Level II of the CFA® 2025 Exam

Study Notes - Quantitative Methods

Offered by AnalystPrep

Last Updated: Feb 28, 2025

Table of Contents

Reading 1: Basics of Multiple Regression and Underlying Assumptions

Los 1 (a) Describe the types of investment problems addressed by multiple linear regression and the regression process

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

Typically, a multiple regression model takes the following form:

$$Y_i = b_0 + b_1X_{1,i} + b_2X_{2,i} + \ldots + b_kX_{k,i} + \epsilon_i$$

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.

ϵ_i = Error term.

n = Number of observations.

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. On the other hand, 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.

Researchers can use multiple regression to test existing theories, identify relationships between variables, or forecast.

The researcher must specify the model to determine the good-fit criteria for the regression

model, including an independent variable. Once the regression model has been specified, it must be estimated and analyzed to ensure it satisfies all the key assumptions.

It is equally noteworthy that a researcher can use multiple regression to test existing forecasting theories. Alternatively, multiple regression can further be used to identify relationships between variables after the model is tested and deemed acceptable for out-of-sample performance.

A single factor cannot adequately explain or forecast the complex world of investments. Due to their complexity, statistical tests and fundamental justification are critical in the exhaustive explanation of financial and economic relations.

There are several ways to use multiple regression, including:

- A company's profitability, growth, revenue, and market share are variables that an investor is interested in. These variables can predict if a company will run into financial difficulties.
- An analyst seeking to determine how a company's stock price and trading volume change daily can correlate them. An analyst can use linear regression to determine the relationship between variables.

Question

Which of the following *most* accurately detects whether the underlying assumptions of multiple linear regression models are satisfied?

A. Scatter plots.
B. Residual plots.
C. Diagnostic plots.

Solution

The correct answer is **C**.

Diagnostic plots for multiple regression show the prediction errors against predicted values. Therefore, they are useful in the determination of the gaps that a researcher should address in their data to improve the accuracy of their predictions. Using diagnostic plots, a researcher can determine whether the assumptions of multiple linear regression are valid.

A is incorrect. The scatterplot is useful for detecting nonlinear relationships between dependent and independent variables.

B is incorrect. A residual plot is an effective tool for detecting violations of homoskedasticity and error independence.

LOS 1 (b) 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.

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

Chase can express the multiple regression equation as follows:

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

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 (USDX).

Furthermore, the positive real rate of interest coefficient implies that an increase in the price of USDX accompanies the real interest rate.

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

The **intercept term** is defined as the value of the dependent variable when the independent variables are zero. On the other hand, each **slope coefficient** is the estimated change in the value of the dependent variable for a one-unit change in the value of the respective independent variable, keeping the other independent variables constant. Slope coefficients are also called **partial slope coefficients**.

Continuing with this example:

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

Where:

P = Price of the US Dollar index.

INF = Annual inflation rate.

IR = Annual real rate of interest.

The regression equation is interpreted as follows:

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.

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.

He obtains the following results from the regression of ROC on profit margin (%), sales, and debt ratio.

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 variable(s) is (are) most likely statistically and significantly different from zero at the 5% significance level, assuming the sample size is 25?

A. Profit margin.
B. Sales and profit margin.
C. Sales and debt ratio.

Solution

The correct answer is **A**.

An independent variable is statistically significant if its p-value is less than the significance level, in this case, 5% or 0.05. Therefore, only the profit margin is statistically and significantly different from zero at the 5% significance level.

B and C are incorrect. At a 5% significance level, only the profit margin is statistically significantly different from zero.

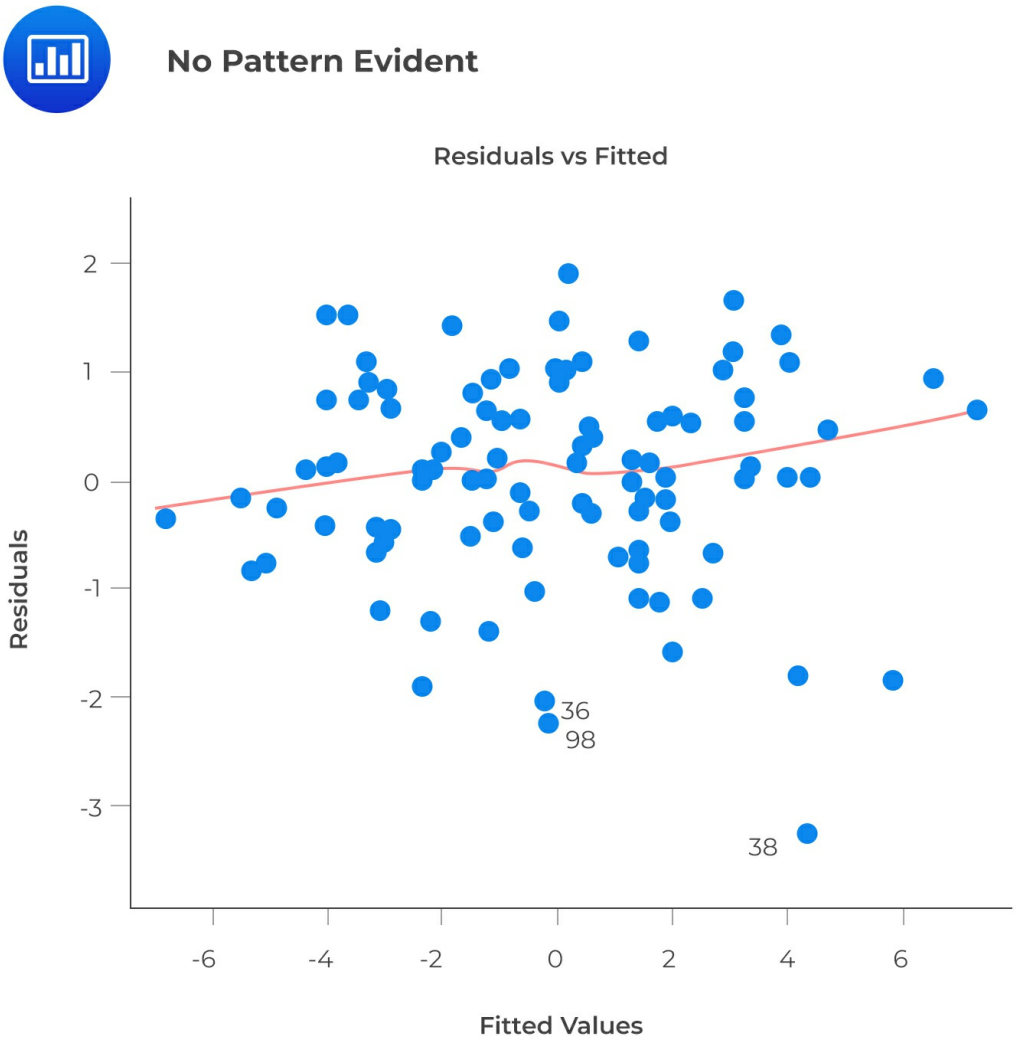
Los 1 (c) Explain the assumptions underlying a multiple linear regression model and interpret residual plots indicating potential violations of these assumptions

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.

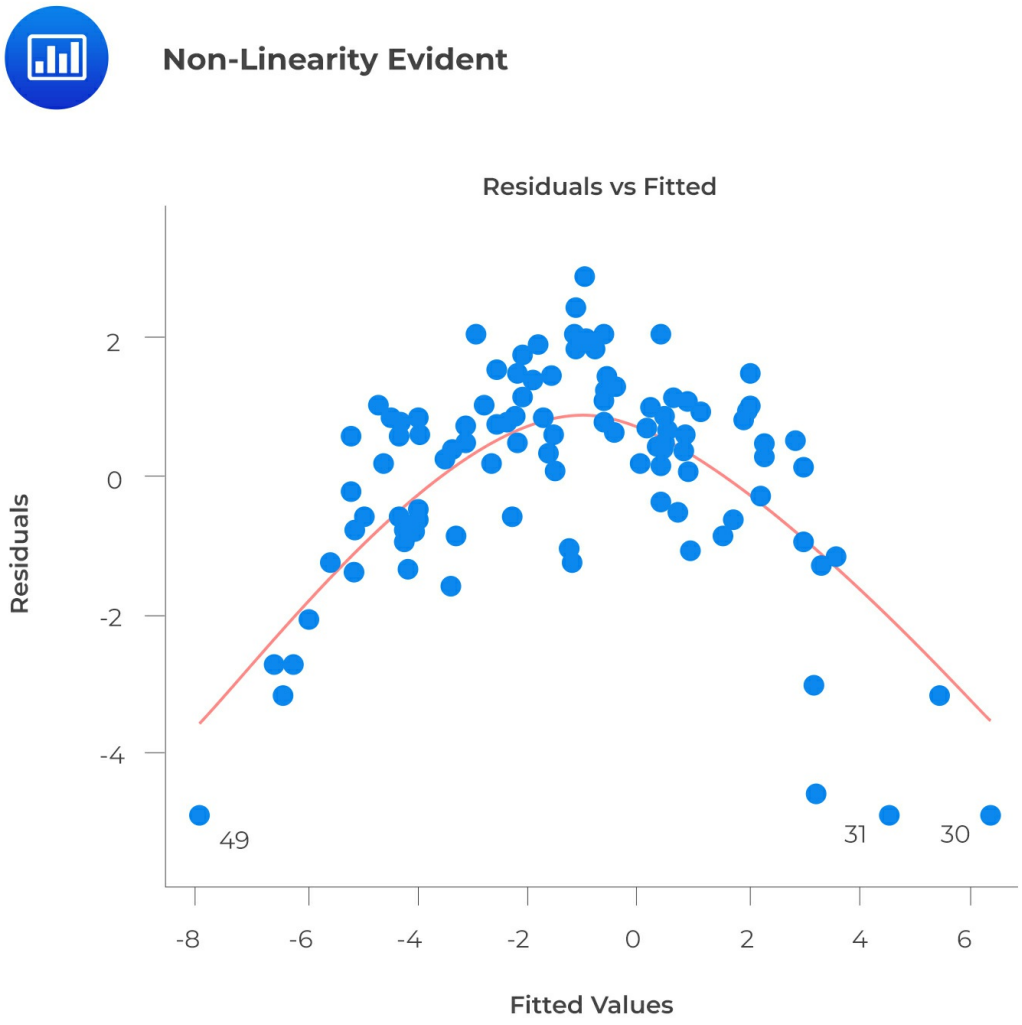
The following assumptions are investigated, and their outcomes are analyzed (if violated) as follows:

i. ***Linearity***: The regression algorithm would mathematically fail 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.

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.



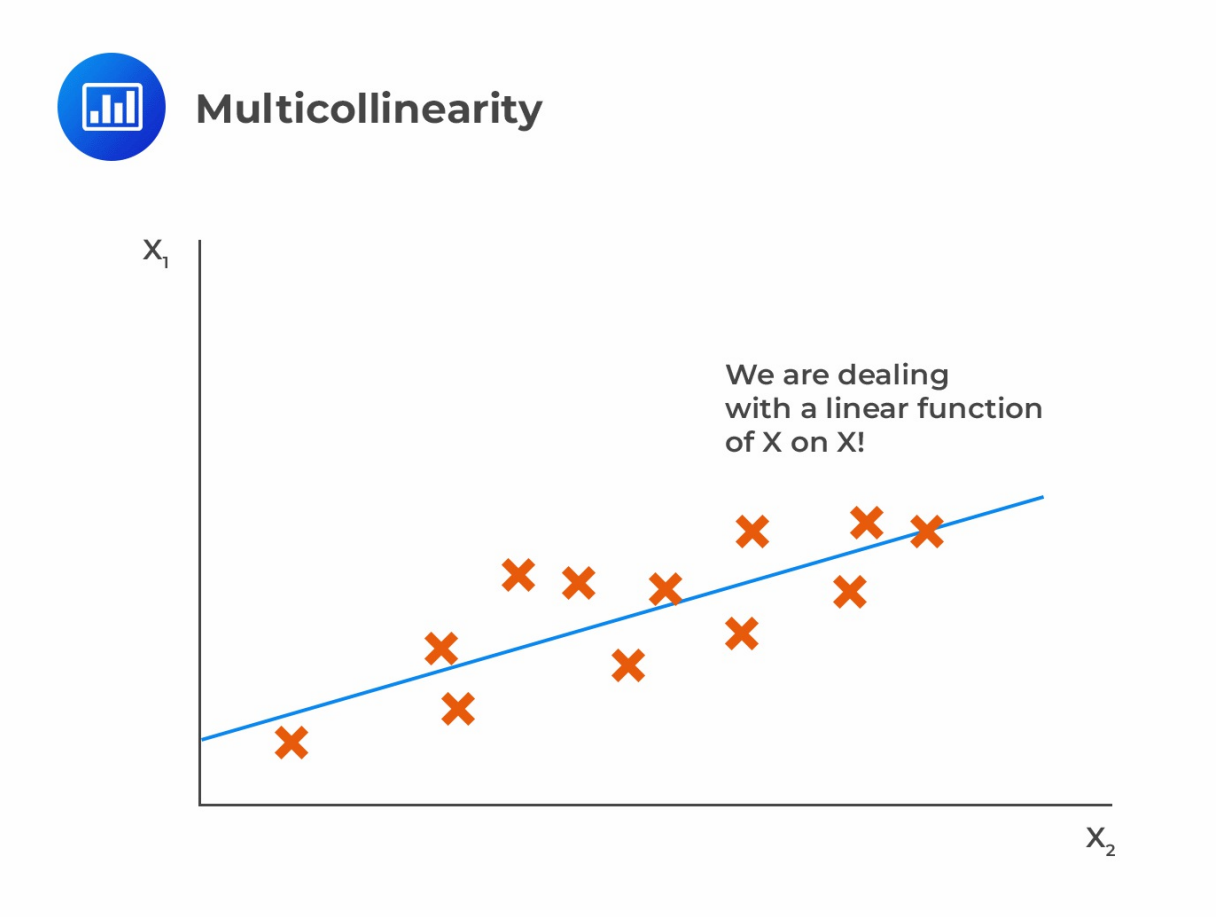
ii. ***Autocorrelation***: A model's accuracy is drastically reduced when correlation exists in error terms. A time series model is characterized by this interaction, in which the next instant depends on the previous one.

Correlated error terms tend to underestimate the true standard errors. Consequently, the estimated standard errors are likely higher than they should be. Note that confidence intervals and prediction intervals are narrowed if this occurs. The probability of the actual coefficient value being contained in a 95% confidence interval is lower than 0.95 if the confidence interval is narrower.

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.

iii. ***Multicollinearity***: A moderately or highly correlated set of independent variables will exhibit this phenomenon. In a model with correlated variables, it is challenging to determine the true relationship between a predictor and response variable.

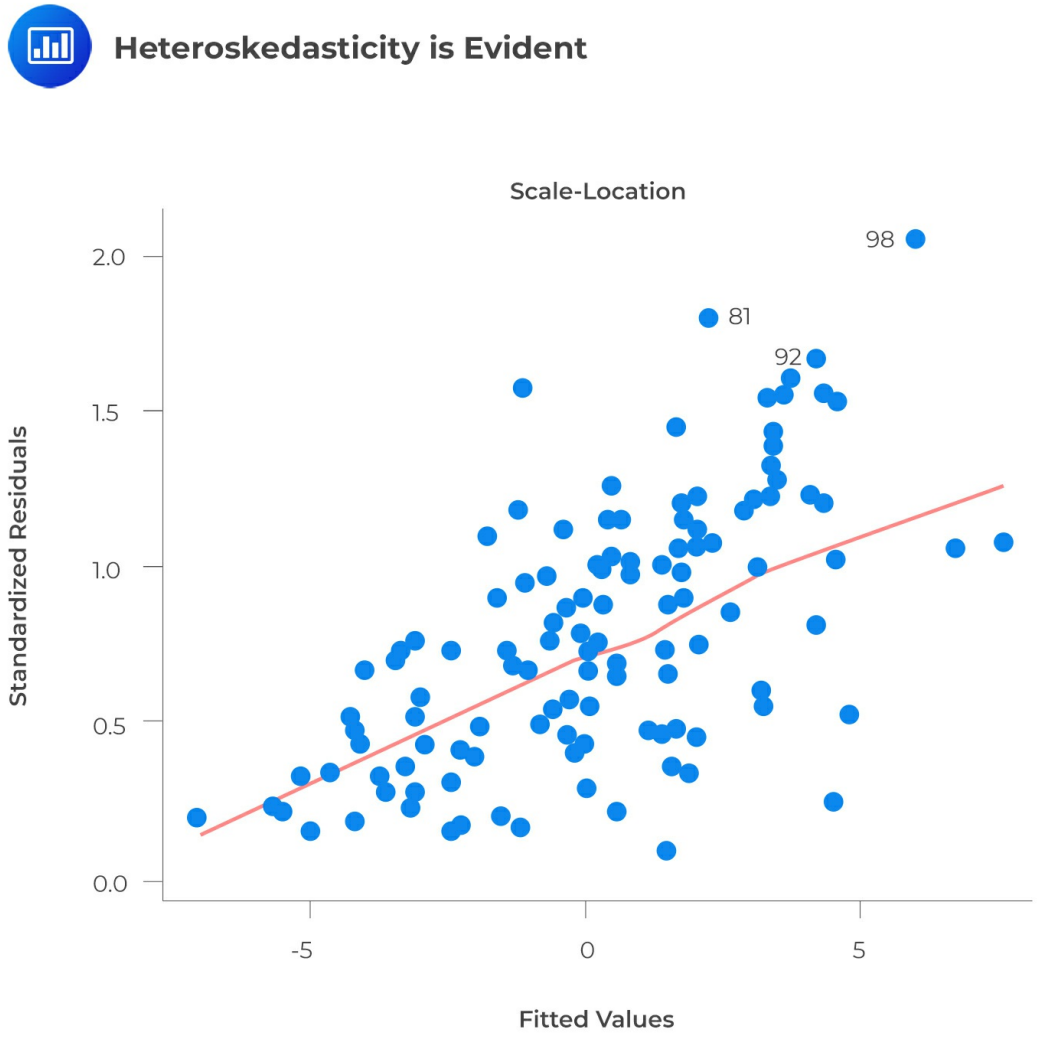


The difficulty lies in determining which variable contributes to the response's prediction. Furthermore, standard errors tend to increase when correlated predictors are present. A large standard error also leads to wider confidence intervals, resulting in less accurate slope parameters.

Correlated predictors have a different regression coefficient depending on which other predictors are included in a model. A variable that strongly / weakly affects a target variable will result in an incorrect conclusion.

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.

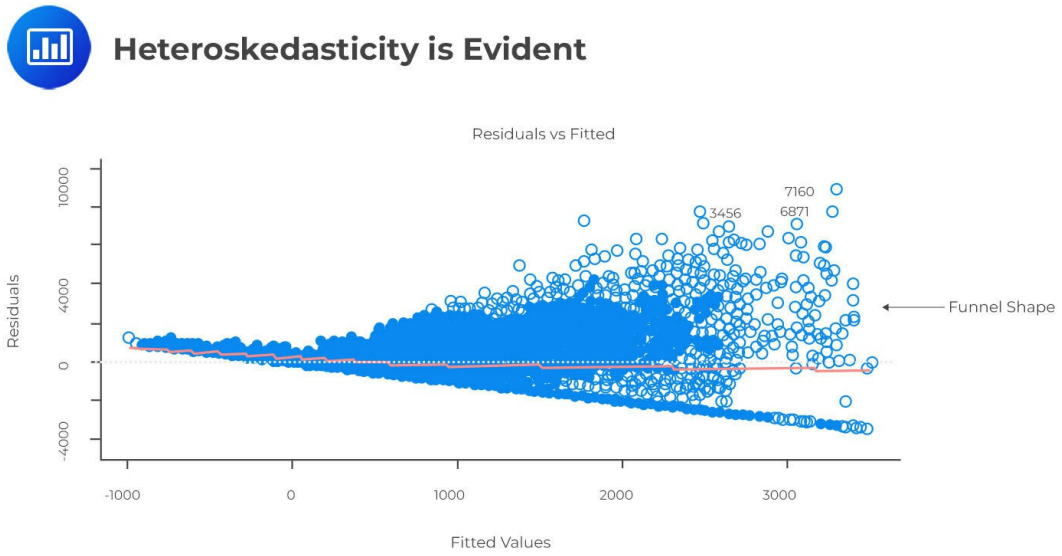
iv. ***Heteroskedasticity***: In heteroskedasticity, the error terms have non-constant variances.



Homoskedasticity is Evident

An outlier or extreme leverage value will typically lead to non-constant variance. These values disproportionately affect a model's performance because they are given too much weight.

Whenever heteroskedasticity 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.

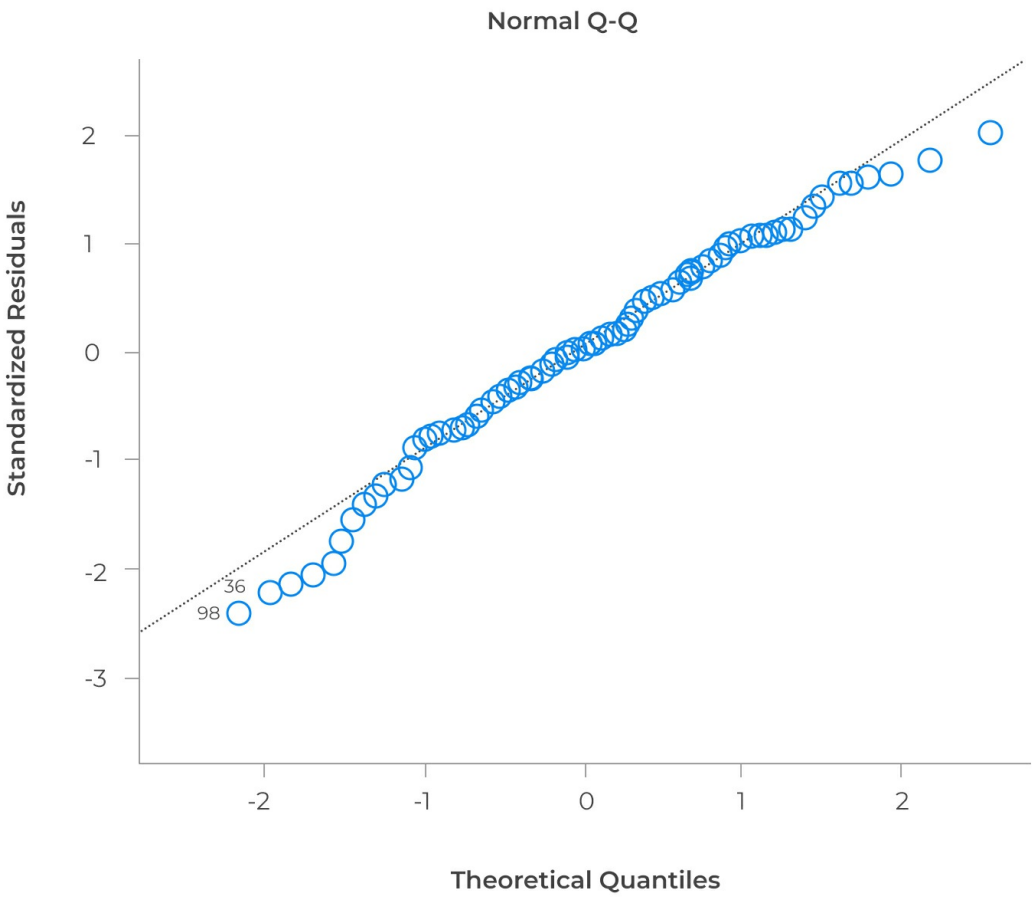


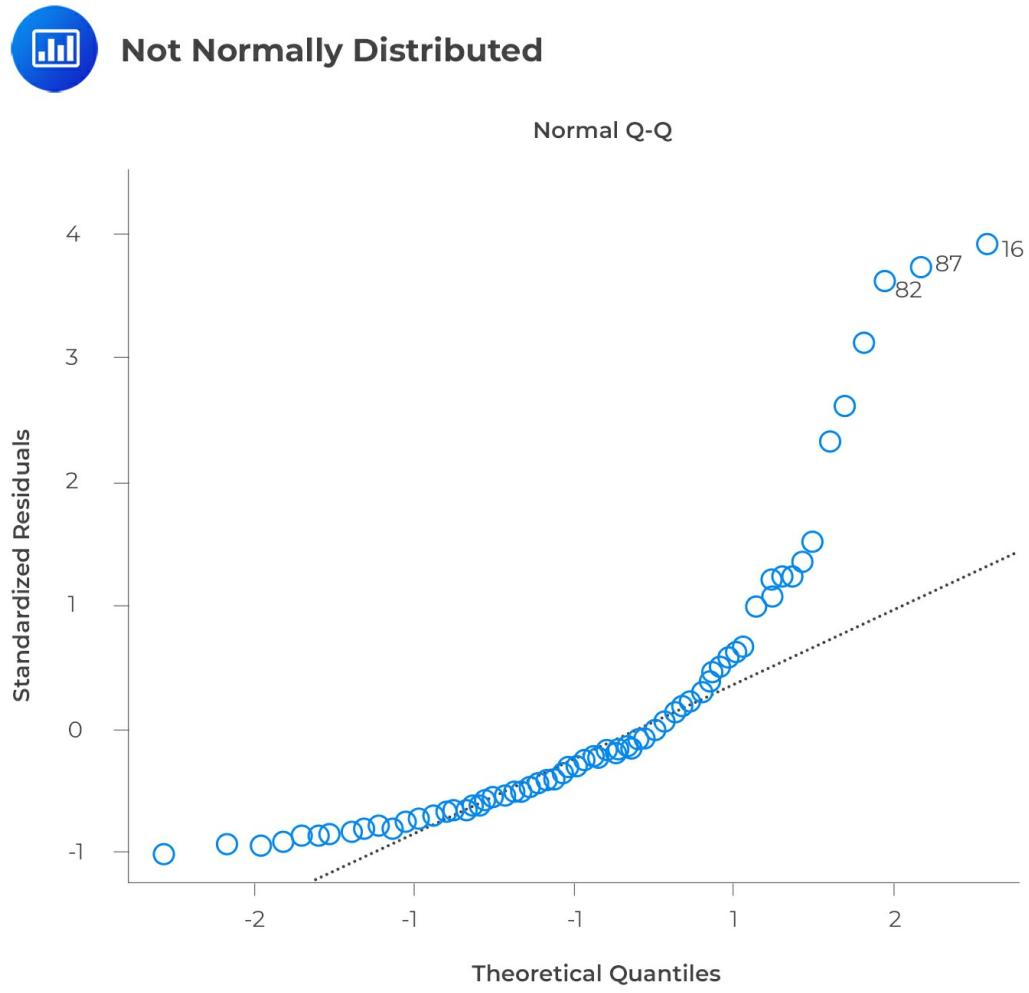
Aside from detecting homoskedasticity, the above plot is used to determine variance equality. As you can see, the residuals are spread out along the range of predictors. It uses standardized residual values instead of residuals versus fitted values.

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 to make a better model.

   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.

Normal Distribution Evident





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. Note that the straight line deviates when there is no normality in the errors.

Question

Which of the following is least likely an assumption of the multiple linear regression model?

A. The independent variables are not random.
B. The error term is correlated across all observations.
C. The expected value of the error term, conditional on the independent variables, is equal to zero.

Solution

The correct answer is **B**.

The error term is **uncorrelated** across all observations.

$$E(\epsilon_i \epsilon_j) = 0 \quad \forall\ i \neq j$$

Other assumptions of the classical normal multiple linear regression model include the following:

i. The independent variables are not random. Additionally, there is no exact linear relationship between two or more independent variables.
ii. The error term is normally distributed.
iii. The expected value of the error term, conditional on the independent variables, is equal to 0.
iv. The variance of the error term is the same for all observations.
v. A linear relation exists between the dependent variable and the independent variables.

A and C are incorrect. They both indicate the correct assumptions of the multiple linear regression model.

Reading 2: Evaluating Regression Model Fit and Interpreting Model Results

Los 2 (a) Evaluate how well a multiple regression model explains the dependent variable by analyzing ANOVA table results and measures of goodness of fit

R-squared (R^2) measures how well an estimated regression fits the data. It is also known as the **coefficient of determination** and can be formulated as:

$$R^2 = \frac{\text{Sum of regression squares}}{\text{Sum of squares total}} = \frac{\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}$$

Where:

n = Number of observations.

Y_i = Dependent variable observations.

$\hat{Y}_i$ = Dependent variables predicted value to the independent variable.

$\bar{Y}$ = Dependent variable mean.

In the presence of independent variables, R^2 will either increase or remain constant. However, R^2 cannot be used to measure the goodness of fit of a model as it will not decrease with the addition of independent variables.

Limitations of R^2

- It is impossible to determine the statistical significance of the coefficients from R^2.
- A bias in the predicted coefficients or estimates cannot be determined with R^2.
- When a model is good, it has a high R^2; when it is bad, it has a low R^2, usually due to overfitting and biases in the model.

An **overfitted regression model** is one with too many independent variables to the number of

observations in a sample. Overfitting may produce coefficients that do not reflect the true relationship between the independent and dependent variables.

Multiple regression software packages usually produce an **adjusted** R^2 ($\bar{R}^2$) as an alternative measure of goodness of fit. Using adjusted R^2 in regression is beneficial since it does not automatically increase when more independent variables are included, given that it adjusts for degrees of freedom.

$$\bar{R}^2 = 1 - \left[\frac{\frac{\text{Sum of squares error}}{n-k-1}}{\frac{\text{Sum of squares total}}{n-1}} \right]$$

Therefore, the relationship between $\bar{R}^2$ and R^2 can mathematically be derived as follows:

$$\bar{R}^2 = 1 - [(\frac{n-1}{n-k-1}) \ \ (1 - R^2)]$$

Note that:

- If $k \geq 1$ then $R^2 >$ adjusted R^2 the result is that adjusted R^2 can be negative while R^2 is zero at minimum.

When including a new variable in the regression, the following should be taken into consideration:

- $\bar{R}^2$ increases when the coefficient t-statistic is $> |1.0|$.
- $\bar{R}^2$ decreases when the coefficient t-statistic is $< |1.0|$.
- At typical significance levels, 5% and 1%, a t-statistic with an absolute value of 1.0 does not indicate that the independent variable is different from zero. Therefore, the adjusted R^2 doesn't demonstrate that it will increase significantly.

ANOVA Table

One of the outputs of multiple regression is the ANOVA table. The following shows the general

structure of an Anova table.

ANOVA	Df (degrees of freedom)	SS (Sum of squares)	MSS (Mean sum of squares)
Regression	k	RSS (Explained variation)	MSR
Residual	$n-(k+1)$	SSE (Unexplained variation)	MSE
Total	$n-1$	SST (Total variation)	

We can use the information in an ANOVA table to determine R^2, the F-statistic, and the standard error estimates (SEE) as expressed below:

$$R^2 = \frac{SSR}{SST}$$

$$F = \frac{MSR}{MSE}$$

$$SSE = \sqrt{MSE}$$

Where:

$$MSR = \frac{RSS}{k}$$
$$MSE = \frac{SSE}{n-k-1}$$

Example: Interpreting Regression Output

Consider the following regression results generated from multiple regression analysis of the price of the US Dollar index on the inflation rate and real interest rate.

ANOVA

	df	SS	Significance F
Regression	2	432.2520	0.0179
Residual	7	200.6349	
Total	9	632.8869	

	Coefficients	Standard Error
Intercept	81	7.9659
Inflation rates	−276	233.0748
Real interest Rates	902	279.6949

Given the above information, the regression equation can be expressed as:

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

Where:

P = Price of USDX.

INF = Inflation rate.

IR = Real interest rate.

R^2 and adjusted R^2 can also be calculated as follows:

$$R^2 = \frac{SSR}{SST} = \frac{432.2520}{632.8869} = 0.6830 = 68.30\%$$

$$\text{Adjusted } R^2 = 1 - (\frac{n-1}{n-k-1})(1 - R^2) = 1 - \frac{10-1}{10-2-1}(1 - 0.6830)$$
$$= 0.5924 = 59.24\%$$

It's important to note the following:

- Multiple regression does not provide a straightforward explanation of adjusted R^2 in terms of the variance explained by the dependent variable, as is the case in simple regression.
- Adjusted R^2 does not indicate whether a regression coefficient's predictions are true or biased. Residual plots and other statistics are required to determine whether or not the predictions are accurate.

- To assess the significance of the model's fit, we use the F-Statistic and other goodness-of-fit metrics from the ANOVA rather than R^2 and adjusted R^2.

Question

Which of the following is *most appropriate* for adjusted R^2?

A. It is always positive.
B. It may or may not increase when one adds an independent variable.
C. It is non-decreasing in the number of independent variables.

Solution

The correct answer is **B**.

The value of the adjusted R^2 increases only when the added independent variables improve the fit of the regression model. Moreover, it decreases when the added variables do not improve the model fit sufficiently.

A is incorrect: The adjusted R^2 can be negative if R^2 is low enough. However, multiple R^2 is always positive.

C is incorrect: The adjusted R^2 **can decrease** when the added variables do not improve the model fit by a good enough amount. However, multiple R^2 is non-decreasing in the number of independent variables. For this reason, it is less reliable as a measure of goodness of fit in regression with more than one independent variable than in a one-independent variable regression.

Los 2 (b) Formulate hypotheses on the significance of two or more coefficients in a multiple regression model and interpret the results of the joint hypothesis tests

In multiple regression, the intercept in simple regression represents the expected value of the dependent variable when the independent variable is zero, while in multiple regression, it's the expected value when all independent variables are zero. The interpretation of slope coefficients remains the same as in simple regression.

Tests for single coefficients in multiple regression are similar to those in simple regression, including one-sided tests. The default test is against zero, but you can test against other hypothesized values.

In some cases, you might want to test a subset of variables jointly in multiple regression, comparing models with and without specific variables. This involves a joint hypothesis test where you restrict some coefficients to zero. To test the slope against a hypothesized value other than zero we will need to:

- You can conduct the test by either modifying the hypothesized parameter value, B_j, in the test statistic or
- by comparing B_j with the confidence interval boundaries derived from the regression coefficient output.

At times, we may want to collectively test a subset of variables within a multiple regression. To illustrate this concept and set the stage, let's say we aim to compare the regression outcomes for a portfolio's excess returns using Fama and French's three-factor model (MKTRF, SMB, HML) with those using their five-factor model (MKTRF, SMB, HML, RMW, CMA). Given that both models share three factors (MKTRF, SMB, HML), the comparison revolves around assessing the necessity of the two additional variables: the return difference between the most profitable and least profitable firms (RMW) and the return difference between firms with the most conservative and most aggressive investment strategies (CMA). The primary goal in determining the superior model lies in achieving simplicity by identifying the most effective independent variables in