Multiple Linear Regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \epsilon, \qquad \mathbb{E}(\epsilon | X_1, X_2, ..., X_k) = 0$$

- The estimated sample regression equation is given by

$$\widehat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + ... + \hat{\beta}_k X_k, \qquad \text{residuals: } e_i = y_i - \widehat{y}_i$$

- The OLS estimator minimizes the sum of squared errors (SSE):

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i} - ... - \hat{\beta}_k X_{ki})^2$$

- The following relationship always holds:

$$SST = SSR + SSE \quad \Leftrightarrow \quad \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\widehat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$$

THE UNIVERSITY OF MELBOURNE

Goodness of Fit

- The (sample) coefficient of determination is given by

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

  - $0 \leq R^2 \leq 1$. The closer R^2 is to 1, the better the model fits the data.
  - However, R^2 can be inflated if (i) sample size (n) is small relative to the number of independent variables (k), or (ii) if more independent variables are added to the model.

- The adjusted coefficient of determination is given by

$$\bar{R}^2 = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)} = 1 - \frac{n-1}{n-k-1}(1-R^2)$$

  - Adjusted R^2 penalises the addition of independent variables that do not improve the model fit.
  - $\bar{R}^2 < R^2$ unless the model fit is perfect. Then, $\bar{R}^2 = R^2 = 1$.

THE UNIVERSITY OF MELBOURNE

Assumptions of Linear Regression Model

- (LR1) random sample of $n > k+1$ independent but identically distributed (iid) observations.
- (LR2) random error has zero conditional expected value, i.e., $\mathbb{E}(\epsilon_i | X_1, X_2, \dots, X_k) = 0$.
- (LR3) random error has constant conditional variance, i.e., $Var(\epsilon_i | X_1, X_2, \dots, X_k) = \sigma^2$.
- (LR4) conditional covariance between random errors is zero, i.e., $\mathbb{E}(\epsilon_i \epsilon_j | X_1, X_2, \dots, X_k) = 0$ for $i \neq j$.
- (LR5) independent variables are not perfectly multicollinear (i.e., no exact linear relationship among the independent variables).
- (LR6) conditional distribution of the random error is normal, i.e., $\epsilon_i \sim N(0, \sigma^2)$.

Under LR1-LR5, the OLS estimator is BLUE (Best Linear Unbiased Estimator).