Simple Linear Regression – Assumptions of the Simple Linear Regression Model

1. The value of y for each value of x is: $y_i = \beta_1 + \beta_2 x_i + e_i$
2. The expected value of the random error is: $E(e_i) = 0$
3. The variance of the random error e is: $\text{var}(e_i) = \sigma^2$
4. The covariance between any pair of random errors e_i and e_j is: $\text{cov}(e_i, e_j) = 0$
5. The variable x is not random $E\left(\left(E(y_i) = E\left(\frac{y_i}{x_i}\right)\right)\right)$ and must take at least two different variables
6. We often assume the errors are normally distributed $e_i \sim Normal(0, \sigma^2)$

Ordinary Least Squares Principle

The estimates b_1 and b_2 are chosen so as to make the sum of squared residuals

$$\sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} (y_i - b_1 - b_2 x_i)^2$$

$$= \sum_{i=1}^{n} \left\{ y_i^2 - 2b_1 y_i - 2b_2 x_i y_i + 2b_1 b_2 x_i + b_1^2 + b_2^2 x_i^2 \right\}$$

as small as possible

Estimators

$$b_1 = \bar{y} - b_2 \bar{x}$$

$$b_2 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\left(\Sigma(x_i - \bar{x})^2\right)} = \sum_{i=1}^{N} w_i y_i \qquad w_i = \frac{x_i - \bar{x}}{\Sigma(x_i - \bar{x})^2}$$

Linearity of Least-Square Estimates

The estimates are linear because they are a linear function of the y_i.

Unbiased Estimates

$$E(b_2) = \mathrm{E}\left(\beta_2 + \sum_{i=1}^{N} w_i e_i\right) = \beta_2 + \sum_{i=1}^{N} w_i E(e_i) = \beta_2$$

$$E(b_1) = E(\bar{y} - b_2 \bar{x}) = E\left(\beta_1 + \beta_2 \bar{x} + \frac{1}{N}\Sigma e_i - b_2 \bar{x}\right)$$

$$= \beta_1 + \beta_2 \bar{x} + \frac{1}{N}\Sigma E(e_i) - E(b_2 \bar{x}) = \beta_1$$

Variances

$$\text{var}(b_1) = \frac{\Sigma x_i^2}{N}\text{var}(b_2); \quad \hat{\text{var}}(b_1) = \frac{\Sigma x_i^2}{N}\hat{\text{var}}(b_2)$$

$$\text{var}(b_2) = \frac{\sigma^2}{\Sigma(x_i - \bar{x})^2}; \quad \hat{\text{var}}(b_2) = \frac{\hat{\sigma}^2}{\Sigma(x_i - \bar{x})^2}$$

$$\hat{\sigma}^2 = \frac{\Sigma \hat{e}_i^2}{N-2}$$

Gauss Markov Theorem

Under assumptions 1-5, the estimators b_1 and b_2 have the smallest variance of all linear and unbiased estimators of β_1 and β_2.

They are the **Best Linear Unbiased Estimators** – **BLUE**.

Goodness-of-Fit: The Coefficient of Determination

R^2: the proportion of variation in y explained by x within the regression model.

R^2 is a descriptive measure.

The objective of regression analysis <u>is not to maximise</u> R^2.

$$R^2 = \frac{SSR}{SST} = 1 - \left(\frac{SSE}{SST}\right)$$
$$SST = SSE + SSR$$

Total Sum of Squares

$SST = \Sigma(y_i - \bar{y})^2$

A measure of total variation in y about the sample mean

Explained Sum of Squares

$SSR = \Sigma(\hat{y}_i - \bar{y})^2$

That part of total variation in y about the sample mean that is explained by or due to regression

Sum of Squared Residuals

$SSE = \Sigma \hat{e}_i^2$

That part of total variation in y about its mean that is not explained by the regression

100(1–α)% Confidence Interval

$P\left(b_2 - t_c \text{se}(b_2) \le \beta_2 \le b_2 + t_c \text{se}(b_2)\right) = 1 - \alpha$

Least Squares Prediction

$\hat{y}_0 = b_1 + b_2 x_0$

Forecast Error

$f = y_0 - \hat{y}_0 = (\beta_1 + \beta_2 x_0 + e_0) - (b_1 + b_2 x_0)$

Variance of the Forecast Error

$$\text{var}(f) = \sigma^2\left[1 + \frac{1}{N} + \frac{(x_0 - \bar{x})^2}{\Sigma\left((x_i - \bar{x})^2\right)}\right]$$

Prediction Interval

The 100(1–α)% prediction interval as: $\hat{y}_0 \pm t_c se(f)$