Chapter 1A: Review of Simple Linear Regression (SLR)

(1) LR Model

- Error term (e): the variation in the dependent variable due to other variables that are not accounted in the model
- Assumptions:
  - $E(e) = 0$
  - e and independent variables are uncorrelated
  - $e \sim N(0, \sigma^2)$
- $Y = \beta_0 + \beta_1 x_1 + e$ with $e \sim N(0, \sigma^2)$. By assumption, Y has a mean of $\beta_0 + \beta_1 x_1$ and is normally distributed with a standard deviation (SD) of σ
- The 95% prediction interval (PI) is: $(\hat{\beta}_0 + \hat{\beta}_1 x_1) \pm 1.96 \underbrace{\sigma}_{RMSE}$

(2) Error SD

$\sigma = \sigma_Y \sqrt{(1 - r^2_{X,Y})}$	σ : error SD of the errors σ_Y : SD of Y r^2_{xy} : correlation of X and Y

- To use this formula, the homoscedasticity assumption needs to hold
- $\hat{\sigma}$ = RMSE on Stata

(3) Residuals

- Residuals are estimated values of the errors
- We can evaluate the residuals, but we do not observe the errors

$$residual = \hat{e} = Y - \hat{Y}$$

  - Fitted value is the composite term: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$ → estimate of the regression line for a given value of x_1 (part of the observation that is explained by the independent variables)
  - Error is the random part that is not explained

(4) Standardised residuals

- If the regression model assumptions are correct → $e \sim N(0, \sigma^2)$. Thus $\frac{e}{\sigma}$ will have a standard normal distribution and should lie between ± 2.6
- e is unknown → use the estimates of the errors ($\hat{e}$)

$$Standardised\ residual = \frac{\hat{e}}{\hat{\sigma}}$$

- It allows us to check for: *(i) normality in the errors, and (ii) outliers*
- Problem with using standardised residual to detect outliers: outliers near the 'end' of the line may swing the line towards it and lessen the residual, making it harder to detect

(5) Outliers

- For observations with large positive standardised residuals (i.e. outliers) there is a departure from normality
- Outliers may distort the whole fit of the regression line
- Only **include observations** with **studentized residuals between $\pm$ 3** and a **normalized leverage of less than 2**

(6) Studentized or *t* residuals

- A *better* version of the standardised residuals
- To obtain the *i*th studentized residual, we use all data except the *i*th case to estimate the β_0, β_1 and σ
- If the *i*th observation is an outlier, omitting that observation when estimating β_0, β_1 and σ allows us to get a better idea if the *i*th observation is an outlier

(7) R^2

- R : the absolute value of the correlation between X and Y

$1 - R^2 = \frac{\sigma^2}{Var(Y)}$	Proportion of the variation in Y that is not explained by X
$\sqrt{(1 - R^2)} = \frac{\sigma}{\sqrt{Var(Y)}}$	Ratio of the SD of the prediction error to the SD of Y

(8) Leverage

- Leverage is the potential ability of a data point to affect the estimated regression line
- The larger the distance of the X value from the center of the X's the higher the leverage
- An outlier with a high leverage have more influence on the regression line than a low leverage outlier

(9) Checking normality of the errors

- It is done because we do not have the errors – we can only examine the residuals for normality
- *Normal probability plot of the studentized residual* → *if the observations come from a normal distribution*, the normal probability plot will be a scatter about the 45° line

(10) Normal quantile plot

- Plots the quantiles of the studentized residuals vs the quantiles of a standard normal
- Similar to the normal probability plot – it *aims to detect non-normality in the data*
- If the data sample comes from a standard normal distribution, there should be a scatter about a straight 45° line in the normal quantile plot

(11) Prediction interval (PI)

- PI is an in-sample prediction and it is an interval around the fitted value

$$PI = fitted \pm 1.96 * \underbrace{stdf}_{\substack{SD\ of\ the \\ forecast\ errors}}$$

- Meanwhile, the confidence interval (CI) provides an estimate of an unknown parameter, given a particular significance level

(12) Adjusted R^2 (R^2_a)

- R^2 tends to overestimate how well the models fit the population → R^2_a more closely reflect the goodness of fit of the model in the population

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

$$R^2_a = 1 - \frac{\frac{ESS}{n-p-1}}{\frac{TSS}{n-1}}$$

$$R^2_a = 1 - \frac{s^2}{s^2_y}$$

ESS: residual or error sum of squares

$$ESS = \sum_{i=1}^{n} \hat{e}_i^2$$

TSS: total sum of squares

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2 = (n-1)Var(Y)$$

RSS: regression sum of squares

$$RSS = TSS - ESS$$

$$RSS = \sum_{i=1}^{n} fitted_i^2$$

n: number of observations

p: number of explanatory variables

- R^2_a penalises for extra variables through the term $\frac{n-1}{n-p-1}$, which increases with p. Thus, ESS decreases as we add more variable
- R^2_a adjusts estimates of variances for degree of freedom; i.e. $\frac{ESS}{n-p-1}$ is an unbiased estimator of σ^2 and $\frac{TSS}{n-1}$ is an unbiased estimator of $Var(y)$
- $1 - R^2_a$ is the ratio of the unbiased estimate of σ^2 from the regression to the unbiased estimate of $Var(y)$. Smaller estimate is better

Chapter 1B: Regression theory for SLR

(1) SLR and least squares

$$y_i = \beta_0 + \beta_1 x_i + e_i, \quad i = 1, \ldots, n$$

Assuming that $e_i \sim N(0, \sigma^2)$ and independent

- Without loss of generality, we can write: $y_i = \beta_0 + \beta_1 (x_i - \bar{x}) + e_i$
  Because $y_i = \underbrace{(\beta_0 - \beta_1 \bar{x})}_{New\ intercept} + \beta_1 x_i + e_i$
- To estimate β_0 and β_1 by least squares, minimise and satisfy the FOCs

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 (x_i - \bar{x})\right)^2$$

$$\frac{\partial Q(\beta_0, \beta_1)}{\partial \beta_0} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 (x_i - \bar{x})\right) = 0$$

$$\sum_{i=1}^{n} y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} (x_i - \bar{x}) = 0$$

$$\because \sum_{i=1}^{n} (x_i - \bar{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x} = n\bar{x} - n\bar{x} = 0$$

$$\therefore \hat{\beta}_0 = \frac{\sum_{i=1}^{n} y_i}{n} = \bar{y}$$