7.1 We consider the duration of breast feeding of women following birth. The `breastfeeding` data contains the following variables:

- `duration`: duration of breast feeding (in weeks)
- `delta`: indicator variable, `1` for completed breastfeeding, `0` for right-censored data
- `race`: race of mother, one of white (`1`), black (`2`), or other (`3`)
- `poverty`: is mother revenue below poverty line? either yes (`1`) or no (`0`)
- `smoke`: smoking status of mother at birth of child, either yes (`1`) or no (`0`)
- `alcohol`: alcohol-drinking status of mother at birth of child, either yes (`1`) or no (`0`)
- `agemth` age of mother at birth of child
- `ybirth` year of child's birth
- `yschool`: years of school of the mother
- `pc3mth`: binary indicator, `0` if mother sought prenatal care in first three months of pregnancy, `1` otherwise.

(a) Estimate and plot the survival curves for the two levels of the smoker status variable `smoke` using the Kaplan–Meier estimator. Comment on the plotted survival curves.

Solution
The largest duration in each group is observed, so the estimated product-limit survival curve drops to zero. The confidence intervals of the curves seemingly overlap, except between 25 and 50 weeks.

(b) Using the output of (a), what is the estimated probability that a mother who is a non-smoker will breastfeed for more than 36 weeks? What about for a mother who is a smoker?

Solution
The estimated probability of survival is 0.1458 for non-smokers and 0.0909 for smokers.

(c) What is the median and mean number of weeks that a mother who is a non-smoker will breastfeed? What about for a mother who is a smoker?

Solution
For non-smokers, the median and mean duration are 12 weeks and 18 weeks, indicating right-skewness. For smokers, these are 8 weeks and 14.093 weeks

(d) Test whether the survival curves for non-smoker and smoker mothers are the same.

Solution
The log rank statistic is 10.09; relative to a χ^2_1 null distribution, we reject the null hypothesis that the two survival curves for smoker and non-smokers are identical (p-value of 0.0015).

(e) Fit a Cox proportional hazards model to evaluate the impact of the poverty status, smoker status, age of the mother and years of schooling on the hazard. Write down the equation of the estimated hazard function and interpret the estimated $\widehat{\boldsymbol{\beta}}$ parameters.

Solution
The hazard is

$$h(t;\boldsymbol{x}) = h_0(t)\exp(\beta_1\texttt{poverty} + \beta_2\texttt{smoke} + \beta_3\texttt{agemth} + \beta_4\texttt{yschool})$$

The effect of age of the mother and poverty on hazard are not significant at the 5% level.

- $\exp(\widehat{\beta}_1) = 0.843$, so *ceteris paribus*, poor women have a 15.7% lower risk of breastfeeding than women who are not poor, so breastfeeding times are longer.
- $\exp(\widehat{\beta}_2) = 1.218$; smokers have 21.8% higher risk than non-smokers, everything else being equal. Thus, the survival times are lower for smokers.
- For every additional year of the mother *ceteris paribus*, the estimated hazard is 1.019 times higher.

Compiled 02/12/2020 at 20:05

- For each additional year of schooling, *ceteris paribus*, the hazard decreases by 6% (so women with longer studies have longer duration).

7.2 A shoe store in Montreal wishes to know how long it takes before products are sold. The dataset `shoes` contains the following variables:

- `status`: categorical variable, 0 for sales, 1 if the article is still in stock, 2 if destocked.
- `time`: storage time of the article (in months).
- `price`: sale price, rounded to the nearest dollar.
- `gender`: binary variable for gender, 0 for men shoes, 1 for women shoes.

Our ojective is to estimate the survival time of items in stock.

(a) What does censoring represent in this example?

Solution

Pairs of shoes that are destocked or still in stock are right-censored. The former is an example of non-random censoring (every pair of shoe is kept until 40 months have lapsed, after which they are discarded). Since the event of interest is sale, these observation are right-censored.

(b) Estimate the survival function of the stocking time using Kaplan–Meier estimator and report the estimated quartiles of the survival time.

Solution

The quartiles are 4, 7 and 11 months.

(c) Fit a Cox proportional hazard model for stocking time as a function of shoes gender and sale price. and report the estimated coefficients, $\widehat{\boldsymbol{\beta}}$. Which of the covariates impacts hazard of staying in store, if any?

Solution

The parameter estimates (standard errors) are

- $\widehat{\beta}_{\texttt{gender}} = -0.164(0.026)$; the hazard ratio for women is $\exp(-0.164) = 0.848$ times that of men, for a pair of shoes of the same price. Women shoes stay longer in store.
- $\widehat{\beta}_{\texttt{price}} = -0.0136(9.59 \times 10^{-4})$; the risk ratio for two pairs of shoes of the same gender whose price difference is 1\$ is 0.986, meaning that the hazard ratio decreases by 12.7% for a pair which is 10\$ lower than another, *ceteris paribus*.

Both coefficients are statistically significant (Wald tests p-values are less than 10^{-4}). The model is also globally significant, with a likelihood ratio statistic for $\mathscr{H}_0 : \beta_{\texttt{gender}} = \beta_{\texttt{price}} = 0$ worth 248, leading to a negligible p-value.

Compiled 02/12/2020 at 20:05