Fundamentals of Statistical Modeling (VT20)

Andrea Discacciati
Karolinska Institutet
Stockholm, Sweden

Lab 5

Load the dataset and the `mlci` command

{{1}}

Exercise 1

We consider again the oral cancer dataset (see Lab 4). We measure time to death (y) due to oral cancer ($d = 1$) (continuous line in the graphs) or to other causes ($d = 2$) (dashed line in the graphs). Just to get started, we exclude the censored observations ($d = 0$) from our estimation procedure (note the `if d != 0` at the end of the `mlexp` command).

We model the joint distribution $f(y, d)$ through the conditional expansion

$$f(y, d) = f(d|y)f(y).$$

We consider a log-normal distribution for $f(y)$ and a bernoulli distribution for $f(d|y)$. Estimate the model's parameters. Remember to constrain the bounded parameters. How do we interpret $\exp(\gamma_1)$?

{{2}}

Plot the estimated distributions $\hat{f}(y)$ and $\hat{f}(d|y)$. Interpret the plots.

{{3}}

Exercise 2

Some of the times (y) are actually right-censored. Estimate the model's parameters by modifying the likelihood accordingly to take this into account (last equation on slide 137). Remember to constrain the bounded parameters.

{{4}}

Plot the estimated distributions $\hat{f}(y)$ and $\hat{f}(d|y)$.

{{5}}

Extra 1

Plot the estimated cumulative incidence functions $\hat{F}(y, d = 1)$ and $\hat{F}(y, d = 2)$ (see slide 140) and overlay them to their nonparametric counterparts obtained using Stata's `stcrreg` command.

$$\hat{F}(y, 1) = \int_0^y \hat{f}_{Y,D}(u, d = 1)du = \int_0^y \hat{f}_Y(u)\hat{f}_{D|Y}(d = 1|u)du$$

and

$$\hat{F}(y,2) = \int_0^y \hat{f}_{Y,D}(u, d=2)du = \int_0^y \hat{f}_Y(u) f_{D|Y}(d=2|u)du$$

{{6}}

Extra 2

We now model the joint distribution $f(y,d,x)$ through conditional expansion.

$$f(y,d,x) = f(d|y,x)f(y|x)f(x),$$

The variable x is tumor grade at diagnosis: low ($x=0$, blue in the graphs) or high ($x=1$, red in the graphs). We consider a log-normal distribution for $f(y|x)$ and a bernoulli distribution for both $f(d|y,x)$ and $f(x)$.

Estimate the model's parameters. Remember to constrain the bounded parameters. How do we interpret the model's parameters?

{{7}}

Plot the estimated distributions $\hat{f}(d|y,x)$, $\hat{f}(y|x)$, and $\hat{f}(x)$.

{{8}}

Exercise 3

We recruited 2,784 subjects in Sweden at the time of their first myocardial infarction. We took a blood sample and measured LDL cholesterol (mmol/L) on a first follow-up visit, 1 month after the MI (variable `ldl1`). We then measured LDL cholesterol again 6 months after the MI (second follow-up visit) (variable `ldl2`). Plot the sample histogram of the 2 variables. What can we say about them?

{{9}}

We model the marginal distributions of the 2 variables: $f(ldl_1), f(ldl_2)$. We consider skew-normal models. This means that $Z = G(Y)$ follows a standard skew-normal distribution.

$$G(y) = (y - \mu)/\sigma$$
$$g(y) = 1/\sigma$$
$$f_Z(z) = 2F_N(\alpha z)f_N(z)$$

where $F_N(z)$ and $f_N(z)$ are the standard normal CDF and PDF, respectively.

Estimate the 2 models' parameters and plot the densities over the sample histograms. Are the data suggesting that the skewness parameter α is different from 0?

{{10}}

Exercise 4

We consider the joint distribution of ldl_1 and ldl_2

$$f(ldl_1, ldl_2) = f(ldl_2|ldl_1)f(ldl_1)$$

We assume that $f(ldl_1)$ and $f(ldl_2|ldl_1)$ are skew-normal. Allow all parameters of $f(ldl_2|ldl_1)$ to depend on ldl_1. Estimate the model's parameters. Remember to constrain the bounded parameters.

{{11}}

Draw a scatterplot of ldl_2 versus ldl_1. Do the results above agree with the plot? Make a qualitative assessment.

Plot the estimated conditional denisity $\hat{f}(ldl_2|ldl_1)$ for ldl_1 values of 2, 3, and 5 mmol/L. Again, make a qualitative assessment of the plot.

{{12}}