National Taiwan University
Institute of Industrial Engineering
Jakey BLUE

Data Analytics

109-2 Homework #04

Due at 23h59, March 28, 2021; files uploaded to NTU-COOL

1. (10%) Given a simple linear regression model: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i, i = 1, \dots, n$, where $\epsilon_i \sim_{iid} \mathrm{N}(\mu, \sigma^2)$ Show that:
   a. $\mathrm{cov}(\hat{\beta}_0, \hat{\beta}_1) = -\bar{x}\sigma^2/S_{xx}$
   b. $\mathrm{cov}(\bar{y}, \hat{\beta}_1) = 0$

2. (10%) Show that the regression sum of squares can be calculated as:
$$SS_R = \left(\sum_{i=1}^{n} \hat{y}_i^2\right) - n\bar{y}^2$$

3. (10%) The matrix, $\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$, derived in multiple regression is usually defined as $\mathbf{H}$. Show that:
   a. $\mathbf{H}$ is idempotent, i.e., $\mathbf{HH} = \mathbf{H}$ and $(\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H}) = \mathbf{I} - \mathbf{H}$
   b. $\mathrm{V}(\hat{\mathbf{y}}) = \sigma^2\mathbf{H}$

4. (10%) Investigate and explain why R^2 cannot be larger than 1 or smaller than 0. (Do not copy directly from the source you found, but explain in your own words.)

5. (15%) In a multiple regression model: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, it is critical to know if $(\mathbf{X}^T\mathbf{X})^{-1}$ exists. The diagonal elements of $(\mathbf{X}^T\mathbf{X})^{-1}$ in correlation form, i.e., $\mathbf{X}$ is normalized, are often called Variance Inflation Factors (VIFs), and they are important multicollinearity diagnostic. VIF for the j^{th} regression coefficient is expressed as
$$\mathrm{VIF}_j = \frac{1}{1 - R_j^2},$$
where R_j^2 is the coefficient of multiple determination obtained from regressing $\mathbf{x}_j$ on the other regressor variables ($\mathbf{x}_1$ to $\mathbf{x}_p$, except $\mathbf{x}_j$). Calculate all the VIFs in the autompg dataset and discuss your observation.