WORKSHEET- VI

***Q.1: ans**. – a.*

***Q.2: ans**. – b.*

***Q.3: ans**. – b.*

***Q.4: ans**. – c.*

***Q.5: ans**. – b.*

***Q.6: ans**. – d.*

***Q.7: ans**. – c.*

***Q.8: ans**. – a.*

***Q.9: ans**. – a.*

***Q.10: ans**. R-squared and adjusted R-squared enable investors to measure the performance of a mutual fund against that of a benchmark. Investors may also use them to calculate the performance of their portfolio against a given benchmark.*

In the world of investing, R-squared is expressed as a percentage between 0 and 100, with 100 signaling perfect correlation and zero no correlation at all. The figure does not indicate how well a particular group of securities is performing. It only measures how closely the returns align with those of the measured benchmark. It is also backwards-looking—it is not a predictor of future results.

Adjusted R-squared can provide a more precise view of that correlation by also taking into account how many independent variables are added to a particular model against which the stock index is measured. This is done because such additions of independent variables usually increase the reliability of that model—meaning, for investors, the correlation with the index.

KEY TAKEAWAYS

- *R-squared and the adjusted R-squared both help investors measure the correlation between a mutual fund or portfolio with a stock index.*
- *Adjusted R-squared, a modified version of R-squared, adds precision and reliability by considering the impact of additional independent variables that tend to skew the results of R-squared measurements.*
- *The predicted R-squared, unlike the adjusted R-squared, is used to indicate how well a regression model predicts responses for new observations.*
- *One misconception about regression analysis is that a low R-squared value is always a bad thing.*

WORKSHEET- VI

Practice trading with virtual money

Find out what a hypothetical investment would be worth today.

SELECT A STOCK

TSLA

TESLA INC

AAPL

APPLE INC

NKE

NIKE INC

AMZN

AMAZON.COM, INC

WMT

WALMART INC

SELECT INVESTMENT AMOUNT

$

SELECT A PURCHASE DATE

CALCULATE

R-Squared

R-squared (R^2) is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model. R-squared explains to what extent the variance of one variable explains the variance of the second variable. So, if the R^2 of a model is 0.50, then approximately half of the observed variation can be explained by the model's inputs.

An R-squared result of 70 to 100 indicates that a given portfolio closely tracks the stock index in question, while a score between 0 and 40 indicates a very low correlation with the index.1 Higher R-squared values also indicate the reliability of beta readings. Beta measures the volatility of a security or a portfolio.

WORKSHEET- VI

While R-squared can return a figure that indicates a level of correlation with an index, it has certain limitations when it comes to measuring the impact of independent variables on the correlation. This is where adjusted R-squared is useful in measuring correlation.

R-Squared is just one of many tools traders should have in their arsenals. Investopedia's Technical Analysis Course provides a comprehensive overview of technical indicators and chart patterns with over five hours of on-demand video. It covers all of the most effective tools and how to use them in real-life markets to maximize risk-adjusted returns.

Adjusted R-Squared

Adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model. The adjusted R-squared increases when the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected. Typically, the adjusted R-squared is positive, not negative. It is always lower than the R-squared.

Adding more independent variables or predictors to a regression model tends to increase the R-squared value, which tempts makers of the model to add even more variables. This is called overfitting and can return an unwarranted high R-squared value. Adjusted R-squared is used to determine how reliable the correlation is and how much it is determined by the addition of independent variables.2

In a portfolio model that has more independent variables, adjusted R-squared will help determine how much of the correlation with the index is due to the addition of those variables. The adjusted R-squared compensates for the addition of variables and only increases if the new predictor enhances the model above what would be obtained by probability. Conversely, it will decrease when a predictor improves the model less than what is predicted by chance.

Key Differences

The most obvious difference between adjusted R-squared and R-squared is simply that adjusted R-squared considers and tests different independent variables against the stock index and R-squared does not. Because of this, many investment professionals prefer using adjusted R-squared because it has the potential to be more accurate. Furthermore, investors can gain additional information about what is affecting a stock by testing various independent variables using the adjusted R-squared model.

R-squared, on the other hand, does have its limitations. One of the most essential limits to using this model is that R-squared cannot be used to determine whether or not the coefficient estimates and predictions are biased. Furthermore, in multiple linear regression, the R-squared can not tell us which regression variable is more important than the other.

Adjusted R-Squared vs. Predicted R-Squared

WORKSHEET- VI

The predicted R-squared, unlike the adjusted R-squared, is used to indicate how well a regression model predicts responses for new observations. So where the adjusted R-squared can provide an accurate model that fits the current data, the predicted R-squared determines how likely it is that this model will be accurate for future data.

R-Squared vs. Adjusted R-Squared Examples

When you are analyzing a situation in which there is a guarantee of little to no bias, using R-squared to calculate the relationship between two variables is perfectly useful. However, when investigating the relationship between say, the performance of a single stock and the rest of the S&P500, it is important to use adjusted R-squared to determine any inconsistencies in the correlation.

If an investor is looking for an index fund that closely tracks the S&P500, they will want to test different independent variables against the stock index such as the industry, the assets under management, how long the stock has been available on the market, and so on to ensure they have the most accurate figure of the correlation.

Special Considerations

R-Squared and Goodness-of-Fit

The basic idea of regression analysis is that if the deviations between the observed values and the predicted values of the linear model are small, the model has well-fit data. Goodness-of-fit is a mathematical model that helps to explain and account for the difference between this observed data and the predicted data. In other words, goodness-of-fit is a statistical hypothesis test to see how well sample data fit a distribution from a population with a normal distribution.

Low R-Squared vs. High R-Squared Value

One misconception about regression analysis is that a low R-squared value is always a bad thing. This is not so. For example, some data sets or fields of study have an inherently greater amount of unexplained variation. In this case, R-squared values are naturally going to be lower. Investigators can make useful conclusions about the data even with a low R-squared value.

In a different case, such as in investing, a high R-squared value—typically between 85% and 100%—indicates the stock or fund's performance moves relatively in line with the index. This is very useful information to investors thus a higher R-squared value is necessary for a successful project.

.

***Q.11: ans**. - Linear regression is a type of linear model that is considered the most basic and commonly used predictive algorithm. This can not be dissociated from its simple, yet effective architecture. A linear model assumes a linear relationship between input variable(s) x and an output variable y. a linear model with n number of features. w is considered the coefficient (or weights) assigned to each feature - an indicator of their significance to the outcome y. For example, we assume that temperature is a larger*

driver of ice cream sales than whether it's a public holiday. The weight assigned to temperature in our linear model will be larger than the public holiday variable.

The goal for a linear model then becomes to optimize the weight (b) via the cost function in equation 1.2. The cost function calculates the error between predictions and actual values, represented as a single real-valued number. The cost function is the average error across n samples in the dataset, This is a regularization technique used in feature selection using a Shrinkage method also referred to as the penalized regression method. Lasso is short for Least Absolute Shrinkage and Selection Operator, which is used both for regularization and model selection. If a model uses the L1 regularization technique, then it is called lasso regression.

Lasso Regression for Regularization

In this shrinkage technique, the coefficients determined in the linear model from equation 1.1. above are shrunk towards the central point as the mean by introducing a penalization factor called the alpha α (or sometimes lamda) values.

***Q.12: ans**. - A variance inflation factor (VIF) is a measure of the amount of multicollinearity in regression analysis. Multicollinearity exists when there is a correlation between multiple independent variables in a multiple regression model. This can adversely affect the regression results. Thus, the variance inflation factor can estimate how much the variance of a regression coefficient is inflated due to multicollinearity.*

KEY TAKEAWAYS

- *A variance inflation factor (VIF) provides a measure of multicollinearity among the independent variables in a multiple regression model.*
- *Detecting multicollinearity is important because while multicollinearity does not reduce the explanatory power of the model, it does reduce the statistical significance of the independent variables.*
- *A large VIF on an independent variable indicates a highly collinear relationship to the other variables that should be considered or adjusted for in the structure of the model and selection of independent variables.*

Understanding a Variance Inflation Factor (VIF)

A variance inflation factor is a tool to help identify the degree of multicollinearity. Multiple regression is used when a person wants to test the effect of multiple variables on a particular outcome. The dependent variable is the outcome that is being acted upon by the independent variables—the inputs into the model. Multicollinearity exists when there is a linear relationship, or correlation, between one or more of the independent variables or inputs.

WORKSHEET- VI

The Problem of Multicollinearity

Multicollinearity creates a problem in the multiple regression model because the inputs are all influencing each other. Therefore, they are not actually independent, and it is difficult to test how much the combination of the independent variables affects the dependent variable, or outcome, within the regression model.

While multicollinearity does not reduce a model's overall predictive power, it can produce estimates of the regression coefficients that are not statistically significant. In a sense, it can be thought of as a kind of double-counting in the model.

In statistical terms, a multiple regression model where there is high multicollinearity will make it more difficult to estimate the relationship between each of the independent variables and the dependent variable. In other words, when two or more independent variables are closely related or measure almost the same thing, then the underlying effect that they measure is being accounted for twice (or more) across the variables. When the independent variables are closely-related, it becomes difficult to say which variable is influencing the dependent variables.

Small changes in the data used or in the structure of the model equation can produce large and erratic changes in the estimated coefficients on the independent variables. This is a problem because the goal of many econometric models is to test exactly this sort of statistical relationship between the independent variables and the dependent variable.

Tests to Solve Multicollinearity

To ensure the model is properly specified and functioning correctly, there are tests that can be run for multicollinearity. The variance inflation factor is one such measuring tool. Using variance inflation factors helps to identify the severity of any multicollinearity issues so that the model can be adjusted. Variance inflation factor measures how much the behavior (variance) of an independent variable is influenced, or inflated, by its interaction/correlation with the other independent variables.

Variance inflation factors allow a quick measure of how much a variable is contributing to the standard error in the regression. When significant multicollinearity issues exist, the variance inflation factor will be very large for the variables involved. After these variables are identified, several approaches can be used to eliminate or combine collinear variables, resolving the multicollinearity issue.

Formula and Calculation of VIF

The formula for VIF is:

Where R_i^2 represents the unadjusted coefficient of determination for regressing the i^{th} independent variable on the remaining ones.

***Q.13: ans**. - Structured Query Language(SQL) as we all know is the database language by the use of which we can perform certain operations on the existing database and also we can use this language to*

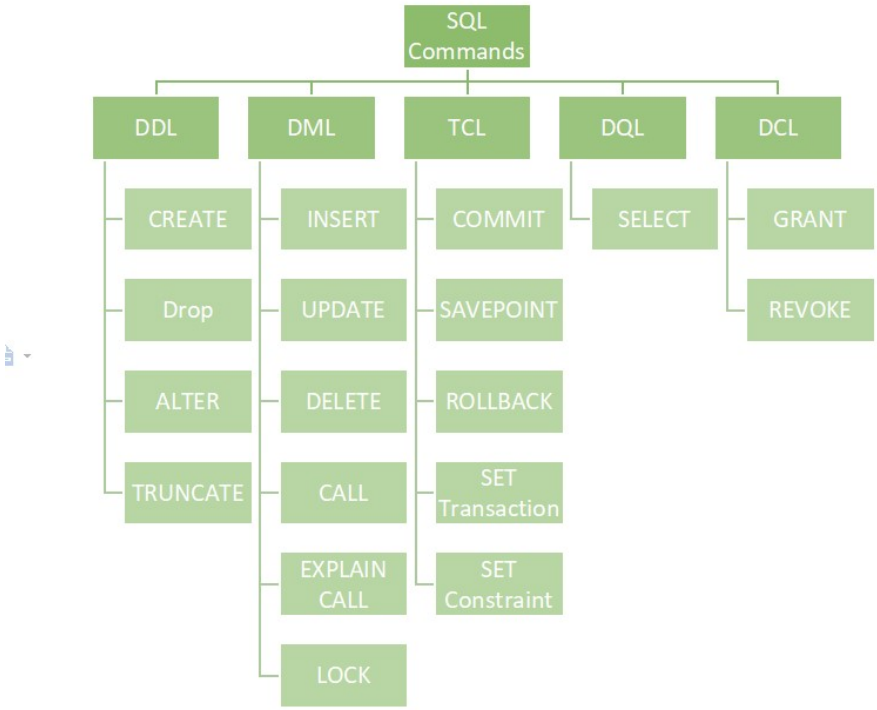
WORKSHEET- VI

create a database. SQL uses certain commands like Create, Drop, Insert, etc. to carry out the required tasks.

These SQL commands are mainly categorized into five categories as:

1. *DDL – Data Definition Language*
2. *DQL – Data Query Language*
3. *DML – Data Manipulation Language*
4. *DCL – Data Control Language*
5. *TCL – Transaction Control Language*

Now, we will see all of these in detail.



DDL (Data Definition Language):

DDL or Data Definition Language actually consists of the SQL commands that can be used to define the database schema. It simply deals with descriptions of the database schema and is used to create and modify the structure of database objects in the database. DDL is a set of SQL commands used to create, modify, and delete database structures but not data. These commands are normally not used by a general user, who should be accessing the database via an application.

WORKSHEET- VI

List of DDL commands:

- *CREATE: This command is used to create the database or its objects (like table, index, function, views, store procedure, and triggers).*
- *DROP: This command is used to delete objects from the database.*
- *ALTER: This is used to alter the structure of the database.*
- *TRUNCATE: This is used to remove all records from a table, including all spaces allocated for the records are removed.*
- *COMMENT: This is used to add comments to the data dictionary.*
- *RENAME: This is used to rename an object existing in the database.*

DQL (Data Query Language):

DQL statements are used for performing queries on the data within schema objects. The purpose of the DQL Command is to get some schema relation based on the query passed to it. We can define DQL as follows it is a component of SQL statement that allows getting data from the database and imposing order upon it. It includes the SELECT statement. This command allows getting the data out of the database to perform operations with it. When a SELECT is fired against a table or tables the result is compiled into a further temporary table, which is displayed or perhaps received by the program i.e. a front-end.

List of DQL:

- *SELECT: It is used to retrieve data from the database.*

DML(Data Manipulation Language):

The SQL commands that deals with the manipulation of data present in the database belong to DML or Data Manipulation Language and this includes most of the SQL statements. It is the component of the SQL statement that controls access to data and to the database. Basically, DCL statements are grouped with DML statements.

List of DML commands:

- *INSERT : It is used to insert data into a table.*
- *UPDATE: It is used to update existing data within a table.*
- *DELETE : It is used to delete records from a database table.*
- *LOCK: Table control concurrency.*
- *CALL: Call a PL/SQL or JAVA subprogram.*

WORKSHEET- VI

- *EXPLAIN PLAN: It describes the access path to data.*

DCL (Data Control Language):

DCL includes commands such as GRANT and REVOKE which mainly deal with the rights, permissions, and other controls of the database system.

List of DCL commands:

- *GRANT: This command gives users access privileges to the database.*
- *REVOKE: This command withdraws the user's access privileges given by using the GRANT command.*

TCL (Transaction Control Language):

Transactions group a set of tasks into a single execution unit. Each transaction begins with a specific task and ends when all the tasks in the group successfully complete. If any of the tasks fail, the transaction fails. Therefore, a transaction has only two results: success or failure. You can explore more about transactions here. Hence, the following TCL commands are used to control the execution of a transaction:

- *COMMIT: Commits a Transaction.*
- *ROLLBACK: Rollbacks a transaction in case of any error occurs.*
- *SAVEPOINT: Sets a save point within a transaction.*
- *SET TRANSACTION: Specifies characteristics for the transaction.*

.

***Q.14: ans**. - Many random variables have distributions that are asymptotically Gaussian but may be significantly non-Gaussian for small numbers. For example the Poisson Distribution, which describes (among other things) the number of unlikely events occurring after providing a sufficient opportunity for a few events to occur. It is pretty non-Gaussian unless the mean number of events is very large. The mathematical form of the distribution is still Poisson, but a histogram of the number of events after many trials with a large average number of events eventually looks fairly Gaussian.*

For me, the best examples come from my field of research (astrophysical data analysis). For example, something that comes up all the time is that we detect stars in astronomical images and solve for their celestial coordinates. My current project uses images about 1.5 degrees on a side and typically detects 60 to 80 thousand stars per image, with the number well modeled as a Poisson Distribution, assuming that the image is not of a star cluster surrounded by mostly empty space. That's about 8 or 9 stars per square arcminute. If we cut out "postage stamps"from the image that are half an arcminute per side, then the mean number of detected stars in them is about 2. If we do that for (say) 1000 postage stamps

and make a histogram of the number of detected stars in them, it will not look very Gaussian, but as we increase the size of the postage stamps, it becomes asymptotically Gaussian.

What generally never becomes Gaussian, however, is the Uniform Distribution. A histogram of the stars' right ascensions or declinations (the azimuthal and elevation angles used in astronomy) looks a lot like a step function, i.e., flat within the image boundaries. The positions are not uniformly spaced, but they are distributed in the same way as a uniformly distributed random variable for any size postage stamp, including the entire image.

Another example is the location of the centers of raindrop ripples on a pond; they are not uniformly spaced in (say) the east-west direction, but they are uniformly distributed.

The simplest example is the distribution of numbers that show up on the top of a fair die after a large number of throws. Each number from 1 to 6 will occur with approximately equal frequency. Increasing the number of throws will not tend to produce a bell-shaped histogram, in fact the fractional occurrence will approach a constant 1/6 over the possible numbers.

***Q.15: ans**. For me, the best examples come from my field of research (astrophysical data analysis). For example, something that comes up all the time is that we detect stars in astronomical images and solve for their celestial coordinates. My current project uses images about 1.5 degrees on a side and typically detects 60 to 80 thousand stars per image, with the number well modeled as a Poisson Distribution, assuming that the image is not of a star cluster surrounded by mostly empty space. That's about 8 or 9 stars per square arcminute. If we cut out "postage stamps"from the image that are half an arcminute per side, then the mean number of detected stars in them is about 2. If we do that for (say) 1000 postage stamps and make a histogram of the number of detected stars in them, it will not look very Gaussian, but as we increase the size of the postage stamps, it becomes asymptotically Gaussian.*

What generally never becomes Gaussian, however, is the Uniform Distribution. A histogram of the stars' right ascensions or declinations (the azimuthal and elevation angles used in astronomy) looks a lot like a step function, i.e., flat within the image boundaries. The positions are not uniformly spaced, but they are distributed in the same way as a uniformly distributed random variable for any size postage stamp, including the entire image.

Many random variables have distributions that are asymptotically Gaussian but may be significantly non-Gaussian for small numbers. For example the Poisson Distribution, which describes (among other things) the number of unlikely events occurring after providing a sufficient opportunity for a few events to occur. It is pretty non-Gaussian unless the mean number of events is very large. The mathematical form of the distribution is still Poisson, but a histogram of the number of events after many trials with a large average number of events eventually looks fairly Gaussian.

For me, the best examples come from my field of research (astrophysical data analysis). For example, something that comes up all the time is that we detect stars in astronomical images and solve for their celestial coordinates. My current project uses images about 1.5 degrees on a side and typically detects 60 to 80 thousand stars per image, with the number well modeled as a Poisson Distribution, assuming

that the image is not of a star cluster surrounded by mostly empty space. That's about 8 or 9 stars per square arcminute. If we cut out "postage stamps"from the image that are half an arcminute per side, then the mean number of detected stars in them is about 2. If we do that for (say) 1000 postage stamps and make a histogram of the number of detected stars in them, it will not look very Gaussian, but as we increase the size of the postage stamps, it becomes asymptotically Gaussian.

What generally never becomes Gaussian, however, is the Uniform Distribution. A histogram of the stars' right ascensions or declinations (the azimuthal and elevation angles used in astronomy) looks a lot like a step function, i.e., flat within the image boundaries. The positions are not uniformly spaced, but they are distributed in the same way as a uniformly distributed random variable for any size postage stamp, including the entire image.

Another example is the location of the centers of raindrop ripples on a pond; they are not uniformly spaced in (say) the east-west direction, but they are uniformly distributed.

The simplest example is the distribution of numbers that show up on the top of a fair die after a large number of throws. Each number from 1 to 6 will occur with approximately equal frequency. Increasing the number of throws will not tend to produce a bell-shaped histogram, in fact the fractional occurrence will approach a constant 1/6 over the possible numbers.