

b) Let $\hat{\tilde{\alpha}}$ and $\hat{\tilde{\beta}}$ be the estimators in (3) with T_i replaced with $\tilde{T}_i$, i.e.

$$\hat{\tilde{\beta}} = \frac{cov_n(A, \tilde{T})}{var_n(A)}, \qquad \hat{\tilde{\alpha}} = E_n[\tilde{T}] - \hat{\tilde{\beta}} E_n[A].$$

Argue that these estimators in general depend on the censoring distribution.

c) The approaches in a) and b) shows two naive estimators that fail to estimate the regression coefficients (4). We will briefly see how the approaches compare in simulations. The file `simulate.RData` (see under this week on Moodle) contains simulations of the 'complete' observations $\{(T_i, T_i^*, A_i, D_i)\}_{i=1}^n$, where $T_i \perp\!\!\!\perp T_i^* | A_i$. This corresponds to a random censoring scenario where we (somehow) have access to the death times T_i for all individuals, even after they are censored. In `R`, use the `lm` function to obtain the regression coefficients under the approaches in a) and d). Compare with the coefficients (3).

d) Consider the following regression model for the censored times

$$\tilde{T}_i = \tilde{\alpha} + \tilde{\beta} A_i + \tilde{\epsilon}_i \tag{5}$$

and suppose $E[\tilde{\epsilon}_i | A_i] = 0$. Assuming the model assumptions behind (2) and (5) hold, show that $\tilde{\alpha} \leq \alpha$ and $\tilde{\alpha} + \tilde{\beta} \leq \alpha + \beta$. [4]

3. **Log-rank test for MP-6 vs placebo example.** In this exercise we will perform a log-rank test to test the null hypothesis

$$\alpha_1(t) = \alpha_2(t), \qquad t \in [0, \tau],$$

where $\alpha_1(t)$ is the hazard in the MP-6 group, $\alpha_2(t)$ is the hazard in the placebo group, and $\tau = 35$. As in exercise 4, we load the above data set for the two groups into `R`:

```
placebo = data.frame(time=c(1, 1, 2, 2,3, 4, 4, 5, 5, 8,8,8, 8, 11, 11, 12, 12,
15, 17, 22, 23), event=1)
mp = data.frame(time=c(6, 6, 6, 6, 7, 9, 10, 10, 11, 13, 16,17, 19 ,20, 22 ,23
,25, 32, 32, 34, 35), event=c(1,1,1,0,1,0,1,0,0,1,1,0,0,0,1,1,0,0,0,0,0)).
```

a) Create a column in the `placebo` data set called `group`, and set the value to `0`. Do the same in the `mp` data set, but set the value to `1`.

b) Merge the data set `placebo` with `mp` and store the merged data set in `combinedData`.[5]

c) Perform the log-rank test using the `survdiff` function. You may use the command `survdiff(Surv(time,event==1)~group,data=combinedData)`. This command displays the log-rank test using the default "weight function" $L(t) = \frac{Z_1(t)Z_2(t)}{Z_\bullet(t)}$.

d) Interpret the output. What is the value of the test statistic, and what is its distribution? What is the p-value?

e) Estimate the restricted mean survival function $R_1(t), R_2(t)$ in both groups for $t = 23$. Derive a test statistic for the null hypothesis

$$H_0 : R_1(23) = R_2(23),$$

and perform the test.

[4] Hint: look at conditional expectations of T_i and $\tilde{T}_i$.

[5] You may use the `rbind` function to do this.