Appendix 1
Explanation of R Squared and Adjusted R Squared

R Squared and Adjusted R Squared

Lines of best-fit can be defined mathematically using a simple regression model equation such as the one below:

$$Y = \alpha + \beta X + \mathcal{E}$$

Where,

Y is the dependent variable;
α is the intercept;
β is the coefficient on the explanatory variable;
X is the explanatory variable;
$\mathcal{E}$ is the error term

The final term in the equation above is the error term, and this refers to the distance between a data point on an XY plot and the true regression line. Thus, this equation can be rearranged as follows to illustrate this:

$$\mathcal{E} = Y - \alpha - \beta X$$

The purpose of this study has been to estimate a regression equation and its line of best-fit. Consequently, the regression model is an *estimate* of the *true* relationship amongst the variables. The error terms in estimated models are referred to as *residuals*, to indicate that they are estimates and not true values. The residual will therefore be the distance between a data point on an XY plot and the fitted (or estimated) regression line. OLS is used to estimate regression models and thereby to estimate values for the coefficients α and β. This estimation is done by estimating values for α and β such that the value of the sum of squared residuals (SSR) is minimised, where

$$SSR = \sum \mathcal{E}^2$$

It should be noted $\mathcal{E}$ in the above equation refers to the residuals and not the errors. OLS estimation can be done manually but is generally carried out using statistical software.

It is convenient at this stage to discuss R-squared, or R^2. Is has just been said that OLS estimation attempts to find the line of best-fit by minimising the SSR. It may be the case, however, that the best-fitting line is not actually a very good fit. It is therefore necessary to have a means of measuring how good a fit the estimated regression line is. This is frequently done by using R-squared. In fact, in this study adjusted R-squared has been used. Adjusted R-squared is mildly different to R-squared and will be explained after R-squared has been explained.

The dependent variable will have taken on many different values during the sample period (as many values as there are points on the XY plot). Using these different values, it is possible to calculate a mean, a variance and a standard deviation. The latter two measures describe how much variation there is in the values of the dependent variable. A similar term to variance is used when calculating R-squared: it is the total sum of squares (TSS) as defined below:

$$TSS = \sum(Y_i - \bar{Y})^2$$

Where Y_i refers to an individual observation of Y (the dependent variable) and $\bar{Y}$ refers to the mean value for Y (the sum of all the observations of the dependent variable divided by the number of observations). TSS is thus a measure of the variability of Y. The total variability of Y can be broken down into two components as shown below:

$$TSS = RSS + SSR$$

Where,

$$RSS = \sum(\hat{Y}_i - \bar{Y})^2$$

$\hat{Y}_i$ refers to an estimated value for Y. The total variation in Y, therefore, can be broken down into the part that can be explained by the independent variables in the regression model (this is the RSS part) and how much cannot be explained by the regression model (the SSR part). Thus, R-squared is defined as follows:

$$R^2 = RSS/TSS$$

Or, equivalently:

$$R^2 = 1 - SSR/TSS$$

R-squared can only take on a value equal to or between 0 and 1. A value of 1 implies that the line of best-fit is a perfect fit, whilst a value of 0 implies that the regression model in no way explains the variation in the dependent variable. Obviously, the higher the R-squared then the better if we are attempting to find regression models that explain the relationship between independent and dependent variables. To provide further insight into R-squared, a regression line that is a perfect fit will effectively have no residuals, such that the value of SSR will be zero. Plugging a zero value for SSR into the R-squared equation will yield a value of 1 for R-squared, thus proving that an R-squared of 1 denotes a perfectly fitting line of best fit, with no residuals.

There is a problem with the R-squared measure in that it does not take into account the inclusion of additional *irrelevant* variables into the regression equation. Adding an additional explanatory variable into a regression model will always increase R-squared even if that variable is not statistically significant. This is because R-squared is a measure of best-fit and adding a new variable to the model cannot make the fit any worse. Even if the coefficient on the new variable is zero (meaning that it has no significance whatsoever) the inclusion of this new variable cannot make the fit any worse. Generally, adding a new variable will increase R-squared. Adjusted R-squared caters for this problem in that its value does not always rise when a new variable is added to the regression equation. This is why adjusted R-squared has been the measure used in this study in preference to R-squared.