Level II
of the
CFA® Program

Quantitative Methods

BASICS OF MULTIPLE REGRESSION AND UNDERLYING ASSUMPTIONS

Learning Outcome Statements

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

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

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

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

Multiple Linear Regression's Uses

Multiple linear regression describes the **variation of the dependent variable** by using two or more independent variables. When used properly, it can **improve predictions**, but if used incorrectly, it can create spurious relationships that can undermine predictions.

Typically, a multiple regression model takes the following form:

$$Y_i = b_0 + b_1 X_{1,i} + b_2 X_{2,i} + \cdots + b_k X_{k,i} + \epsilon_i$$

Where:

$Y\_i =$ Dependent variable

$b\_0 =$ Intercept term

$b_1, b_2, \ldots, b_k =$ Slope coefficients

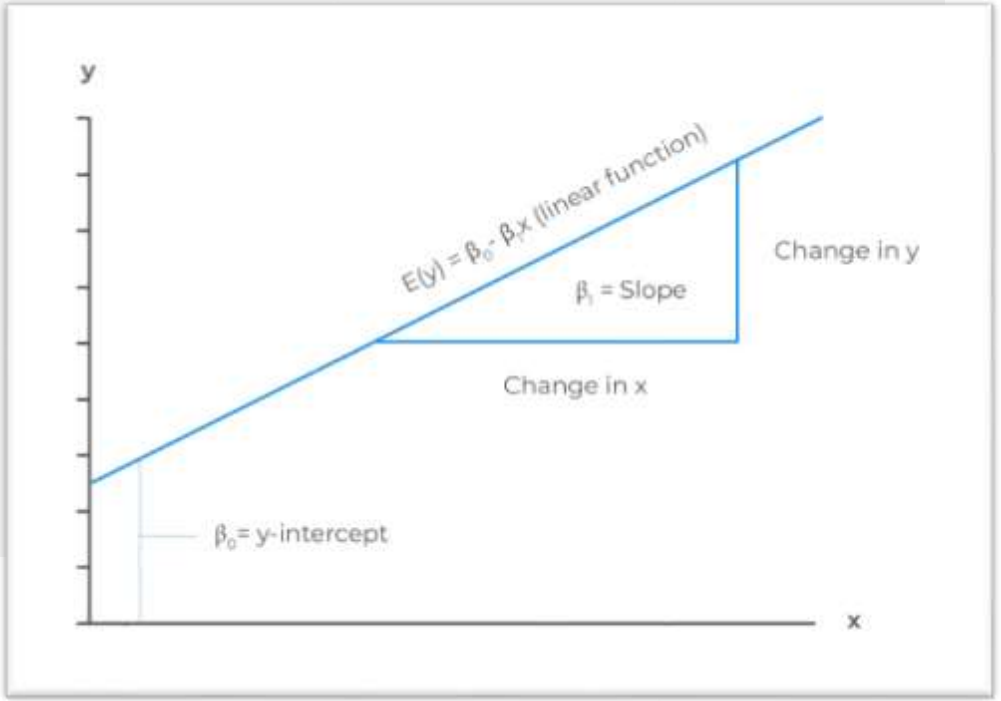
$X_{1,i}, X_{2,i}, \ldots, X_{k,i} =$ Independent variables

$\epsilon_i =$ Error term

n = Number of observations

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

- A regression equation has ***k* slope coefficients** and ***k+1* regression coefficients**.
- The **intercept term** is defined as the value of the dependent variable when the independent variables are zero.
- The slope coefficient is defined as the **estimated change in the dependent variable given a one-unit change in the independent variable,** keeping the other independent variables constant.



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

Multiple Regression Uses

- Multiple regression can be used to **test existing theories**, **identify relationships** between variables, or **forecast**.
- A single factor cannot adequately explain or forecast the complex world of investments. Due to their complexity, **statistical tests and fundamental justification** are necessary to explain financial and economic relations fully.
- There are several ways to use multiple regression, including:

A company's **profitability**, **growth**, **revenue**, and market share are variables that an investor wants to know can **predict if it will face financial difficulties**.

A company's **stock price** and **trading volume** can be correlated by an analyst seeking to determine how they change daily.

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

Example: Multiple Regression in Investment World

- James Chase, an investment analyst, wants to determine the **impact of inflation rates** and **real rates** of interest on the **price of the US Dollar index** (USDX).
- Chase uses the multiple regression model below:

$$P = b_0 + b_1 INF + b_2 IR + \epsilon_t$$

Where:

P = Price of USDX

INF = Inflation rate

IR = Real rate of interest

ϵ_t=Error term

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

The **regression of the price of USDX** on inflation and real interest rates generates the following results:

	Coefficients	Standard Error	t Stat	P-value
Intercept	81	7.9659	10.1296	0.0000
Inflation rates	**-276**	233.0748	-1.1833	0.2753
Real interest Rates	**902**	279.6949	3.2266	**0.0145**

The multiple regression equation can be expressed as:

$$P = 81 - 276INF + 902IR$$

The regression coefficient estimate of the inflation rate is **negative**. This indicates that an **increase in the inflation rates** causes a **decrease in the price of the US Dollar index**.

Furthermore, the **positive real rate of interest coefficient** implies that an increase in the real interest rate is accompanied by an **increase in the price of USDX**.

The t-statistic indicates that **only the real interest rate variable is significant at the 5% significance level**.

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

The **regression of the price of USDX** on inflation and real interest rates generates the following results:

	Coefficients	Standard Error	t Stat	P-value
Intercept	81	7.9659	10.1296	0.0000
Inflation rates	-276	233.0748	-1.1833	0.2753
Real interest Rates	902	279.6949	3.2266	0.0145

The multiple regression equation can be expressed as:

$$P = 81 - 276INF + 902IR$$

The **intercept term of 81** implies that the price of **USDX is $81** when both the inflation rate and real interest rate are 0.

A 1% increase in the inflation rate leads to a **$276 decrease in the price of USDX**, keeping real interest rates constant.

On the other hand, a 1% increase in the real interest rate leads to a **$902 increase in the price of USDX**, keeping the inflation rate constant.

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

Question

Adil Suleman, CFA, wishes to establish the **possible drivers of a company's percentage return on capital** (ROC). Suleman identifies performance measures such as the **profit margin** (%), **sales**, and **debt ratio** as possible drivers of ROC.

SUMMARY OUTPUT				
	Coefficients	**Standard Error**	**t Stat**	**P-value**
Intercept	8.6531	0.9174	9.4323	0.0000
Sales	0.0009	0.0005	1.7644	0.0922
Debt ratio	0.0229	0.0165	1.3880	0.1797
Profit Margin (%)	0.2996	0.0564	5.3146	0.0000

Which independent variables is (are) ***most likely*** **statistically significantly different from zero** at the 5% significance level, assuming the sample size is 25?

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

Question

Adil Suleman, CFA, wishes to establish the **possible drivers of a company's percentage return on capital** (ROC). Suleman identifies performance measures such as the **profit margin** (%), **sales**, and **debt ratio** as possible drivers of ROC.

SUMMARY OUTPUT				
	Coefficients	**Standard Error**	**t Stat**	**P-value**
Intercept	8.6531	0.9174	9.4323	0.0000
Sales	0.0009	0.0005	1.7644	0.0922 > 0.05
Debt ratio	0.0229	0.0165	1.3880	0.1797 > 0.05
Profit Margin (%)	0.2996	0.0564	5.3146	**0.0000 < 0.05**

Only the **profit margin** is statistically significantly different from zero at the 5% significance level.

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

The following assumptions are used to build multiple regression models:

- The relationship between the dependent variable, and the independent variables, is **linear**.
- The independent variables are **not random**.
- There is **no definite linear relationship between two or more independent variables**. A high correlation between two or more independent variables is known as **multicollinearity**.
- The **expected value of the error term**, conditional on the independent variables, is **equal to 0**.
- The **variance of the error term** is **equal** for all observations. This is known as the **homoskedasticity** assumption.
- The **error term is uncorrelated** across all observations.
- The **error term is normally distributed**.

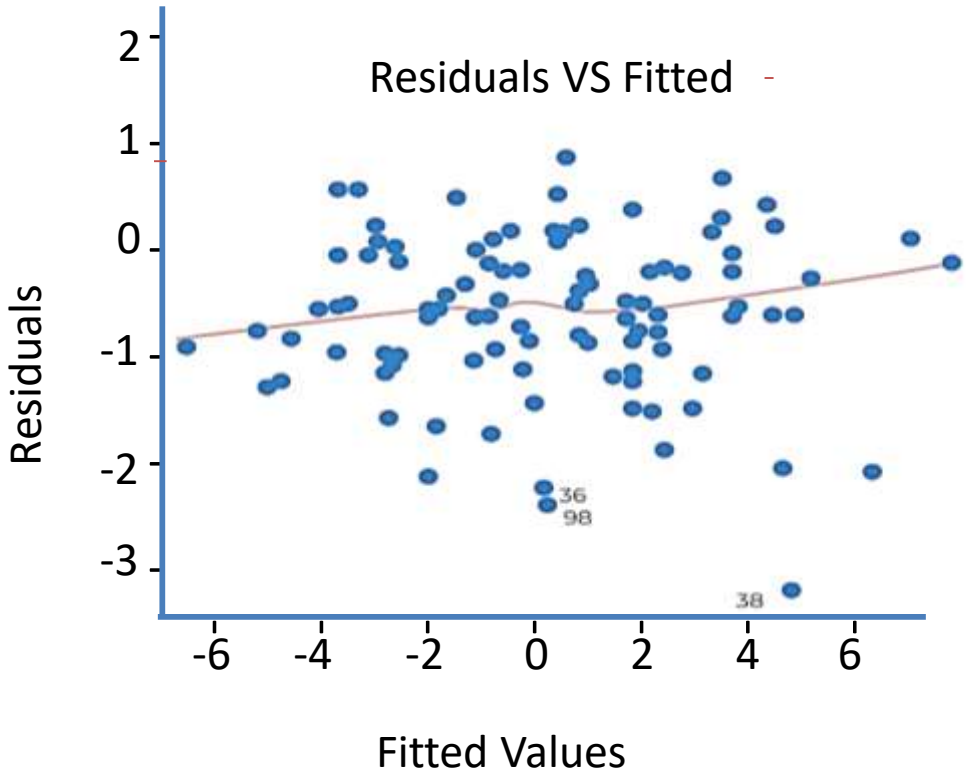
LOS : Explain the assumptions underlying a multiple linear regression model and interpret residual plots indicating potential violations of these assumptions.

The following assumptions are investigated, and their outcomes are analyzed (if violated) as follows:

I. Linearity: The regression algorithm would fail mathematically to capture the trend when fitted to a nonlinear, non-additive data set.

A prediction based on unobserved data will also be incorrect. The model can capture the nonlinear effect by including polynomial terms.

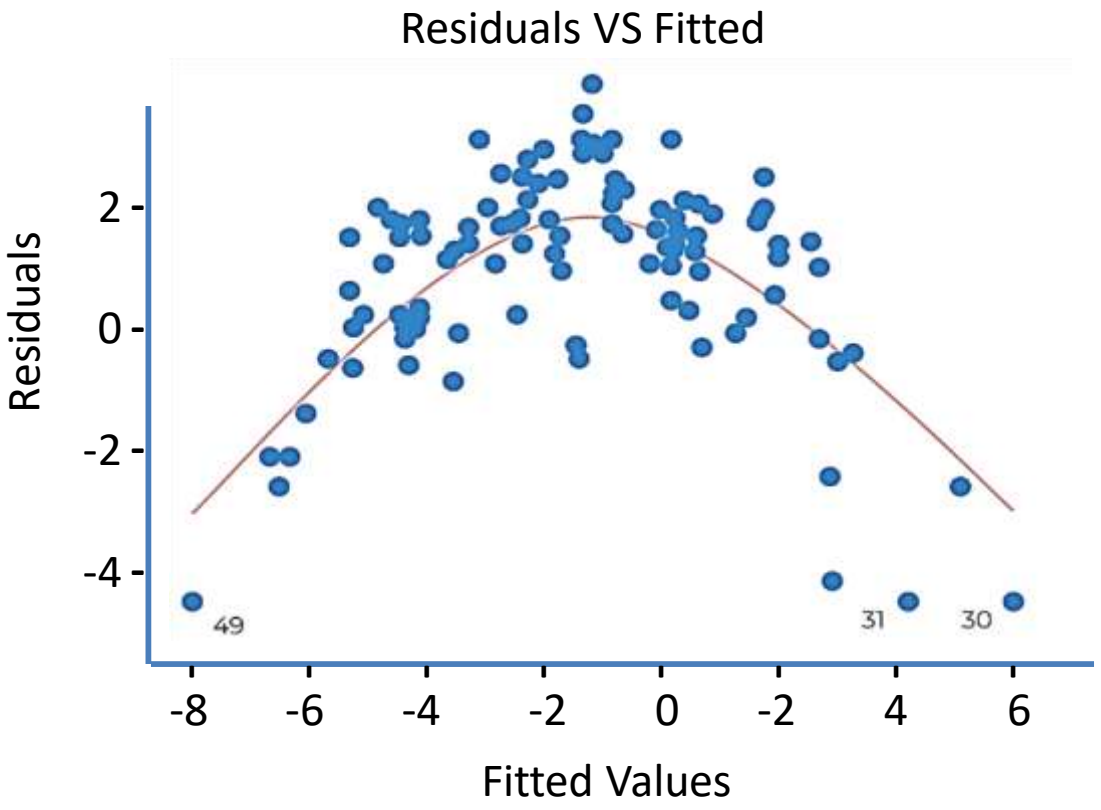
No Pattern Evident



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

This plot may show **non-linearity if any pattern** (such as a parabolic shape) appears. In other words, the **model fails to capture nonlinear effects**.

No Linearity Evident



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

II. **Autocorrelation:** A model's accuracy is drastically reduced when correlation exists in error terms. A time series model is typically characterized by this interaction, in which the next instant depends on the previous one.

Correlated error terms tend to underestimate the true standard errors, so the estimated standard errors are likely higher than they should be

To check for autocorrelation, calculate **Durbin-Watson (DW) statistics**. The value must be **between 0 and** 4.

If DW = 2 implies no autocorrelation, $0 < DW < 2$ implies positive autocorrelation, while $2 < DW < 4$ indicates negative autocorrelation.

The residual values can also be plotted against time to identify **seasonal or correlated patterns**.

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

III. **Multicollinearity:** It becomes challenging to determine the **true relationship between a predictor and response variable in a model with correlated variables**.

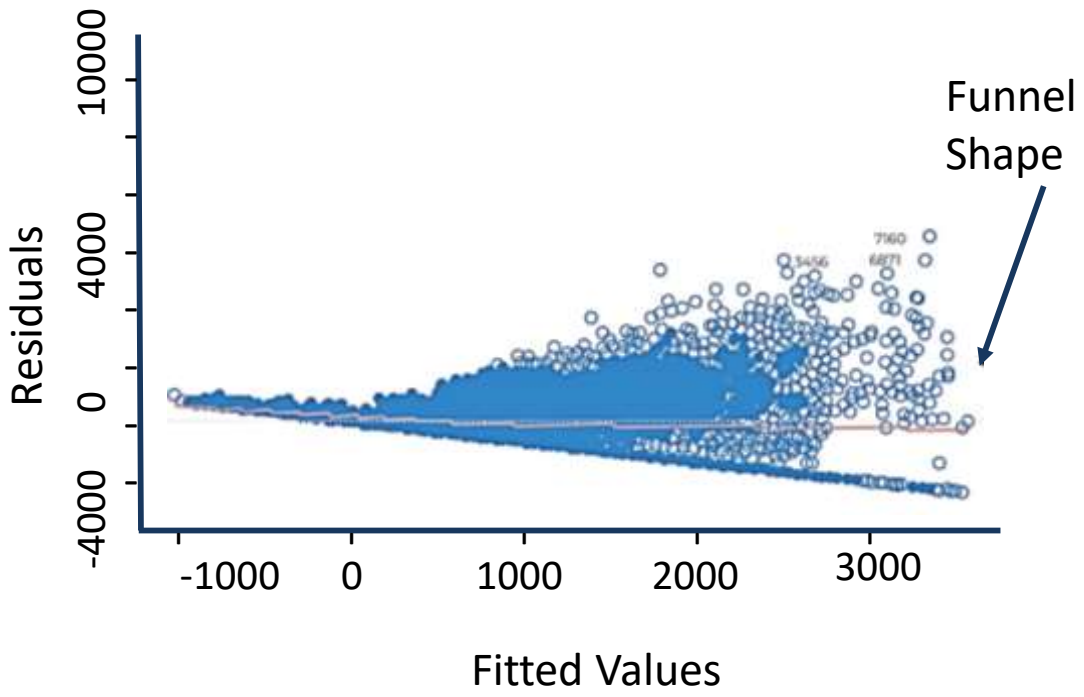
Standard errors tend to increase when correlated predictors are present. A large standard error also leads to wider confidence intervals, resulting in **less accurate slope parameters**.

Scatter plots can visualize correlations between variables to check for multicollinearity. VIF factor can also be used. VIF value <= 4 suggests no multicollinearity, whereas a value of **>= 10 implies serious multicollinearity.**

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

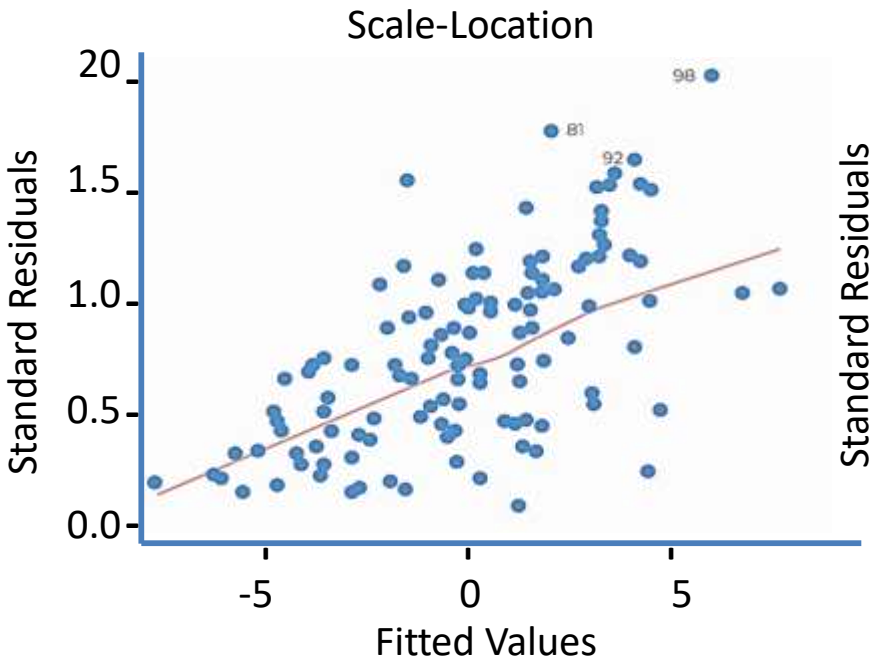
IV. **Heteroskedasticity:** In heteroskedasticity, the error terms have **non-constant variances**. An outlier or extreme leverage value will typically lead to non-constant variance. These values disproportionately affect the model's performance because they are **given too much weight**.

Whenever this phenomenon occurs, confidence intervals for out-of-sample predictions become unrealistically large or small. If a residuals versus fitted values plot exhibits heteroskedasticity, the plot will show **a funnel shape**. Alternatively, you can conduct a **Breusch-Pagan / Cook-Weisberg or White general test**.

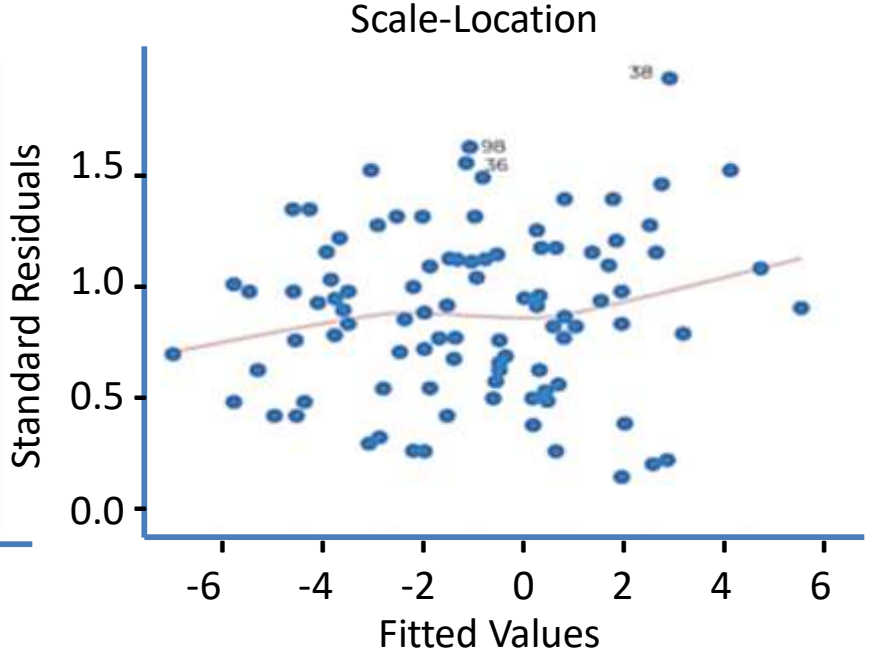


LOS : Explain the assumptions underlying a multiple linear regression model and interpret residual plots indicating potential violations of these assumptions.

Heteroskedasticity is evident



Homoskedasticity is evident



As well as detecting homoskedasticity, the above plot is used to determine variance equality. As you can see, the residuals are spread out along the range of predictors. It uses **standardized residual values instead of residuals versus fitted values**.

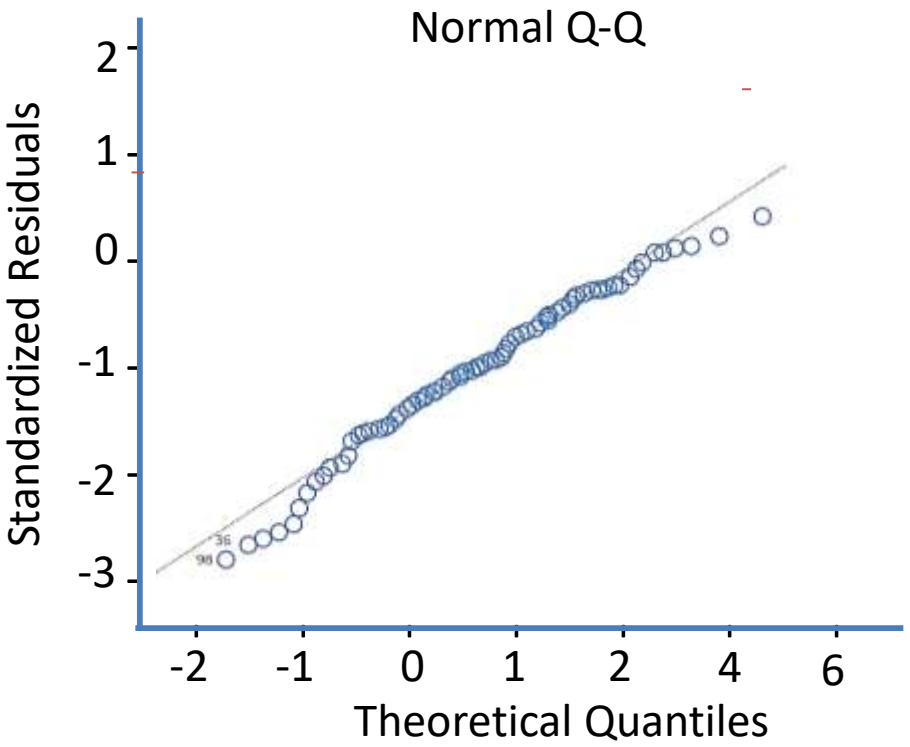
LOS : Explain the assumptions underlying a multiple linear regression model and interpret residual plots indicating potential violations of these assumptions.

Normal Distribution Evident

V. Normal Distribution of error terms: Confidence intervals may be too wide or narrow if the error terms are not normally distributed.

Using **least squares** to minimize coefficients becomes difficult once confidence intervals become unstable.

If non-normal distributions are present, there will probably be some **unusual data points that will need to be closely examined**.

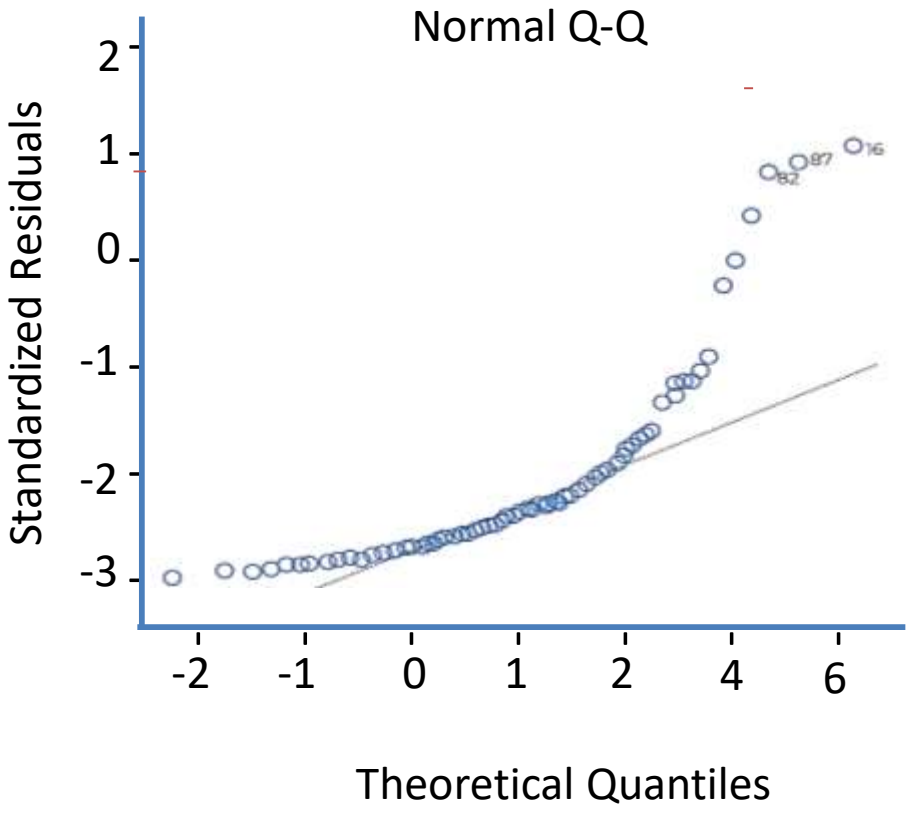


The best way to determine the normal distribution of error terms is to **plot a QQ plot**. **Kolmogorov-Smirnov** and **Shapiro-Wilk tests** can also be used to test for normality.

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

The q-q or quantile-quantile plot is used to verify the assumption that a data set follows a normal distribution. We can determine **whether the data follows a normal distribution with this plot**. There would be a fairly straight line on the plot if that were the case. Observe that the **straight line deviates when there is no normality in the errors**.

Non Normal Distributed



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

- The **main assumptions** that underly the multiple regression models are:

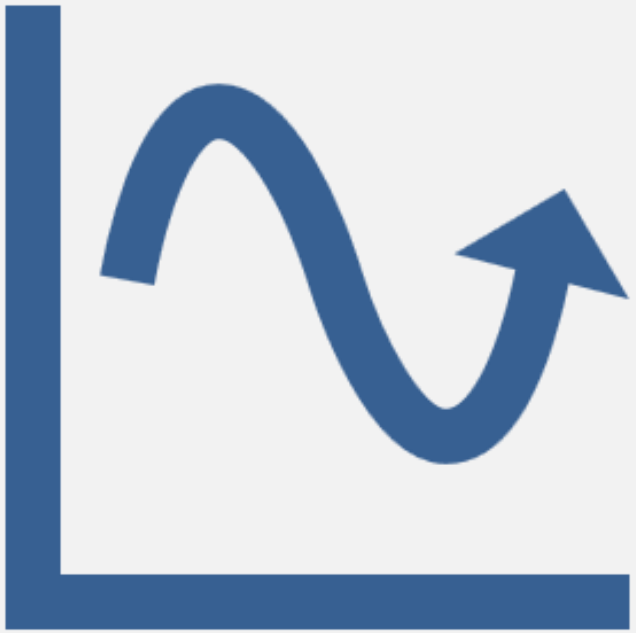
  i. Homoskedasticity.

  ii. Independence of the independent variables.

  iii. Independence of errors.

  iv. Normality.

  v. Linearity.



Learning Outcome Statements

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

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

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