Development Statistics

S12 Regression

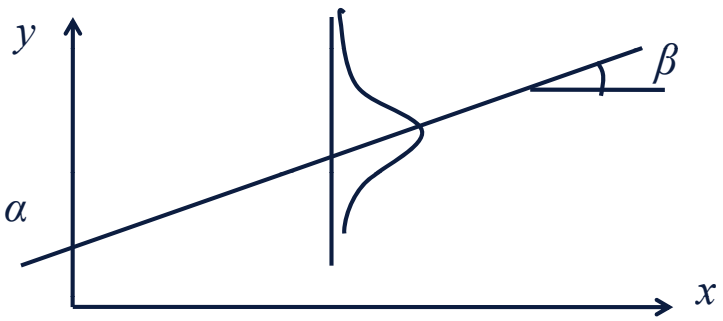
Fujikawa, Kiyoshi
Nagoya University, GSID

Linear regression model

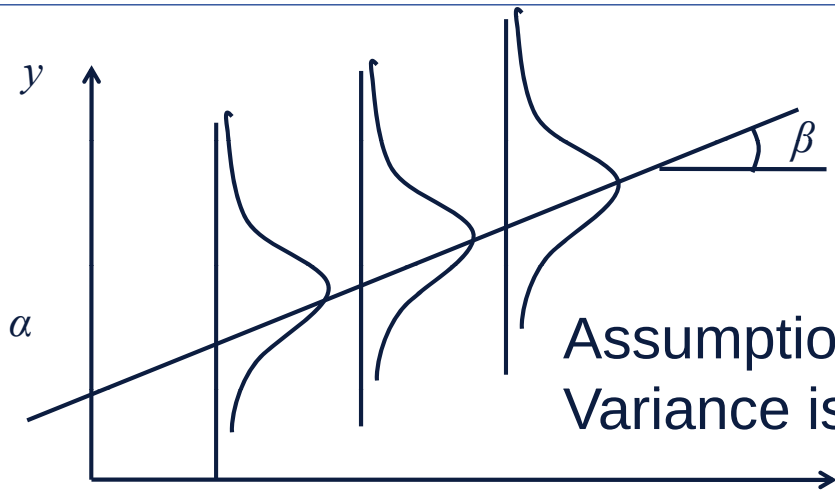
- Basic model

$$y_i = \alpha + \beta x_i + u_i$$

- α and β : *coefficient parameters*



Distribution of y and u



Assumption:
Variance is the same

$$y_i \sim N(\alpha + \beta x_i, \sigma^2) \qquad u_i \sim N(0, \sigma^2)$$

Residual

- Regression line

$$\hat{y}_i = a + bx_i$$

- Difference between the regression line and the observation is called *residual*

$$y_i - \hat{y}_i = y_i - (a + bx_i) = e_i$$

Least Square Method

- Square sum of the residual

$$s = \sum e_i^2 = \sum \left(y_i - (a + bx_i)\right)^2$$

- Minimize square sum of the residual

$$\frac{\partial s}{\partial a} = 2\sum \left[\left(y_i - (a + bx_i)\right)(-1)\right] = 0$$

$$\frac{\partial s}{\partial b} = 2\sum \left[\left(y_i - (a + bx_i)\right)(-x_i)\right] = 0$$

Constant term parameter a (upper eqation)

- *Regression line* goes through the center of the sample

$$\sum y_i - na - b\sum x_i = 0$$

$$\bar{y} - a - b\bar{x} = 0$$

Slope parameter b (lower equation)

$$\sum x_i y_i - a\sum x_i - b\sum x_i^2 = 0$$

$$\sum x_i y_i - (\bar{y} - b\bar{x})\sum x_i - b\sum x_i^2 = 0$$

$$\sum x_i (y_i - \bar{y}) - b\sum x_i (x_i - \bar{x}) = 0$$

$$\sum (x_i - \bar{x})(y_i - \bar{y}) - b\sum (x_i - \bar{x})(x_i - \bar{x}) = 0$$

Regression coefficient

- Slope parameter b

$$b = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

- Constant term parameter a

$$\bar{y} = a + b\bar{x} \Rightarrow a = \bar{y} - b\bar{x}$$

Regression coefficient (variance parameter)

- The main purpose of regression analysis is to check the significance of "b"
- Here also the variance is important !
- s^2 : *unbiased estimator* of σ^2
- s : *standard error*

$$s^2 = \frac{\sum_i e_i^2}{n-2}$$

- Sample size is n

Linear regression model (matrix)

- Basic model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

- Sample size is n
- The number of parameters is k

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{21} & \cdots & x_{k1} \\ 1 & x_{22} & \cdots & x_{k2} \\ 1 & \vdots & \ddots & \vdots \\ 1 & x_{2n} & \cdots & x_{kn} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

Distribution of y and u

- Basic model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

- Error term u follows iid Normal
- Identically Independent Distribution

$$\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}) \qquad \mathbf{u} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$$

Regression coefficient

- Coefficient parameter

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{y})$$

- s^2 : *unbiased estimator* of σ^2
- s : *standard error*

$$s^2 = \frac{\sum_i e_i^2}{n-k}$$

- Sample size is n
- The number of parameters is k

R square

- Measure of fitness of the equation

$$R^2 = \frac{\sum_i (y_i - \bar{y}_i)^2 - \sum_i e_i^2}{\sum_i (y_i - \bar{y}_i)^2} = 1 - \frac{\sum_i e_i^2}{\sum_i (y_i - \bar{y}_i)^2}$$

- But, R^2 gets larger when the number of variables increases

R square

- Measure of fitness of the equation

$$R^2 = \frac{\sum_i (y_i - \bar{y}_i)^2 - \sum_i e_i^2}{\sum_i (y_i - \bar{y}_i)^2} = 1 - \frac{\sum_i e_i^2}{\sum_i (y_i - \bar{y}_i)^2}$$

- But, R^2 gets larger when the number of variables increases

$$\bar{R}^2 = 1 - \frac{\sum_i e_i^2 / (n-k)}{\sum_i (y_i - \bar{y}_i)^2 / (n-1)}$$

Distribution of “b” 1

- Coefficient vector b follows the normal distribution

$$\mathbf{b} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$$

- “**b**” is the BLUE
  - *Best* : the smallest variance
  - *Linear* : Linear function of y
  - *Unbiased*

Distribution of “b” 2

- Coefficient vector b follows the normal distribution

$$\mathbf{b} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X}))$$

- As to one coefficient bi

$$\frac{b_i - \beta_i}{SD(b_i)} \sim N(0,1)$$

Distribution of "b" 3

- But, the variance of the error term is unknown → you cannot use this!
- When SE is used instead of SD, the following statistics follows t-distribution

$$\frac{b_i - \beta_i}{SE(b_i)} \sim t(n-k)$$

T-value of "b"

- If $\beta i = 0$, this variable is meaningless
- Assuming $\beta i = 0$

$$\frac{b_i}{SE(b_i)} \sim t(n-k)$$

- This statistic is called t-value

$$t(b_i) = \frac{b_i}{SE(b_i)}$$

You can confirm

- If the number of parameters is 2
- "t-value test of bi" is same as "Test of the correlation of x and y"

$$t(b_i) = \frac{b_i}{SE(b_i)} \sim t(n-2)$$

$$T(r) = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t(n-2)$$