Fundamentals of Statistical Modeling (VT21)

Andrea Discacciati
Karolinska Institutet
Stockholm, Sweden

Lab 2

Load the dataset and the `mlci` command

{{1}}

Install the `qplot` command (you need to be connected to the Internet)

{{2}}

Install the `rcsgen` command (you need to be connected to the Internet)

{{3}}

Exercise 1

This dataset contains information on the blood concentration of a biomarker (y) in a random sample of 1432 subjects. Take a look at the histogram. What can we say about the distribution of this biomarker?

Plot also the histogram of $\log(y)$. How does the distribution of the biomarker after logarithmic transform look like?

{{4}}

{{5}}

Exercise 2

We assume that $f(y)$ is gamma (see Lab 1). Estimate the parameters α and β using the `gammaden()` function. Fix the location parameter g (the third argument of the `gammaden()` function) to be equal to 0. Constrain α and β to be positive.

Note: the parameters α and β are not interpretable. We can reparametrise the gamma distribution so that one parameter is equal to its mean. This is described in the Extra material for Lab 2.

{{6}}

Plot the estimated density $\hat{f}(y)$ over the sample histogram

{{7}}

Exercise 3

We assume that $f(y)$ is log-normal distributed. That is, we assume that the biomarker is standard normal distributed after we apply the transform

$$G(y) = (\log(y) - \mu)/\sigma$$

The derivative of $G(y)$ with respect to y is

$$G'(y) = g(y) = 1/(y\sigma).$$

Estimate the parameters μ and σ. Constrain σ to be positive.

{{8}}

Compare the likelihood with that from the gamma model

Plot the estimated density $\hat{f}(y)$ over the sample histogram

{{9}}

Exercise 4

We make the transform $G(y)$ more flexible using polynomials. Consider the transform

$$G(y) = (\log(y) + \eta \log(y)^2 - \mu)/\sigma$$

The derivative of $G(y)$ with respect to y is

$$G'(y) = g(y) = (1 + 2\eta \log(y))/(\sigma y)$$

Estimate the parameters μ, σ, η. Constrain σ to be positive.

{{10}}

Plot the estimated density $\hat{f}(y)$ over the sample histogram

{{11}}

Exercise 5

Instead of a quadratic term, we add two restricted cubic splines transforms of $\log(y)$: $V_2(\log(y))$ and $V_3(\log(y))$. We consider the transform

$$G(y) = (\log(y) + \eta_1 V_2(\log(y)) + \eta_2 V_3(\log(y)) - \mu)/\sigma$$

The derivative of $G(y)$ with respect to y is

$$G'(y) = g(y) = (1 + \eta_1 v_2(\log(y)) + \eta_2 v_3(\log(y)))/(\sigma y)$$

Estimate the parameters $\mu, \sigma, \eta_1, \eta_2$. Constrain σ to be positive. Jointly test the 2 parameters η_1, η_2 to assess whether adding the 2 RCS transforms improves the fit of this model with respect to the "basic" log-normal model (see Exercise 3).

{{12}}

Plot the estimated density $\hat{f}(y)$ over the sample histogram

{{13}}

Exercise 6

Let's assess the goodness-of-fit of the log-normal model with RCS transforms (see Exercise 5) and of the log-normal model (see Excercise 3) using a quantile plot.

{{14}}

Extra: Exercise 7 (more on transforms of random variables)

We now assume that $f(y)$ is gamma-distributed after square root transform.

$$G(y) = \sqrt{y}$$

The derivative is

$$G'(y) = g(y) = 0.5/\sqrt{y}$$

Estimate the parameters α and β using the `gammaden()` function. Fix the location parameter g to be equal to 0. Constrain α and β to be positive. Compare the likelihood with that form the log-normal and gamma models

{{15}}

Plot the estimated density $\hat{f}(y)$ over the sample histogram. Visually compare the estimated density from the lognormal + splines model with the density from the gamma model after square root transform. What do you conclude?

{{16}}

Extra: Exercise 8 (more on goodness of fit)

Let's go back to the normal distributed variable (Exercise 1, Lab 1).

{{17}}

Assume that $f(y_n)$ is normal and estimate the parameters μ and σ. Generate the transform $u = \hat{F}(y_n)$. Draw the estimated quantile plot using the `qplot` command.

{{18}}

Assume now that $f(y_n)$ is exponential and estimate the parameter λ. Generate the transform $u = \hat{F}(y_n)$. Draw the estimated quantile plot using the `qplot` command.

{{19}}

What can you conclude about the goodness of fit of the normal and exponential model?

Extra: Exercise 9 (binary variables)

Assume that y_{ber} follows a Bernoulli distribution. We want to estimate the probability of "success" ($y_{ber} = 1$). Estimate the probability η while constraining it to be bounded between 0 and 1. First, write down the likelihood by hand. Then, use the `binomialp()` function.

Are the results you obtain identical to those obtained from logistic regression?

{{20}}