LM01 Basics of Multiple Regression and Underlying Assumptions

This document should be read in conjunction with the corresponding learning module in the 2024 Level II CFA® Program curriculum. Some of the graphs, charts, tables, examples, and figures are copyright 2023, CFA Institute. Reproduced and republished with permission from CFA Institute. All rights reserved.

Required disclaimer: CFA Institute does not endorse, promote, or warrant the accuracy or quality of the products or services offered by IFT. CFA Institute, CFA®, and Chartered Financial Analyst® are trademarks owned by CFA Institute.

Version 1.0

1. Introduction

Multiple linear regression is used to model the linear relationship between one dependent variable and two or more independent variables.

When constructing regression models, most of the heavy computational work is done by statistical software such as: Excel, Python, R, SAS, and STATA. They can estimate the model parameters and produce related statistics. An analyst's primary role is to specify the model correctly and to interpret the output from statistical software.

2. Uses of Multiple Linear Regression

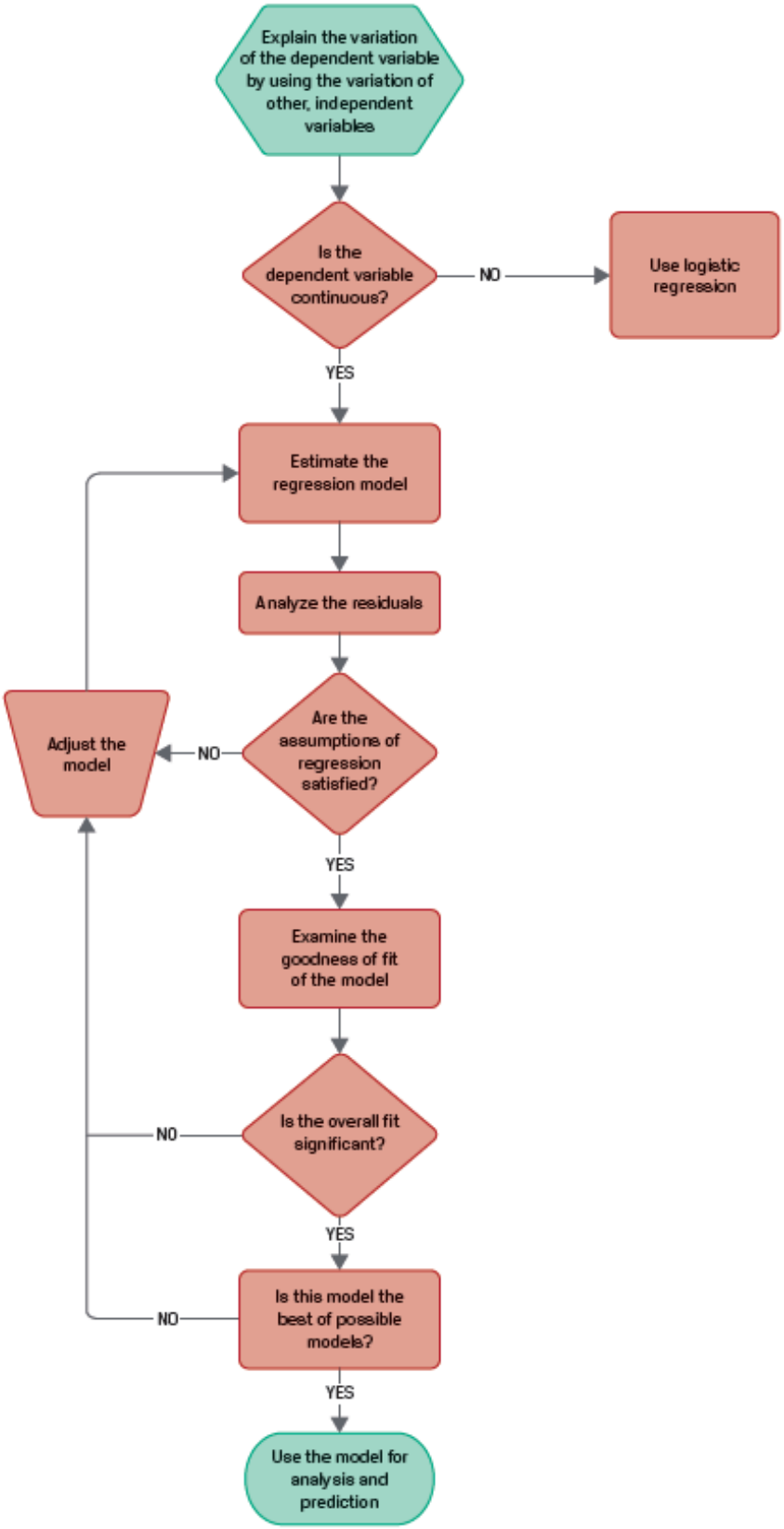
In practice, multiple regression can be used:

- 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.

Exhibit 2 from the curriculum outlines a general process of regression analysis.

- We first start by determining if the dependent variable is continuous (e.g. returns) or discrete (e.g. takeover target or not a takeover target). For continuous variables, traditional regression models can be used. For discrete variables, a logistic regression model is needed.
- We then estimate the regression model and analyze the residuals to see if any key underlying regression assumptions are violated. If violations occur, the model has to be adjusted.
- Next, we examine a model's 'goodness of fit' to check if the overall fit is significant. The model has to be adjusted until it meets the analyst's criteria.
- After a model has been deemed acceptable (i.e. the regression assumptions are satisfied, the overall fit is significant, and the model is the best model of the possible models) we can use it for analysis and forecasting.

Instructor's Note: These steps are covered in detail later in this reading and in subsequent readings.



3. The Basics of Multiple Regression

A multiple linear regression model has the general form:

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

where:
Y_i = the *i*th observation of the dependent variable
$X_{1i}, \ldots, X_{ki}$ = the ith observation of the independent variables
b_0 = the intercept term
$b_1, \ldots, b_k$ = slope coefficients for each of the independent variables
ε_i = the error term of the *i*th observation
n = the number of observations

A regression equation has one intercept coefficient and k slope coefficients (also called partial regression coefficients), where k is equal to the number of independent variables.

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. 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.

The intercept coefficient b_0 represents the expected value of Y if all independent variables are zero. In our example, if X_1 and X_2 are each zero, then the expected value of Y is 0.2.

Example:
(*This is the Knowledge Check example from Sec 3 in the curriculum.*)

An institutional salesperson has just read the research report in which you estimated a regression of monthly excess returns on a portfolio, RETRF, against the Fama–French three factors:

- MKTRF, the market excess return;
- SMB, the difference in returns between small- and large-capitalization stocks; and
- HML, the difference in returns between value and growth stocks.

All returns are stated in whole percentages (that is, 1 for 1%), and the estimated regression equation is

RETRF = 1.5324 + 0.5892MKTRF + −0.8719SMB + −0.0560HML.

Before this salesperson meets with her client firm, she asks you to do the following regarding your estimated regression model:

---

1. Interpret the intercept.

Solution

If the market excess return, SMB, and HML are each zero, then we expect a return on the portfolio of 1.534%.

---

2. Interpret each slope coefficient.

Solution

Each slope coefficient is interpreted assuming the other variables are held constant.

- For MKTRF, if the market return increases by 1%, we expect the portfolio's return to increase by 0.5892%.
- For SMB, if the size effect returns increase by 1%, we expect the portfolio's return to decrease by 0.8719%.
- For HML, if the value effect returns increase by 1%, we expect the portfolio's return to decrease by 0.056%.

---

3. Calculate the predicted value of the portfolio's return if

MKTRF = 1, SMB = 4, and HML = −2.

Solution

Given the expected values of the independent variables, the expected return on the portfolio is:

$R = 1.534 + 0.5892(1) - 0.8719(4) - 0.0560(-2) = -1.2524$.

4. 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.

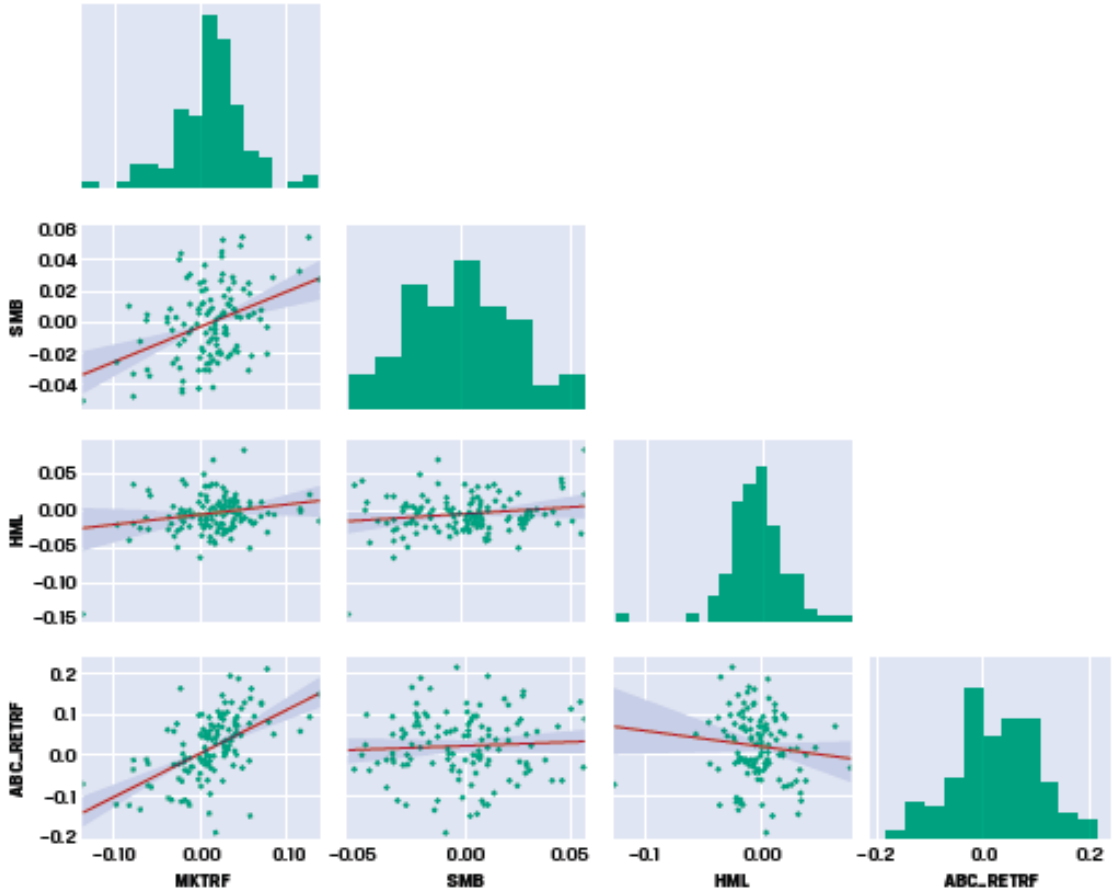
Regression software produces diagnostic plots which can help detect if these assumptions are violated. Commonly used diagnostic plots are discussed below.

Scatterplots of dependent and independent variables: A scatterplot matrix (also referred to as a pairs plot) is useful for detecting non-linear relationships.

For example, consider a model that explains the excess return of ABC stock using market excess return (MKTRF), size (SMB) and value (HML) as explanatory variables.

$$ABC\_RETRF_t = b_0 + b_1MKTRF_t + b_2SMB_t + b_3HML_t + \varepsilon_t$$

The regression software uses 10-years of monthly data and produces the following scatterplot matrix.



The bottom row shows the scatter plot between Y and each of the three independent variables. We can draw the following conclusions:

- There is a positive relationship between ABC_RETF and the market factor, MKTRF.
- There seems to be no apparent relation between ABC_RETRF and the size factor, SMB.
- There is a negative relationship between ABC_RETF and the value factor, HML.

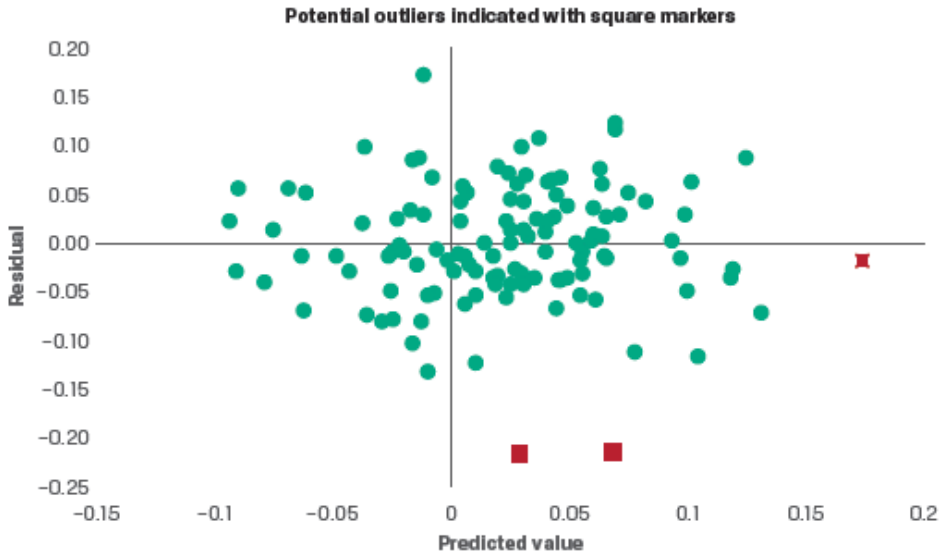
Instructor's Note: If you observe a non-linear relationship between Y and any independent variable (e.g. a curve instead of a straight line) then the regression assumption of 'linearity' has been violated.

Next, we look at the scatterplot between the independent variables. The relatively flat line for the SMB-HML pair indicates that SMB and HML have little to no correlation. This is a desirable characteristic between explanatory variables.

Instructor's Note: If you observe a strong relationship between two independent variables then the regression assumption of 'independence of independent variables' has been violated.

Scatterplots of residuals: This plot is useful for detecting violations of homoskedasticity and independence of errors. It can also help identify outliers in our data.

We first look at the scatterplot of residuals against the dependent variable.



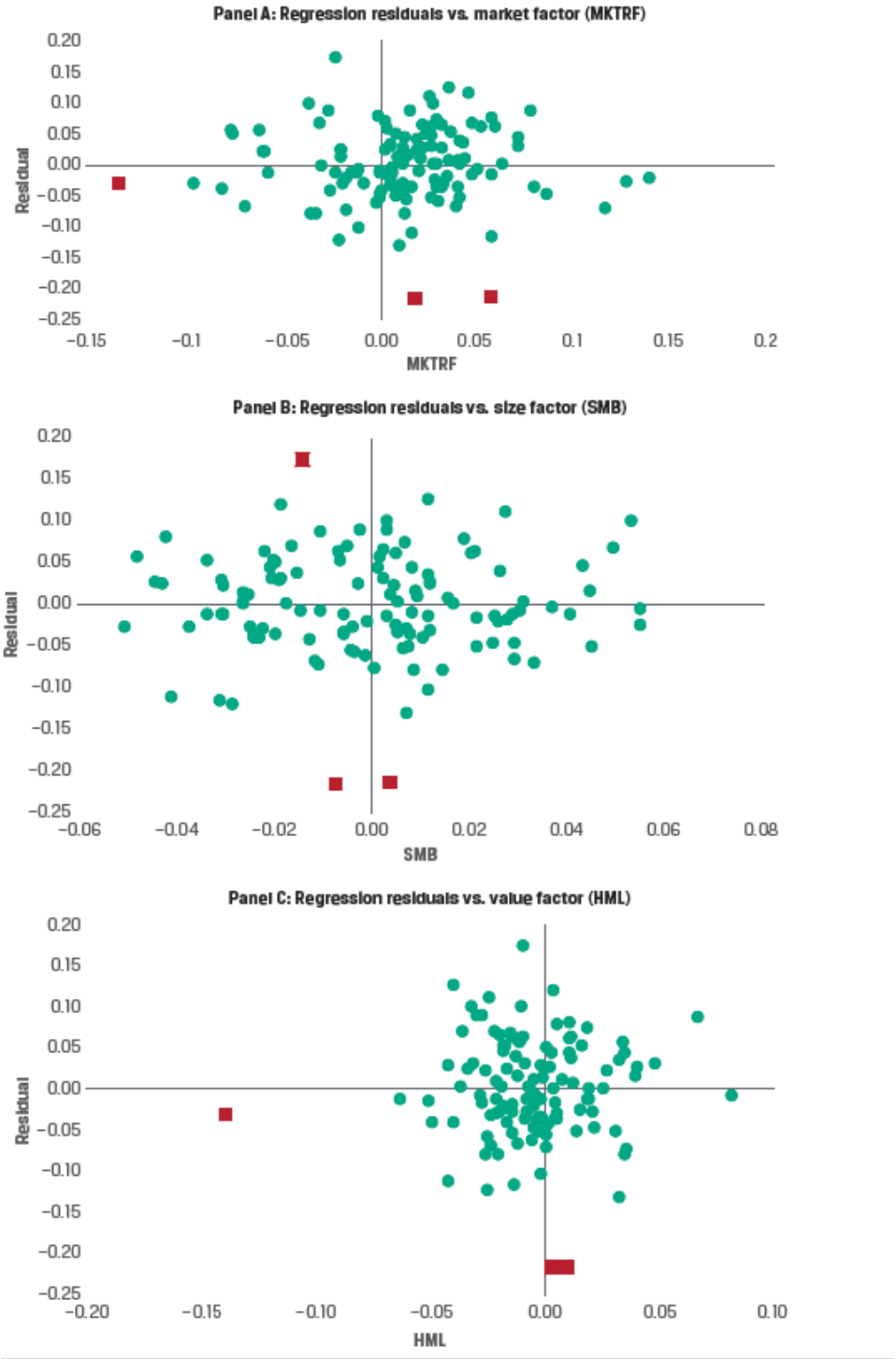
Potential outliers indicated with square markers

As indicated by the line centered near residual value 0.00, a visual inspection does not reveal any directional relationship between the residuals and the predicted values from the regression model. This indicates that the residuals behave in an independent manner and that the regression's errors have a constant variance and are uncorrelated with each other.

The square markers – months 7, 25, and 95 indicate potential outliers. This data can be used to look for shocks caused by factors not considered in the model that occurred at these points in time.

Instructor's Note: If you observe a strong relationship then the regression assumptions of 'homoskedasticity' and 'independence of errors' has been violated.

Next, we look at the scatterplot of the regression residuals versus each of the three independent variables.



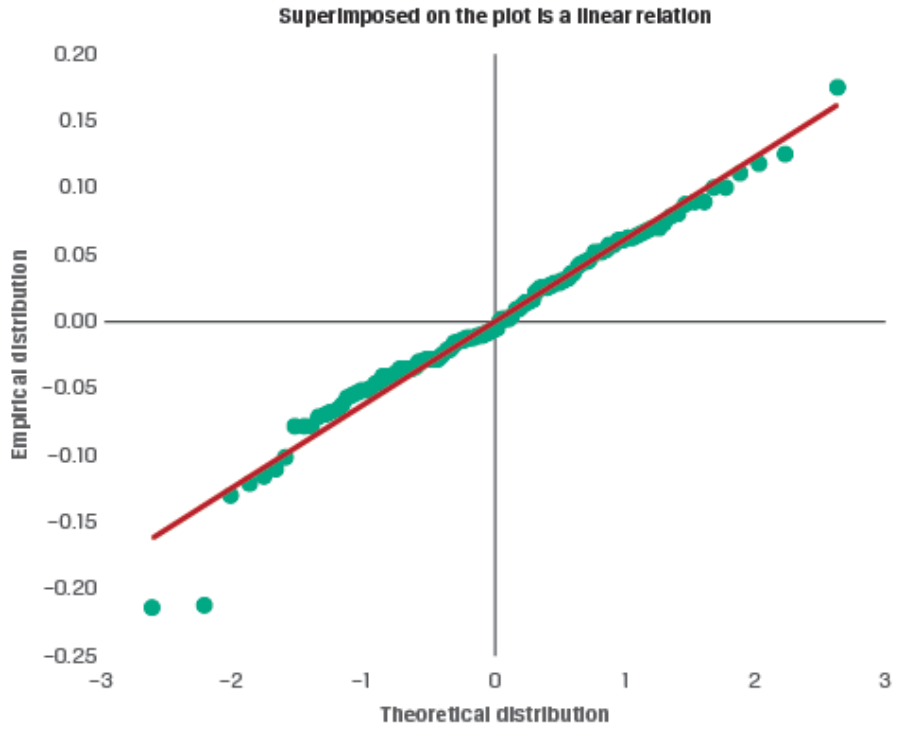
A visual inspection reveals no directional relationship between the residuals and the explanatory variables, implying no violation of the multiple linear regression assumptions.

Also, the same three potential outliers identified in the 'residual v/s dependent variable' plot are also apparent in these scatter plots as indicated by the square markers. So, we can conclude that these are indeed outliers.

Instructor's Note: If you observe a significant relationship between the residuals and an independent variable then the model is misspecified.

Normal Q-Q plot: A normal Q-Q plot is used to visualize the distribution of a variable by comparing it to a normal distribution. If the variable is normally distributed, it should align along the diagonal. We can use this plot to check if the model's residuals are normally distributed.

A Q-Q plot for the residuals of our regression model is presented below:



Superimposed on the plot is a linear relation

Apart from the three outliers, all other observations are very close to the diagonal line. Hence, we can conclude that the regression model error term is close to being normally distributed.

Instructor's Note: If you observe that observations move away from the diagonal then the regression assumption of 'normality' is violated. Deviations from the diagonal past the ± 2 standard deviations mark indicate that the distribution is 'fat-tailed' – a commonly observed feature of financial data.

Summary

LO: Describe the types of investment problems addressed by multiple linear regression and the regression process.

Multiple regression can be used:

- To explain the relationships between financial variables.
- To test existing theories.
- To make forecasts.

LO: Formulate a multiple linear regression model, describe the relation between the dependent variable and several independent variables, and interpret estimated regression coefficients.

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.

LO: Explain the assumptions underlying a multiple linear regression model and interpret residual plots indicating potential violations of these assumptions.

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.

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.

