Rangkuman Bab 11

Muhammad Akmal
13517028

Simple Linear Regression & Correlation.

Linear Relation

A reasonable form of a relationship between the response Y and the regressor x is the linear relationship: $Y = \beta_0 + \beta_1 x$.

β_0 is the intercept and β_1 is the slope.

Simple Linear Regression Model

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

$\beta_0 \rightarrow$ unknown intercept

$\beta_1 \rightarrow$ unknown slope

$\varepsilon \rightarrow$ a random variable that is assumed to be distributed with $E(\varepsilon) = 0$ and $var(\varepsilon) = \sigma^2$

Fitted Regression Line

- An important aspect of regression analysis is to estimate the parameters β_0 and β_1
- Suppose we denote the estimates b_0 for β_0, b_1 for β_1, then $\hat{y}$ is the predicted / fitted value.
- True regression line: $Y = \beta_0 + \beta_1 x$.
- Estimated ~~off~~ fitted line: $\hat{y} = b_0 + b_1 x$.
- We expect that the fitted line should be closer to the true regression line when a large amount of data are available

Residual: Error in Fit

- A residual is essentially an error in the fit of the model $\hat{y} = b_0 + b_1 x$
- Given a set of regression data and a fitted model, $\hat{y}_i = b_0 + b_1 x_i$, the ith residual e_i is given by:

$$e_i = y_i - \hat{y}_i, \quad i = 1, 2, \ldots, n.$$

$$y_i = b_0 + b_1 x_i + e_i$$

- If a set of n residuals is large, then the fit of the model is not good. Small residuals are a sign of a good fit.

Least Square Estimators (LSE)

- Least squares: minimization procedure for estimating the parameters. We shall find b_0 and b_1, the estimators of β_0 and β_1, so that the sum of the squares of the residuals / errors (SSE) is a minimum.

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

$$b_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Correlation

- Correlation coefficient attempts to measure the strength of relationship between two variables, X and Y.
- Pada bab 4.2, kovariansi dua variabel random X dan Y dengan rataan μ_X dan μ_Y adalah

$$\sigma_{XY} = E(XY) - \mu_X \mu_Y$$

- Variabel random X dan Y dengan kovariansi σ_{XY} dan simpangan baku msg² σ_X dan σ_Y, koefisien korelasi ρ_{XY}.

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

- $-1 \le \rho_{XY} \le 1$

Koefisien korelasi ρ_{XY}

- Koefisien korelasi mengukur asosiasi antar dua variabel, sedangkan gradien menunjukkan arah garis (b_1) pada $\hat{y}_i = b_0 + b_1 x_i + e_i$
- $\rho_{XY} = 0 : b_1 = 0$
- $\rho_{XY} = 1$ if $b_1 > 0$
- $\rho_{XY} = -1$ if $b_1 < 0$

Correlation and Regression

- Regression line $y_i = b_0 + b_1 x_i + e_i$
- The value correlation coefficient ρ is 0 when $b_1 = 0$, which results when there essentially is no linear regression, the regression line is horizontal and any knowledge of x is useless in predict Y.
- The value $\rho = 1$ if $b > 0$ and value $\rho = -1$ if $b < 0$
- Thus a value of $\rho = +1$ implies a perfect linear relationship with a positive slope, while a value of $\rho = -1$ results from a perfect linear relationship with a negative slope.

The Sample Correlation Coefficient (r)

$$SSE = S_{yy} - 2b_1 S_{xy} + b_1^2 S_{xx} = S_{yy} - b_1 S_{xy}$$

di mana

$$S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2, \quad S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}),$$

$$S_{yy} = \sum_{i=1}^{n} (x_i - \bar{x})^2, \text{ maka,}$$

$$r = b_1 \sqrt{\frac{S_{xx}}{S_{yy}}} = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left[\sum_{i=1}^{n} (x_i - \bar{x})^2\right]\left[\sum_{i=1}^{n} (y_i - \bar{y})^2\right]}}$$

$$b_0 = \frac{\sum_{i=1}^{n} y_i - b_1 \sum_{i=1}^{n} x_i}{n} = \bar{y} - b_1 \bar{x}$$

Properties LSE, Model $Y = \beta_0 + \beta_1 X + \varepsilon$.

1. $\sum e_i = 0$.
2. ε_i berdistribusi normal (mean $\mu = 0$ dan variansi σ^2)
3. $SSE = \sum (e_i)^2$ minimum.
4. Taksiran bo dan bi tak bias.
5. Jika $X = \bar{x}$ rataan, maka $\hat{y} = \bar{y}$

A Measure of Quality of Fit: Coefficient of Determination, R^2

- Besaran $R^2 \rightarrow$ koefisien determinasi adalah suatu ukuran proporsi dari variasi model fitted (regresi) dan variasi variable response.
- Variasi model fitted $= SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$
- Variasi variabel response $= SST =$ sum ~~of~~ Square of Total. $\sum_{i=1}^{n} (y_i - \bar{y}_i)^2$
- $R^2 = \left(1 - \frac{SSE}{SST}\right)$
- Nilai koefisien determinasi antara 0 dan 1
- $R^2 = 1$ artinya garis regresi fit sempurna, $SSE = 0$
- Nilai $R^2 = 0$ artinya garis regresi tidak fit sempurna, $SSE = SST$ (hampir sama)

* Some Useful Transformations to Linearize

Functional Form Relating y to x.	Power Transformation	Form of Simple Linear Regression.
Exponential: $y = \beta_0 e^{\beta_1 x}$	$y^* = \ln y$	Regress y^* against x
Power: $y = \beta_0 x^{\beta_1}$	$y^* = \log y$; $x^* = \log x$	Regress y^* against x^*
Reciprocal: $y = \beta_0 + \beta_1 \left(\frac{1}{x}\right)$	$x^* = \frac{1}{x}$	Regress y against x^*
Hyperbolic: $y = \dfrac{x}{\beta_0 + \beta_1 x}$	$y^* = \frac{1}{y}$; $x^* = \frac{1}{x}$	Regress y^* against x^*

Multiple Linear Regression (MLR)

- Persamaan : $Y = b_0 + b_1 x_1 + \dots + b_n x_n + \varepsilon$

MLR dengan Matriks

$$y = Xb + \varepsilon.$$

$$\varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ 1 & x_{13} & \vdots & \cdots & \vdots \\ \vdots & \vdots & & & \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$$

$$SSE = \varepsilon^T \varepsilon = (y - Xb)^T (y - Xb)$$

$$b = (X^T X)^{-1} X^T y$$