R handout: Estimation and Transformation

Yuhui Yao

1. METHODOLOGY

1. Box-Cox transformation and regression

   (a) Motivation

   In practice, most of our data do not follow any normal distribution. The first attempt we can use is to transform this data into a normal distribution and then follow the traditional methods based on the normal distribution.

   (b) Issue

   This transformation change the physical unit. We need to transform it back to the original unit.

   (c) We still learn the "break-into-commercial" sample from the data, *Crime in Vancouver*. Suppose the original sample is X and the transformed sample is Y. The transformation used here is the squared-root transformation. We have this assumption

   $$Y_i = \sqrt{X_i}$$

   where we assume Y_i follows a normal distribution with mean μ_i and variance σ^2. Also, let $\mu_i = at_i^2 + bt_i + c$. The matrix form of the normal distribution

   $$f_{\underline{Y}}(\underline{y}) = (2\pi)^{-n/2}(|\underline{\Sigma}|)^{-1/2}e^{-\frac{1}{2}(\underline{y}-\underline{\mu})^T\underline{\Sigma}^{-1}(\underline{y}-\underline{\mu})}$$

   The log-likelihood function

   $$lnL(\underline{\mu}, \underline{\Sigma}, \theta; \underline{y}) = -\frac{n}{2}ln(2\pi) - \frac{1}{2}ln(|\underline{\Sigma}|) - \frac{1}{2}(\underline{y} - \underline{\mu})^T\underline{\Sigma}^{-1}(\underline{y} - \underline{\mu})$$

   where the vector $\underline{y} = \{y_1, y_2, ..., y_n\}^T$, $y_i = \sqrt{x_i}$, the mean vector $\underline{\mu} = \{\mu_1, \mu_2, ..., \mu_n\}^T$ and the n by n covariance matrix $\underline{\Sigma}$ with diagonal elements σ^2 and off-diagonal elements 0. And this is the target we need to maximize. After estimating the parameters, we need to transform it back to our original unit.

   $$Y_i \sim N(\mu_i, \sigma^2) \Rightarrow \sqrt{X_i} \sim N(\mu_i, \sigma^2)$$

$$\Rightarrow \frac{\sqrt{X_i}}{\sigma} \sim N(\mu_i, 1) \Rightarrow \frac{X_i}{\sigma^2} \sim \chi^2_{1,\mu_i} \xrightarrow{D} N(1+\mu_i, 2(1+\mu_i))$$

where χ^2_{1,μ_i} is the noncentral chisquare distribution with 1 degree of freedom and noncentral parameter μ_i.

$$Thelog - likehood$$

2. Another model based on Poisson.

References

[1] Kaggle, *Crime in Vancouver*, Link: https://www.kaggle.com/wosaku/crime-in-vancouver 2017.