STAT 3008 Applied Regression Analysis
Tutorial 2. Simple Linear Regression

XU Yongze

Sep 16 & 17& 19, 2013

Contents

1 Basic Concepts

let $\mu_1, ..., \mu_n$ be n random variables. $a_0, ..., a_n$ be n+1 constant.

1.1 Expectation

$$E(\mu_i) = \int \mu_i f(u_i) du_i = \sum \mu_i f(\mu_i) \tag{1.1}$$

$$E(a_0 + \sum_{i=1}^{n} a_i \mu_i) = a_0 + \sum_{i=1}^{n} a_i E(\mu_i) \tag{1.2}$$

1.2 Variance

$$Var(\mu_i) = E([\mu_i - E(\mu_i)]^2) \tag{1.3}$$

$$Var(a_0 + \sum_{i=1}^{n} a_i \mu_i) = \sum_{i=1}^{n} {a_i}^2 Var(u_i) \qquad \mu_1, ..., \mu_n \quad \text{are} \quad \text{independent} \tag{1.4}$$

1.3 Covariance

$$Cov(\mu_i, \mu_j) = Cov(\mu_j, \mu_i) = E[\mu_i - E(\mu_i)][\mu_j - E(\mu_j)] \tag{1.5}$$

$$Cov(\mu_i, \mu_i) = Var(\mu_i) \tag{1.6}$$

$$Cov(a_0 + a_1\mu_1, a_2 + a_3\mu_2) = a_1 a_3 Cov(\mu_1, \mu_2) \tag{1.7}$$

$$Var(a_0 + \sum_{i=1}^{n} a_i \mu_i) = \sum_{i=1}^{n} {a_i}^2 Var(i) + 2\sum_{i<j} a_i a_j Cov(\mu_i, \mu_j) \tag{1.8}$$

1.4 Correlation

$$\rho(\mu_i, \mu_j) = \frac{Cov(\mu_i, \mu_j)}{\sqrt{Var(\mu_i)Var(\mu_j)}} \tag{1.9}$$

2 Basic Simple Linear Regression

♢ The model is defined as:

$$y_i = \beta_0 + \beta_1 x_i + e_i, \quad e_i \sim \text{i.i.d.} N(0, \sigma^2) \tag{2.1}$$

- x_i is known
- another expression of the model: $E(Y|X = x) = \beta_0 + \beta_1 x, Var(Y|X = x) = \sigma^2$.

◇ Parameters of Interest: $\beta_0, \beta_1, \sigma^2$
β_0 : y-intercept(value of y when x = 0)
β_1 : slope (change of y if x changes by 1 unit)
σ^2 : variance of the random error

◇ Estimation:
Ordinary Least Square Method: "least square" is a criterion, an idea to find the "best fit". It doesn't matter what the distribution of y_i is and what the specific model is.
$\hat{y_i} = \hat{\beta}_0 + \hat{\beta}_1 x_i$
Residual for case i is:
$\hat{e_i} = y_i - \hat{y_i} = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$
RSS(Residual Sum of Square) = $\sum \hat{e_i}^2$
Minimize the RSS:

$$(\beta_0, \beta_1) = argmin_{\beta_0,\beta_1} RSS(\beta_0, \beta_1)$$

$$\frac{\partial \mathbf{RSS}}{\partial \beta_\mathbf{0}}|_{\beta_0=\hat{\beta_0}} = 0 \tag{2.2}$$

$$\frac{\partial \mathbf{RSS}}{\partial \beta_\mathbf{1}}|_{\beta_1=\hat{\beta_1}} = 0 \tag{2.3}$$

Solve the equations:

$$\hat{\beta_1} = \frac{SXY}{SXX} \tag{2.4}$$

$$\hat{\beta_0} = \bar{y} - \hat{\beta_1}\bar{x} \tag{2.5}$$

with:
$\bar{y} = \frac{1}{n}\sum_i y_i$
$\bar{x} = \frac{1}{n}\sum_i x_i$
$SYY = \sum_i (y_i - \bar{y})^2 = \sum_i {y_i}^2 - n\bar{y}^2$
$SXX = \sum_i (x_i - \bar{x})^2 = \sum_i {x_i}^2 - n\bar{x}^2$
$SXY = \sum_i (x_i - \bar{x})(y_i - \bar{y}) = \sum_i x_i y_i - n\bar{x}\bar{y}$
$\widehat{RSS} = SYY - \frac{SXY^2}{SXX} = SYY - \hat{\beta_1}^2 SXX$

◇ Properties of estimators:

- $E(\hat{\beta_0}) = \beta_0, E(\hat{\beta_1}) = \beta_1, E(\hat{\sigma}^2) = \sigma^2$

- $\sum \hat{e}_i = 0$ (But $\sum e_i \neq 0$.)
- $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$ (The fitted line passes through $(\bar{x}, \bar{y})$.)

The second and third points are only for the model $y_i = \beta_0 + \beta_1 x_i + e_i, e_i \sim$ i.i.d.$N(0, \sigma^2)$. Otherwise they are not always true.

Example: a model $y_i = \beta x_i + e_i, e_i \sim$ i.i.d.$N(0, \sigma^2)$ (assuming that $\beta_0 = 0$),

$$RSS = \sum (y_i - \beta x_i)^2.$$
$$\frac{\partial RSS}{\partial \beta} = -2 \sum x_i (y_i - \beta x_i) = 0 \Rightarrow \hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2}.$$

No equation to ensure $\sum (y_i - \hat{y}_i) = \sum (y_i - \hat{\beta} x_i) = 0$. When $\sum (y_i - \hat{\beta} x_i) \neq 0$, $\sum y_i \neq \hat{\beta} \sum x_i$, i.e., $\bar{y} \neq \hat{\beta} \bar{x}$, the fitted line does NOT pass through $(\bar{x}, \bar{y})$.

3 Exercises

3.1 Exercise 2 Q1 2.1.2, 2.1.3

3.2 Exercise 2 Q2

3.3 Exercise 3 Q6

3.4 Exercise 2 Q4

4 Appendix

To calculate $E(\hat{\beta}_0), Var(\hat{\beta}_0)$ and $Cov(\hat{\beta}_0, \hat{\beta}_1)$ for the model $y_i = \beta_0 + \beta_1 x_i + e_i, e_i \sim$ i.i.d.$N(0, \sigma^2)$, which was not done in the lecture for reasons of time, is a good practice of using the basic concepts and culculation rules of expectations, variance and covariance.

4.1 $E(\hat{\beta}_0)$

$$E(\hat{\beta}_0) = E(\bar{y}) - \bar{x} E(\hat{\beta}_1) = \frac{1}{n} \sum E y_i - \bar{x} \beta_1 = \frac{1}{n} \sum (\beta_0 + \beta_1 x_i) - \frac{1}{n} \beta_1 \sum x_i = \beta_0. \tag{4.1}$$

4.2 $Var(\hat{\beta}_0)$

$$\because Cov(\bar{y}, \hat{\beta}_1) = 0,$$
$$\therefore Var(\hat{\beta}_0) = Var(\bar{y}) + \bar{x}^2 Var(\hat{\beta}_1) = \frac{1}{n} Var(y_i) + \bar{x}^2 \frac{\sigma^2}{SXX} = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{SXX} \right). \tag{4.2}$$

4.3 $Cov(\hat{\beta}_0, \hat{\beta}_1)$

$$\because Cov(\bar{y}, \hat{\beta}_1) = 0,$$

$$\therefore Cov(\hat{\beta}_0, \hat{\beta}_1) = Cov(\bar{y} - \hat{\beta}_1 \bar{x}, \hat{\beta}_1) = -\bar{x} Var(\hat{\beta}_1) = -\sigma^2 \frac{\bar{x}}{SXX}. \tag{4.3}$$