1. INTRODUCTION

Multiple linear regression is a method used to model the linear relationship between a dependent variable and ***more than one*** independent variables.

Multiple linear regression lets analysts estimate using more complex models with multiple explanatory variables.

Multiple linear regression can be used to predict future returns, improve portfolio construction, and understand security drivers. However, it can produce spurious results and poor predictions if used incorrectly.

How the model works

- The analyst specifies the dependent and independent variables and uses software to estimate the model and generate statistics.
- The software does the estimation and the analyst's task is to interpret the output.

2. USES OF MULTIPLE LINEAR REGRESSION

Financial and economic relations are complex and require models with multiple explanatory variables that must pass rigorous statistical and theoretical scrutiny.

Multiple regression can be used to identify relationships between variables to test current or future theories.

For example, examining multiple factors such as political stability, economic conditions, and ESG considerations on stock returns.

The Regression Process

The objective is to use the variation of two or more independent variables to explain the variation in the dependent variable.

Find out whether the dependent ***variable is continuous***? If the answer is:

- No, use logistic regression
- Yes, use traditional regression model

⬇

Estimate the regression model. Examine the model to see ***if it meets the key assumptions***? If the answer is:

- No, adjust the model
- Yes, examine the model's goodness-of-fit

⬇

Determine whether the ***overall fit is significant***? If the answer is:

- No, adjust the model.
- Yes, determine if this model is the best of all possible models?

⬇

Check ***if the model is the best of all models***? If the answer is:

- No, adjust the model
- Yes, use the model for analysis and prediction.

Refer to: Exhibit 2: The Regression Process' from the CFA Institute's Curriculum.

Practice: Questions under 'Knowledge Check' from the CFA Institute's Curriculum.

3. THE BASICS OF MULTIPLE REGRESSION

Regression analysis is a tool used to examine whether a variable is useful to explain another variable.

Multiple regression model predicts the value of a dependent variable based on the value of two or more independent variable.

Variation represents the difference between observation and its expected value i.e., how much estimatedI ith value differs from the actual ith value.

A multiple linear regression model has the following general form:

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

where,

Y_i = i^{th} *observation of dependent variable Y*
X_{ki} = i^{th} *observation of* k^{th} *independent variable X*
β_0 = *intercept term*
β_k = *slope coefficient of* k^{th} *independent variable*
ε_i = *error term of ith observation*
n = number of observations
k = total number of independent variables

- ***A slope coefficient,*** β_j ***is known as partial regression coefficients or partial slope coefficients.*** It 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 term*** (β_0) is the value of the dependent variable when the independent variables are all equal to zero.
- A regression equation has k slope coefficients and one intercept i.e., k + 1 regression coefficients.

Practice: Questions under 'Knowledge Check' from the CFA Institute's Curriculum.

4. ASSUMPTIONS UNDERLYING MULTIPLE LINEAR REGRESSION

The five key assumptions of the multiple linear regression model are:

1) Linearity
2) Homoskedasticity
3) Independence of errors
4) Normality
5) Independence of independent variables

1 Assumption 1: Linearity

'Relation between the dependent variable Y and the independent variables (X_1, X_2, . . . ,X_k *)is linear.'*

2 Assumption 2: Homoskedasticity

'The variance of residuals is the same for all observations. It is known as Homoskedasticity (same scatter) assumption.'

3 Assumption 3: Independence of errors

'The observations (pairs of Xs and Ys) are independent of each other, which implies the residuals are uncorrelated across observations. (i.e. ***no serial correlation***).'

4 Assumption 4: Normality

'The regression residuals (error term) must be normally distributed.'

5 Assumption 5: Independence of independent variables

a) Independent variables (X_1, X_2, . . . ,X_k) are not random.
b) No exact linear relation exists between two or more of the independent variables.

Note:

- When an exact linear relationship exists between two or more independent variables, linear regression ***cannot*** be estimated.
- Furthermore, when two or more independent variables are highly correlated, the model can be estimated but its interpretation is ***problematic***.

Normal Q-Q Plot

A Q-Q plot is used to compare the distribution of a variable to a normal distribution.

A Q-Q plot is used in regression to compare the model's standardized residuals to a theoretical standard normal distribution. The residuals should align along the diagonal if they are normally distributed.

Refer to: Exhibit 8: Normal Q-Q Plot of Regression Residuals 'from the CFA Institute's Curriculum.

Practice: Questions under 'Knowledge Check' 'and end-of-chapter questions from the CFA Institute's Curriculum and FinQuiz Question-bank.