LM01 Basics of Multiple Regression and Underlying Assumptions

Uses of multiple linear regression

Multiple regression allows us to determine the effect of more than one independent variable on a particular dependent variable.

Multiple regression can be used:

- To explain the relationships between financial variables: e.g., the relationship between inflation, GDP growth rates and interest rates.
- To test existing theories – e.g., are equity returns impacted by a stock's market cap and value/growth factors.
- To make forecasts – e.g., using variables such as financial leverage, profitability, revenue growth, and changes in market share to predict whether a company will face financial distress.

The basics of multiple linear regression

A multiple linear regression model has the general form:

$$Y_i = b_0 + b_1X_{1i} + b_2X_{2i} + \cdots + b_kX_{ki} + \varepsilon_i \text{ , } i = 1, 2, \ldots n$$

The slope coefficient b_j measures how much the dependent variable Y changes when the independent variable, X_j, changes by one unit holding all other independent variables constant.

The intercept coefficient b_0 represents the expected value of Y if all independent variables are zero.

For example, consider the following regression equation:

$Y = 0.2 + 0.6X_1 + 0.5\ X_2 + \epsilon$

If X_1 changes by 1 unit and X_2 remains constant, then Y will change by 0.6 units. Similarly, if X_1 remains constant and X_2 changes by 1 unit, then Y will change by 0.5 units. If X_1 and X_2 are each zero, then the expected value of Y is 0.2.

Assumptions underlying multiple linear regression

The five main assumptions underlying multiple regression models are:

1. Linearity: The relationship between the dependent variable and the independent variables is linear.
2. Homoskedasticity: The variance of the regression residuals is the same for all observations.
3. Independence of errors: The observations are independent of one another. This

implies the regression residuals are uncorrelated across observations.

4. Normality: The regression residuals are normally distributed.
5. Independence of independent variables:

   5a. Independent variables are not random.

   5b. There is no exact linear relation between two or more of the independent variables or combinations of the independent variables.

Commonly used diagnostic plots

Scatterplots of dependent and independent variables are used to check if the assumptions of 'linearity' and 'independence of independent variables' have been violated.

Scatterplots of residuals is used to check if the assumptions of 'homoskedasticity' and 'independence of errors' have been violated.

A 'Q-Q' plot of residuals is used to check if the assumption of 'normality' has been violated.