Log Transformation in Simple Linear Regression (3/3)

a_statistician

March 10, 2017

Suppose the log transformation is performed on the response variable Y, i.e.,

$$Z = \log(Y).$$

Then the following simple linear regression between Z and covariate X is fit:

$$Z = \alpha + \beta X + \epsilon$$

α is the intercept; β is the slope; and ϵ is the error term which is random and follows a normal distribution with mean 0 and variance σ^2.

The slope β can be explained as the mean change of Z when covariate X increases by 1 unit. Similarly, $\exp(\beta)$ is the ratio of means of Y given X increased by 1 unit. It means

$$\frac{\mathrm{E}(Y|X = x + 1)}{\mathrm{E}(Y|X = x)} = \exp(\beta)$$

The slope β can be estimated by least square method, indicated by $\hat{\beta}$. Under the general assumptions of linear model, we know that $\hat{\beta}$ is an unbiased estimate of β. It means

$$\mathrm{E}(\hat{\beta}) = \beta.$$

Furthermore, under the assumption that the error term ϵ follows a normal distribution with mean 0 and variance σ^2, we have

$$\hat{\beta} \sim \mathrm{N}\left(\beta, \frac{\sigma^2}{\sum(x_i - \bar{x})^2}\right)$$

Because $\exp(\beta)$ represents the ratio of means of $Y|X = x + 1$ vs $Y|X = x$, it is natural to consider $\exp(\hat{\beta})$ as the estimate of the ratio of mean of $Y|X = x + 1$ vs $Y|X = x$. But we need to know that $\exp(\hat{\beta})$ is the biased estimate of $\exp(\beta)$ as explained below: Because $\hat{\beta}$ follows a normal distribution, $\exp(\hat{\beta})$ follows a lognormal distribution. It means

$$\mathrm{E}(\exp(\hat{\beta})) = \exp\left(\beta + \frac{\sigma^2}{2\sum(x_i - \bar{x})^2}\right) \neq \exp(\beta).$$

See the linked properties of lognormal distribution. Maybe we can use

$$\exp\left(\hat{\beta} - \frac{\hat{\sigma}^2}{2\sum(x_i - \bar{x})^2}\right)$$

where $\hat{\sigma}^2$ is the estimate of σ^2, to estimate $\exp(\beta)$ such that the bias is reduced. Of course, $\exp(\hat{\beta})$ is an asymptoticly unbiased estimate of $\exp(\beta)$, because $\frac{\sigma^2}{2\sum(x_i-\bar{x})^2}$ converges to zero as sample size goes to infinity. But if the sample size is small and variance of error term is large, this bias is not ignorable.

In fact, for any unbiased estimate $\hat{\theta}$ of θ, $f(\hat{\theta})$ is the biased estimate of $f(\theta)$ given that second derivative of $f(\theta)$ is not zero at true value of θ and the variance of $\hat{\theta}$ is not zero. $f(\hat{\theta})$ is asymptotically unbiased estimate of $f(\theta)$ if the variance of $\hat{\theta}$ converges to zero, which is true for most of the estimates. But for a small sample size, need to be careful.

Note: Maybe you wonder where the variance term in $\frac{\mathrm{E}(Y|X=x+1)}{\mathrm{E}(Y|X=x)}$ is, because Y follows lognormal distribution. In fact, we have

$$\mathrm{E}(Y|X = x + 1) = \exp(\alpha + \beta(x + 1) + \sigma^2/2)$$

$$\mathrm{E}(Y|X = x) = \exp(\alpha + \beta x + \sigma^2/2)$$

and

$$\frac{\mathrm{E}(Y|X = x + 1)}{\mathrm{E}(Y|X = x)} = \exp(\beta)$$