Chpater 2 : Linear Regression

Coefficient of determination and Detection of Collinearity and Confidence and Prediction interval

JuHyun Kang

February 23, 2025

The Three Sisters of Newton
School of Mathematics, Statistics and Data Science
Sungshin Women's Universit

Outline

1. Coefficient of Determination and the Detection of Collinearity

2. Confidence and Prediction Intervals

Using Matrices W and H

- We define a matrix $W \in \mathbb{R}^{N\times N}$ such that all the elements are $1/N$
  $Wy \in \mathbb{R}^N$ are $\bar{y} = Wy = \frac{1}{N}\sum_{i=1}^{N} y_i$ for $y_1, \cdots, y_N \in \mathbb{R}$

$$Wy = \begin{bmatrix} 1/N & \cdots & 1/N \\ \vdots & \cdots & \vdots \\ 1/N & \cdots & 1/N \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} (y_1 + \cdots y_N)/N \\ \vdots \\ (y_1 + \cdots y_N)/N \end{bmatrix} = \bar{y}$$

- We can express that $\hat{y} = Hy$

RSS, ESS, and TSS

- Residual Sum of Squares RSS

$$\text{RSS} = ||\widehat{y} - y||^2 = ||Hy - y||^2 = ||(H - I)y||^2 = ||(I - H)y||^2$$

- Explained Sum of Squres ESS

$$\text{ESS} \triangleq ||\widehat{y} - \bar{y}||^2 = ||Hy - Wy||^2 = ||(H - W)y||^2$$

- Total Sum of Squres TSS

$$\text{TSS} \triangleq ||y - \bar{y}||^2 = ||y - Wy||^2 = ||(I - W)y||^2$$

Derivation of TSS = RSS + ESS

- **Proof** by TSS - RSS - ESS = 0

$$||y-\bar{y}||^2 - ||y-\hat{y}||^2 - ||\hat{y}-\bar{y}||^2 = 0$$

$$\Rightarrow (y-\bar{y})^T(y-\bar{y}) - (y-\hat{y})^T(y-\hat{y}) - (\hat{y}-\bar{y})^T(\hat{y}-\bar{y}) = 0$$

$$\Rightarrow y^Ty - y^T\bar{y} - \bar{y}^Ty + \bar{y}^Ty - y^Ty + y^T\hat{y} + \hat{y}^Ty - \hat{y}^T\hat{y} - \hat{y}^T\hat{y} + \hat{y}^T\bar{y} + \bar{y}^T\hat{y} - \bar{y}^T\bar{y} = 0$$

$$\Rightarrow -2\bar{y}^Ty + 2\hat{y}^Ty + 2\bar{y}^T\hat{y} - 2\hat{y}^T\hat{y} = 0$$

$$\Rightarrow -\bar{y}^T(y-\hat{y}) + \hat{y}^T(y-\hat{y}) = 0$$

$$\Rightarrow (\hat{y}-\bar{y})^T(y-\hat{y}) = 0$$

Coefficient of determination and the Correlation Coefficient

- Coefficient of determination

$$R^2 = \frac{\text{ESS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

- Correlation between the covariates and response

$$\hat{\rho} \triangleq \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2}}$$

Relationship Between R^2 and the Correlation Coefficient

- If $p = 1$, R^2 coincides with the square of the sample-based correlation coefficient

$$\frac{\text{ESS}}{\text{TSS}} = \frac{\hat{\beta}_1^2||x-\bar{x}||^2}{||y-\bar{y}||^2} = \left\{ \frac{\sum_{i=1}^N (x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^N (x_i-\bar{x})^2} \right\}^2 \frac{\sum_{i=1}^N (x_i-\bar{x})^2}{\sum_{i=1}^N (y_i-\bar{y})^2}$$

$$= \frac{\left\{\sum_{i=1}^N (x_i-\bar{x})(y_i-\bar{y})\right\}^2}{\sum_{i=1}^N (x_i-\bar{x})^2] \sum_{i=1}^N (y_i-\bar{y})^2} = \hat{\rho}^2$$

- We sometimes use a variant of the coefficient of determination which is the adjusted coefficient of determination

$$1 - \frac{\text{RSS}/(N-p-1)}{\text{TSS}/(N-1)}$$

Variance Inflation Factors

- VIF measures the redundancy of each covariate when the other covariates are present
$$\mathrm{VIF} \triangleq \frac{1}{1 - R^2_{X_j | X_{-j}}}$$
- The minimum value of VIF is one, and we say that the collinearity of covariate is strong when its VIF value is large

Outline

1 Coefficient of Determination and the Detection of Collinearity

2 Confidence and Prediction Intervals

Confidence Interval

- We have showed how to obtain the estimate $\hat{\beta}$ of $\beta \in \mathbb{R}^{p+1}$, confidence interval of $\hat{\beta}$ as follows

$$\beta_i = \hat{\beta}_i \pm t_{N-p-1}(\alpha/2)\mathrm{SE}(\hat{\beta}_i), \quad \text{for} \;\; i = 0, 1, \cdots, p$$

- Confidence interval of $x_* \hat{\beta}$ for another point $x_* \in \mathbb{R}^{p+1}$
  - The average

$$E[x_* \hat{\beta}] = x_* E[\hat{\beta}]$$

  - The variance

$$V[x_* \hat{\beta}] = x_* V(\hat{\beta}) x_*^T = \sigma^2 x_* (X^T X)^{-1} x_*^T$$

- We define

$$\hat{\sigma} \triangleq \sqrt{\mathrm{RSS}/(N-p-1)}, \quad \mathrm{SE}(x_* \hat{\beta}) \triangleq \hat{\sigma} \sqrt{x_* (X^T X)^{-1} x_*^T}$$

Confidence and Prediction Intervals in Regression

- C $\sim t_{N-p-1}$

- variance in the difference between $x_*\hat{\beta}$ and $y_* \triangleq x_*\beta + \varepsilon$

$$V[x_*\hat{\beta} - (x_*\beta + \varepsilon)] = V[x_*(\hat{\beta} - \beta)] + V[\varepsilon] = \sigma^2 x_*(X^TX)^{-1}x_*^T + \sigma^2$$

- Similarly, we can derive the following

$$P \triangleq \frac{x_*\hat{\beta} - y_*}{\mathrm{SE}(x_*\hat{\beta} - y_*)} = \frac{x_*\hat{\beta} - y_*}{\sigma(1 + \sqrt{x_*(X^TX)^{-1}x_*^T})} \Big/ \sqrt{\frac{\mathrm{RSS}}{\sigma^2} \Big/ (N-p-1)} \sim t_{N-p-1}$$

- The confidence and prediction intervals

$$x_*\beta = x_*\hat{\beta} \pm t_{N-p-1}(\alpha/2)\hat{\sigma}\sqrt{x_*(X^TX)^{-1}x_*^T}$$
$$y_* = x_*\hat{\beta} \pm t_{N-p-1}(\alpha/2)\hat{\sigma}\sqrt{1 + x_*(X^TX)^{-1}x_*^T}$$

Q & A

Q & A

Thank you :)