Ministry of High Education and Scientific Research
University Mohamed Khider of Biskra
Faculty of Economics, Commerce and Management Sciences
Department of Commerce

Time Series Analysis

Pedagogical publication presented for the 1st year Master degree students finance & international Trade

Presented by: Prof. Yasmina Guechari

Academic year 2023/2024

Introduction

Data can take many forms, it could represent measurements, survey responses, financial transactions, sensor readings, or any other information relevant to the domain of interest. Economic data refers to information concerning economic activities within a specific region or secton; this data could include metrics such as GDP (Gross Domestic Product), inflation rates, unemployment rates, stock prices, commodity prices, and more. Understanding economic data is crucial for making informed decisions in various domains, from policy-making to investment strategies. Time series analysis is a statistical technique used to analyze data points collected, recorded, or observed over a period of time.

Time series analysis provides a powerful framework for understanding and interpreting data collected over time. By employing techniques such as simple and multiple linear regression, enabling informed decision-making and predictive modeling across a wide range of fields and applications. Understanding the characteristics of time series data and employing appropriate analytical techniques are essential for deriving meaningful conclusions and actionable insights from temporal datasets.

The present work serves as a comprehensive guide to understanding the principles, methodologies, and applications that underpin the field of time series analysis; this is by addressing the following axes:

- **Introduction to Time Series Analysis**
- **Simple & Multiple Linear Regression Model**
- **Stationarity Autocorrelation and Partial Autocorrelation**

Table of Content

Part one:

Introduction to time series data

Time series data represents observations or measurements collected, recorded, or observed sequentially over a period of time. Unlike cross-sectional data, which captures information at a single point in time, time series data provides insights into how phenomena evolve over time. Understanding time series data is essential for making informed decisions, identifying patterns, trends, and seasonal variations, as well as for forecasting future values. This type of data often exhibits temporal dependencies, where each observation is influenced by previous observations, making it distinct from other forms of data analysis.

In this chapter, we will explore the fundamental concepts of time series data, including its characteristics, components, We will also discuss various applications across different domains.

Chapter 1: Economic Data

An economic dataset typically comprises a collection of structured data points that represent various economic indicators, variables, or metrics over a specific period. These datasets can be sourced from government agencies, international organizations, research institutions, or private sector organizations. The structure of an economic dataset can vary depending on factors such as the specific economic indicators being measured, the frequency of data collection, and the intended use of the dataset.

1-1- Concept of economic data

An economic data set refers to a structured collection of quantitative information related to various aspects of economic activity within a particular region, industry, or market. The majority of the economic data of a country or a region comes from government agencies such as the Central Banks, or international organizations, such as the World Bank, International Monetary Fund, Data comes from market research along with various published reports. Economic data is typically historical data collected and curated over a given period (Explorium, 2023). Economic data Attributes differs based on countries and cover the essential economic indicators such as Gross Domestic Product (GDP), GDP growth rate, Government debt to GDP, Gross national expenditure, National wealth, Balance of trade, Interest rate, Inflation rate, Unemployment rate, International Trade, Exports, Industrial Production; the data attributes may also encompass prices, taxes, employment data, retail sales, housing starts, home sales, and other relevant business and trade data, typically published by government or trade organizations (Explorium, 2023). Economic data is typically expressed in numerical terms, allowing for precise measurement and analysis of economic phenomena. These measurements may include monetary values, quantities, rates, indices, and percentages. From the above definition we can conclude some key features of economic data:

- ❖ **Wide Scope**: Economic data covers diverse aspects of economic activity ; it provides a comprehensive view of the economy's performance and structure.
- ❖ **Regular Collection**: Economic data is collected on a regular basis through various sources, such as government agencies, international organizations, central banks, research institutions, and surveys of households and businesses. Data collection may occur at different frequencies, ranging from daily, monthly, quarterly, to annually.
- ❖ **Time Series Nature**: Economic data is often organized into time series, where observations are recorded over successive time periods. This allows for the analysis of trends, patterns, and cyclical movements in economic variables over time.
- ❖ **Cross-Sectional Dimension**: Economic data may also have a cross-sectional dimension, where observations are collected for different entities or units (e.g., countries, regions, industries, households, firms) at a specific point in time. Cross-sectional data enables comparisons and analyses across different groups or categories.
- ❖ **Statistical Analysis**: Economic data is subject to statistical analysis and interpretation using various quantitative techniques, including descriptive statistics, regression analysis, time series analysis, and econometric modeling. These analytical tools help economists, policymakers, businesses, and researchers identify relationships, test hypotheses, and make predictions based on the data.

1-2- Structure of economic data

The structure of an economic dataset can vary depending on the frequency of data collection, and the number of sections. A statistical model used to analyze the economic relationship must be compatible with the type of data, so it is important to know the type of data we are using before formulating the model. The most important types of data that the econometrician or economist who conducts correlation studies is interested in are:

i. **Time Series Data:** time series data set consists of observations on a variable or several variables over time. Examples of time series data include stock prices, money supply, consumer price index, gross domestic product, annual homicide rate……. The chronological ordering of observations in a time series conveys potentially important information (Wooldridge, 2012).

Example: Monthly oil price and USD exchange rates. Here's a simplified representation:

Table 1: Time series data example

time period	oil price	exchange rate
Mar-19	63,79	0,885152
Apr-19	68,58	0,890132
May-19	66,83	0,893524
Jun-19	59,76	0,884948
Jul-19	61,48	0,891499
Aug-19	57,67	0,897508
Sept-19	60,04	0,907199
Oct-19	57,27	0,903815
Nov-19	60,4	0,904806
Dec-19	63,35	0,900067

Source: author’ elaboration based on the data from: https://fred.stlouisfed.org/series/DCOILBRENTEU

In this example, the time series data consists of monthly oil price (in US Dollar) recorded over year 2019, and Monthly US Dollar exchange rates (USD/EURO exchange rates). Each column in the table represents a time series data (Monthly oil prices data and monthly USD exchange rates data).

ii. **Cross-Section Data:** consist of data from several units observed at the same time or in the same time period. The data may be single observations from a sample survey or from all units in a population. Examples of cross-section data are the Household Budget Survey for the year 2022, The Manufacturing Statistics for the year 2000, the Population Census for the year 2011 (Griliches, 1986).

The difference between Cross-Sectional data and time series data: the cross-sectional data show spatial variation: Variation across units (individuals, households, firms, ...), while Time-Series data show temporal variation: Variation over periods (years, months, weeks, seconds, ...) (Griliches, 1986)

Example: Cross-sectional data refers to information collected at a single point in time across different entities or units. Here's an example of cross-sectional data:

Let's say you're conducting a survey to study the average income levels of individuals in a city. You randomly select 500 people from different areas of the city and ask them about their annual income. Each person's response constitutes one data point. The data collected in this survey represents cross-sectional data because it captures the income levels of individuals at a specific point in time, without tracking changes over time.

Here's a simplified representation of the cross-sectional data collected:

Table 2: Cross- Sectional data example

Person ID	**Income (USD)**
1	50,000
2	65,000
3	40,000
...	...
500	55,000

Source: author' elaboration

This is an example of cross-sectional data because it collects information on income across different individuals (or units) at a single point in time.

iii. Panel Data

Panel data, also known as longitudinal data or cross-sectional time-series data, refers to a type of dataset that combines both cross-sectional and time-series dimensions. In other words, it contains observations on multiple individuals or entities (cross-sectional units) observed over multiple time periods (Wooldridge, 2012).

Example: imagine you have data on five countries' GDP growth rates over three years. Each country represents a cross-sectional unit, and the years represent the time-series dimension. This setup allows researchers to analyze both within-unit variations over time and between-unit variations at a particular point in time.

Table 3: Panel data example

Country	Year	GDP Growth Rate (%)
USA	2019	2.3
USA	2020	-3.5
USA	2021	4.1
Canada	2019	1.9
Canada	2020	-2.1
Canada	2021	3.0
UK	2019	1.5
UK	2020	-4.0
UK	2021	2.2
Germany	2019	0.8
Germany	2020	-1.8
Germany	2021	3.5
Japan	2019	0.7
Japan	2020	-5.0
Japan	2021	2.8

Source: author' elaboration based on the data from: https://fred.stlouisfed.org/series/GDPC1MD

In this dataset: Each row represents the GDP growth rate of a specific country in a particular year. The data allows researchers to analyze both within-unit variations over time (e.g., how GDP growth rates change within each country across different years) and between-unit variations at a particular point in time (e.g., comparing GDP growth rates between different countries in the same year).

With this panel data, researchers can conduct various analyses, such as examining trends in GDP growth rates over time for each country, comparing the performance of different countries in specific years, or investigating factors influencing GDP growth across countries.

Panel data analysis is particularly useful for studying phenomena where both individual heterogeneity and time dynamics are important, such as economic growth, labor market dynamics, consumer behavior, and many others. Common techniques for analyzing panel data include fixed effects models, random effects models, pooled OLS regression, and various types of dynamic panel data models. Panel data enables researchers to study both the time dimension (changes within entities over time) and the cross-sectional dimension (differences across entities at a given point in time), providing richer insights than cross-sectional or time series data alone.

The key feature of panel data that distinguishes them from a pooled cross section is that the *same* cross-sectional units (individuals, firms, or counties) are followed over a given time period.

Chapter 2: What is Time Series Data?

Time series data refers to a sequence of observations or measurements collected and recorded at successive, equally spaced intervals over time ; these observations are typically taken at regular intervals depending on the context of the data being analyzed. Time series data is widely used in various fields for analyzing trends, forecasting future values, and understanding the underlying patterns and dynamics over time.

2-1- Concept of Time Series data

Time series data: is the data collected for a single entity at multiple points in time (Stock & Watson, 2020).

i. Frequency of Time Series Data

The frequency of time series data refers to the regularity or interval at which observations are recorded or measured. It indicates how often data points are collected within a given unit of time. The choice of frequency depends on the nature of the phenomenon being studied and the purpose of the analysis (Shumway & Stoffer , 2011).

Common frequencies for time series data include:

- ❖ **Annual**: Data points are recorded or measured once per year. This frequency is often used for macroeconomic indicators such as annual GDP growth rates, population counts, or annual financial statements.

- **Quarterly**: Data points are recorded or measured once per quarter, typically at the end of each quarter. Many economic indicators, financial reports, and government statistics are reported quarterly, such as quarterly GDP, unemployment rates, or corporate earnings.
- **Monthly**: Data points are recorded or measured once per month, usually at the end of each month. Monthly data is common in economic and financial time series, including monthly sales, employment data, or consumer price indices.
- **Weekly**: Data points are recorded or measured once per week, typically at the end of each week. Weekly data may be used for tracking stock prices, retail sales, or other high-frequency phenomena.
- **Daily**: Data points are recorded or measured once per day, usually at the same time each day. Daily data is commonly used for financial market data- stock prices, website traffic.

The choice of frequency depends on factors such as the **availability of data**, the **level of detail required for analysis**, **the underlying dynamics of the phenomenon**, and **the computational resources available for processing the data**. The higher frequency data provide more detailed insights.

2-2- Example of Time Series data

Here's an example of time series data representing monthly sales for a XYZ company over a period of one year:

Table 4: Time series data example

Time Period	Sales
jan 2023	15000
Feb 2023	18000
Mar 2023	20000
Apr 2023	22000
May 2023	25000
Jun 2023	28000
Jul 2023	30000
Aug 2023	32000
Sep 2023	34000
Oct 2023	38000
Nov 2023	40000
Dec 2023	45000

Source: author’ elaboration

- Each row represents a specific month in the year 2023.
- The first column represents the month and year.
- The second column represents the sales figures for that month, measured in dollars ($).

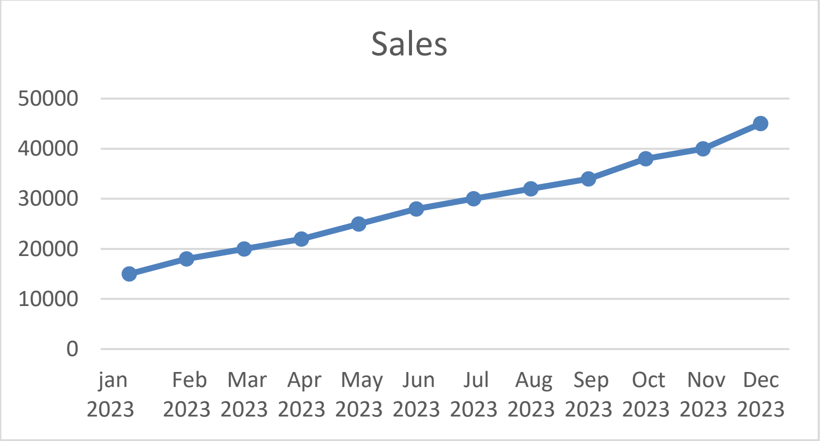
This time series dataset captures the variation in sales over a year for the company XYZ. Analysis of this data could reveal seasonal patterns, trends, or anomalies in sales performance over time, which could inform business decisions such as inventory management, marketing strategies, and financial planning.

2-3- Time Series Data graph

here are several types of time series graphs, each suitable for different types of data analysis and visualization purposes. Here are some common types used to represent the data of one variable:

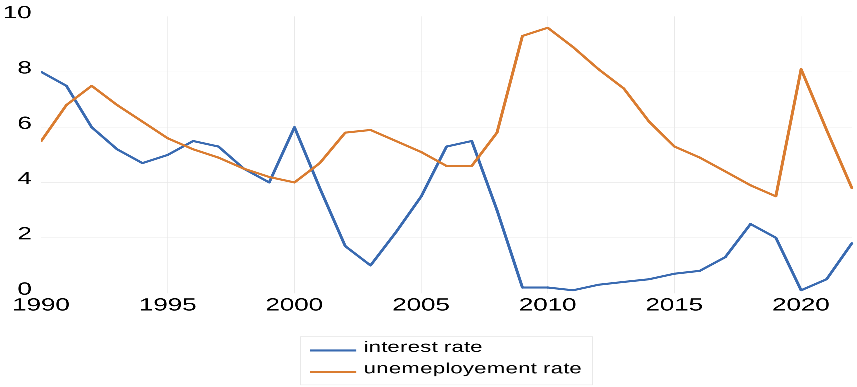
- ❖ **Line Graph**:

Figure 1: example of time series graph (line graph)



Source: Excel output based on the data from table 4

Figure 2: example of time series graph (line graph)



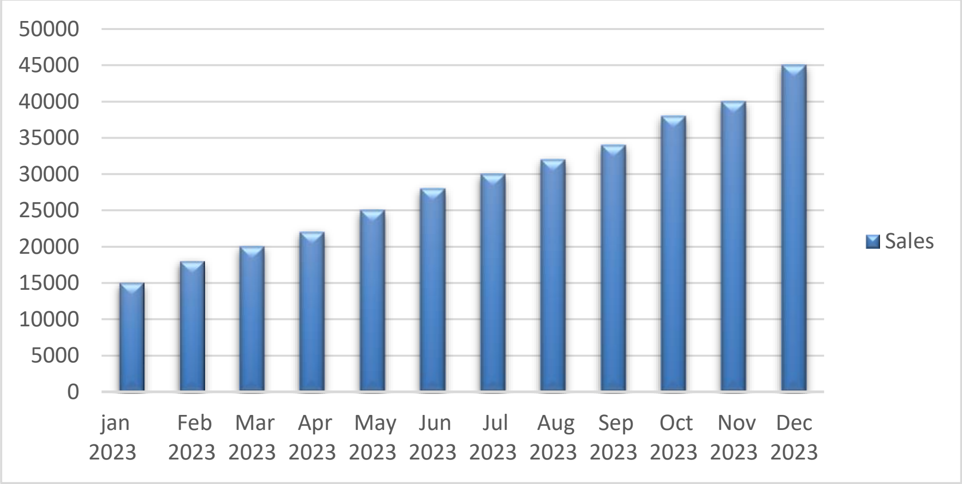
Source : EViews output based on the data from https://fred.stlouisfed.org/series/REAINTRATREARAT10Y; https://fred.stlouisfed.org/series/UNRATE

The line graphs is suitable and widely used method for presenting time series data due to their ability to convey trends, patterns, and relationships over time in a clear and understandable manner.

❖ **Bar Graph :**

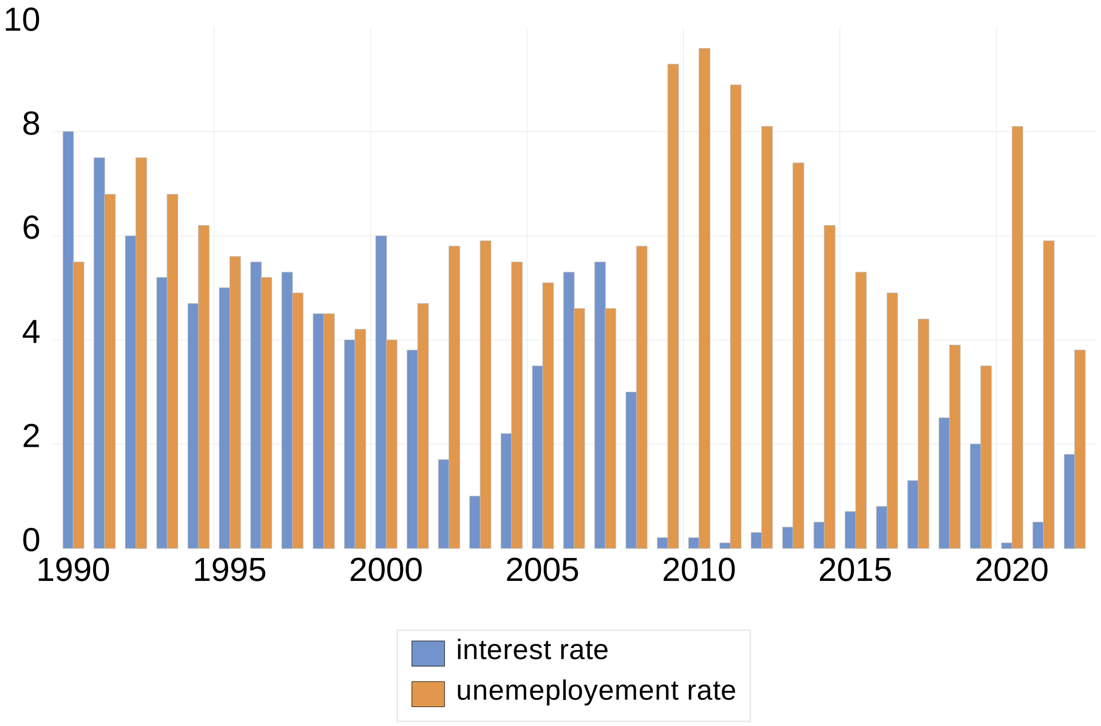
- Uses bars to represent the values of different time periods.
- Suitable for comparing values across different time periods or categories.

Figure 3: example of time series graph (Bar graph)



Source: Excel output based on the data from table 4

Figure 4: example of time series graph (Bar graph)



Source: EViews output base on the data from https://fred.stlouisfed.org/series/REAINTRATREARAT10Y; https://fred.stlouisfed.org/series/UNRATE

Bar graphs are a versatile tool for visualizing categorical data, comparing groups, and illustrating proportions or frequencies within each category. They are particularly useful when the data is discrete or when comparing multiple categories or groups.

Other type of graph representation, these graphs are particularly useful when analyzing the data characteristics and the relationships between two variables.

- **Scatter Plot**:
  - Displays individual data points as dots on a graph.
  - Useful for identifying relationships or correlations between two variables over time.
- **Stacked Area Graph**:
  - Similar to an area graph, but multiple series are stacked on top of each other.

  - Useful for comparing the contributions of different categories to the total over time.

- **Box Plot**:
  - Shows the distribution of a variable's values over time using quartiles.
  - Useful for identifying variability and outliers in the data.
- **Seasonal Decomposition Plot**:
  - Decomposes a time series into its trend, seasonal, and residual components.
  - Useful for understanding the underlying patterns and seasonal effects in the data.
- **Autocorrelation Plot**:
  - Displays the correlation of a time series with its lagged values.
  - Useful for identifying patterns and periodicity in the data.

These are just a few examples of the types of time series graphs commonly used in data analysis and visualization. The choice of graph depends on the specific objectives of the analysis and the characteristics of the data under study.

2-4- How to create a graph in Excel and EViews

To create a time series data graph, you'll typically need some data first. Once you have the data, you can use various tools and programming languages like EViews, Stata, SPSS or even Excel to create your graph.

Creating a graph in Excel is relatively straightforward. Here's a step-by-step guide to create a basic time series graph using Excel (Moore, McCabe, & Craig's, 2009):

i. **Enter Your Data**:
   - Enter your time series data into an Excel spreadsheet. Typically, you'll have two columns: one for the dates (or time periods) and another for the corresponding values.

ii. **Select Data**:
   - Highlight the data you want to include in the graph, including both the dates and values.

iii. **Insert Chart**:
   - Go to the "Insert" tab on the Excel ribbon.
   - Click on the "Line" or "Line with Markers" chart icon in the Charts group. This will create a basic line graph.

- If you prefer a different type of graph (such as a line or bar or scatter plot or area chart), you can choose that from the "Charts" dropdown menu.

iv. **Adjust Chart Options**:

- Once the chart is inserted, you can customize it further.
- Right-click on various elements of the chart (e.g., axis labels, data series) to access options for formatting.
- Experiment with different chart styles, colors, and layouts until you achieve the desired look.

v. **Add Titles and Labels**:

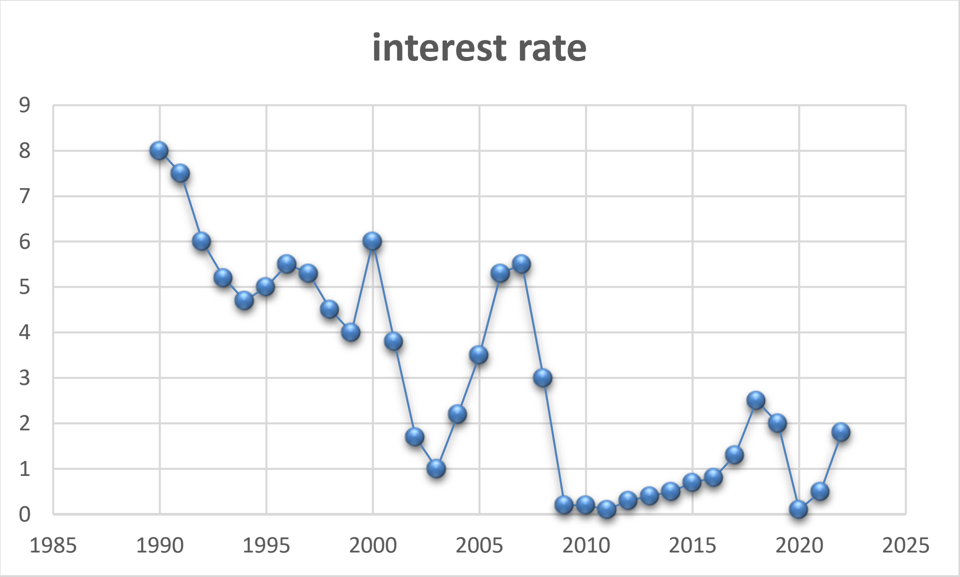
- Make sure to add a title to your chart that describes what the chart is showing.
- Label the x-axis (horizontal axis) with your time periods or dates, and label the y-axis (vertical axis) with the corresponding values.

vi. **Finalize and Save**:

- Once you're satisfied with your chart, you can save your Excel file.

Example:

Figure 5: Time series Data Graph

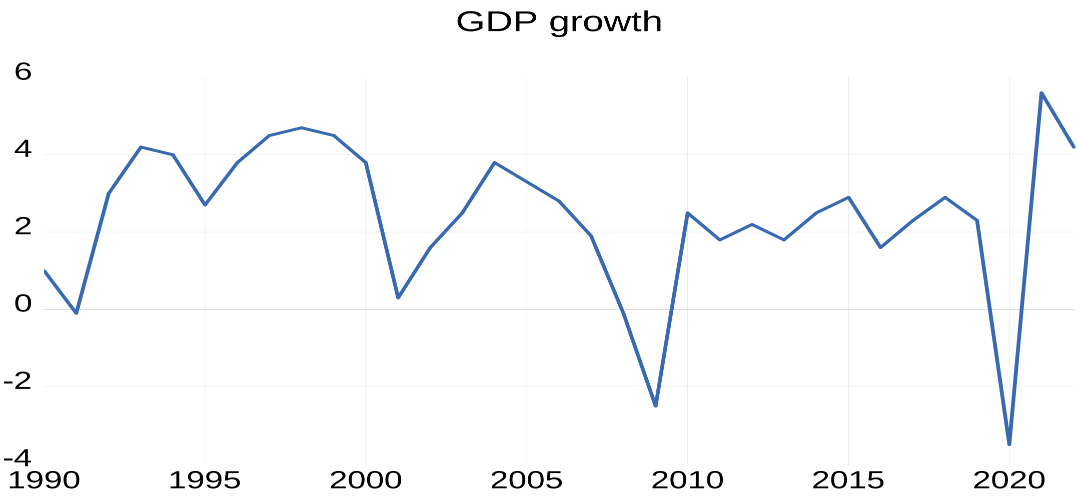


Source : Excel output based on the data from https://fred.stlouisfed.org/series/GDP

To create a graph in EViews, you can follow these steps (EViews, 2020):

1. **Open Your Data**: Start by opening your dataset in EViews.
2. **Select the Series**: Identify the series or variables you want to graph. In EViews, these are typically listed in the workfile window.
3. **Navigate to Quick Graph**: Once you have selected the series you want to graph, you have a couple of options to create a graph:
    - You can go to the "Quick" menu at the top and select "Graph..."
    - Alternatively, you can right-click on the selected series and choose "Quick/Graph..." from the context menu.
4. **Configure Graph Options**: In the "Quick Graph" dialog box, you can configure various options for your graph:
    - Choose the type of graph you want to create (line plot, scatter plot, bar chart, etc.) from the "Type" dropdown menu.
    - Select additional options such as titles, axis labels, legend, line styles, colors, etc.
5. **Preview and Create the Graph**: After configuring the options, you can preview your graph in the "Quick Graph" dialog box. Make sure everything looks correct.
    - If you're satisfied with the preview, click the "OK" button to create the graph.
6. **View and Edit the Graph**: Once the graph is created, it will appear in a new window within the EViews interface. You can further customize the graph by right-clicking on various elements (axis labels, data series, etc.) and accessing options for formatting.
    - You can also use the toolbar at the top of the graph window to access additional options for customization.
7. **Save or Export the Graph**: Finally, you can save it within your EViews workfile or export it to an image file (such as PNG or JPEG) by right-clicking on the graph and selecting the appropriate option from the context menu.

Figure 6: example of time series graph



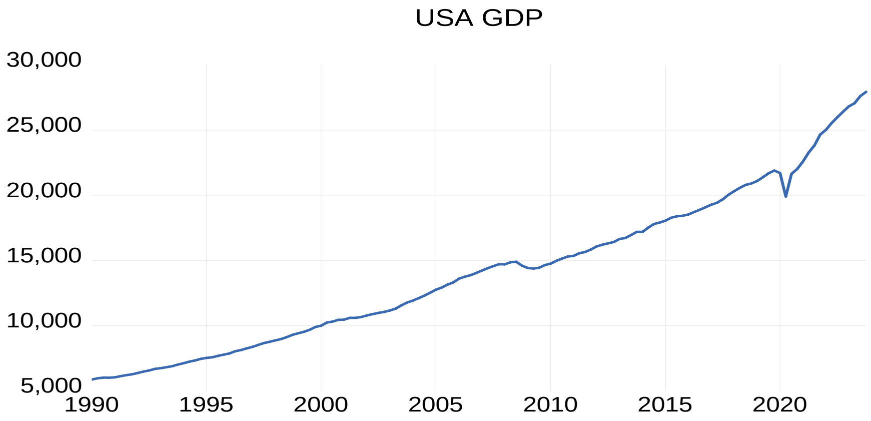
Source: E-views output base on the data from https://fred.stlouisfed.org/series/GDP

Chapter 3: Time series data Characteristics/Properties

3-1- Trend & Seasonality

i. **Trend**: The trend component of a time series represents the long-term movement or directionality of the data over time. It captures the underlying, sustained pattern of growth or decline in the data ; Trends can be upward (indicating growth), downward (indicating decline), or relatively flat (Field A. , 2005) .

Figure 7: Example of Trend component



Source: EViews output based on the data from https://fred.stlouisfed.org/series/GDP

The figure above reveals the presence of a discernible trend in the graph indicates a consistent pattern or directionality in the data, suggesting the existence of a underlying trend component. This trend is a continuous increase over time and across data points.

So, the trend Component signifies the **long-term movement** in a series; often influenced by factors such as population growth, price inflation and general economic development

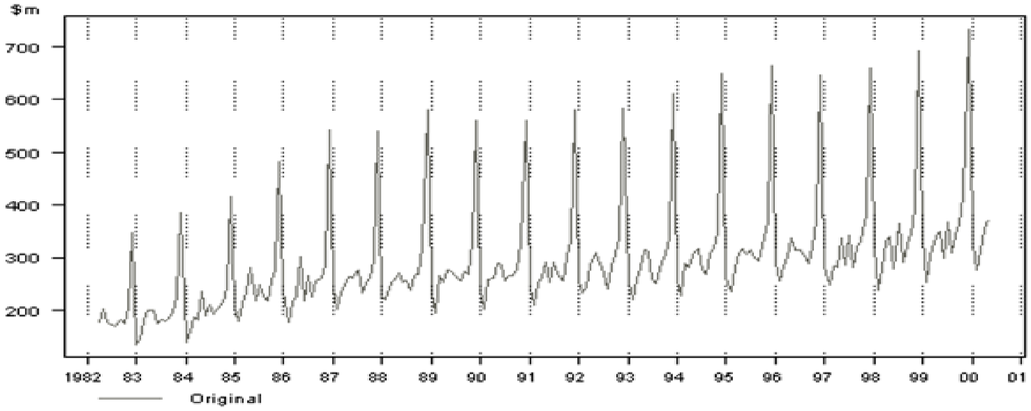
ii. **Seasonality**: Seasonality refers to regular and predictable patterns that repeat at fixed intervals within a time series data. These patterns are often related to calendar or seasonal factors, such as monthly, quarterly, or yearly cycles. Seasonal fluctuations may occur due to factors like weather, holidays, or cultural events. (Field A. , 2005)

Seasonality in data refers to recurring patterns or fluctuations that follow a consistent pattern within specific time frames, such as days, weeks, months, or quarters. These patterns are typically influenced by regular, predictable factors such as weather, holidays, or cultural events (Montgomery , Jennings, & Kulahci, 2015).

Example: Ice Cream Sales

In regions with distinct seasons, ice cream sales typically exhibit seasonality. During warmer months, such as spring and summer, ice cream sales tend to increase due to higher temperatures and increased demand for cold treats. Conversely, during colder months, like fall and winter, ice cream sales typically decrease as the demand decreases with the dropping temperatures. This seasonal pattern repeats annually, reflecting the influence of weather and seasonal factors on consumer behavior.

Figure 8 : Example of seasonality



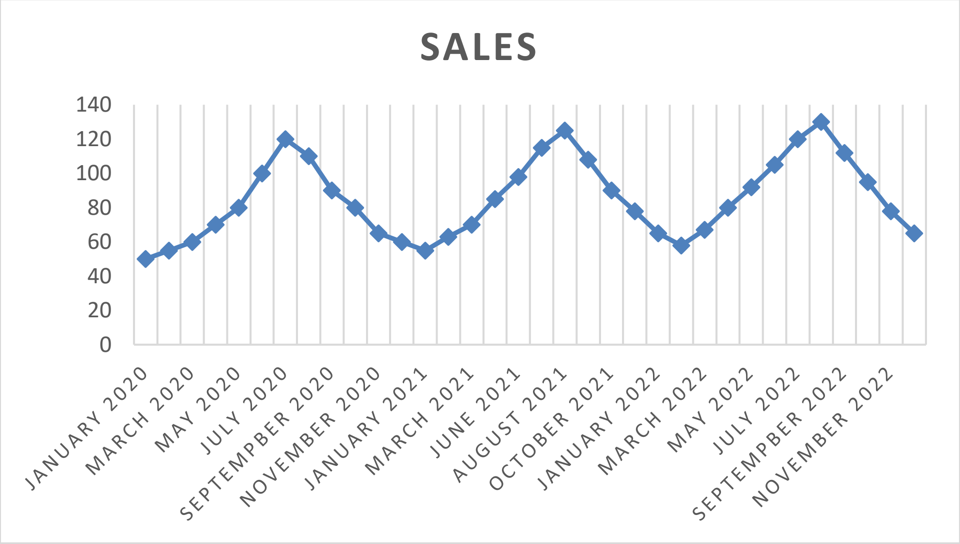
Source: https://www.slideserve.com/thomascatherine/chapter-2-time-series-powerpoint-ppt-presentation

Seasonality in a time series can be identified by regularly spaced peaks and troughs which have a consistent direction and approximately the same magnitude every year. The above diagram depicts a strongly seasonal series. There is an obvious large seasonal increase in December retail sales due to Christmas shopping; in this example, the magnitude of the seasonal component increases over time.

Let's consider an hypothetical example of monthly ice cream sales data to illustrate seasonality:

Month: Sales (in thousands of dollars)

Figure 8 : seasonality component



In this example, you can observe a seasonal pattern in ice cream sales data. Sales tend to increase during warmer months (e.g., June, July, August) and decrease during colder months (e.g., December, January), reflecting the influence of weather on consumer behavior. This recurring pattern within the data is characteristic of seasonality.

Understanding the trend and seasonality in time series data is crucial for making informed decisions in various fields such as economics, finance, sales forecasting, and environmental monitoring, among others.

3-2- Serial Correlation

Serial correlation, also known as autocorrelation, refers to the correlation between observations of a time series with themselves at different points in time. In other words, it measures the extent to which the values of a variable at different time periods are correlated with each other (Hamilton, 1994).

So, the Serial correlation refers to the correlation between a variable and a lagged version of itself; in other words, it examines the relationship between an observation and previous observations in a time series. Unlike regular correlation, which examines the relationship between two different variables, serial correlation assesses the relationship between a variable and its own past values.

Positive serial correlation (where successive observations tend to be similar) and negative serial correlation (where successive observations tend to be dissimilar) are both possible. (Wooldridge, 2012)

3-3- Exogenous & Endogenous properties

i. Exogenous:

Definition: An exogenous property refers to a characteristic or factor that originates from outside the system being analyzed. It is typically considered as an independent variable that influences the system but is not influenced by other variables within the system.

Reference: This definition aligns with the usage in economics, where exogenous variables are often considered as external factors that affect economic systems (Pindyck & Rubinfeld, 2011)

In the context of time series data analysis, the term "exogenous" is typically used to describe variables that are external to the system being analyzed. These exogenous variables are considered as inputs or factors that influence the behavior of the time series but are not influenced by the series itself (Shumway & Stoffer, 2017).

In econometric models, exogenous variables are often referred to as "explanatory variables" or "independent variables". These variables are often manipulated or controlled by the researcher and are not affected by other variables in the model.

Examples of exogenous variables include factors like government policies, external economic conditions, or demographic characteristics that are assumed to influence outcomes but are not themselves influenced by the other variables in the model.

ii. Endogenous:

In the realm of time series analysis, an "endogenous" property typically refers to variables or factors that are internal to the system being analyzed. These endogenous variables are influenced by other variables within the system and can be considered dependent variables in the modeling process (Shumway & Stoffer, 2017).

In econometric models, endogenous variables are often referred to as "dependent variables". These variables are typically the focus of analysis and are influenced by changes in the exogenous variables.

Examples of endogenous variables include outcomes like prices, quantities, or economic growth rates that are influenced by factors such as government policies, consumer behavior, or market conditions.

3-4- Cyclical variation and Irregular variation

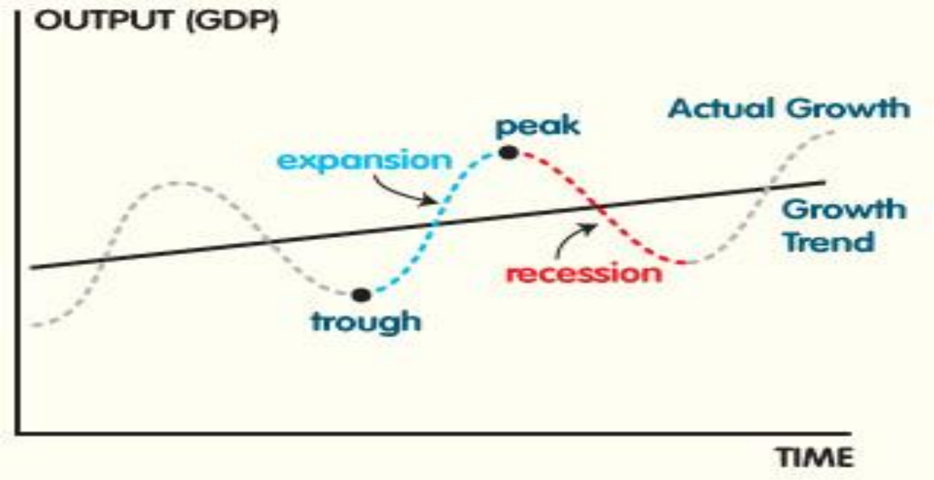
i. Cyclical Variation:

Cyclical variation as one of the fundamental components of time series data. Cyclical variation in time series data refers to the recurring patterns or fluctuations that occur over an extended period, typically spanning several years, and are not tied to any specific seasonality or trend. These cyclical movements represent the ebb and flow of economic activity or other phenomena, characterized by alternating periods of expansion and contraction (Chatfield, 2003).

From this definition we conclude that the cyclical variation refers to the periodic, repetitive fluctuations in time series data that occur over a longer time frame than seasonal variations. Unlike seasonal variations, which occur at fixed intervals such as weekly, monthly, or yearly, cyclical variations can have irregular durations and amplitudes ; it typically represents economic cycles, business cycles, or other long-term patterns of expansion and contraction in an economy or system.

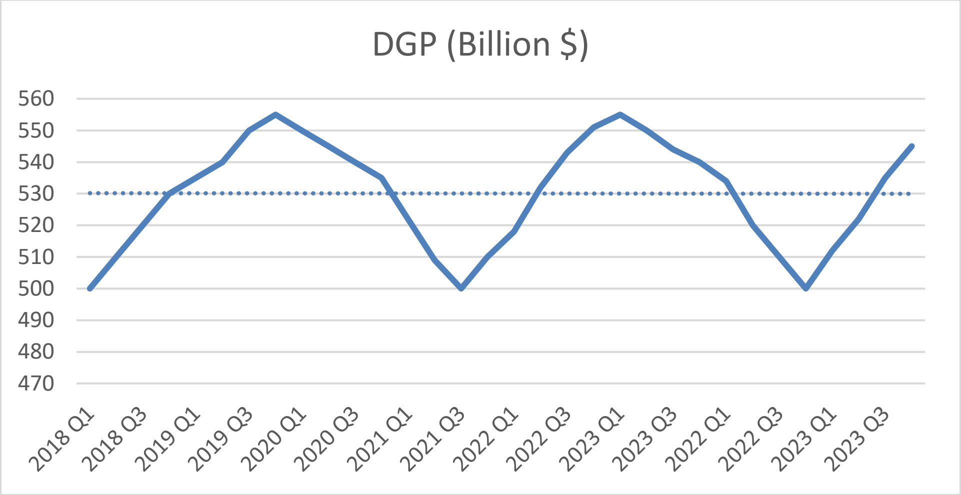
Examples of cyclical variations in economic time series data are economic recessions and booms. Let's consider an example of cyclical variation in the quarterly GDP (Gross Domestic Product) data of a country over a period of several years.

Figure 10 : Cyclical variation



Source : http://moirabaricollegeonline.co.in/attendence/classnotes/files/1608735484.pdf

Figure 11: Cyclical variation component



Source: author elaboration based on hypothetical data

In this example, we can observe a cyclical pattern in the GDP data:

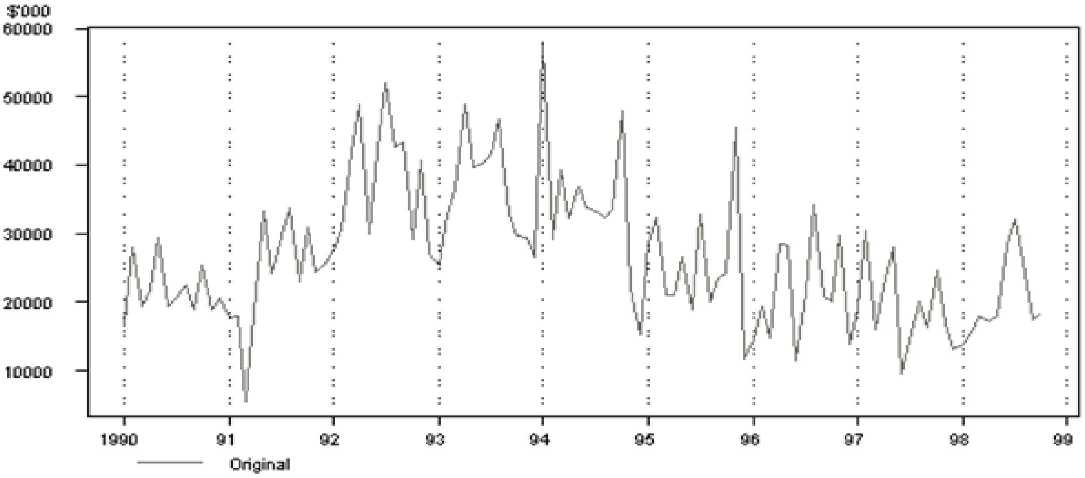
- Expansion Phase (2018-2019): GDP steadily increases from $500 billion in Q1 2018 to $555 billion in Q4 2019, indicating economic growth.
- Contraction Phase (2020): GDP starts declining from Q1 2020 ($550 billion) to Q4 2020 ($535 billion), signaling a slowdown or recessionary phase.

This alternating pattern of expansion and contraction in the GDP data over multiple quarters represents the cyclical variation in the economy. It reflects the natural fluctuations in economic activity driven by factors such as business cycles, investment cycles…….

ii. Irregular Variation:

Irregular variation in time series data refers to the random or unpredictable fluctuations that cannot be attributed to any systematic or identifiable pattern, trend, seasonality, or cyclical behavior. These irregular components are often considered as noise or random disturbances in the data that cannot be modeled or forecasted using traditional time series analysis techniques (Brockwell & Davis, Introduction to Time Series and Forecasting, 2016).

Figure 12 : Irregular variation component



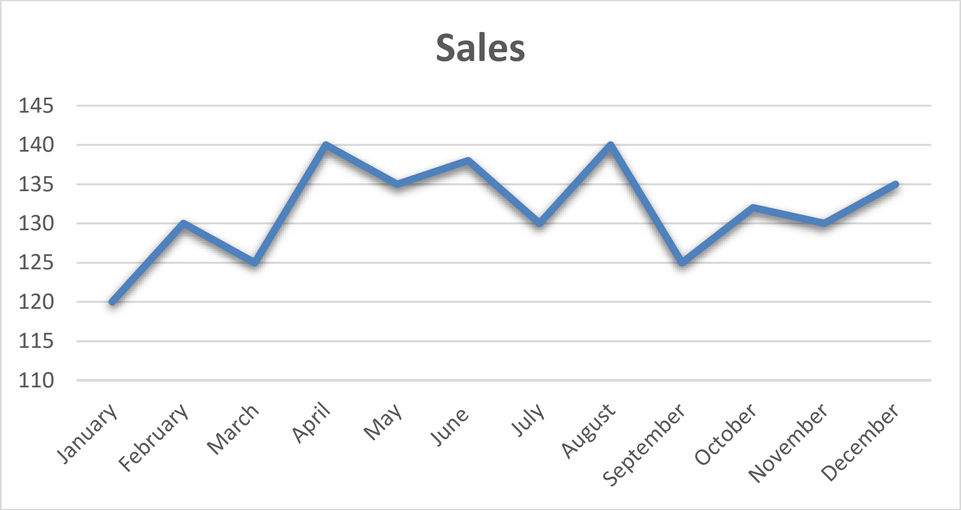
Source: https://www.slideserve.com/thomascatherine/chapter-2-time-series-powerpoint-ppt-presentation

Irregular variations are often caused by random or unforeseen events, measurement errors, or other unpredictable factors that affect the data but do not follow any discernible pattern.

Example: assume the data of monthly sales for a retail store XYZ:

time period	January	February	March	April	May	June	July	August	September	October	November
Sales	120	130	125	140	135	138	130	140	125	130	135

Figure 11: Example of irregular variation



Source: author elaboration based on hypothetical data

In this example, the sales data exhibit irregular variation, as there is no apparent pattern, trend, seasonality, or cyclical behavior. The fluctuations in sales from month to month seem random and unpredictable, indicating the presence of irregular components or noise in the data.

For instance, the sales in February increase from January, then decrease in March, rise again in April, decline in May, and slightly increase in June…. These fluctuations cannot be attributed to any systematic factors and may result from random variations in customer behavior, external events, or other unpredictable factors affecting sales.

This example illustrates how irregular variation manifests as random fluctuations in time series data, making it challenging to model or forecast using traditional time series analysis techniques.

In summary, cyclical variation represents long-term, periodic fluctuations in time series data related to economic or systemic cycles, while irregular variation represents unpredictable and erratic fluctuations caused by random or unforeseen factors. Understanding these components helps analysts better interpret and model time series data to extract meaningful insights.

Practice Questions

Exercise One: *choose the right answer for the following questions*

1. **What is the concept of economic data?**
   a) Data related to economic theories
   b) Data related to the study of the economy and its components
   c) Data related to stock market fluctuations
   d) Data related to political trends
2. **What does the structure of economic data refer to?**
   a) The physical format of the data storage
   b) The organization and arrangement of economic data
   c) The economic theories underlying the data
   d) The statistical techniques used to analyze the data
3. **What is Time Series data?**
   a) Data collected over time at regular intervals
   b) Data collected from multiple sources
   c) Data collected from a single point in time
   d) Data collected for a specific economic sector
4. **What does the term "frequency" refer to in Time Series Data?**
   a) The rate at which data is collected
   b) The number of data points in the series
   c) The periodicity of the data collection
   d) The amplitude of the data fluctuations
5. **Which of the following is an example of Time Series data?**
   a) Annual GDP growth rates for a country
   b) A snapshot of unemployment rates across different countries
   c) Monthly sales figures of a company
   d) Population data of a city from a single year
6. **What are trend and seasonality in Time Series data?**
   a) Trend is the long-term movement of data, while seasonality is the short-term cyclicality.

b) Trend is the short-term cyclicality, while seasonality is the long-term movement of data.

c) Trend and seasonality both refer to the long-term movement of data.

d) Trend and seasonality both refer to the short-term cyclicality of data.

7. **What is serial correlation in Time Series data?**

a) The correlation between different time series datasets

b) The correlation between consecutive observations in a single time series dataset

c) The correlation between different economic indicators

d) The correlation between different variables in an econometric model

8. **What do exogenous and endogenous properties refer to in Time Series data?**

a) Exogenous properties are determined within the system being analyzed, while endogenous properties are external influences.

b) Exogenous properties are external influences, while endogenous properties are determined within the system being analyzed.

c) Exogenous and endogenous properties both refer to external influences on the data.

d) Exogenous and endogenous properties both refer to internal characteristics of the data.

9. **What is cyclical variation and irregular variation in Time Series data?**

a) Cyclical variation refers to long-term movements, while irregular variation refers to short-term fluctuations.

b) Cyclical variation refers to short-term fluctuations, while irregular variation refers to long-term movements.

c) Cyclical variation and irregular variation both refer to short-term fluctuations in the data.

d) Cyclical variation and irregular variation both refer to long-term movements in the data.

Exercise two

1. **What type of data involves observations collected at multiple time points?**

a) Cross-sectional data

b) Time series data

c) Panel data

d) Longitudinal data

2. **Which type of data involves observations on multiple entities over time?**

a) Cross-sectional data

b) Time series data

c) Panel data

d) Longitudinal data

3. **If a dataset contains information on individuals or entities at a single point in time, what type of data is it?**

   a) Time series data

   b) Panel data

   c) Cross-sectional data

   d) Longitudinal data

4. **What type of data is typically used to analyze trends, seasonality, and cyclical patterns?**

   a) Cross-sectional data

   b) Time series data

   c) Panel data

   d) Longitudinal data

5. **If a dataset contains information on the same individuals or entities over several time periods, what type of data is it?**

   a) Time series data

   b) Cross-sectional data

   c) Longitudinal data

   d) Panel data

6. **What type of data is often used to study the effects of both time-invariant and time-varying variables?**

   a) Cross-sectional data

   b) Time series data

   c) Panel data

   d) Longitudinal data

7. **Which type of data is useful for analyzing individual-level changes over time?**

   a) Cross-sectional data

   b) Time series data

   c) Panel data

   d) Longitudinal data

8. **If a dataset contains information on different individuals or entities, each observed at a single point in time, what type of data is it?**

   a) Time series data

   b) Panel data

   c) Cross-sectional data

   d) Longitudinal data

9. **What type of data is commonly used in econometrics and social sciences to study both within-group and between-group variations?**

   a) Time series data

   b) Panel data

   c) Cross-sectional data

   d) Longitudinal data

10. **If a dataset contains information on individuals or entities at different time points, but each observation is independent of others, what type of data is it?**

    a) Time series data

    b) Panel data

    c) Cross-sectional data

    d) Longitudinal data

Answers

Exercise One

question	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
Answer	*b*	*b*	*a*	*c*	*a*	*a*	*b*	*b*	*a*

Exercise two

question	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Answer	*b*	*c*	*c*	*b*	*c*	*c*	*d*	*c*	*b*	*a*

Part two:
Simple & Multiple Linear Regression

Linear regression is a fundamental statistical technique used to understand the relationship between one or more independent variables and a dependent variable. It's widely employed in various fields, including economics, social sciences, engineering, and business, to analyze and predict outcomes based on input variables.

Chapter 1: What is a Simple Linear Regression?

Simple linear regression is a fundamental statistical method used to model the relationship between two continuous variables. It is a powerful tool for understanding and predicting the behavior of one variable based on the value of another. In simple linear regression, we aim to find the best-fitting straight line that represents the relationship between the predictor variable (also known as the independent variable or the explanatory variable) and the response variable (also known as the dependent variable).

1-1- Concept of Simple Linear Regression

i. **Definition:** "Simple linear regression is a statistical method used to model the relationship between two continuous variables by fitting a straight line to the observed data points. It aims to predict the value of one variable (the dependent or response variable) based on the value of another variable (the independent or predictor variable)." (Montgomery, Peck, & Vining, 2012)

ii. Model building

Regression analysis is a statistical technique for investigating and modeling the relationship between variables ; regression analysis may be the most widely used statistical technique (Montgomery, Peck, & Vining, 2012).

As an example of a problem in which regression analysis may be helpful, suppose that an economist starts with economic theory or intuition that suggests a relationship between GDP growth and unemployment. He assumes that, when the economy is growing rapidly, businesses tend to hire more workers to meet increased demand for goods and services, leading to a decrease in the unemployment rate. Conversely, during economic downturns or recessions, GDP growth slows down, leading to a rise in unemployment as businesses cut back on production and lay off workers. The economist collects the data observes the unemployment rate and GDP growth for a given country economy over thirteen years; the 13 observations are plotted in Figure 1. This graph is called a scatter diagram. This display clearly suggests a relationship between unemployment rate and GDP growth; in fact, the impression is that the data points generally, but not exactly, fall along a straight line (straight-line relationship).

If we let x represent the Unemployment rate and y represent GDP growth rate, then the equation of a straight line relating these two variables is

$Y = \beta_0 + \beta_1 X$ ……………………………………….*(Eq.1)*

where:

- Y : is the Response variable.
- X : is the predictor variable.
- β_0: is the intercept (the value of Y when X=0).
- β_1 : is the slope (the change in Y for a one-unit change in X).

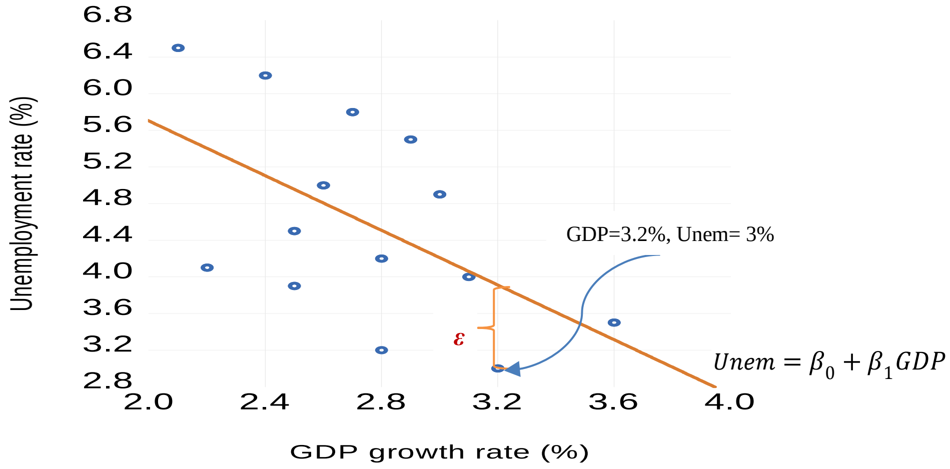
The data points do not fall exactly on a straight line, so Eq.1 should be modified to account for this. Let the difference between the observed value of y and the straight line ($\beta_0 + \beta_1 X$) be an error ε.

It is convenient to think of ε as a statistical error; that is, it is a random variable that accounts for the failure of the model to fit the data exactly. The error may be made up of the effects of other variables on unemployment rate, measurement errors, and so forth. Thus, a more plausible model for the unemployment rate data data is

$Y = \beta_0 + \beta_1 X + \varepsilon$ ……………………………………*(Eq.2)*

The Eq.2 is called a linear regression model. Customarily x is called the independent (predictor or regressor) variable and y is called the dependent (response) variable. Because Eq.2 involves only one regressor variable, it is called a simple linear regression model .

Figure 12: Scatter Plot (linear regression)



Source: EViews output base on the data from https://fred.stlouisfed.org/series/REAINTRATREARAT10Y; https://fred.stlouisfed.org/series/UNRATE

In the scatter plot, each point represents a value from the table above (the actual value of the dependent variable, Unemployment), while the regression line indicates the predicted values of Unemployment calculated using the Ordinary Least Squares (OLS) technique. Each data point on the scatter plot represents an observation or data pair (GDP, Unemployment). The position of the point corresponds to the actual values of the variables (value in the table); For example, the value of GDP is (3.2%) it's corresponding actual value of Unemployment is (3%) it is below the regression line and the predicted value of unemployment corresponding to the GDP = 3.2% is 3.9%; the difference between the actual value and the predicted value of the unemployment represent the error of estimation.

iii. **Variables**: In simple linear regression, we have two main variables (Kutner M. , Nachtsheim, Neter, & LI, 2005):

- **Predictor Variable (X)**: also known as the independent variable it used to predict or explain the behavior of another variable. It is denoted as X.
- **Response Variable (Y)**: This is the variable we are trying to predict or understand based on the predictor variable. It is also known as dependent variable, it is denoted as Y.

iv. **Objective:** The objective of simple linear regression is to estimate the values of β_0 and β_1 that minimize the sum of squared differences between the observed values of YY and the predicted values by the regression line.

v. **Assumptions:** Simple linear regression assumes that there is a linear relationship between the predictor variable and the response variable ; it relies on several assumptions, including (Gujarati & Porter, 2009): sq

- **Linearity**: The relationship between the independent variables (predictors) and the dependent variable (outcome) is linear. This means that the change in the dependent variable for a one-unit change in the independent variable is constant.
- **Independence** : The observations in the dataset are independent of each other. In other words, there should be no correlation between the residuals (errors) of the regression model. Violation of this assumption can lead to biased and inefficient estimates.
- **Homoskedasticity**: Also known as constant variance, this assumption states that the variance of the residuals is constant across all levels of the independent variables. In simpler terms, the spread of the residuals should remain roughly the same as the values of the independent variables change.
- **Normality of Residuals**: The residuals should be normally distributed. This means that the distribution of the residuals should follow a bell-shaped curve when plotted. Deviation from normality can affect the reliability of statistical tests and confidence intervals.
- **No Multicollinearity**: In multiple linear regression (with more than one independent variable), there should be no multicollinearity, which means that the independent variables should not be highly correlated with each other. Multicollinearity can lead to inflated standard errors and misleading interpretations of the coefficients.
- **No Autocorrelation**: In time-series data, the residuals should not exhibit autocorrelation, meaning that there should be no pattern in the residuals over time. Autocorrelation can indicate that the model is not adequately capturing some underlying dynamics in the data.

vi. **Application:** Simple linear regression is a powerful statistical technique for modeling and understanding the relationship between two continuous variables. By estimating the parameters of a linear equation, it allows us to make predictions, identify patterns, and draw insights from data. Simple linear regression has numerous applications across various fields, including (Kutner M. , Nachtsheim, Neter, & LI, 2005):

- **Economics:** Predicting the relationship between factors like income and expenditure.
- **Finance:** Analyzing the impact of interest rates on stock prices.

- **Healthcare:** Studying the relationship between patient age and health outcomes.
- **Marketing:** Understanding the effect of advertising spending on sales.

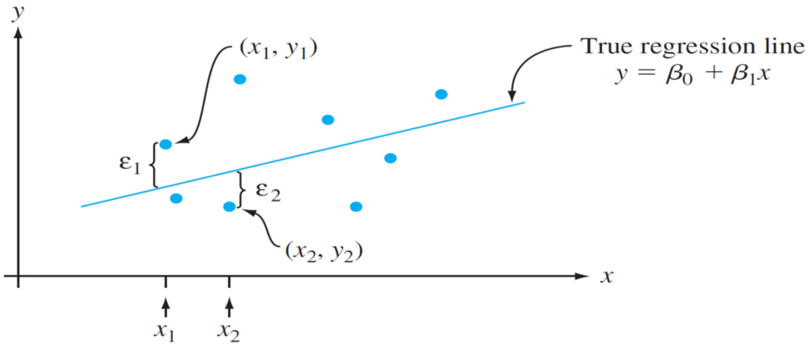
1-2- How Perform a Simple Linear Regression

Performing a simple linear regression involves fitting a linear model to a set of data points to find the best-fitting line that describes the relationship between the independent variable and the dependent variable. Here are the general steps to perform a simple linear regression (SANFORD, 2014):

i. **Collect Data**: Gather a set of data points that you believe exhibit a linear relationship between the independent and dependent variables.

ii. **Visualize the Data**: Plot the data points on a scatter plot to get an initial sense of the relationship between the variables.

Figure 13: scatter plot



Source: https://numerary.readthedocs.io/en/latest/linear-regression.html

Each data point on the scatter plot represents an observation or data pair (x, y). The position of the point corresponds to the real or the actual values of the variables it represents. The regression line is a line that best fits the observed data points in the scatter plot. It's an estimate of the relationship between the independent variable (x) and the dependent variable (y) based on the observed data. The regression line represents the best estimate of the relationship between x and y based on the observed data.

The error term $\varepsilon(\varepsilon_1, \varepsilon_2)$ represents the deviation or the difference between the observed value of the dependent variable (y) and the value predicted by the regression line. In other words, it quantifies the discrepancy between the actual data points and the model's predictions.

- **Example:** according to the figure 12, the data point lies above the regression line (x_1, y_1), the error term (ε_1) for that point is positive, indicating that the actual value of y is higher than the predicted value. Conversely, if a data point lies below the regression line like (x_2, y_2), the error term $(\varepsilon_2)$is negative.
- Significance: The error term captures the variability in the dependent variable that is not explained by the independent variable(s) included in the regression model. It is an essential component of the model's residual analysis and helps assess the model's goodness of fit.

iii. Estimate the parameters (Slope and intercept)

a. Estimate Slope (β1): To estimate the slope (β1) in a simple linear regression, you typically use the method of least squares. This method aims to minimize the sum of the squared differences between the observed values and the values predicted by the regression line ; Here's a basic outline of the steps:

1. Calculate the mean of the independent variable (X) and the mean of the dependent variable (Y).
2. Compute the deviations of each data point from their respective means: $(\boldsymbol{X_i} - \boldsymbol{\bar{X}})$ for X and $(\boldsymbol{Y_i} - \boldsymbol{\bar{Y}})$ for Y.
3. Compute the product of the deviations for each data point : $\boldsymbol{X_i - \bar{X}, Y_i - \bar{Y}}$
4. Compute the sum of the products of the deviations: Σ(($\boldsymbol{X_i} - \boldsymbol{\bar{X}}$) * ($\boldsymbol{Y_i} - \boldsymbol{\bar{Y}}$)).
5. Compute the sum of the squared deviations of X: $\sum_{i=1}^{N}(\boldsymbol{X_i} - \boldsymbol{\bar{X}})^2$
6. Estimate the slope (β1) using the formula:

$$\boldsymbol{\hat{\beta}} = \frac{\sum_{i=1}^{N}(\boldsymbol{X_i} - \boldsymbol{\bar{X}})(\boldsymbol{Y_i} - \boldsymbol{\bar{Y}})}{\sum_{i=1}^{N}(\boldsymbol{X_i} - \boldsymbol{\bar{X}})^2}$$

This formula calculates the slope by dividing the sum of the products of the deviations by the sum of the squared deviations of X (Gujarati & Porter, 2009).

So, first compute Means: Calculate the mean of the independent variable (x) and the dependent variable (y).

Mean: The mean of a set of values is the sum of all values in the dataset divided by the total number of values (Kaplan & Saccuzzo, 2013). It is calculated following this formula:

$$\boldsymbol{\bar{X}} = \frac{\sum_{i-1}^{N} \boldsymbol{X_i}}{\boldsymbol{N}} \qquad \boldsymbol{\bar{Y}} = \frac{\sum_{i=1}^{N} \boldsymbol{Y_i}}{\boldsymbol{N}}$$

Second, Compute the deviation, compute the deviation of the value of X from the mean value, and the deviation of Y value from the mean value:

$$\boldsymbol{X_i - \bar{X}, Y_i - \bar{Y}}$$

Variance & Covariance: Calculate the covariance between x and y and the variance of x.

1. **Variance ($\boldsymbol{\sigma^2}$)** is a statistical measure that quantifies the spread or dispersion of a set of data points. It is calculated as the average of the squared differences from the mean. The variance of a set of values is the average of the squared differences between each value and the mean of the dataset (Gonick & Smith, 1993). It is calculated as follow:

$$\boldsymbol{\sigma^2} = \frac{\mathbf{1}}{\boldsymbol{N}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{N}}(\boldsymbol{X_i - \bar{X}})^{\mathbf{2}}$$

2. **Covariance (cov)** Covariance is a statistical measure that quantifies the degree to which two variables change together. It indicates the direction of the linear relationship between two variables. If the covariance is positive, it means that as one variable increases, the other variable tends to increase as well. If the covariance is negative, it means that as one variable increases, the other variable tends to decrease. However, covariance alone does not provide information about the strength of the relationship.measures how much two variables change together. Covariance measures the degree to which two variables vary together. It is calculated as the average of the products of the deviations of each variable from their respective means (Keller & Warrack, 2005). It is calculated following this formula:

$$\boldsymbol{Cov} = \frac{\mathbf{1}}{\boldsymbol{N}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{N}}(\boldsymbol{X_i - \bar{X}})(\boldsymbol{Y_i - \bar{Y}})$$

Interpretation: The slope of the regression line represents the change in the dependent variable for a one-unit change in the independent variable. In other words, it quantifies the average change in regressor variable for each unit change inresponse variable.

Example: If the slope of the regression line is 2, it means that for every one-unit increase in the independent variable, the dependent variable increases, on average, by 2 units. If the slope of the example in the graph (12) is – 1.5 it means one unit **increase** in GDP growth rate, the Unemployment rate **decrease** by 1.5 unit.

Significance: The slope provides insight into the direction and strength of the relationship between the variables. A positive slope indicates a positive correlation (as x increases, y increases), while a negative slope indicates a negative correlation (as x increases, y decreases and vice versa).

b. Estimate Intercept: To estimate the intercept (β_0) in a simple linear regression, you use the method of least squares. The intercept represents the value of the dependent variable (Y) when the independent variable (X) is zero. The formula to estimate the intercept is:

$$\boldsymbol{\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X}}$$

Where:

- ❖ $\boldsymbol{\widehat{\beta}_0}$: is the estimated intercept.
- ❖ $\boldsymbol{\overline{Y}}$ **:** is the mean of the dependent variable Y.
- ❖ $\boldsymbol{\widehat{\beta}_1}$ is the estimated slope.
- ❖ $\boldsymbol{\overline{X}}$: is the mean of the independent variable X.

This formula calculates the intercept by subtracting the product of the estimated slope and the mean of X from the mean of Y (Montgomery, Peck, & Vining, 2012).

Example: If the intercept of the regression line is 5, it means that when x is zero, the predicted value of y is 5.

Significance: The intercept provides information about the baseline level of the dependent variable when the independent variable is absent or has no effect. It may or may not have practical meaning depending on the context of the data.

1-3- Least - Squares Estimation of the Parameters

Simple linear regression is a statistical technique used to model the relationship between two variables: one independent variable (X) and one dependent variable (Y). The goal is to estimate the parameters of the linear equation that best describes the relationship between these variables.

The Ordinary Least Squares (OLS) Method: is the most widely used method for estimating the parameters of a linear regression model. OLS minimizes the sum of the squared differences between the observed and predicted values of the dependent variable. The coefficients are estimated by finding the values that minimize the sum of squared residuals (Montgomery, Peck, & Vining, 2012).

For a simple linear regression, conducting Ordinary Least Squares (OLS) estimation involves the following steps:

1. **Specify the Model**: Define the simple linear regression model. It typically takes the form:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

   - *Y* is the dependent variable.
   - *X* is the independent variable.
   - β_0 is the intercept (the value of *Y* when *X*=0).
   - β1 is the slope (the change in *Y* for a one-unit change in *X*).
   - ε represents the error term (which denotes the deviation of predicted Y value from actual Y value).

2. **Collect Data**: Gather data on the dependent variable (*Y*) and the independent variable (*X*) for each observation in your dataset.
3. **Formulate the hypothesis:** Before estimating the coefficient, we first establish the hypotheses related to the slope coefficient (β_1)

$$\begin{cases} H_0: \beta_1 = 0, here\ is\ a\ significant\ relationship\ betwween\ unemployment\ and\ GDP \\ H_a: \beta_1 \neq 0, here\ is\ a\ significant\ relationship\ betwween\ Unemployment\ and\ GDP \end{cases}$$

4. **Estimate Parameters**: Compute the OLS estimates for β_0 and β_1 using this formula

   $$\boldsymbol{\hat{\beta}} = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{N}(X_i - \bar{X})^2} \quad ,$$

   $$\boldsymbol{\hat{\beta}_0} = \bar{Y} - \hat{\beta}_1 \bar{X}$$

   Or, software such as Excel, EViews, Stata……..
5. **Assess the Model**: Evaluate the goodness of fit of the estimated model. You can compute statistics such as the coefficient of determination (R^2) to assess how well the model fits the data.

6. **Interpret Results**: Interpret the estimated coefficients $\widehat{\boldsymbol{\beta}}_{\mathbf{0}}$and $\widehat{\boldsymbol{\beta}}_1$ in the context of your problem. For example, if $\widehat{\boldsymbol{\beta}}_1 > 0$, it indicates a positive relationship between X and Y, while if $\widehat{\boldsymbol{\beta}}_1 < 0$, it indicates a negative relationship.

7. **Use the Model**: Once you have estimated the model and assessed its validity, you can use it to make predictions about Y for new values of X, or to test hypotheses about the relationship between X and Y.

These steps outline the process of conducting OLS estimation for a simple linear regression model. There are several ways to find the Ordinary least square parameters

i. **The closed-form solution for simple linear regression** involves directly computing the coefficients of the regression equation using formulas derived from the sample data (Doe , 2024).

Example: Suppose you're analyzing the relationship between the Unemployment rate, and GDP growth rate. To find the impact of GDP growth rate on unemployment rate, you have conducted a linear regression analysis using OLS method and the following data.

Table 5: the USA GDP growth rate and Unemployment rate

Time period	GDP growth rate (%)	Unemployment rate (%)
2010	2,1	6,5
2011	2,4	6,2
2012	2,7	5,8
2013	2,9	5,5
2014	2,6	5
2015	2,5	4,5
2016	3	4,9
2017	2,8	4,2
2018	2,2	4,1
2019	2,5	3,9
2020	3,1	4
2021	3,6	3,5
2022	3,2	3
2023	2,8	3,2

Source : https://fred.stlouisfed.org/series/REAINTRATREARAT10Y; https://fred.stlouisfed.org/series/UNRATE

First- model specification According to the example, it is evident that the dependent variable is the unemployment rate (*Unemp*) and the independent variable is the GDP growth rate (*GDP*), the regression equation is as follow:

$\boldsymbol{Unemp = \beta_0 + \beta_1 GDP}$ **..........................(Eq. 3)**

Second- coefficient estimation

- **We estimate** the slope (β1) using the formula:

$$\boldsymbol{\widehat{\beta} = \frac{\sum_{i=1}^{N}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{N}(X_i - \overline{X})^2}}$$

Time period	GDP growth rate (X)	Unemployment rate (%) (Y)	$X_i - \bar{X}$	$Y_i - \bar{Y}$	$(X_i - \bar{X})(Y_i - \bar{Y})$	$(X_i - \bar{X})^2$
2010	2,1	6,5	-0,643	1,907	-1,2260	0,4133
2011	2,4	6,2	-0,343	1,607	-0,5510	0,1186
2012	2,7	5,8	-0,043	1,207	-0,0527	0,0028
2013	2,9	5,5	0,157	0,907	0,14362	0,0257
2014	2,6	5	-0,143	0,407	-0,0586	0,0204
2015	2,5	4,5	-0,243	-0,093	0,0230	0,0590
2016	3	4,9	0,257	0,307	0,0794	0,0661
2017	2,8	4,2	0,057	-0,393	-0,0226	0,0033
2018	2,2	4,1	-0,543	-0,493	0,2683	0,2957
2019	2,5	3,9	-0,243	-0,693	0,1687	0,0590
2020	3,1	4	0,357	-0,593	-0,2127	0,1286
2021	3,6	3,5	0,857	-1,093	-0,9372	0,7357
2022	3,2	3	0,457	-1,593	-0,7286	0,2090
2023	2,8	3,2	0,057	-1,393	-0,0807	0,0033
Sum	35,657	59,707	0	0	-3,1857	2,1343
Mean	2,7428	4,5928	/////	/////	////////////////////	/////////

$$\boldsymbol{\widehat{\beta} = \frac{-3,1857}{2,1343} = -1,4926}$$

- To estimate the intercept (β_0) in a simple linear regression, you use the formula to estimate the intercept is:

$$\boldsymbol{\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1\overline{X}}$$

- ❖ $\overline{\boldsymbol{Y}}$ **:** is the mean of the dependent variable Y, from the table the value of $\overline{\boldsymbol{Y}} = \mathbf{4,5928}$
- ❖ $\widehat{\boldsymbol{\beta}}_{\mathbf{1}}$ is the estimated slope, $\widehat{\boldsymbol{\beta}}_{\mathbf{1}} = \mathbf{-1,4926}$

$\overline{\boldsymbol{X}}$: is the mean of the independent variable X, from the table above the value of $\overline{\boldsymbol{X}}$ **=2,7428**

$$\widehat{\boldsymbol{\beta}}_{\mathbf{0}} = \mathbf{4,5928 - (-1,4926 * 2,7428) = 8,6867}$$

So, the regression equation that represent the impact of GDP on unemployment

$$\boldsymbol{Unemp} = \mathbf{8,6867 - 1,4926}\ \boldsymbol{GDP}$$

Third- Interpretation: According to this regression, one unit increase (decrease) in GDP growth, the Unemployment rate increase (decrease) by 1,6745 unit.

ii. Estimating the Coefficient of Simple Linear Regression (β_0, β_1) using Excel

To estimate the coefficient of simple linear regression using excel software we proceed through the following steps (Moore, McCabe, & Craig's, 2009):

1. **prepare Your Data**:
   - Organize your data into two columns: one for the independent variable (X) and one for the dependent variable (Y).
2. **Open Excel**:
   - Enter your independent variable (X) values in one column and your dependent variable (Y) values in another column.
3. **Open Data Analysis Tool**:
   - Go to the "Data" tab.
   - Click on "Data Analysis" in the Analysis group.
4. **Select Regression**:
   - In the Data Analysis dialog box, scroll down and select "Regression."
   - Click "OK."
5. **Set Up Regression Dialog Box**:
   - In the Regression dialog box, you'll need to specify the input range for your independent variable (X) and the dependent variable (Y).
   - Click on the input range selector button next to "Input Y Range" and select the range containing your dependent variable (Y) data.

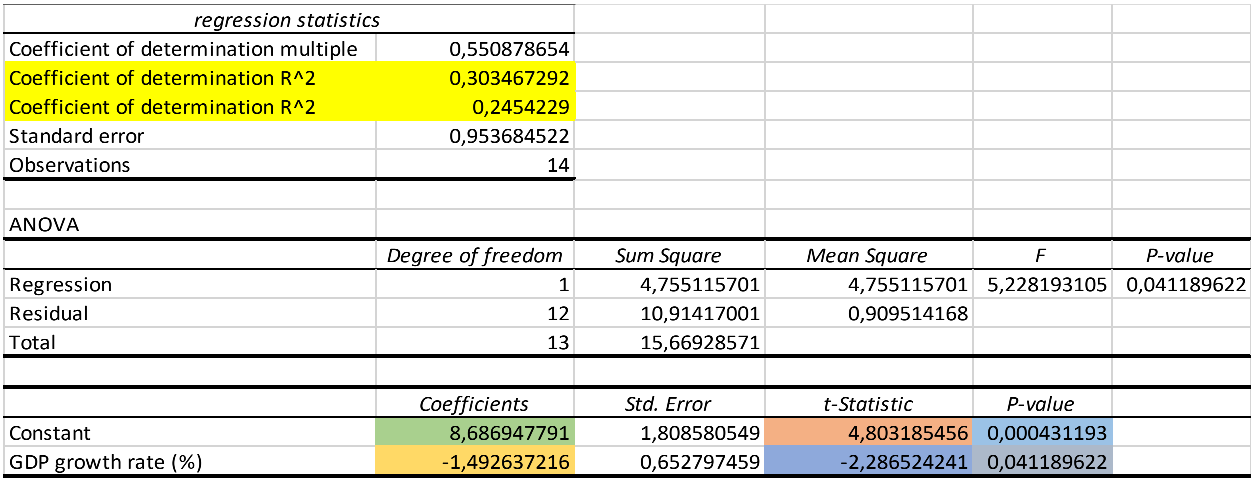
- Click on the input range selector button next to "Input X Range" and select the range containing your independent variable (X) data.
- Optionally, you can choose to output the regression results to a new worksheet or specify a location for the output.
- Check the box next to "Labels" if your data has headers.
- You can also choose to include additional statistics like residuals, confidence level, and more.

6. **Run the Regression**:
   - Click "OK" to run the regression analysis.
7. **Interpret Results**:
   - Excel will generate a new worksheet with the regression output.
   - Review the coefficients to find the slope (β_1) and y-intercept (β_0).

The following table show the simple regression results related to data from the table 5, and the regression equation

$$Unemp = \beta_0 + \beta_1 GDP$$

Table 6: Simple regression equation results



regression statistics					
Coefficient of determination multiple	0,550878654				
Coefficient of determination R^2	0,303467292				
Coefficient of determination R^2	0,2454229				
Standard error	0,953684522				
Observations	14				
ANOVA					
	Degree of freedom	*Sum Square*	*Mean Square*	*F*	*P-value*
Regression	1	4,755115701	4,755115701	5,228193105	0,041189622
Residual	12	10,91417001	0,909514168		
Total	13	15,66928571			
	Coefficients	*Std. Error*	*t-Statistic*	*P-value*	
Constant	8,686947791	1,808580549	4,803185456	0,000431193	
GDP growth rate (%)	-1,492637216	0,652797459	-2,286524241	0,041189622	

Source: Excel Output base on the data from the table 5

Based on Table 6, the intercept (constant) is estimated to be $\hat{\beta}_0$ = 8.6869; This indicates that when all predictor (independent) variables are at zero, the unemployment rate is expected to be 8.6869.

Additionally, the slope is estimated to be $\hat{\beta}_1$= -1.4926, This suggests that for every one unit increase in GDP growth rate, the unemployment rate is expected to decrease by 1.4926 units, and conversely, for every one unit decrease in GDP growth rate, the unemployment rate is expected to increase by 1.4926 units.

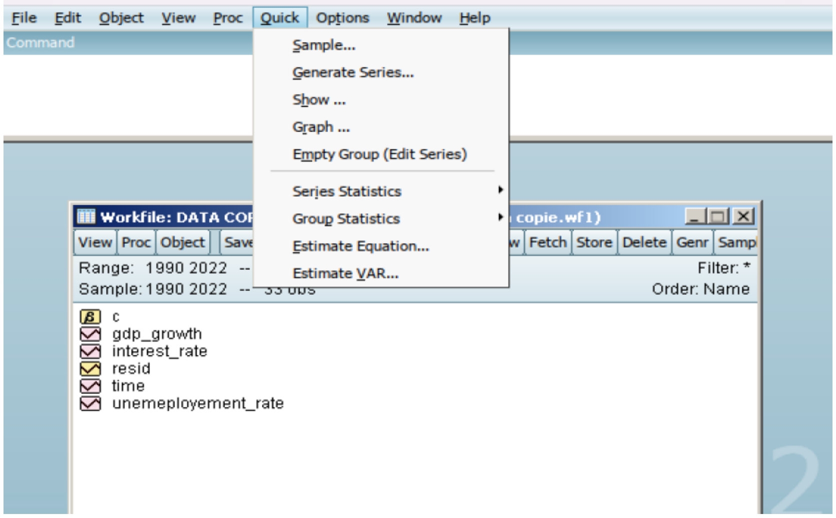
$$\boldsymbol{Unemp = 8,6869 - 1,4926\ GDP}$$

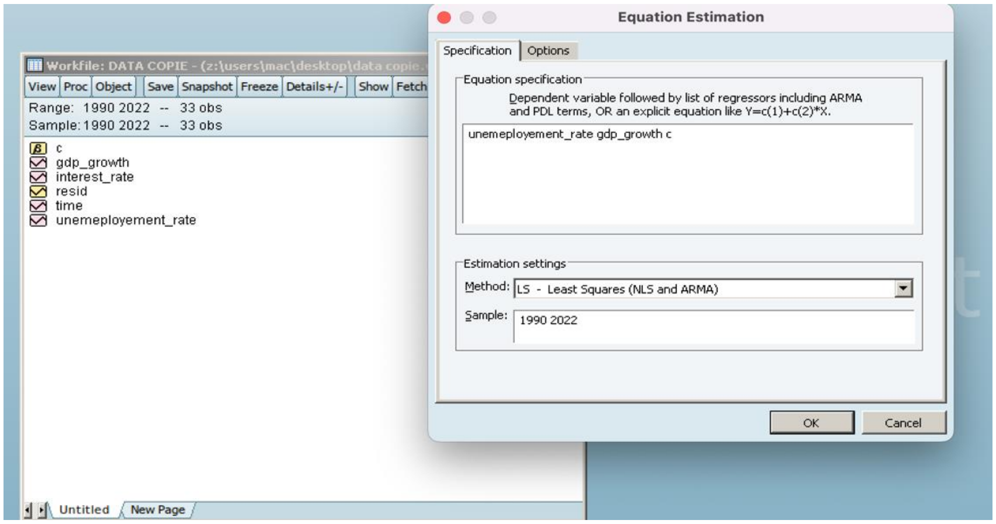
iii. How Estimating the Parameters of Simple Linear Regression (β_0, β_1) using EViews Software

Estimating the Ordinary Least Squares (OLS) method in EViews software typically involves the following steps (EViews, 2020):

1. **Load Data**: Import your dataset into EViews or create a new workfile containing your data.
2. **Specify the Regression Equation**: Define the regression equation by specifying the dependent variable and the independent variable.
3. **Run Regression**: Use the "Quick" menu or the command window to run the regression analysis. You can do this by selecting "Estimate Equation" or by using the appropriate command (e.g., "OLS").
4. **View Results**: After running the regression, EViews will provide output containing information about the estimated coefficients, standard errors, t-statistics, p-values, and other relevant statistics. These step are detailed in the following figures

Figure 13 : step of performing simple linear regression in EViews





Range: 1990 2022 -- 33 obs
Sample: 1990 2022 -- 33 obs

c
gdp_growth
interest_rate
resid
time
unemployement_rate

Equation specification

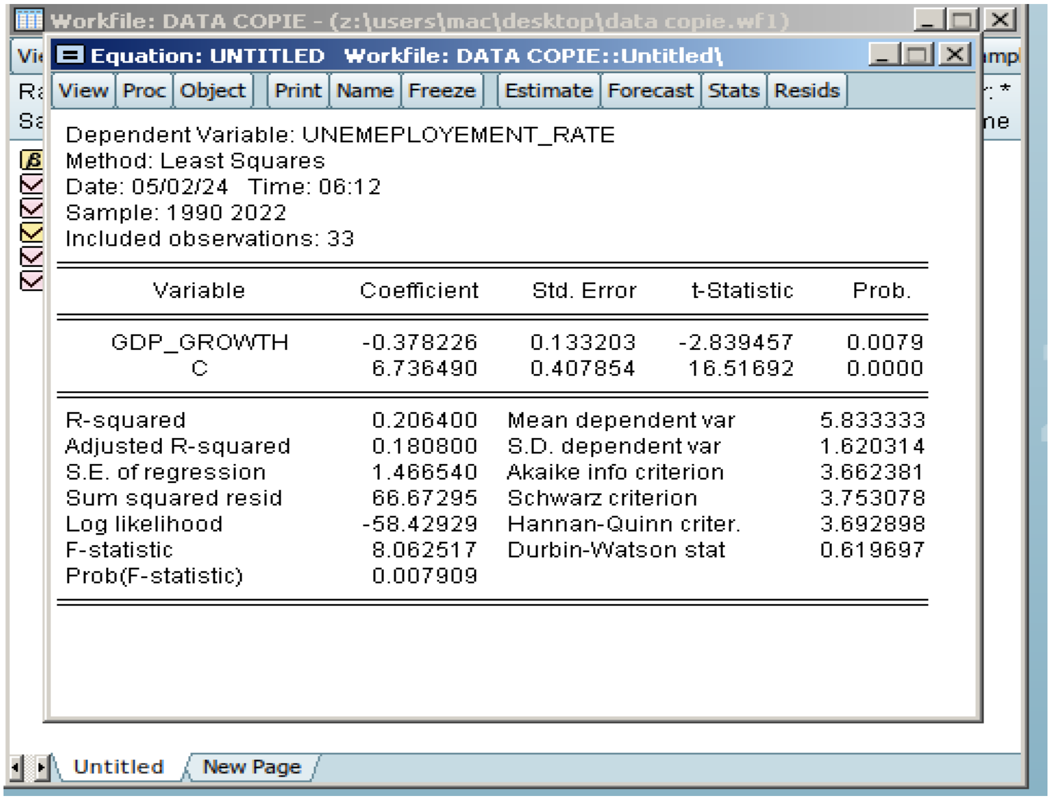
Dependent variable followed by list of regressors including ARMA and PDL terms, OR an explicit equation like Y=c(1)+c(2)*X.

```
unemployement_rate gdp_growth c
```

Estimation settings

Method: LS - Least Squares (NLS and ARMA)

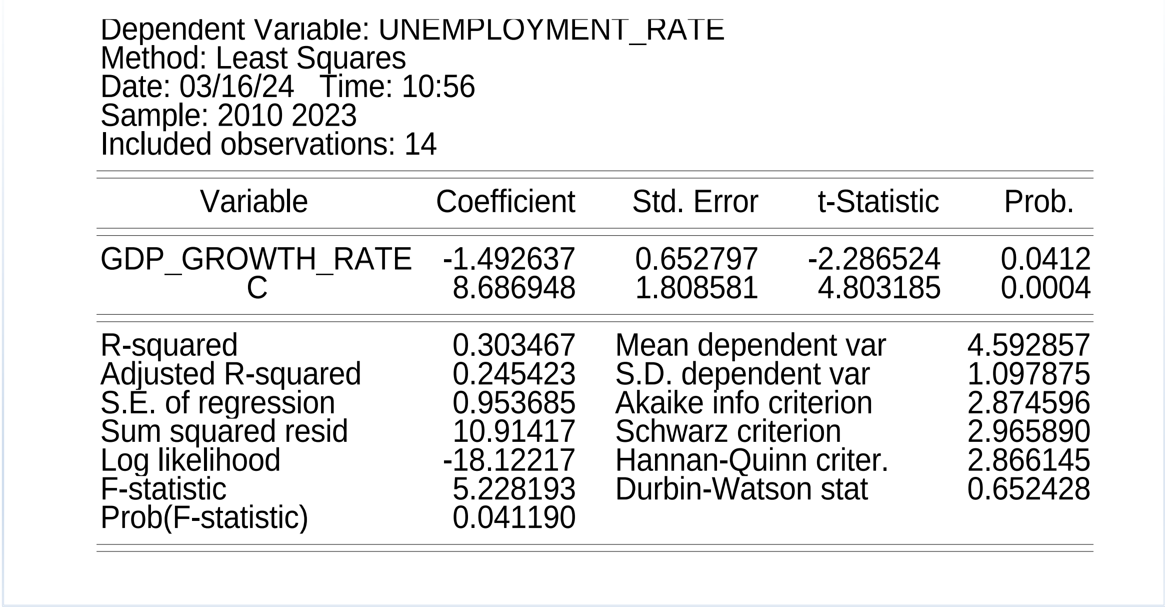
Sample: 1990 2022


Workfile: DATA COPIE - (z:\users\mac\desktop\data copie.wf1)
Equation: UNTITLED Workfile: DATA COPIE::Untitled\
View
Proc
Object
Print
Name
Freeze
Estimate
Forecast
Stats
Resids
Untitled
New Page

```
Dependent Variable: UNEMPLOYEMENT_RATE
Method: Least Squares
Date: 05/02/24   Time: 06:12
Sample: 1990 2022
Included observations: 33

        Variable          Coefficient    Std. Error    t-Statistic    Prob.

     GDP_GROWTH           -0.378226      0.133203      -2.839457     0.0079
          C                6.736490      0.407854      16.51692      0.0000

R-squared                  0.206400    Mean dependent var       5.833333
Adjusted R-squared         0.180800    S.D. dependent var       1.620314
S.E. of regression         1.466540    Akaike info criterion    3.662381
Sum squared resid          66.67295    Schwarz criterion        3.753078
Log likelihood            -58.42929    Hannan-Quinn criter.     3.692898
F-statistic                8.062517    Durbin-Watson stat       0.619697
Prob(F-statistic)          0.007909
```

Source: EViews output

Table 7: Simple regression equation results



```
Dependent Variable: UNEMPLOYMENT_RATE
Method: Least Squares
Date: 03/16/24   Time: 10:56
Sample: 2010 2023
Included observations: 14

      Variable        Coefficient    Std. Error    t-Statistic     Prob.

GDP_GROWTH_RATE       -1.492637      0.652797      -2.286524      0.0412
       C               8.686948      1.808581       4.803185      0.0004

R-squared              0.303467   Mean dependent var       4.592857
Adjusted R-squared     0.245423   S.D. dependent var       1.097875
S.E. of regression     0.953685   Akaike info criterion    2.874596
Sum squared resid      10.91417   Schwarz criterion        2.965890
Log likelihood        -18.12217   Hannan-Quinn criter.     2.866145
F-statistic            5.228193   Durbin-Watson stat       0.652428
Prob(F-statistic)      0.041190
```

Source: EViews 12 Output base on the data from the table 5

The table 7, displays the simple linear regression estimates based on the least square method; the table provide output containing information about the estimated Parameters β_0, β_1 including their values, standard errors, t-statistics, p-values, as well as additional pertinent statistics such as R-squared and F-statistic.

Based on the value from the table 7, we can formulate the simple linear regression equation for the model depicting the relationship between the unemployment rate and GDP growth rate as follows:

$$\boldsymbol{Unemp = 8,6869 - 1,4926\ GDP}\text{.................}\textbf{\textit{(Eq. 4)}}$$

$$\textbf{(1.8085)} \quad \textbf{(0.6527)}$$

The value of slope β_1 is -1.4926 and its standard error is 0.6527; the value of intercept β_0 is 8.6869 and its standard deviation is 1.8085.

iv. Hypothesis testing on the Slope and intercept

The test procedure for $H_0: \beta_1 = 0$ may be developed from two approaches. The first approach simply makes use of the t statistic, the second is use the p-value approach.

a. Use of t-Tests

Suppose that we wish to test the hypothesis that the slope equals a constant, say β_1. Using t-tests to test hypotheses is a common statistical technique, especially in the context of hypothesis testing in regression analysis. Here's a step-by-step guide on how to use t-tests to test hypotheses:

1) **The appropriate hypotheses are** :

$$\begin{cases} H_0\colon \beta_1 = 0, here\ is\ a\ significant\ relationship\ betwween\ unemployment\ and\ GDP \\ H_a\colon \beta_1 \neq 0, here\ is\ a\ significant\ relationship\ betwween\ Unemployment\ and\ GDP \end{cases}$$

2) **Determine the Critical Value:** Determine the critical value for the t-test based on the desired level of significance (α, for example 5%) and degrees of freedom (df). This critical value corresponds to the cutoff point beyond which you would reject the null hypothesis. The degree of freedom is equal to the number of observations (n) minus the number of parameters (2), *df = n-2*

3) **Calculate the Test Statistic:** Calculate the test statistic for each coefficient in regression model using the formula:

$$t - Statistic = \frac{Coefficient}{Standard\ error}$$

4) **Compare T-Statistic to Critical Value**: Compare the calculated test statistic to the critical value from the t-distribution table.

If the absolute value of the t-statistic is greater than the critical value, means reject the null hypothesis in favor of the alternative hypothesis.

If the absolute value of the t-statistic is less than the critical value, means fail to reject (accept) the null hypothesis in favor of the alternative hypothesis.

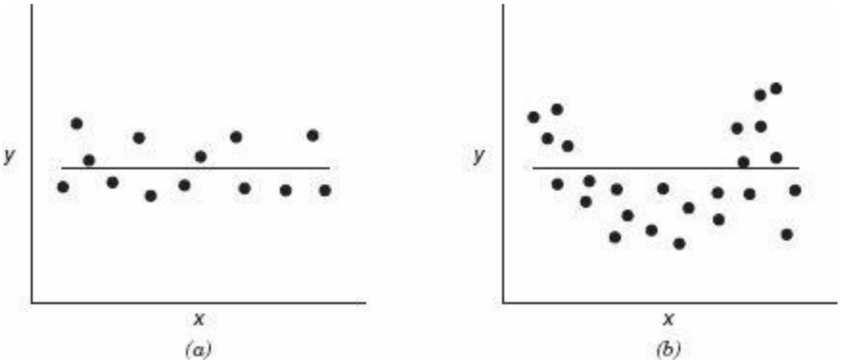
5) **Interpret Results**: If you reject the null hypothesis, it suggests that there is sufficient evidence to conclude that the corresponding coefficient is statistically significant. Conversely, if you fail to reject the null hypothesis, it indicates that there is not enough evidence to support the significance of that coefficient.

6) **Make Inferences:** Based on the results of the t-tests, make inferences about the significance of the coefficients in your regression model. These inferences can help you understand the relationships between the independent variables and the dependent variable.

These hypotheses relate to the significance of regression. Failing to reject $H_0\colon \beta_1 = 0$ implies that there is no linear relationship between x and y, this is illustrated in the figure 14.

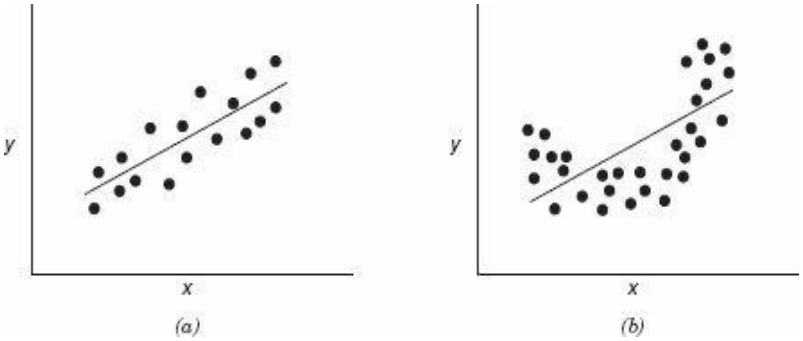
Alternatively, if $H_0\colon \beta_1 = 0$ is rejected, this implies that x is of value in explaining the variability in y. This is illustrated in figure 15.

Figure 14: Scatter plot when the hypothesis $H_0\colon \beta_1 = 0$ is accepted



Source: (Montgomery, Peck, & Vining, 2012, p. 62) https://ocd.lcwu.edu.pk/cfiles/Statistics/Stat-503/IntroductiontoLinearRegressionAnalysisbyDouglasC.MontgomeryElizabethA.PeckG.GeoffreyViningz-lib.org.pdf

Figure 15: Scatter plot when the hypothesis $H_0\colon \beta_1 = 0$ is accepted



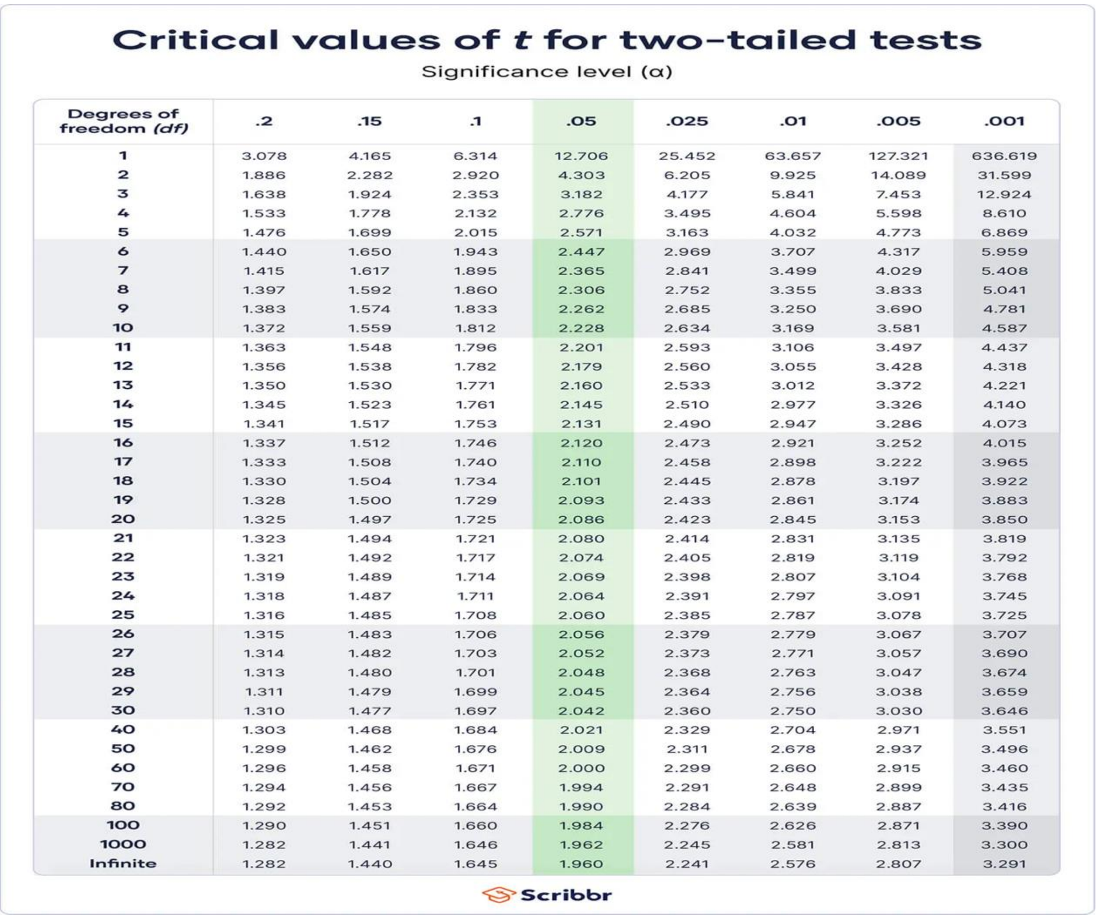
Source: (Montgomery, Peck, & Vining, 2012, p. 63) https://ocd.lcwu.edu.pk/cfiles/Statistics/Stat-503/IntroductiontoLinearRegressionAnalysisbyDouglasC.MontgomeryElizabethA.PeckG.GeoffreyViningz-lib.org.pdf

We test for significance of regression in the Unemployment and GDP growth rate regression model of Example above (Eq. 4). The estimate of the slope is $\hat{\beta}_1$= -1.4926 and the standard error of the slope is 0.6527 Therefore, the t-Statistic calculated is : $t - Stat_{cal} = \frac{-1.4926}{0.6527} =$ - 2.2868 If we choose α = 0.05, and based on our example the degree of freedom is n-2 =14-2 = 12; the critical value of tstatistic critical (value of t-statistic from the t-distribution table) is $t_{crit} = 2.179$. Thus, the absolute value of t-statistic calculated is greater than the t-statistic critical, we would reject

$H_0 : \beta_1 = 0$ and conclude that there is a linear relationship between Unemployment rate and GDP growth rate.

The Excel output and EViews output in table 6 &7 respectively give the standard errors of the slope and intercept along with the t-statistic for testing H_0: $\beta_1 = 0$. Notice that the results shown in this table for the slope essentially agree with the manual calculations.

Table 8: Student's *t* table



Critical values of *t* for two-tailed tests

Significance level (α)

Degrees of freedom *(df)*	.2	.15	.1	.05	.025	.01	.005	.001
1	3.078	4.165	6.314	12.706	25.452	63.657	127.321	636.619
2	1.886	2.282	2.920	4.303	6.205	9.925	14.089	31.599
3	1.638	1.924	2.353	3.182	4.177	5.841	7.453	12.924
4	1.533	1.778	2.132	2.776	3.495	4.604	5.598	8.610
5	1.476	1.699	2.015	2.571	3.163	4.032	4.773	6.869
6	1.440	1.650	1.943	2.447	2.969	3.707	4.317	5.959
7	1.415	1.617	1.895	2.365	2.841	3.499	4.029	5.408
8	1.397	1.592	1.860	2.306	2.752	3.355	3.833	5.041
9	1.383	1.574	1.833	2.262	2.685	3.250	3.690	4.781
10	1.372	1.559	1.812	2.228	2.634	3.169	3.581	4.587
11	1.363	1.548	1.796	2.201	2.593	3.106	3.497	4.437
12	1.356	1.538	1.782	2.179	2.560	3.055	3.428	4.318
13	1.350	1.530	1.771	2.160	2.533	3.012	3.372	4.221
14	1.345	1.523	1.761	2.145	2.510	2.977	3.326	4.140
15	1.341	1.517	1.753	2.131	2.490	2.947	3.286	4.073
16	1.337	1.512	1.746	2.120	2.473	2.921	3.252	4.015
17	1.333	1.508	1.740	2.110	2.458	2.898	3.222	3.965
18	1.330	1.504	1.734	2.101	2.445	2.878	3.197	3.922
19	1.328	1.500	1.729	2.093	2.433	2.861	3.174	3.883
20	1.325	1.497	1.725	2.086	2.423	2.845	3.153	3.850
21	1.323	1.494	1.721	2.080	2.414	2.831	3.135	3.819
22	1.321	1.492	1.717	2.074	2.405	2.819	3.119	3.792
23	1.319	1.489	1.714	2.069	2.398	2.807	3.104	3.768
24	1.318	1.487	1.711	2.064	2.391	2.797	3.091	3.745
25	1.316	1.485	1.708	2.060	2.385	2.787	3.078	3.725
26	1.315	1.483	1.706	2.056	2.379	2.779	3.067	3.707
27	1.314	1.482	1.703	2.052	2.373	2.771	3.057	3.690
28	1.313	1.480	1.701	2.048	2.368	2.763	3.047	3.674
29	1.311	1.479	1.699	2.045	2.364	2.756	3.038	3.659
30	1.310	1.477	1.697	2.042	2.360	2.750	3.030	3.646
40	1.303	1.468	1.684	2.021	2.329	2.704	2.971	3.551
50	1.299	1.462	1.676	2.009	2.311	2.678	2.937	3.496
60	1.296	1.458	1.671	2.000	2.299	2.660	2.915	3.460
70	1.294	1.456	1.667	1.994	2.291	2.648	2.899	3.435
80	1.292	1.453	1.664	1.990	2.284	2.639	2.887	3.416
100	1.290	1.451	1.660	1.984	2.276	2.626	2.871	3.390
1000	1.282	1.441	1.646	1.962	2.245	2.581	2.813	3.300
Infinite	1.282	1.440	1.645	1.960	2.241	2.576	2.807	3.291

Scribbr

Source : (Shaun Turney, 2022) https://www.scribbr.com/statistics/students-t-table/

b. Use of P-value

The significance level (α) is the threshold for the p-value to reject the null hypothesis. Commonly used values are 0.1, 0.05, 0.01, etc. The p-value is the probability of obtaining a test statistic as extreme as, or more extreme than, the one observed in the sample, assuming that the null hypothesis is true.

- If the p-value is less than the significance level (α), reject the null hypothesis.
- If the p-value is greater than or equal to the significance level (α), fail to reject the null hypothesis.

Example: let's continue with our example of Unemployment and GDP growth rate regression result

Assume that we choose a significance level of 0.05. After conducting the study and analyzing the data, the p-value for the test for significance of parameter $\beta_1$is reported as P-value = 0.0412 Since the p-value (0.0412) is less than the significance level (0.05), we reject the null hypothesis. Therefore, we conclude that there is evidence to support the claim that the GDP growth rate has a statistically significant impact on Unemployment rate

v. Estimating the Deviation – Error

In simple linear regression, the deviation or error is typically estimated using the residuals. Residuals are the differences between the observed values of the dependent variable (Y) and the values predicted (fitted value) by the regression line ($\hat{Y}$). Here's how you can estimate the deviation or error:

a. **Fit the Regression Line**: First, you fit the regression line to your data using the least squares method. The equation of the regression line is typically represented as :

$$\hat{Y} = \beta_0 + \beta_1 X + \varepsilon \text{ --------------- } \varepsilon = \hat{Y} - (\beta_0 + \beta_1 X)$$

b. **Calculate Residuals**: For each data point (x_i,y_i), calculate the residual (ε_i) as the difference between the observed value of the dependent variable (y_i) and the predicted value ($\hat{y}_i$)from the regression line.

❖ **Manual calculation:** we can calculate the deviation or the residual as follow:

Let's continue with our example of Unemployment and GDP growth rate regression

$$\boldsymbol{Unemp = 8,6869 - 1,4926\ GDP}$$

and the data from the table 5

Table 9: Table of residual

Time period	GDP growth rate (X)	Unemployment rate (%) (Y) --- observed value	Predicted value ($\hat{Y}$)	Deviation Residual ($Y - \hat{Y}$)
2010	2,1	6,5	5,55244	**0,94756**
2011	2,4	6,2	5,10466	**1,09534**
2012	2,7	5,8	4,65688	**1,14312**
2013	2,9	5,5	4,35836	**1,14164**
2014	2,6	5	4,80614	**0,19386**
2015	2,5	4,5	4,9554	**-0,4554**
2016	3	4,9	4,2091	**0,6909**
2017	2,8	4,2	4,50762	**-0,30762**
2018	2,2	4,1	5,40318	**-1,30318**
2019	2,5	3,9	4,9554	**-1,0554**
2020	3,1	4	4,05984	**-0,05984**
2021	3,6	3,5	3,31354	**0,18646**
2022	3,2	3	3,91058	**-0,91058**
2023	2,8	3,2	4,50762	**-1,30762**

Source: Author' elaboration based on the data from table 5 and the regression equation (Eq. 4)

$\hat{Y}_1 = 8{,}6869 - 1{,}4926\ (2{,}1)$ = *5,55244*. Then we continue in the same way for all value in the table. After we got all the Ŷ, we proceed to calculate the deviation

$\hat{\varepsilon}_1 = \boldsymbol{Y_1} - \boldsymbol{\hat{Y}_1}$; *6,5-5,55244=0,94756*

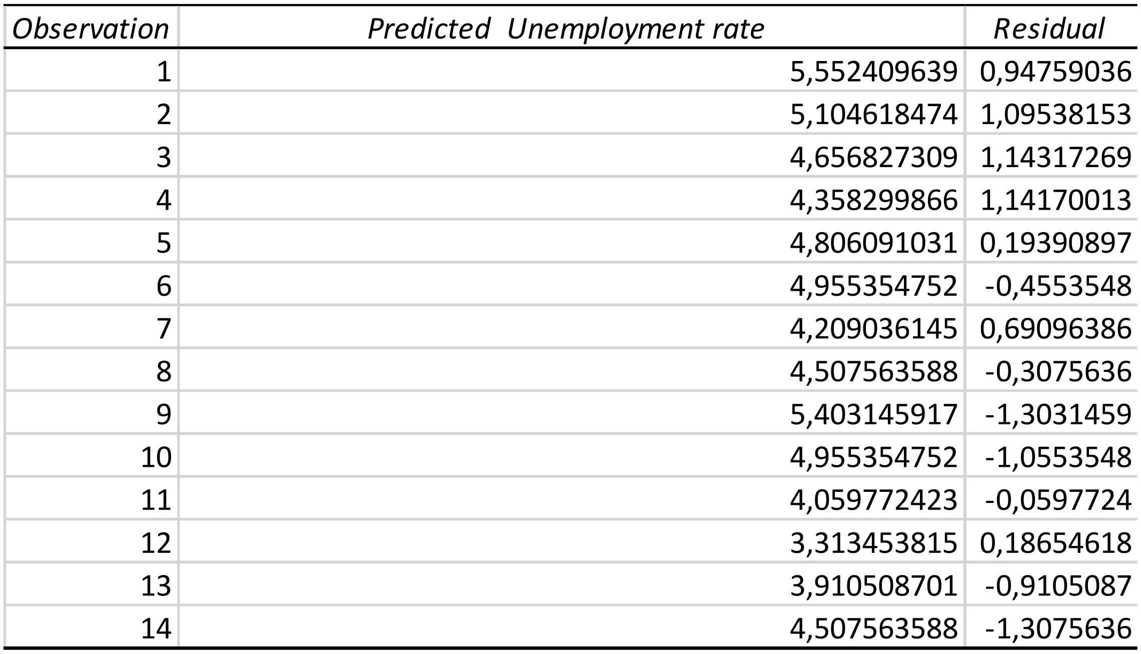
❖ **Calculating residuals in simple linear regression using Excel** involves a straightforward process. Here is the steps to follow:

- Open Excel and input the dataset. For simple linear regression, you typically have two columns: one for the independent variable (X) and one for the dependent variable (Y).
- Go to the "Data" tab and click on "Data Analysis"
- Choose "Regression" from the list of analysis tools and click "OK."
- In the Regression dialog box, input the "Input Y Range" (dependent variable) and "Input X Range" (independent variable).
- Check the box for "Residuals" and click "OK" to run the regression analysis.
- Once the regression analysis is complete, Excel will output the results in a new worksheet.

- You'll find the residuals listed in a column alongside other regression statistics such as the coefficients, standard error, t-statistic, and p-value.

Example: Calculate the residuals for a regression model predicting the unemployment rate based on GDP growth in Excel software.

Table 10: Residual analysis



Observation	*Predicted Unemployment rate*	*Residual*
1	5,552409639	0,94759036
2	5,104618474	1,09538153
3	4,656827309	1,14317269
4	4,358299866	1,14170013
5	4,806091031	0,19390897
6	4,955354752	-0,4553548
7	4,209036145	0,69096386
8	4,507563588	-0,3075636
9	5,403145917	-1,3031459
10	4,955354752	-1,0553548
11	4,059772423	-0,0597724
12	3,313453815	0,18654618
13	3,910508701	-0,9105087
14	4,507563588	-1,3075636

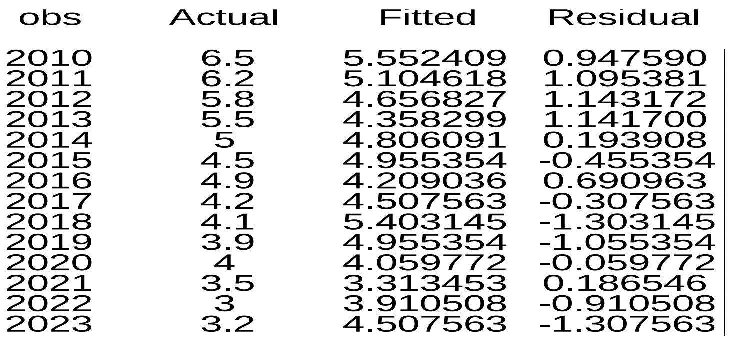
Source: Excel output based on data from table 5

❖ **Calculating residuals in simple linear regression using EViews software** involves a straightforward process. Here is the steps to follow:

- Open EViews and load your dataset containing the variables for the simple linear regression analysis.
- Go to "Quick" > "Estimate Equation" from the main menu.
- In the "Specification" tab of the Equation Estimation dialog box, specify your dependent variable (Y) and independent variable (X).
- Choose the appropriate estimation method (usually Least Squares) and click "OK" to run the regression analysis.
- After EViews completes the regression analysis, it will display the regression results in a new window.

- In the regression results window, you can find various statistics including the coefficients, standard errors, t-statistics, p-values, and most importantly, the residuals.
- To view the residuals specifically, locate the "Residuals" tab in the regression results window.
- The residuals will be listed alongside the corresponding observations or data points.
- **Example:** Calculate the residuals for a regression model predicting the unemployment rate based on GDP growth in EViews Software.

Table 11: Residual table



obs	Actual	Fitted	Residual
2010	6.5	5.552409	0.947590
2011	6.2	5.104618	1.095381
2012	5.8	4.656827	1.143172
2013	5.5	4.358299	1.141700
2014	5	4.806091	0.193908
2015	4.5	4.955354	-0.455354
2016	4.9	4.209036	0.690963
2017	4.2	4.507563	-0.307563
2018	4.1	5.403145	-1.303145
2019	3.9	4.955354	-1.055354
2020	4	4.059772	-0.059772
2021	3.5	3.313453	0.186546
2022	3	3.910508	-0.910508
2023	3.2	4.507563	-1.307563

Source: EViews output based on data from table 5

1-4- Model validity

Model Validity and Explanatory power in simple linear regression can be measured by several metrics. The most commonly used metric is the coefficient of determination (R^2), Adjusted R^2, Sum Squared Error (SSE**),** F-statistic and Correlation Coefficient (r). These metrics provide different perspectives on the explanatory power of the model and are often used together to gain a comprehensive understanding of how well the model fits the data and explains the relationship between the variables (Chatterjee. & Hadi, 2012).

i. **Sum Squared Error (SSR):** SSR measures the average deviation of the observed values from the regression line. A lower SEE indicates a better fit of the model to the data.

The significance of regression can be tested through an analysis-of-variance (ANOVA) approach. This technique involves breaking down the total variability in the response variable to make inferences about the significance of the predictor variable. To achieve this breakdown, we start with this equation (Montgomery, Peck, & Vining, 2012):

$y_i - \bar{y} = (y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$ ……………………………………………Eq. 3.1

Squaring both sides of Eq. 3.1 and summing over all n observations produces

$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + 2\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ …………………*Eq.3.2*

Note that the second term on the right-hand side of the expression Eq. 3.2 can be rewritten as:

$$2\sum_{i=1}^{n}(\hat{y}_i - \bar{y})(y_i - \hat{y}_i) = 2\sum_{i=1}^{n}\hat{y}_i(y_i - \hat{y}_i) - 2\bar{y}\sum_{i=1}^{n}(y_i - \hat{y}_i)$$

$$= 2\sum_{i=1}^{n}\hat{y}_i\,\varepsilon_i - 2\bar{y}\sum_{i=1}^{n}\varepsilon_i = 0$$

Since the sum of the residuals is always zero and the sum of the residuals weighted by the corresponding fitted value $\hat{y}_i$is also zero. Therefore, the Eq. 3.2 will be

$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ ………………………………………………*Eq. 3.3*

The left-hand side of the Eq.3.3 is the corrected sum of squares of the observations, SST, which measures the total variability in the observations. The two components of SST measure, respectively, the amount of variability in the observations y_i accounted for by the regression line (SSR) and the residual variation left unexplained by the regression line (SSE). We recognize $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ as the residual or Error Sum of Squares (SSE) from the Eq.3.3. It is customary to call $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$the Regression Sum of Squares (SSR). The Eq.3.3. is the fundamental analysis-of-variance identity for a regression model. Symbolically, we usually write it as follow :

$SST = SSE + SSR$ ……………………………………………………………Eq.3.4

Example: use the GDP growth rate and Unemployment rate data

Table12: analysis of variance

Time period	GDP growth rate (X)	Unemployment rate (%) (Y) --- observed value	Predicted value ($\hat{Y}$)	$y_i - \hat{y}_i$	$(y_i - \hat{y}_i)^2$	$\hat{y}_i - \bar{y}$	$(\hat{y}_i - \bar{y})^2$
2010	2,1	6,5	5,55244	0,94756	0,897869	0,9595828	0,9207992
2011	2,4	6,2	5,10466	1,09534	1,199769	0,5118028	0,2619421
2012	2,7	5,8	4,65688	1,14312	1,306723	0,0640228	0,0040989
2013	2,9	5,5	4,35836	1,14164	1,303341	-0,234497	0,0549889
2014	2,6	5	4,80614	0,19386	0,03758	0,2132828	0,0454895
2015	2,5	4,5	4,9554	-0,4554	0,207389	0,3625428	0,1314373
2016	3	4,9	4,2091	0,6909	0,477342	-0,383757	0,1472695
2017	2,8	4,2	4,50762	-0,30762	0,094630	-0,085237	0,0072653
2018	2,2	4,1	5,40318	-1,30318	1,698278	0,8103228	0,6566231
2019	2,5	3,9	4,9554	-1,0554	1,113869	0,3625428	0,1314373
2020	3,1	4	4,05984	-0,05984	0,0035808	-0,533017	0,2841072
2021	3,6	3,5	3,31354	0,18646	0,034767	-1,279317	1,6366523
2022	3,2	3	3,91058	-0,91058	0,8291559	-0,682277	0,4655021
2023	2,8	3,2	4,50762	-1,30762	1,709870	-0,085237	0,0072653
$\bar{y}$		4,5928		SSE	10,91417	SSR	4,7548786

Source : Author' elaboration based on the data from table 5

$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 10{,}91417$

$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = 4{,}7548786$

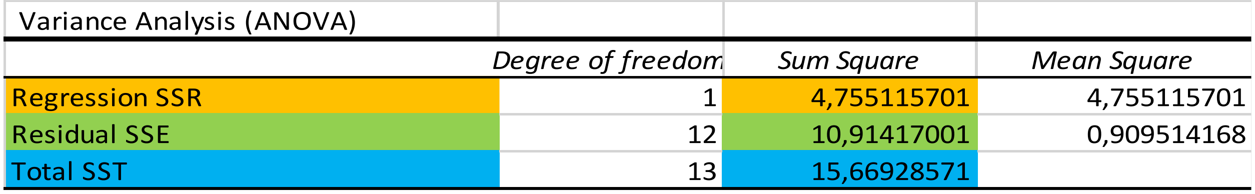
$SST = SSR + SSE = 15{,}6690487$

According to the results we conclude that the 4,7548 of the variation in Unemployment rate is explained by GDP growth rate; and 10,91417 of the variation in Unemployment rate is left unexplained by GDP growth rate. This calculation is made by hand; in the following how, we calculate SSR, SST and SSE using Excel:

- In the Data Analysis dialog box, scroll down and select "Regression."Click "OK."
- In the Regression dialog box, you'll need to specify the input range for your independent variable (X) and the dependent variable (Y).
- Click on the input range selector button next to "Input Y Range" and select the range containing your dependent variable (Y) data.
- Click on the input range selector button next to "Input X Range" and select the range containing your independent variable (X) data.

- Optionally, you can choose to output the regression results to a new worksheet or specify a location for the output.
- Check the box next to "Labels" if your data has headers.
- You can also choose to include additional statistics like residuals, confidence level, and more.
- Click "OK" to run the regression analysis.
- Excel will generate a new worksheet with the regression output, and the Variance analysis results (as shown in the following table- table 13)

Table 13: Variance Analysis



Variance Analysis (ANOVA)			
	Degree of freedom	*Sum Square*	*Mean Square*
Regression SSR	1	4,755115701	4,755115701
Residual SSE	12	10,91417001	0,909514168
Total SST	13	15,66928571	

Source: Excel output based on the data from table 5

We can also, find the value of SSE from the results of table 7, Sum squared resid. =10.91417

ii. Coefficient of determination R²

The coefficient of determination, often denoted as R^2 (R-squared), is a statistical measure that quantifies the proportion of the variance in the dependent variable (Y) that is explained by the independent variable(s) (X). In the context of linear regression, it represents the goodness-of-fit of the regression model. Here's a concise definition of the coefficient of determination:

a. **Definition: The coefficient of determination**, ***R^2***, is a statistical measure indicating the proportion of the variance in the dependent variable that is predictable or explained by the independent variable(s) in a regression model. It ranges from 0 to 1, where 0 indicates that the model does not explain any of the variability in the dependent variable, and 1 indicates that the model perfectly explains all of the variability." (Montgomery, Peck, & Vining, 2012) Here's how it's calculated and interpreted:

b. **Calculation of R^2:**

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \text{ ....(I)}$$

Where: SSE is the sum of squared residuals (differences between observed and predicted values).

- SST is the total sum of squares, representing the total variability in the dependent variable.
- SSR is the sum squared regression

c. Interpretation of R^2:

- R^2 ranges from 0 to 1.
- R^2=0 indicates that the independent variable does not explain any of the variability in the dependent variable.
- R^2=1 indicates that the independent variable explains all of the variability in the dependent variable.
- A higher R^2 value suggests that the model provides a better fit to the data, as it explains a larger proportion of the variance.

Example: from the linear regression result from excel (table 6) and OLS result from the EViews 12 software (table 7), the R^2 is equal 0.3034 (30.34%) this means that the independent variable GDP growth rate has a moderate explanatory power; this result indicates that 30.34% of the variation in Unemployment rate is explained by the GDP growth rate.

We can calculate the R^2 using the formula (I) and the results (SSR, SSE and SST) from the (section 3-1): $R^2 = \frac{4,7551}{15,6692} = 1 - \frac{10,9141}{15,6692}$ = 0,3034 = 30,34%

- ❖ **Adjusted R^2**: Adjusted R^2 adjusts for the number of predictors in the model and is often preferred when comparing models with different numbers of predictorsThe adjusted R-squared value typically ranges between 0 and 1, where a higher value indicates a better fit of the model to the data. Adjusted R-squared takes into account the number of predictors in the model, penalizing excessive complexity that might lead to overfitting. (Gujarati & Porter, 2009). The formula to calculate adjusted R^2 is (Gujarati & Porter, 2009):

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

Where:

- $R2$ is the coefficient of determination
- n is the sample size (number of observations)
- k is the number of predictors (in simple linear regression, k=1)

Example:

Let's say $R2$=0.3034 and the sample size (n) is : 14

Adjusted R^2=$1-\frac{14-1}{14-1-1}(1-0{,}3034)=0{,}2453$

So, the adjusted R^2 is 0.2453.

iii.Correlation coefficient

a. **Definition:** The correlation coefficient (often denoted as r) measures the strength and direction of the linear relationship between the independent variable (X) and the dependent variable (Y). It quantifies the degree to which changes in one variable are associated with changes in the other variable (Moore & McCabe, 2017). The correlation coefficient ranges between -1 and 1, where (Moore & McCabe, 2017):

- r=1: Perfect positive correlation. As X increases, Y increases proportionally.
- r=−1: Perfect negative correlation. As X increases, Y decreases proportionally.
- r=0: No linear correlation. There is no predictable linear relationship between X and Y.
- The closer r is to 1 or -1, the stronger the linear relationship between X and Y.
- The sign of r (+ or -) indicates the direction of the relationship: positive if r is positive, and negative if r is negative.

i. **Calculation of Correlation Coefficient (*r*)** (Moore & McCabe, 2017)**:**

The correlation coefficient r is computed using the following formula:

$$r=\frac{n(\sum XY)-(\sum X)(\sum Y)}{\sqrt{(n\sum X^2-(\sum X)^2)(n\sum Y^2-(\sum Y)^2)}}$$

Where:

- X and Y are individual data points.
- and $\bar{X},\bar{Y}$ are the means of X and Y, respectively.

In summary, the correlation coefficient provides valuable insight into the strength and direction of the linear relationship between variables in simple linear regression. the correlation coefficient can also provide insight into the explanatory power of the model, especially in simple linear regression where there is only one independent variable.

Example: let’s use the data of table 5, using the formula above we find that the correlation coefficient is: - 0,5508787, This result suggests a negative correlation between the GDP growth rate and the unemployment rate. Given that the coefficient is considerably distant from 1, we infer that the two variables are not strongly correlated.

The correlation coefficient can also be computed using Excel software by following these steps:

1. **Organize the Data**: Make sure the data is arranged in two columns, with one variable in each column. For example, you might have GDP growth rate in column A and unemployment rate in column B.
2. **Enter the CORREL Function**: In an empty cell where you want the correlation coefficient to appear, type "=CORREL(". Then, select the range of cells containing the data for the first variable (e.g., GDP growth rate), and then select the range of cells containing the data for the second variable (e.g., unemployment rate). Your formula should look something like this:

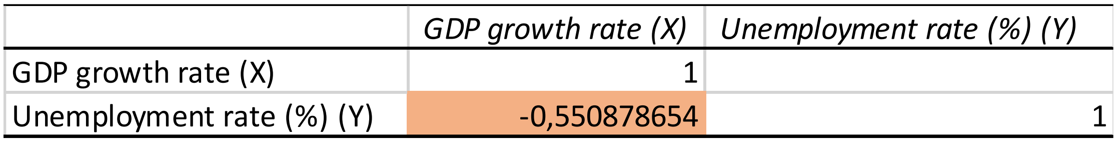
=CORREL(A2:A14, B2:B14   A2: A14 mean the observation from cell A2 till A14

3. **Press Enter**: Once the formula is entered, press Enter. Excel will calculate the correlation coefficient.

We can also follow these steps to calculate the correlation coefficient using Excel software:

- In the Data Analysis dialog box, scroll down and select "Correlation analysis"Click "OK."
- In the Correlation coefficient dialog box, you'll need to specify the input range for your independent variables.
- Optionally, you can choose to output the regression results to a new worksheet or specify a location for the output.
- Check the box next to "Labels" if your data has headers.
- Click "OK" to run the regression analysis.
- Excel will generate a new worksheet with the correlation coefficient output (as shown in the following table- table 14)

Table 14 : the correlation coefficient



	GDP growth rate (X)	*Unemployment rate (%) (Y)*
GDP growth rate (X)	1	
Unemployment rate (%) (Y)	-0,550878654	1

Source: Excel output, based on the data from table 5

iv. Test of overall significance of the model

a. **Definition: F-statistic** is a statistical measure used in regression analysis to assess the overall significance of the regression model. It compares the variability explained by the regression model to the variability not explained by the model. In other words, it evaluates whether the regression model as a whole provides a better fit to the data compared to a model with no independent variables (Montgomery, Peck, & Vining, 2012).

In simple linear regression, the test of overall significance (F-test), assesses whether the linear relationship between the independent variable (X) and the dependent variable (Y) is statistically significant. This test evaluates whether the regression model as a whole explains a significant amount of variance in the dependent variable.

b. F-test Hypotheses :

- **Null Hypothesis (**H_0): The null hypothesis states that there is no linear relationship between the independent and dependent variables (β_1=0).
- **Alternate Hypothesis (**H_a): The alternate hypothesis asserts that there is a linear relationship between the independent and dependent variables (β1≠0).
- **Test Statistic**: The F-statistic is calculated using the ratio of two variances:

The formula for the F-statistic in the context of regression analysis can be expressed as:

$$F - statistic = \frac{Explained\ variability \Big/ Number\ of\ predictors\ (k)}{Unexplained\ variablity \Big/ degree\ of\ freedom\ (n-k-1)}$$

Where*: k is the number of predictor or independent variable (in simple linear regression k=1)*

n: total number of observations.

$$F - statistic = \frac{SSR}{\frac{SSE}{n-2}} = \frac{\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2}{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{n-2}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\text{(II)}$$

Explained Variability (*SSR*): The variance explained by the regression model.

Unexplained Variability (*SSE*): The variance not explained by the regression model, also known as the residual variance.

c. Decision Rule: Compare the calculated F-statistic to the critical F-value from the F-distribution at a chosen significance level (e.g., 0.05).

- If the calculated F-statistic is greater than the critical F-value, reject the null hypothesis and conclude that the model is statistically significant.
- If the calculated F-statistic is less than the critical F-value, fail to reject the null hypothesis and conclude that the model is not statistically significant.

- ❖ We can also use the p-value associated with F-statistic, if the p-value is less than 5% (chosen significant level) than we reject the null hypothesis and conclude that the model is statistically significant.
- ❖ If the p-value associated with F-statistic is greater than 5% (chosen significant level) than we fail to reject the null hypothesis and conclude that the model is not statistically significant.

d. Interpretation: If the null hypothesis is rejected, it indicates that the regression model as a whole is statistically significant, and there is evidence of a linear relationship between the independent and dependent variables.

Example: let's continue with the example of GDP growth rate and Unemployment rate regression, we have found previously that: *SSR= 4,7551 , SSE= 10,9141, n=14, k=1*

So, we can use the formula *(II) then F-statistic*= $\frac{4{,}7551/_{1}}{10{,}9141/_{12}}$= *5,2282*

According to the F-table of Critical Values for Significance Level = 0.05 (see https://statisticsbyjim.com/hypothesis-testing/f-table/), the critical F-value is 4.75; the calculated F-value is greater than the critical F-value so we reject the null hypothesis means that the model is statistically significant.

If we back to the results of linear regression from excel (table 6) and the OLS result from the EViews 12 software (table 7), the F-statistic is equal to 5,22819 and it' associated p-value is 0,0418 =4,18% less than the chosen significance level 5% so we reject the null hypothesis means that the model is statistically significant.

The F-test assesses the joint significance of all coefficients in the regression model.

A significant F-test does not necessarily imply that each individual coefficient is significant.

Practice Questions

Suppose you are given a dataset that contains information about the number of hours student studied and their corresponding exam scores. You want to build a simple linear regression model to establish the relationship between exam scores and the number of hours studied. The dataset is as follows:

Hour studied (X)	Exam Score (Y)
2	60
3	62
4	66
5	71
6	78
8	80
10	86
12	93

Your task is to:

1. Calculate the mean of hours studied ($\bar{x}$) and the mean of exam scores ($\bar{y}$).
2. Using the formulas for the slope and intercept of the simple linear regression line:

- Calculate the slope (β_1) of the regression line.
- Calculate the intercept (β_0) of the regression line.

3. Write the equation of the regression line in the form: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.
4. Use the regression equation to predict the exam score when a student studies for 15 hours, or 0 hour ?
5. Calculate the error or the deviation of estimated (predicted) value from the actual value?
6. Calculate the value of the dependent variable that is unexplained by the independent variable?

Exercise 1

Suppose you are given a dataset that contains information about oil price and interest rate. The dataset is as follows:

Month	Oil Price (Y)	Interest rate (X)
Sep 2018	75.36	2.19
Oct 2018	76.73	2.20
Nov 2018	62.32	2.27
Dec 2018	53.96	2.40
Jan 2019	56.58	2.40
Feb 2019	61.13	2.41
Mar 2019	63.79	2.41
Apr 2019	68.58	2.42
May 2019	66.83	2.39
Jun 2019	59.76	2.38
Jul 2019	61.48	2.40

Aug 2019	57.67	2.13
Sep 2019	60.04	2.04
Oct 2019	57.27	1.83
Nov 2019	60.40	1.55
Dec 2019	63.35	1.55
Jan 2020	61.63	1.55
Feb 2020	53.35	1.58
Mar 2020	32.20	0.55
Apr 2020	21.04	0.05
May 2020	30.38	0.05
Jun 2020	39.46	0.08
Jul 2020	42.07	0.09
Aug 2020	43.44	0.10
Sep 2020	40.60	0.09
Oct 2020	39.90	0.09
Nov 2020	42.30	0.09
Dec 2020	48.73	0.09
Jan 2021	53.60	0.09
Feb 2021	60.46	0.08
Mar 2021	63.83	0.07
Apr 2021	62.95	0.07
May 2021	66.40	0.06
Jun 2021	71.80	0.08
Jul 2021	73.28	0.10
Aug 2021	68.87	0.09
Sep 2021	72.80	0.08
Oct 2021	82.06	0.08
Nov 2021	79.92	0.08
Dec 2021	72.87	0.08
Jan 2022	83.92	0.08
Feb 2022	93.54	0.08
Mar 2022	112.40	0.20
Apr 2022	103.41	0.33
May 2022	110.10	0.77
Jun 2022	116.80	1.21
Jul 2022	105.08	1.68
Aug 2022	95.97	2.33
Sep 2022	88.22	2.56
Oct 2022	90.33	3.08
Nov 2022	87.38	3.78
Dec 2022	78.07	4.10
Jan 2023	80.41	4.33
Feb 2023	80.25	4.57
Mar 2023	76.47	4.65
Apr 2023	82.46	4.83
May 2023	74.12	5.06
Jun 2023	73.26	5.08
Jul 2023	78.98	5.12

Use the excel application or other software to answer the following questions:

1. Calculate the mean of Interest rate ($\bar{x}$) and the mean of oil prices ($\bar{y}$).
2. Using the formulas for the slope and intercept of the simple linear regression line: to calculate (β_0) & (β_1) of the regression line
3. Write the equation of the regression line in the form: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.
4. Calculate the error or the deviation of estimated (predicted) value from the actual value?
5. Calculate the value of the dependent variable that is unexplained by the independent variable?

Chapter 2: Multiple Linear Regression Model

Multiple Linear Regression is a statistical technique used to analyze the relationship between a single dependent variable and two or more independent variables; it allows researchers to understand and quantify the relationship between multiple independent variables and a dependent variable, enabling prediction and hypothesis testing in various fields including economics, social sciences, and engineering.

2-1- The concept and formula of multiple linear regression

i. **Definition:** Multiple Linear Regression is a statistical method used to model the relationship between a single dependent variable and two or more independent variables by fitting a linear equation to observed data. The goal of MLR is to estimate the coefficients of the linear equation, which represents the effect of each independent variable on the dependent variable, while accounting for the combined influence of all independent variables (Gareth , Witten, Hastie, & Tibshirani, 2017).

ii. **Formula of Multiple linear regression**: Instead of fitting a separate simple linear regression model for each predictor, a better approach is to extend the simple linear regression model (Eq. 1) so that it can directly accommodate multiple predictors. We can do this by giving each predictor a separate slope coefficient in a single model. In general, suppose that we have p distinct predictors. Then the multiple linear regression model takes the form (Gareth , Witten, Hastie, & Tibshirani, 2017) :

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots\cdots\cdots \beta_k X_k + \varepsilon$ …………………………..(Eq. 3)

Here, each Xi represents a different independent variable, βi represents the coefficient associated with that independent variable, and $\beta 0$ is the intercept. The error term (ε) captures the variability in Y that cannot be explained by the independent variables.

2-2- Assumption of multiple linear regression

When we fit a linear regression model to a particular data set, many problems may occur. Most common among these are the following:

i. **Non-linearity of the response-predictor relationships:** Non-linearity of the data refers to situations where the relationship between the independent (predictor) and dependent (response) variables cannot be accurately described by a straight line. In linear regression, the assumption is that the relationship between the independent (predictor) variables and the dependent (response) variable is linear, meaning that a change in the independent variable leads to a proportional change in the dependent variable (Montgomery, Peck, & Vining, 2012).

Detecting non-linearity in data is crucial for ensuring the validity of linear regression models. If the relationship between variables is non-linear, attempting to fit a linear model may result in biased estimates and inaccurate predictions.

Residual plots are a useful graphical tool for identifying non-linearity. Given a simple linear regression model, we can plot the residuals, $\varepsilon_i = Y_i - \hat{Y}_i$, versus the predictor X_i. In the case of a multiple regression model, since there are multiple predictors, we instead plot the residuals versus the predicted (or fitted) values $\hat{Y}_i$.

If the residual plot indicates that there are non-linear associations in the data, then a simple approach is to use non-linear transformations of the predictors, such as logX, √X, and X^2, in the regression model. In the later chapters of this book, we will discuss other more advanced non-linear approaches for addressing this issue.

(Gareth , Witten, Hastie, & Tibshirani, 2017).

ii. **The normality assumption** : in multiple linear regression the normality refers to the assumption that the residuals (the differences between the observed and predicted values of the dependent variable) are normally distributed. This assumption is important because it allows for the use of inferential statistical tests and ensures the reliability of parameter estimates. If the residuals are not normally distributed, it may indicate that the model does not accurately capture the underlying relationship between the variables, potentially leading to biased or incorrect conclusions. (Gujarati & Porter, 2009) Therefore, assessing the normality of residuals is a crucial diagnostic step in multiple linear regression analysis.

iii. **Correlation of error terms** : The correlation of error terms refers to the degree to which the errors (residuals) from a regression model are correlated with each other. In regression analysis, it's assumed that the errors are independent of each other. However, if there is correlation among the errors, it violates one of the assumptions of linear regression and can affect the reliability of the model estimates and predictions. Detecting and addressing correlated errors is important for ensuring the validity of the regression analysis.

To address correlated error terms in statistical analyses, several techniques can be employed. First, including relevant variables in the model may help capture some of the variation previously attributed to correlation (Kutner M. , Nachtsheim , Neter , & Li , 2005). Transforming variables, such as logarithmic or power transformations, can also mitigate correlation issues (Draper & Smith, 1998). Generalized Estimating Equations (GEE) allow for specifying correlation structures, while clustered standard errors adjust for intra-group correlation (Hardin & Hilbe, 2012). Time series analysis techniques, like ARIMA models, are effective for temporal correlation (Box , Jenkins,, & Reinsel, Time Series Analysis: Forecasting and Control, 2015). Instrumental Variables (IV) can address endogeneity issues, and residual analysis methods, such as Durbin-Watson tests and ACF plots, diagnose autocorrelation (Angrist & Pischke, 2008). Finally, robust standard errors account for heteroscedasticity and correlation in regression models, ensuring the validity of statistical inferences (Greene, 2017). These techniques collectively offer robust strategies for correcting correlated error terms and enhancing the reliability of statistical analyses.

iv. **Non-constant variance of error terms (Homoskedasticity) : a**nother important assumption of the linear regression model is that the error terms have a constant variance, $\mathrm{Var}(\varepsilon i) = \sigma^2$. However, one can identify non-constant variances in the errors, or heteroscedasticity. When faced with this problem, one possible solution is to transform the response Y using a concave function such as log Y or $\sqrt{Y}$. Such a transformation results in a greater amount of shrinkage of the larger responses, leading to a reduction in heteroscedasticity issue.

v. **Collinearity :** Multicollinearity is a statistical phenomenon that occurs when two or more independent variables in a regression model are highly correlated with each other. In other words, multicollinearity exists when there is a linear relationship between independent variables, making it difficult for the model to determine the individual effects of each independent variable on the dependent variable.

Multicollinearity can lead to several issues in regression analysis, including:

- ❖ Increased standard errors of coefficients, making them less precise.
- ❖ Difficulty in determining the true significance of individual independent variables.

- Unstable coefficient estimates, as small changes in the data can lead to large changes in the estimated coefficients.

Detecting and addressing multicollinearity is essential for ensuring the validity and reliability of regression analysis.

Multicollinearity in regression analysis can be addressed using various techniques.

- **Collect more data**: Sometimes multicollinearity can be mitigated by collecting more data, especially if the correlation between variables is driven by a small sample size. However, this may not always be feasible or practical (Kutner M. , Nachtsheim, Neter, & LI, 2005).
- **Remove one of the correlated variables**: If two or more variables are highly correlated, you may choose to remove one of them from the model. This decision should be based on theoretical understanding, domain knowledge, or the importance of the variables in the context of your analysis (Weisberg , 2014).
- **Combine the correlated variables**: Instead of including highly correlated variables separately in the model, you can create a single composite variable by averaging or summing them. This can help reduce multicollinearity while retaining the relevant information (Johnson & Wichern, 2007).
- **Principal Component Analysis (PCA)**: PCA is a dimensionality reduction technique that can be used to transform correlated variables into a smaller set of uncorrelated variables called principal components. These principal components can then be used as predictors in the regression model (Montgomery , Jennings, & Kulahci, 2015).
- **Partial Least Squares Regression (PLSR)**: PLSR is a regression technique that combines features of principal component analysis and multiple regression. It can handle multicollinearity by identifying new variables (latent variables) that are linear combinations of the original predictors (Varmuza & Filzmoser, 2009).
- **Centering or Standardizing Variables**: Mitigate multicollinearity by centering or standardizing variables (Johnson & Wichern, 2007).

2-3- Estimating of the Model Parameters

In multivariate regression, the goal is to estimate the model parameters that describe the relationship between multiple independent variables and a dependent variable. The parameters are typically estimated using a method such as ordinary least squares (OLS).

i. Estimating the regression Coefficients

Based on the available data, we wish to estimate the parameters $\beta_0, \beta_1, \beta_2, \cdots\cdots \beta_k$. As in the case of simple regression presented in Chapter 2, we use the least squares method, that is, we minimize the sum of squares of the errors. From (Eq.3), the errors can be written as (Chatterjee. & Hadi, 2012) :

$\varepsilon = Y - \beta_0 - \beta_1 X_1 - \beta_2 X_2 - \cdots\cdots\cdots - \beta_k X_k$

If we take the different observations of the data, we could write this equation as follow:

$\varepsilon_i = Y_i - \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i2} - \cdots\cdots\cdots \beta_k X_{ik}$ ……………………… ***(Eq.4)***

i=1,2, 3……………..n (n: number of observations)

The Sum of squares of these errors is :

$S(\beta_0, \beta_1\, , \beta_2 \cdots\cdots \beta_k) = \sum_{i=1}^{n} {\varepsilon_i}^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i2} - \cdots\cdots \beta_k X_{ik})^2$ …..**(Eq.5)**

By direct application of calculus, it can be shown that the least squares estimates $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2 \cdots\cdots\cdots \hat{\beta}_k$ which minimize $S(\beta_0, \beta_1\, , \beta_2 \cdots\cdots \beta_k)$, are given by the solution of a system of linear equations. The estimate $\hat{\beta}_0$ is usually referred to as the *intercept* or *constant*, and β_j as the *estimate* of the (partial) regression coefficient of the predictor X_j. We assume that the system of equations is solvable and has a unique solution. Solving this equation can be by closed-form formula using matrix (Chatterjee. & Hadi, 2012). Suppose we have observed data for *n* cases or units, meaning we have a value of *Y* and all of the regressors for each of the *n* cases. We define (Weisberg , 2014)

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad X = \begin{pmatrix} 1 \ x_{11} & \cdots\cdots & x_{1k} \\ 1 \ x_{21} & \ddots & x_{2k} \\ \vdots & & \vdots \\ 1 \ x_{n1} & \cdots & x_{nk} \end{pmatrix} \quad \cdots\cdots\cdots \textbf{(Eq.6)}$$

The first row of **X** is x_1' which is $(x_{11}, x_{12}, \cdots \cdots x_{1p}$ and the first row of **Y** is y_1 an ordinary number or scalar. The regressors in **X** are in the order intercept. The matrix **X** is n × p and **Y** is n×1. Next, define $\boldsymbol{\beta}$ to be a $(p + 1) \times 1$ vector of unknown regression coefficients,

$$\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2 \cdots \cdots \beta_k)'$$

An equation for the mean function evaluated at $\mathbf{x}_i$ is

$$\begin{aligned} E(Y/X = x_i) &= \beta_i x_i{}' \\ &= \beta_0 + \beta_1 x_{1i} + \cdots \cdots \beta_k x_{ki} \end{aligned}$$

and the mean function in matrix terms is : $E(Y/X) = \boldsymbol{\beta} X$ …………………..(Eq. 7)

where **Y** is the vector of responses, and **X** is the $n \times (k + 1)$ matrix whose *i*th row is x_1'.

We shall not say anything more about the actual process of solving the normal equations. We assume the availability of computer software that gives a numerically accurate solution.

ii. Fitted Values and Residuals

The least squares estimate $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is chosen to minimize the residual sum of squares (SSE) function

$SSE(\beta) = \sum_{i=1}^{n} (y_i - x_i'\, \beta)^2 = (Y - \beta X)'(Y - X\beta)$ ………….*(Eq. 8)*

The OLS estimates can be found from (Eq.8) by differentiation in a matrix analog to the development. The ols estimate is given by the formula (Weisberg , 2014) :

$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$ **…………………………..(Eq.9)**

Using the estimated regression coefficients $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2 \cdots \cdots \cdots \hat{\beta}_p$we write the fitted least squares regression equation as (Chatterjee. & Hadi, 2012)

$\widehat{\boldsymbol{Y}} = \widehat{\boldsymbol{\beta}}_{\mathbf{0}} + \widehat{\boldsymbol{\beta}}_{\mathbf{1}}\boldsymbol{X}_{\mathbf{1}} + \cdots \cdots + \widehat{\boldsymbol{\beta}}_{\boldsymbol{k}}\boldsymbol{X}_{\boldsymbol{k}}$ **……………(Eq. 10)**

For each observation in our data we can compute
$\widehat{\boldsymbol{y}}_{\boldsymbol{i}} = \widehat{\boldsymbol{\beta}}_{\mathbf{0}} + \widehat{\boldsymbol{\beta}}_{\mathbf{1}}\boldsymbol{x}_{\boldsymbol{i}\mathbf{1}} + \cdots \cdots \cdots \widehat{\boldsymbol{\beta}}_{\boldsymbol{k}}\boldsymbol{x}_{\boldsymbol{ik}}$ **…………..(Eq.11)**

These are called the *fitted* values. The corresponding *ordinary* least squares residuals are given by : (Weisberg , 2014) $\varepsilon_i = y_i - \hat{y}_i$ Where *i= 1,2………,n*

$\sum_{i=1}^{n} \boldsymbol{\varepsilon}_{\boldsymbol{i}}^{\mathbf{2}} = \sum_{i=1}^{n} (\boldsymbol{y}_{\boldsymbol{i}} - \widehat{\boldsymbol{y}}_{\boldsymbol{i}})^{\mathbf{2}} = \boldsymbol{SSE}$ ***………………..(Eq.12)***

SSE is the *sum o fsquared residuals.*

2-4- Test the significance of the model

In the context of multiple linear regression analysis, several statistical measures and tests are commonly used to assess the goodness of fit of the model and the significance of the predictors. These include the coefficient of determination (R-squared), adjusted R-squared, and ANOVA (Analysis of Variance) test.

i. Hypothesis Testing about Individual Regression Coefficients

If we invoke the assumption that the error is normally distributed[1] $\varepsilon_i \sim N(0, \sigma^2)$, then we can use the *t* test to test a hypothesis about any *individual* partial regression coefficient. To illustrate the mechanics, consider the regression, (Eq. 3). Let us postulate that

$H_0: \beta_1 = 0$

$H_a: \beta_1 = 0$

The null hypothesis states that, with *X*$_2$*, X*$_3$*, ……X*$_p$ all held constant, *X*$_1$ or has no (linear) influence on *Y*. To test the null hypothesis, we use the t-test given in (Eq. 13). If the computed t value exceeds the critical t value at the chosen level of significance, we may reject the null hypothesis; otherwise, we may not reject it (Gujarati & Porter, 2009).

$$t - stat_{calculated} = \frac{\widehat{\beta}_1}{se(\widehat{\beta}_1)} \quad \text{………………………………}(Eq.13)$$

For our illustrative example, using Eq.13 , assume that we have the following estimating results:

$$S = 110.46 + 3.362 Y_d + 10.197 r + \varepsilon \text{ ……………….}\textbf{(Eq.14)}$$

(0.421) (1.582)

p-value 0.0078 0.00345

$\hat{\beta}_1 = 3.362$ and the $se(\hat{\beta}_1) = 0.421$ then the $t - stat_{cal} = \frac{3.362}{0.241} = 7.9857$

If the chosen significance level $\alpha = 5\%$, and *n= 52* the degree of freedom *df = n-k-1= 52-2-1= 49;* from the table of t-distribution - **Student's t table** (table 8), we find that the $t - stat_{critical} =$ 2.09. the t-stat value calculated is greater than t-stat critical so, we reject the null hypothesis; this suggest that the X$_1$ has a significant linear influence on *Y*.

In practice, one does not have to assume a particular value of α[2] to conduct hypothesis testing. One can simply use the *p* value given in (Eq.14), which in the present case is 0.0078. The interpretation

[1] ε_i follow thenormal distribution with zero mean and constant variance σ^2

[2] α is the threshold at which we are willing to reject the null hypothesis, assuming it to be true.

of this *p* value (i.e., the exact level of significance) is that if the null hypothesis were true, the probability of obtaining a *t* value of as much as 7.9857 or greater (in absolute terms) is only 0.0065 or 0.65 percent, which is indeed a small probability, much smaller than the adopted value of 5%. So, it is allowed to reject the null hypothesis and accept the alternative hypothesis which means that there is evidence of a linear relationship between X_1 and Y.

ii. Coefficient of Determination, R-squared, and Adjusted R-squared, ANOVA test

a. Coefficient of Determination (R-squared):

- R-squared is a measure of the proportion of the variance in the dependent variable (target variable) that is predictable from the independent variables (predictor variables) in the regression model (Chatterjee. & Hadi, 2012).
- Mathematically, R-squared is calculated as the ratio of the explained sum of squares (SSR) to the total sum of squares (SST): $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$.
- Higher values of R-squared generally indicate a better fit of the model to the data, but it is important to consider other factors as well (Kutner M. , Nachtsheim, Neter, & LI, 2005).
- It ranges from 0 to 1, where 0 indicates that the model does not explain any of the variability in the dependent variable, and 1 indicates that the model explains all the variability.

b. Adjusted R-squared:

- Adjusted R-squared is a modification of R-squared that adjusts for the number of predictors in the model. It penalizes the addition of unnecessary predictors that do not improve the model significantly.
- It is particularly useful when comparing models with different numbers of predictors (Kutner M. , Nachtsheim, Neter, & LI, 2005).
- Mathematically, adjusted R-squared is calculated using the formula:

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

  where *n* is the number of observations and *k* is the number of predictors in the model.

c. ANOVA Test (Analysis of Variance):

- The ANOVA test is used to assess the overall significance of the regression model by testing whether at least one of the predictors has a non-zero coefficient. It compares the variance explained by the regression model to the residual variance (unexplained variance) (Gareth , Witten, Hastie, & Tibshirani, 2017).

Given the k-variable regression model (Gujarati & Porter, 2009):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots\cdots\cdots \beta_k X_k + \varepsilon$$

To test the hypothesis

$H_0: \beta_1 = \beta_2 = \cdots\cdots\cdots \beta_k = 0$ *(i.e., all slope coefficients are simultaneously zero);* versus

H_a: Not all slope coefficients are simultaneously zero

The null hypothesis (H0) in ANOVA for regression is that all the regression coefficients are equal to zero, implying that none of the predictors have a significant effect on the dependent variable.

The test statistic for ANOVA is an F-statistic, which follows an F-distribution under the null hypothesis.

$$F = \frac{SSR/df}{SSE/df} = \frac{SSR/k-1}{SSE/n-k-1}$$

If $F > F_{\alpha\ (k-1, n-k-1)}$ reject H0; otherwise you do not reject it, where $F_{\alpha\ (k-1, n-k-1)}$is the critical F value at the α level of significance and (k − 1) numerator df and (n – k-1) denominator degree of freedom (*df*), where k is the number of the parameters to be estimated, on which 1 is the intercept, and n is the number of observation. Alternatively, if the p value of F obtained from the formula of F statistic is sufficiently low (or less the chosen significance level α), then, one can reject H_0, and it can be concluded that the regression model is statistically significant ; If the p-value is greater than the chosen significance level, we fail to reject the null hypothesis, indicating that the model is not statistically significant (Montgomery, Peck, & Vining, 2012).

Example:

$$S = 110.46 + 3.362Y_d + 10.197r + \varepsilon$$

(0.421) (1.582)

SSE = 1592.301 *SST = 13581.137* *n=52* If the significance level is 5%:

From the example we have *n=52, k = 3; SSR = SST – SSE = 13581.137-1592.301= 11987.836*

$$F = \frac{SSR/k-1}{SSE/n-k-1} = \frac{11987.836/3-1}{1592.301/52-3-1} = 188.2156$$

$F_{0.05\ (2,50)} =$

For the four F tables (in the http://www.socr.ucla.edu/Applets.dir/F_Table.html), the rows represent denominator degrees of freedom and the columns represent numerator degrees of freedom. For example, to determine the .05 critical value for an F distribution with 2 and 50 degrees of freedom, look in the 2 column (numerator) and 50 row (denominator) of the F Table for alpha=.05. $F_{0.05\ (2,50)}$=*3.2317;* the $F > F_{0.05\ (2,50)}$ so, we reject H_0 indicating that the regression model is statistically significant.

In Section 3-1 we discussed the test of significance of a single regression coefficient and in section 3-2 we have discussed the joint or overall test of significance of the estimated regression (i.e., all slope coefficients are simultaneously equal to zero). **We reiterate that these tests are different.** Thus, on the basis of the *t* test or confidence interval (of Sec tion 3-1) it is possible to accept the hypothesis that a particular slope coefficient, β*k* , is zero, and yet reject the joint hypothesis that all slope coefficients are zero.

d. An Important Relationship between R^2 and *F*

There is an intimate relationship between the coefficient of determination R^2 and the *F* test used in the analysis of variance. Assume that the disturbances are normally distributed and that the null hypothesis is : $H_0: \beta_1 = \beta_2 = \cdots \beta_k = 0$ *i.e., all slope coefficients are simultaneously zero;* then it follows that

$$F = \frac{SSR/k-1}{SSE/n-k-1} = \frac{SSR}{(SST-SSR)}\frac{n-k-1}{k-1}$$

$$= \frac{SSR/SST}{(SST-SSR)/SST}\frac{n-K-1}{k-1} = \frac{R^2}{1-R^2}\frac{n-k-1}{k-1}$$

$$F = \frac{R^2/k-1}{1-R^2/n-k-1}$$

iii. Testing the Contribution of an Additional Explanatory Variable

When adding an additional explanatory variable to the model, you want to test whether this variable contributes significantly to explaining the variation in the response variable beyond what is already explained by the existing variables.to do so we follow these steps:

- ❖ Fit the original model with the existing explanatory variables.
- ❖ Fit the new model by adding the additional explanatory variable.
- ❖ Conduct an F-test comparing the two models.

If the p-value associated with the F-test is less than the chosen significance level, you conclude that the additional variable contributes significantly to the model.

If the p-value is greater than the chosen significance level, you conclude that the additional variable does not contribute significantly to the model, and you may consider removing it.

The F-test procedure just outlined provides a formal method of deciding whether a variable should be added to a regression model. Often researchers are faced with the task of choosing from several competing models involving the same dependent variable but with different explanatory variables. As a matter of ad hoc choice (because very often the theoretical foundation of the analysis is weak), these researchers frequently choose the model that gives the highest adjusted ($\bar{R}^2$). Therefore, if the inclusion of a variable increases ($\bar{R}^2$). it is retained in the model although it does not reduce SSE significantly. So, if the model incorporating the extra variable exhibits a higher adjusted R-squared value, we can infer that including this variable improves the model's overall fit (Gujarati & Porter, 2009).

Example:

Let's say we have a multiple linear regression model with two explanatory variables: X_1, $X_{2,}$ and a response variable Y. To test the overall significance of the model, we compute the ANOVA table. To test the contribution of an additional explanatory variable say X_3, we compare the original model with the additional variable added to it using an F-test, and the adjusted R-squared.

These tests help ensure that your model is adequately explaining the variability in the response variable and that the variables included in the model are indeed contributing significantly to the prediction.

2-5- Check for linear regression assumption

To check the assumptions of linear regression, we can perform various diagnostic tests and examinations. By systematically checking these assumptions, we can evaluate the validity and reliability of the linear regression model and make appropriate adjustments. The subsequent sections provide an overview of each assumption and detail how to check each of them.

i. Linearity

Examine scatterplots of the independent variables against the dependent variable to ensure that the relationships appear linear. Additionally, you can use partial regression plots or component-residual plots to check for linearity while controlling for other variables (Montgomery, Peck, & Vining, 2012).

We can check for linearity in Excel by creating a scatterplot of each independent variable against the dependent variable. Here's a step-by-step guide (Moore, McCabe, & Craig's, 2009):

a. **O**rganize the data in columns with the independent variables in one column each and the dependent variable in another column.

b. Select the data to plot, including the independent and dependent variables. Then, go to the "Insert" tab in Excel and choose "Scatter" from the Charts group.

c. Select the scatterplot type where the independent variable is on the x-axis and the dependent variable is on the y-axis.

d. Examine the resulting scatterplot. If the relationship between the independent and dependent variables appears to be roughly linear, then the assumption of linearity is met. However, if the points form a pattern that does not resemble a straight line, further investigation may be needed.

e. Add a trendline to the scatterplot to visualize the overall trend. Right-click on any data point in the scatterplot, select "Add Trendline," and choose the appropriate type of trendline (linear, polynomial, etc.). The trendline can help you assess the linearity of the relationship more accurately.

Repeat these steps for each independent variable to check for linearity against the dependent variable. If the relationships appear to be nonlinear, transformations or other modeling techniques may be necessary to address this issue.

ii. Check for the normality

To assess the normality of the data distribution, we can conduct the Jarque-Bera test.

Jarque–Bera (JB) Test of Normality: is an asymptotic, or large-sample, test. It is also based on the OLS residuals. This test first computes the skewness[3] and kurtosis[4] measures of the OLS residuals and uses the following test statistic (Gujarati & Porter, 2009):

$$JB = n\,(\frac{S^2}{6} + \frac{1}{24}(K-3)^2)\text{…………………(Eq. 15)}$$

where n = sample size, S = skewness coefficient, and K = kurtosis coefficient. For a normally distributed variable, $S = 0$ and $K = 3$. Therefore, the JB test of normality is a test of the joint hypothesis that S and K are 0 and 3, respectively. In that case the value of the JB statistic is expected to be 0 (Gujarati & Porter, 2009).

The null hypothesis of the Jarque-Bera test (H_0): Residuals are normally distributed.

The alternative hypothesis of Jarque-Bera test (H_a): Residuals are not normally distributed.

The JB statistic given in (Eq.15) follows the chi-square distribution (appendix A3) with 2 df.

If the **JB calculated exceeds the critical value**, suggesting that the data deviates significantly from normality (Residuals are not normally distributed)

If the **JB calculated is less than or equal to the critical value**, indicating that there is insufficient evidence to reject the null hypothesis of normality (Data is normally distributed)

For our example the JB statistic obtained from GDP growth-inflation rate is 0.576, JB critical is 5.991, $JB_{cal} < JB_{crit}$;

The null hypothesis that the residuals in the present example are normally distributed cannot be rejected (Accept the null hypothesis) means the data is normally distributed

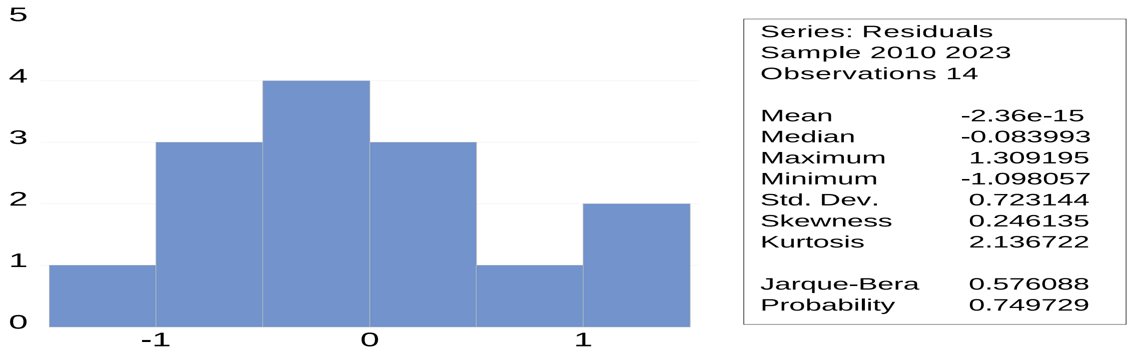
To run a Jarque-Bera test in **EViews,** you can follow these steps(2011 ،السواعي) :

1. Go to the " View" menu in EViews.
2. Navigate to " Diagnostic Tests" and click on " Normality Test "
3. EViews will compute the Jarque-Bera test statistic and its associated p-value.
4. The results will typically include the Jarque-Bera statistic, and p-value.

[3] *Skewness is a statistical measure that describes the asymmetry of a distribution around its mean. It indicates whether the data points are more spread out on one side of the mean than the other. Skewness can be positive, negative, or zero.*

[4] *Kurtosis is a statistical measure used to describe the distribution of data points in a data set. It provides insight into the "tailedness" of the distribution, which refers to the extent to which the tails of the distribution differ from the tails of a normal distribution.*

Figure 16: Histogram of residuals GDP growth- Inflation data



```
Series: Residuals
Sample 2010 2023
Observations 14

Mean           -2.36e-15
Median         -0.083993
Maximum         1.309195
Minimum        -1.098057
Std. Dev.       0.723144
Skewness        0.246135
Kurtosis        2.136722

Jarque-Bera     0.576088
Probability     0.749729
```

Source: EViews 12 output

5. Interpret the results based on the p-value (Gujarati & Porter, 2009):

- If the **p-value is less than the chosen significance level (e.g., 0.05),** we reject the null hypothesis of normality, suggesting that the residuals are not normally distributed.
- If the **p-value is greater than the significance level**, fail to reject the null hypothesis, indicating that the residuals are normally distributed.

According to our results the probaility of the Jarque-Bera is 0.749, which is greater than the 0.05 (the chosen significance level) this mean we fail to reject the null hypothesis (the residuals are normally distributed)

iii. Check for the multicollinearity

We can employ diverse approaches to assess multicollinearity, one of which involves computing variance inflation factors (VIFs).

a. **Definition Variance Inflation Factor (VIF)** measures how much the variance of an estimated regression coefficient increases if the predictors are correlated. This can lead to issues with coefficient estimates, making them unstable or difficult to interpret. It can also inflate standard errors and affect hypothesis testing. Calculate the variance inflation factor (VIF) for each independent variable to assess the degree of multicollinearity ; VIF values greater than 10 or significantly different from 1 indicate multicollinearity (Hair, Black, Babin, & Anderso, 2010).

b. **Calculate VIF**: calculating the VIF for each independent variable. The formula to calculate VIF for each variable is as follows (Hair, Black, Babin, & Anderso, 2010):

$$VIF = 1 / (1 - R^2) \text{ ………………………….(Eq. 16)}$$

Where R^2 is the coefficient of determination obtained by regressing the one independent variable on all other independent variables.

c. **Variance Inflation Factors (VIF) calculation**:

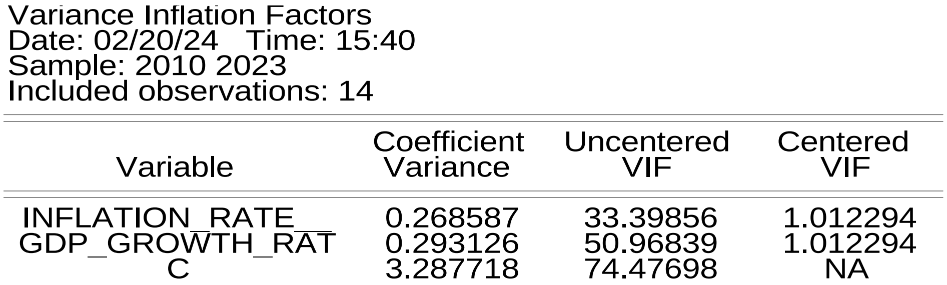
❖ **Calculation VIF in Excel:**

- Excel does not have a built-in function for calculating VIFs, so you'll need to perform regression analyses separately for each independent variable means regression of each independent variable against all other independent variables.
- Use the formula you provided: VIF = 1 / (1 - R²), where R² is the coefficient of determination from a regression of each independent variable against all other independent variables.
- Calculate the VIF for each independent variable to assess multicollinearity.
- After calculating VIFs, look for values greater than 10 or significantly different from 1, indicating multicollinearity.

❖ **Check Multicollinearity in EViews.**

- Run a regression
- After estimating the regression model, go to the "View" menu, select "Coefficient Diagnostics...", and then choose "Variance Inflation Factors".
- In the VIF dialog box, select the variables you want to include in the VIF calculation.
- Click "OK" to generate the VIF values.

Table 15: variance inflation factor results



```
Variance Inflation Factors
Date: 02/20/24   Time: 15:40
Sample: 2010 2023
Included observations: 14
==================================================================
                     Coefficient    Uncentered      Centered
      Variable        Variance         VIF            VIF
==================================================================
 INFLATION_RATE__     0.268587      33.39856       1.012294
 GDP_GROWTH_RAT       0.293126      50.96839       1.012294
        C             3.287718      74.47698          NA
```

Source: EViews output

- The higher the VIF, the higher the possibility that multicollinearity exists. When VIF is higher than 5, there is significant multicollinearity that needs to be corrected.

According to the table (Table 15), the VIF (centred VIF) are all less than 5, means there is no multicolinearity issue in the chose data.

If multicollinearity is detected, consider removing one of the correlated variables or using other techniques such as principal component analysis (PCA) to address it.

iv. Check for the homoskedasticity

To check for homoskedasticity, we can use several diagnostic techniques, and the most commonly used is the Breusch-Pagan Test

a. **Definition: Breusch-Pagan Test** is a statistical test used to determine whether the variance of errors in a regression model is constant (homoskedasticity) or varies with the independent variables (heteroscedasticity). This test is widely employed in econometrics and regression analysis (Breusch & Pagan, 1979).

b. **How to perform the Breusch-Pagan test:** we can follow these steps:

- First, estimate the regression model using ordinary least squares (OLS) or any other regression technique.
- Then, Obtain Residuals
- Square the Residuals
- Run an auxiliary regression where the squared residuals are regressed on the independent variables used in your original regression model.

c. **Breusch-Pagan test Hypothesis**

The null hypothesis (H_0) of the Breusch-Pagan test: the variance of the errors is constant (homoskedasticity).

The alternative hypothesis (H_a): the variance of the errors is not constant (heteroscedasticity).

d. **Interpretation:** to test this hypothesis we examine the significance of the coefficient of determination F- statistics from the auxiliary regression.

- If the *p-value* is statistically significant (i.e., the p-value associated with the f-statistic is less than your chosen significance level, commonly 0.05), you reject the null hypothesis, indicating the presence of heteroscedasticity.
- If the *f-statistic* value is not statistically significant (i.e., the p-value associated with the *f-statistic* is greater than your chosen significance level), you fail to reject the null hypothesis, suggesting homoskedasticity.

- ❖ **Check the Homoskedasticity using Excel software:** to check for homoskedasticity using Excel without relying on a scatter plot, it is suggested to perform a regression of the squared residuals on the independent variables. Here's how to do it:

a. **Obtain Residuals**: First, calculate the residuals from your linear regression model. These are the differences between the actual values of the dependent variable and the predicted values from the main regression model.

b. **Square the Residuals**: Create a new column where you square each residual value obtained from step 1.

c. **Perform Regression**: Use Excel's regression analysis tool to regress the squared residuals on the independent variables. Here's how:
   - Go to the "Data" tab.
   - Click on "Data Analysis".
   - Choose "Regression" from the list of analysis tools and click "OK."
   - In the Regression dialog box:
     - Select the range containing the squared residuals as the Y range.
     - Select the range containing the independent variables as the X range.
     - Check the box for "Labels" if your data has column labels.
     - Choose an output range where you want Excel to output the regression results.
     - Click "OK" to run the regression analysis.

e. **Interpret Results**: In the output, look for the coefficient of determination (R-squared). If the R-squared value is close to zero, it suggests homoskedasticity (constant variance of errors). A significant R-squared value indicates heteroscedasticity (varying variance of errors). In the context of F-statistic, a significant F-statistic may suggest that the variance of the errors is not constant across all levels of the independent variables, indicating potential heteroscedasticity. Conversely, a non-significant F-statistic imply homoskedasticity.

- ❖ **Check the Homoskedasticity using EViews software:**

i. Upload the data in Eviews software

ii. Go to »View » menu at the top of the Eviews windows,

iii. select « open data as equation » click OK

iv. You will get the result of the OLS estimation;

v. Go to " View" menu in EViews.

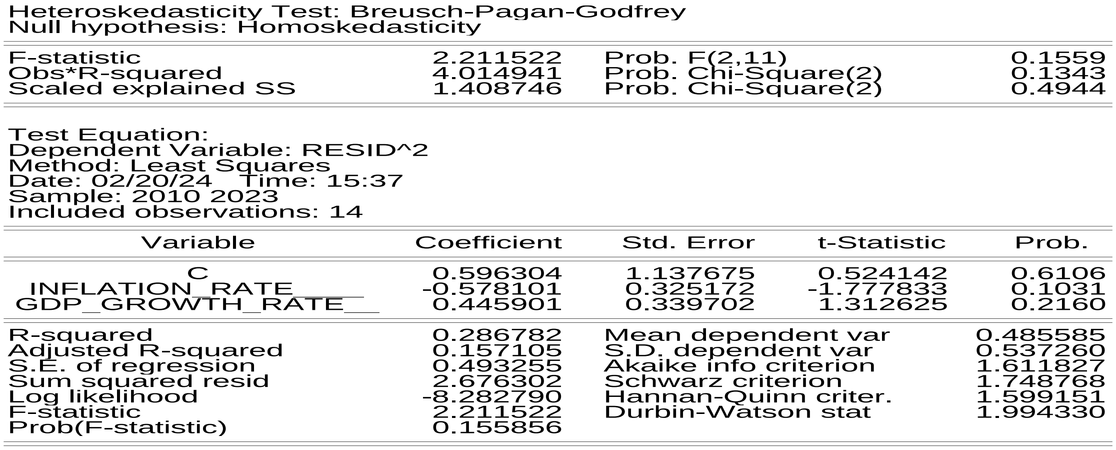
vi. Navigate to " Residual Diagnostic" and click on " Heteroskedasticity Test "

vii. In the Options dialog box, go to the "Test types" select **Breusch-Pagan Test**

viii. Run the Regression: Once you've specified your regression model and test options, click "OK" to run the regression.

ix. Look for the section of the output that corresponds to the Breusch-Pagan test. The test results will include the f- statistic, its associated p-value, and other relevant information.

Table 16: heteroskedasticity test results



```
Heteroskedasticity Test: Breusch-Pagan-Godfrey
Null hypothesis: Homoskedasticity
================================================================================
F-statistic              2.211522    Prob. F(2,11)                      0.1559
Obs*R-squared            4.014941    Prob. Chi-Square(2)                0.1343
Scaled explained SS      1.408746    Prob. Chi-Square(2)                0.4944
================================================================================

Test Equation:
Dependent Variable: RESID^2
Method: Least Squares
Date: 02/20/24   Time: 15:37
Sample: 2010 2023
Included observations: 14
================================================================================
         Variable             Coefficient    Std. Error    t-Statistic    Prob.
================================================================================
             C                  0.596304      1.137675      0.524142     0.6106
   INFLATION_RATE____          -0.578101      0.325172     -1.777833     0.1031
  GDP_GROWTH_RATE__             0.445901      0.339702      1.312625     0.2160
================================================================================
R-squared                0.286782    Mean dependent var              0.485585
Adjusted R-squared       0.157105    S.D. dependent var              0.537260
S.E. of regression       0.493255    Akaike info criterion           1.611827
Sum squared resid        2.676302    Schwarz criterion               1.748768
Log likelihood          -8.282790    Hannan-Quinn criter.            1.599151
F-statistic              2.211522    Durbin-Watson stat              1.994330
Prob(F-statistic)        0.155856
================================================================================
```

Source: EViews output

Interpret Results

Based on the results in the table above (table 16), the regression of squared residuals on the independent variables yields an insignificant R-squared value, it suggests that the assumption of homoskedasticity is fulfilled. In addition, the p-value related to F-statistic is greater than the significance level, we fail to reject the null hypothesis of homoskedasticity, so, there is evidence of homoskedasticity

v. Check for the autocorrelation

i. **Definition:** Autocorrelation, also known as serial correlation, refers to the correlation of a time series with a lagged version of itself. In other words, it measures the degree of similarity between observations at different time points within the same series. It can lead to underestimates of the standard error and can cause one to think predictors are significant when they are not. (Brockwell & Davis, 2016).

ii. **The Durbin-Watson test** is a statistical test used to detect the presence of autocorrelation in the residuals of a regression analysis. Autocorrelation occurs when the residuals (errors) of a regression model are correlated with each other. In other words, it measures whether there is a systematic pattern in the residuals that indicates a violation of the assumption of independence (Montgomery, Peck, & Vining, 2012).

The Durbin-Watson test statistic, denoted by *d*, ranges in value from 0 to 4. A value of *d* near 2 suggests that there is no significant autocorrelation in the residuals. Values of *d* significantly below 2 indicate positive autocorrelation, while values significantly above 2 indicate negative autocorrelation. The formula to calculate the Durbin-Watson statistic is as follows (Montgomery, Peck, & Vining, 2012) :

$$d = \frac{\sum_{t=2}^{n}(\varepsilon_t - \varepsilon_{t-1})^2}{\sum_{t=1}^{n}{\varepsilon_t}^2}$$

ε_t is the residual from the Ordinary Least Squares (OLS) regression.

This formula calculates the numerator and denominator components of the Durbin-Watson statistic. In this formula, the numerator represents the sum of squared differences between consecutive residuals, and the denominator represents the sum of squared residuals.

iii. **Durbin Watson test Hypothesis**: the null and alternative hypotheses for the Durbin-Watson test are as follows:

- **Null Hypothesis (*H*0):** There is no autocorrelation in the residuals.
- ***Alternative Hypothesis (Ha):*** There is autocorrelation in the residuals**.**

A **rule of thumb** is that test statistic values (*d*) in the range of 1.5 to 2.5 are relatively normal. Values outside of this range could be cause for concern. (Field A. , 2009) suggests that values under 1 or more than 3 are a definite cause for concern.

The Durbin Watson test, requires the use of tables which you can find in the appendix (see appendix A, Durban Watson significance A 4)

If the ***d*** is above $\boldsymbol{d_U}$, we fail to reject the null hypothesis, then there is no evidence of serial correlation

If the ***d*** is below the $\boldsymbol{d_L}$ we reject the null hypothesis, then there is evidence that the data is positively autocorrelated (autocorrelation in the residual).

iv. **Calculate the Durbin-Watson Statistic in statistical software**

❖ **Calculate the Durbin-Watson statistic in Excel**

First, perform a regression analysis using dataset. After running the regression, obtain the residuals (ε_i) from the regression output. Residuals are the differences between the observed values of the dependent variable and the values predicted by the regression equation. Square each residual (ε_i^2) and sum them up: $\sum_{i=1}^{n} \varepsilon_i$

Calculate the sum of squared differences between consecutive residuals $\varepsilon_i - \varepsilon_{i-1}$; the sum them up: $\sum_{i=1}^{n}(\varepsilon_i - \varepsilon_i)^2$. **Finally,** Divide the sum of squared differences between residuals by the sum of squared residuals:

If the calculated statistic is close to 2 (2±0.5), there is no significant autocorrelation. If it's significantly less than 2, it suggests positive autocorrelation, and if it's significantly greater than 2, it suggests negative autocorrelation.

Compare to Critical Values: You can compare the calculated Durbin-Watson statistic to critical values to determine whether autocorrelation is significant. For example, for a significance level of 0.05, you can use the following critical values:

Lower critical value: ~1.5

Upper critical value: ~2.5

- If the calculated Durbin-Watson statistic falls outside this range,we reject the null hypothesis which suggests significant autocorrelation. If it is in the range, we fail to reject the null hypothesis which means that there is no autocorrelation in the residuals.

❖ Perform the Durbin-Watson statistic in EViews:

After estimating the regression using OLS and obtaining the results in the table of OLS results, we find the Durbin-Watson statistic

We can also check for autocorrelation in EViews using LM test, we follow these steps:

a. Go to "View" menu.

b. Select "Diagnostic Tests"

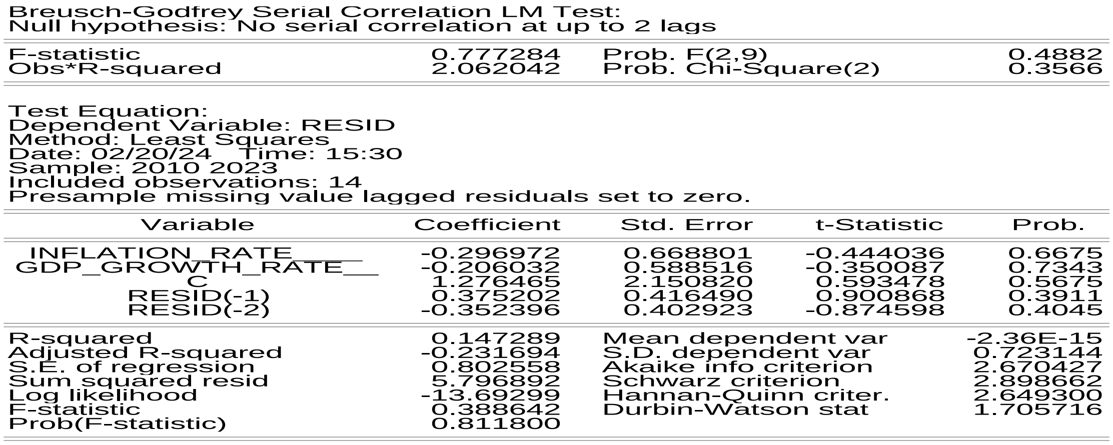
c. Choose " serial correlation LM test ".

d. In the dialog box, select the lag length.

e. Click "OK" to run the test.

The result will appear as follow:

Table 17: the results of autocorrelation (LM Test)



```
Breusch-Godfrey Serial Correlation LM Test:
Null hypothesis: No serial correlation at up to 2 lags

F-statistic              0.777284    Prob. F(2,9)                0.4882
Obs*R-squared            2.062042    Prob. Chi-Square(2)         0.3566

Test Equation:
Dependent Variable: RESID
Method: Least Squares
Date: 02/20/24   Time: 15:30
Sample: 2010 2023
Included observations: 14
Presample missing value lagged residuals set to zero.

        Variable          Coefficient   Std. Error   t-Statistic    Prob.

  INFLATION_RATE____      -0.296972     0.668801     -0.444036     0.6675
 GDP_GROWTH_RATE__        -0.206032     0.588516     -0.350087     0.7343
          C                1.276465     2.150820      0.593478     0.5675
       RESID(-1)           0.375202     0.416490      0.900868     0.3911
       RESID(-2)          -0.352396     0.402923     -0.874598     0.4045

R-squared                0.147289    Mean dependent var       -2.36E-15
Adjusted R-squared      -0.231694    S.D. dependent var        0.723144
S.E. of regression       0.802558    Akaike info criterion     2.670427
Sum squared resid        5.796892    Schwarz criterion         2.898662
Log likelihood         -13.69299    Hannan-Quinn criter.      2.649300
F-statistic              0.388642    Durbin-Watson stat        1.705716
Prob(F-statistic)        0.811800
```

Source : EViews output

Test decision

i. If the p-value of **observed R-squared** is less than the chosen significance level (e.g., 0.05), we may conclude that there is evidence of autocorrelation, indicating that the the error terms are serially correlated.

ii. If the p-value **observed R-squared** is greater than the significance level, we fail to reject the null hypothesis of no serial correlation, means the error are not serially correlated (not autocorrelated).

According to our results, the p-value of obs*R-suared is high greater that the 5% so we fail to reject the null hypothesis this means that there is no evidence of serial correlation, so the independence condition is satisfied

In the following section we provide an example which summarize what was stated in previous sections.

Example

Suppose you're analyzing the relationship between Balance of payment (BoP), Export (EX) and Foreign direct Investment (FDI). To find the impact Export (EX) and the foreign direct investment (FDI) on the balance of payment (BoP), You have conducted a linear regression analysis using OLS method and the following data.

Table 18 : USA Data

Year	USA Exports (USD billions)	USA FDI (USD billions)	Balance of Payments (USD billions)
2000	1,044	140	-811
2001	1,119	143	-775
2002	1,148	160	-718
2003	1,151	168	-668
2004	1,282	170	-594
2005	1,421	179	-530
2006	1,528	207	-484
2007	1,581	214	-470
2008	1,837	228	-469
2009	1,842	231	-469
2010	2,099	235	-449
2011	2,196	239	-423
2012	2,209	240	-421
2013	2,243	291	-390
2014	2,272	319	-345
2015	2,343	345	-328
2016	2,345	348	-306
2017	2,432	391	-293

1. Write the linear regression model that presenting the relationship between the variables under study
2. Formulate the hypothesis of the study?
3. Perform a regression analysis (using Excel or EViews).
4. Test the study hypothesis using p-value and t-statistic (according to the results you will get in Excel or EViews software).
5. Test the overall significance of model (using R^2 and F-statistic).
6. Check if the the assumptions of linear regression are met (specify the hypothesis of each diagnostic test and how you conduct it).
7. If the assumptions are not met, how you can correct the issues?

First, Write the linear regression model that presenting the relationship between the variables under study, since we examine the effect of export and Foreign Direct Investment (FDI) on balance of payment (BoP), so the Balance of payment is the dependent variable, while the FDI and Export are the independent variables

$$BoP = \beta_0 + \beta_1 Export + \beta_2 FDI + \varepsilon$$

Second, the study Hypothesis

$$\begin{cases} H_0: \beta_1 = 0, \textit{there is no significant relationship between export and BoP} \\ H_a: \beta_1 \neq 0, \textit{there is significant relationsip between Export and BoP} \end{cases}$$

$$\begin{cases} H_0: \beta_2 = 0, \textit{there is no significant relationship between FDI and BoP} \\ H_a: \beta_2 \neq 0, \textit{there is significant relationsip betweenFDI and BoP} \end{cases}$$

Third, perform a regression analysis (using Excel or EViews). We begin running the regression in Excel, followed by replicating the analysis in EViews Software.

i. Performing multiple linear regression in Excel, will provide the result which includ coefficients, standard errors, R-squared value, and other relevant statistics (Moore, McCabe, & Craig's, 2009). Here how we perform the multiple linear regression step by step as well as the result the excel show

1. Organize the data in Excel with one column for the dependent variable and multiple columns for the independent variables (the variables that are used to predict the dependent variable).

2. Click on "Data Analysis" in the "Data" tab, then choose "Regression" from the list and click "OK".

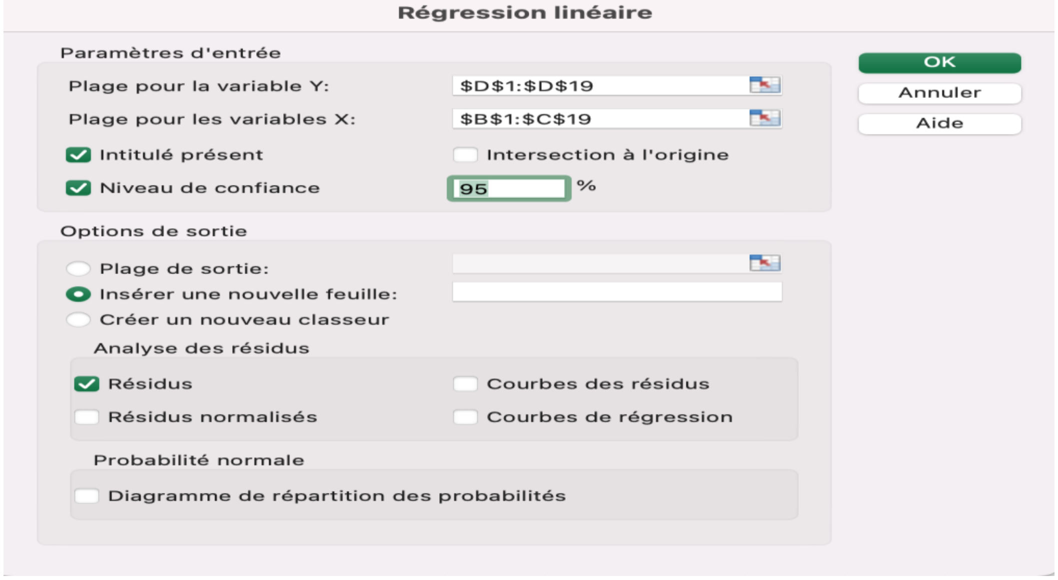
3. In the Regression dialog box, select the input range for your independent variables and the dependent variable. Make sure to include labels if you have them.

4. Check the "Labels" box if your data has column labels. Choose where the output should to be displayed - either a new worksheet or a specific range.

5. Optionally, we can choose additional regression statistics to include in the output, such as residuals, confidence intervals, and ANOVA table.

6. Once we have specified our options, click "OK" to run the regression analysis

Figure 17: Regression dialog box in excel



Régression linéaire

Paramètres d'entrée

Plage pour la variable Y: $D$1:$D$19

Plage pour les variables X: $B$1:$C$19

☑ Intitulé présent □ Intersection à l'origine

☑ Niveau de confiance 95 %

Options de sortie

○ Plage de sortie:

● Insérer une nouvelle feuille:

○ Créer un nouveau classeur

Analyse des résidus

☑ Résidus □ Courbes des résidus

□ Résidus normalisés □ Courbes de régression

Probabilité normale

□ Diagramme de répartition des probabilités

Table 19: Excel Results for multiple liner regression

Summary Output					
Regression Statistics					
Multiple R	0,946538532				
Coefficient de détermination R squared	0,895935192				
Adjusted R Squared	0,882059885				
Standard Error	54,08841461				
Observations	18				
ANOVA					
	Degree of freedom	*Sum Square*	*Mean Square*	*F*	*Significance F*
Regression	2	377809,1511	188904,5755	64,57047383	4,26363E-08
Residual	15	43883,34892	2925,556595		
Total	17	421692,5			
	Coefficients	*Standard Error*	*t Stat*	*P-value*	
Intercept	-1020,122694	48,6997598	-20,94718122	1,61214E-12	
USA Exports	215,3962316	64,70401335	3,328947008	0,004578865	
USA FDI	0,590092426	0,430599904	1,370396093	0,190724595	

Source: excel output based on the data from the table 18

The above output provides the estimated regression coefficients; the intercept, $\beta_0 = -1020{,}12$, the slope for Export variable, $\beta_1 = 215{,}39$, and the slope for FDI, $\beta_2 = 0{,}59$. Therefore, the regression equation is :

$$BoP = -1020{,}12 + 215{,}39Export + 0{,}59FDI$$

(48,69) (64,70) (0,43)

The values between bracket indicate the standard error (S_{β_i}), $S_{\beta_0} = 48{,}69, S_{\beta_1} = 64{,}7, S_{\beta_2} = 0{,}59$

Fourth, Test the study hypothesis using p-value and t-statistic (according to the results we get in Excel):

To test the hypothesis H_0 :β_j = 0, the value of t is computed as $t\ stat(\beta_i) = \beta_i / S_{\beta_i}$. For each coefficient, the value appears in the column marked *t* stat in the table 19. The values are given as $t\ stat(\beta_0) = -20{,}947, t\ stat(\beta_1) = 3{,}328, t\ stat(\beta_2) = 1{,}37$. If we compare these t-statistics to the critical t statistic (= 2.131) provided by the student's t table (see Appendix A1) at 5% level of significance and with degree of freedom 15, we conclude that only the coefficient β_1 is significant, this led us to reject the null hypothesis that $\beta_1 = 0$ indicating that the export has a positive impact on balance of payments. Specifically, one unit increase in the export value the BoP increases by 215,39 unit. Conversely, the coefficient β_2 is insignificant because it's t stat less than the critical t statistic, thereby we fail to reject the null hypothesis which suggesting that the FDI has no impact on BoP.

Another approach to test the hypothesis related to the significance of the parameters is the p-value, The *P*-values for a testing H_0 :β_j = 0 against H_a :$\beta_j \neq 0$ are provided in the column marked *p-value*, which are p-value (β_1) = 0,0045 = 0,45% , and p-value (β_2) $= 0{,}1907 = 19{,}07\%$.

If the p-value is less than the chosen significance level (5% in our case), then we reject the null hypothesis, while if the p-value is greater than the significance level we fail to reject the null hypothesis.

According to our results in the table 19, the p-value of β_1 is less than 5%, so we reject the null hypothesis (confirm the result of t-stat) and we accept the alternative hypothesis, which means the export has significant impact on BoP. On other hand, the p-value related to β_2 is greater than 5%, so we fail to reject the null hypothesis, this suggests that the FDI has no significant impact on BoP.

Fifth, Test the overall significance of model (using R^2 and F-statistic)

Start with the R-Squared, R-squared (R^2) measures the proportion of the variance in the dependent variable that is explained by the independent variables in the model.

Based on the results from the table 19, the R-squared value of 0.8959 means that approximately 89.59% of the variability in the dependent variable (Balance of Payment, BoP) can be explained by the independent variables (Export and FDI) included in the model. In other words, the model accounts for a high percentage of the variation observed in the dependent variable.

This high R-squared value indicates that the independent variables in the model are strongly related to the dependent variable and are effective in predicting its values. The high R-squared value like this suggests that the model fits the data well and provides a good representation of the relationship between the independent and dependent variables

Then, The F-statistic tests whether at least one of the independent variables has a non-zero coefficient in the population regression equation.

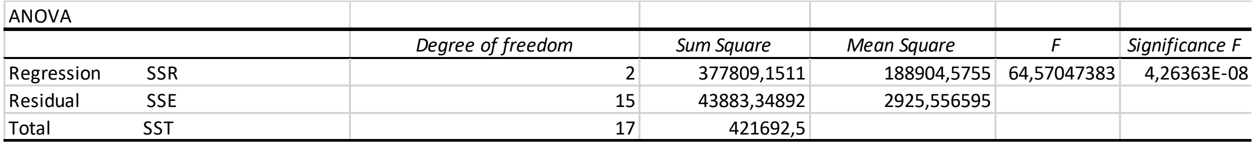
$H_0: \beta_1 = \beta_2 = \cdots\cdots\cdots \beta_k = 0$ *(i.e., all slope coefficients are simultaneously zero);* versus

H_a: Not all slope coefficients are simultaneously zero

We use the formula for the F-statistic, which is the ratio of the explained variance to the unexplained variance, adjusted for the number of parameters estimated (degree of freedom, *df*). The formula is:

$$F = \frac{SSR/df}{SSE/df} = \frac{SSR/k-1}{SSE/n-k-1} \qquad \text{or} \qquad F = \frac{R^2/k-1}{1-R^2/n-k-1}$$

Based on the results of performing multiple regression in Excel Table 19 and according to the results of ANOVA we can calculate the F statistic using the first formula



ANOVA					
	Degree of freedom	*Sum Square*	*Mean Square*	*F*	*Significance F*
Regression SSR	2	377809,1511	188904,5755	64,57047383	4,26363E-08
Residual SSE	15	43883,34892	2925,556595		
Total SST	17	421692,5			

$$F = \frac{377809{,}15/2}{43883{,}34/15} = 188904{,}575/2925{,}556 = 64{,}57$$

Let’s calculate the F statistic using the formula of R-Squared, based on the results of table 19, we have R-Squared = 0,8959

$$F = \frac{0{,}8959/2}{(1-0{,}8959)/15} = \frac{0{,}44795}{0{,}00694} \approx 64{,}55$$

The test statistic for ANOVA is an F-statistic, which follows an F-distribution under the null hypothesis. If $F > F_{0.05\ (2,15)}$ reject H0; otherwise you do not reject it.

$F_{0.05\ (2,15)} = 19{,}43$ (see the appendix F-distribution table)

So, the F calculated (64,57) is greater than the F-critical, (19,43), we reject the null hypothesis indicating that the model is overall significant.

Referring to the F-statistic provided in Table 19 (ANOVA), where the F-Statistic is 64.57 with a p-value of 0.00004%, which is below the significance level, further supports the rejection of the null hypothesis, emphasizing the significance of the F-statistic. This implies that at least one independent variable in the model possesses a non-zero coefficient, and the model effectively explains the variation in the dependent variable.

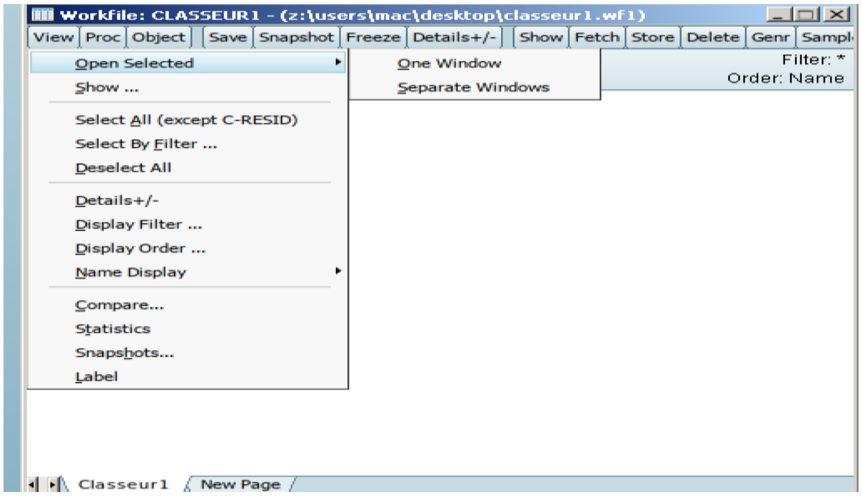
Running the regression in EViews software: before we jump to the question six, we first show how to run the regression in EViews software and answer the question (3, 6 & 7)

The following are the steps for running multiple linear regression using Ordinary Least Squares (OLS) technique in EViews software (EViews, 2020):

i. **Data Preparation** : the data should be properly organized with the dependent variable and all independent variables in separate columns. Load the dataset into EViews.

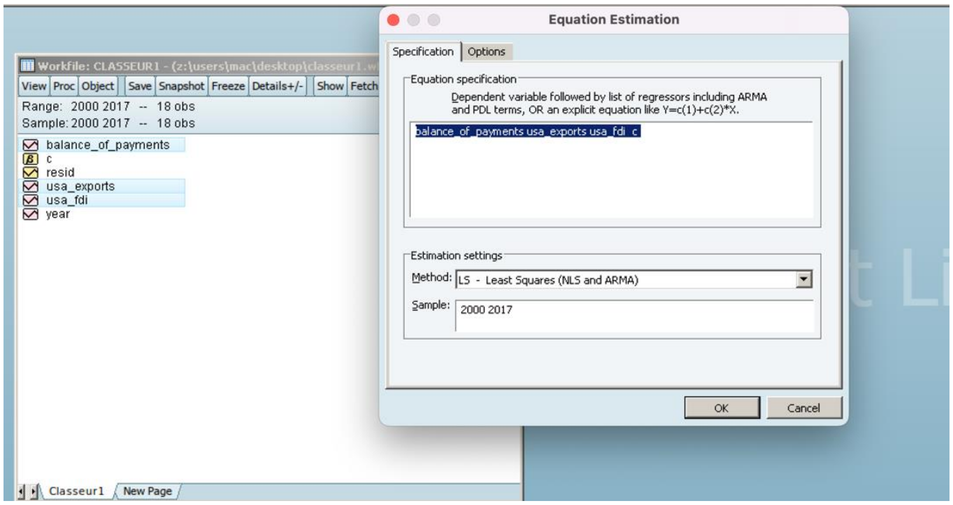
ii. **Open EViews**: Launch EViews software on the computer.

iii. **Open Data File**: Open the data file containing the dataset by going to View> Open selected > one window or separate windows > open equation.



iv. **Specify Regression Equation**:

- In the dialog box that appears, specify the estimation method (example: OLS method) > click Ok



Workfile: CLASSEUR1 - (z:\users\mac\desktop\classeur1.w

Range: 2000 2017 -- 18 obs
Sample: 2000 2017 -- 18 obs

Equation Estimation

Equation specification

Dependent variable followed by list of regressors including ARMA and PDL terms, OR an explicit equation like Y=c(1)+c(2)*X.

```
balance_of_payments usa_exports usa_fdi c
```

Estimation settings

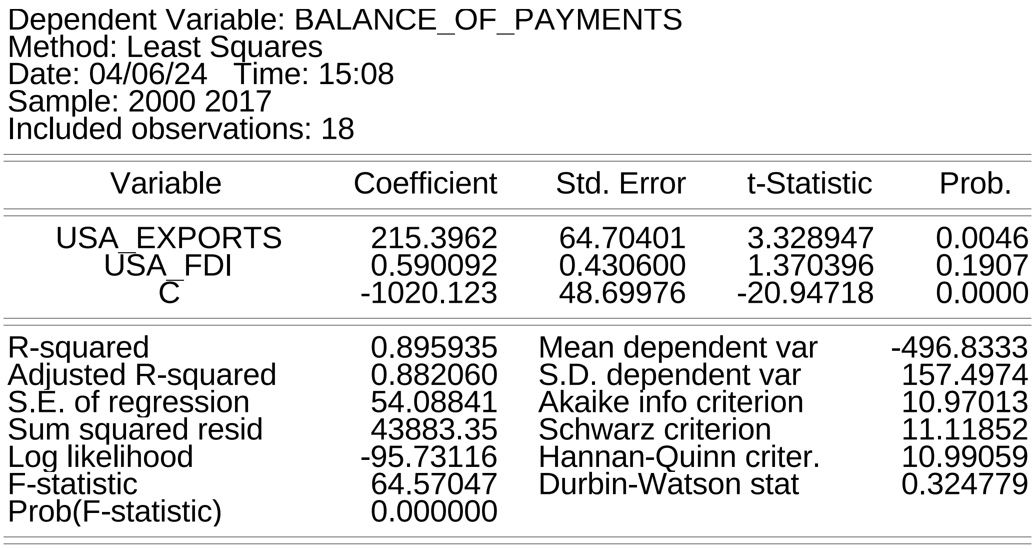
Method: LS - Least Squares (NLS and ARMA)

Sample: 2000 2017

v. **View Results**:

- After the estimation is complete, EViews will display the results in a new window. we will see the coefficients, standard errors, t-statistics, p-values, R-squared, and other relevant statistics.

Table 20: results of the OLS estimation for multiple linear regression



```
Dependent Variable: BALANCE_OF_PAYMENTS
Method: Least Squares
Date: 04/06/24   Time: 15:08
Sample: 2000 2017
Included observations: 18

        Variable          Coefficient    Std. Error    t-Statistic    Prob.

      USA_EXPORTS          215.3962      64.70401      3.328947      0.0046
        USA_FDI            0.590092      0.430600      1.370396      0.1907
           C              -1020.123      48.69976     -20.94718      0.0000

R-squared                  0.895935    Mean dependent var        -496.8333
Adjusted R-squared         0.882060    S.D. dependent var         157.4974
S.E. of regression         54.08841    Akaike info criterion      10.97013
Sum squared resid          43883.35    Schwarz criterion          11.11852
Log likelihood            -95.73116    Hannan-Quinn criter.       10.99059
F-statistic                64.57047    Durbin-Watson stat         0.324779
Prob(F-statistic)          0.000000
```

Source: EViews output based on the data from the example

vi. **Interpret Results**:

To Interpret the results of the regression analysis, we pay attention to the coefficients and their significance levels, as well as the overall fit of the model (R-squared value & F-statistic).

The above output (table 20) provides the estimated regression coefficients; the intercept, $\beta_0 = -1020{,}12$, the slope for Export variable $\beta_1 = 215{,}39$, and the slope for FDI $\beta_2 = 0{,}59$. Therefore, the regression equation is :

$$BoP = -1020{,}12 + 215{,}39 Export + 0{,}59 FDI$$
$$\textit{(48,69)} \quad \textit{(64,70)} \qquad \textit{(0,43)}$$

The values between bracket indicate the standard error (S_{β_i}), S_{β_0} = 48,69, S_{β_1} = 64,7, S_{β_2} = 0,59

Test the study hypothesis using p-value and t-statistic (according to the results we get in Excel):

To test the hypothesis H_0 :β_j = 0,

the value of *t* is computed for each coefficient appears in the column marked *t*-statistic in the **table 20**. The values are given as $t\ stat(\beta_0) = -20{,}947, t\ stat(\beta_1) = 3{,}328, t\ stat(\beta_2) = 1{,}37$.

If we compare these t-statistics to the critical t statistic (= 2.131) provided by the student's t table (see Appendix A1) at 5% level of significance and with degree of freedom 15, we conclude that only the coefficient β_1 is significant, this led us to reject the null hypothesis that $\beta_1 = 0$ indicating that the export has a positive impact on balance of payments. Specifically, one unit increase in the export value the BoP increases by 215,39 unit. Conversely, the coefficient β_2 is insignificant because it's t-statistic is less than the critical t-statistic, thereby we fail to reject the null hypothesis which suggesting that the FDI has no impact on BoP.

Another approach to test the hypothesis related to the significance of the parameters is the p-value, The *P*-values for a testing H_0 :β_j = 0 against H_a :$\beta_j \neq 0$ are provided in the column marked *prob.* which are prob.(β_1) = 0,0046 = 0,46% , and prob.(β_2) $= 0{,}1907 = 19{,}07\%$.

If the p-value is less than the chosen significance level (5% in our case), then we reject the null hypothesis, while if the p-value is greater than the significance level we fail to reject the null hypothesis.

According to our results in the table 20, the prob. of β_1 is less than 5%, so we reject the null hypothesis (confirm the result of t-stat) and we accept the alternative hypothesis, which means the export has significant impact on BoP. On other hand, the prob. related to β_2 is greater than 5%, so we fail to reject the null hypothesis, this suggests that the FDI has no significant impact on BoP.

Fifth, Test the overall significance of model (using R^2 and F-statistic)

Start with the R-Squared, based on the results from the table 20, the R-squared value of 0.8959 means that approximately 89.59% of the variability in the dependent variable (Balance of Payment, BoP) can be explained by the independent variables (Export and FDI) included in the model. In other words, the model accounts for a high percentage of the variation observed in the dependent variable.

The high R-squared value like this suggests that the model fits the data well and provides a good representation of the relationship between the independent and dependent variables

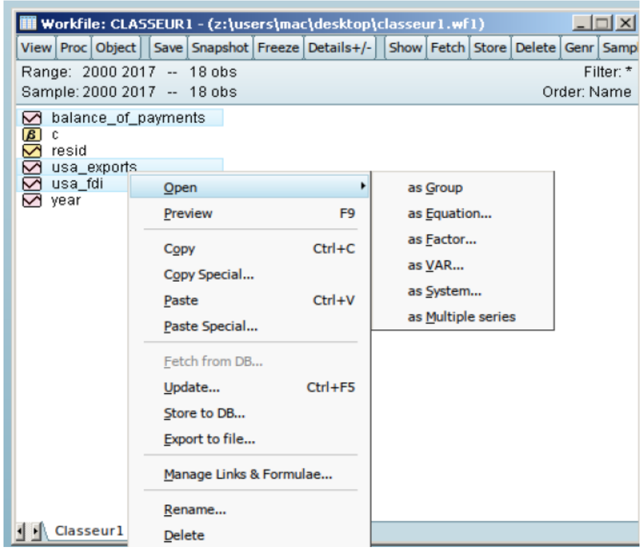
Then, The F-statistic referring to the F-statistic provided in Table 20, where the F-Statistic is 64.57 with a prob. (F-statistic) 0.0000%, which is below the significance level, further supports the rejection of the null hypothesis, emphasizing the significance of the F-statistic. This implies that at least one independent variable in the model possesses a non-zero coefficient, and the model effectively explains the variation in the dependent variable.

7. **Check if the the assumptions of linear regression are met (specify the hypothesis of each diagnostic test and how you conduct it).**

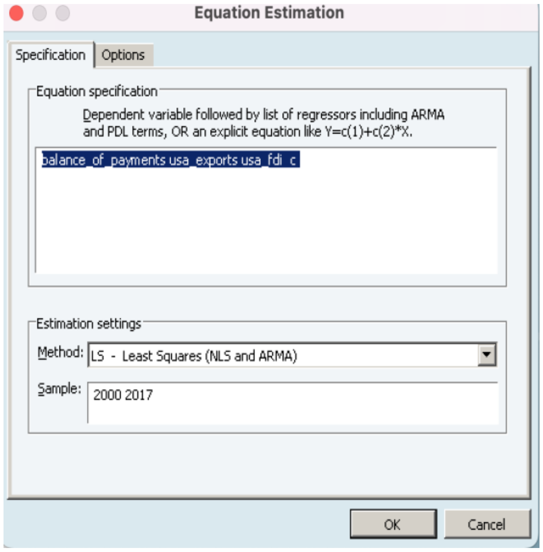
i. **Check for the normality:** To assess the normality of the data distribution, we can conduct the Jarque-Bera test. **Jarque–Bera (JB) Test of Normality:**

The null hypothesis of the Jarque-Bera test (H_0): Residuals are normally distributed.
The alternative hypothesis of Jarque-Bera test (H_a): Residuals are not normally distribute
To run a Jarque-Bera test in **EViews,** you can follow these steps(السواعي، 2011) :

1. After running the regression using OLS technique,



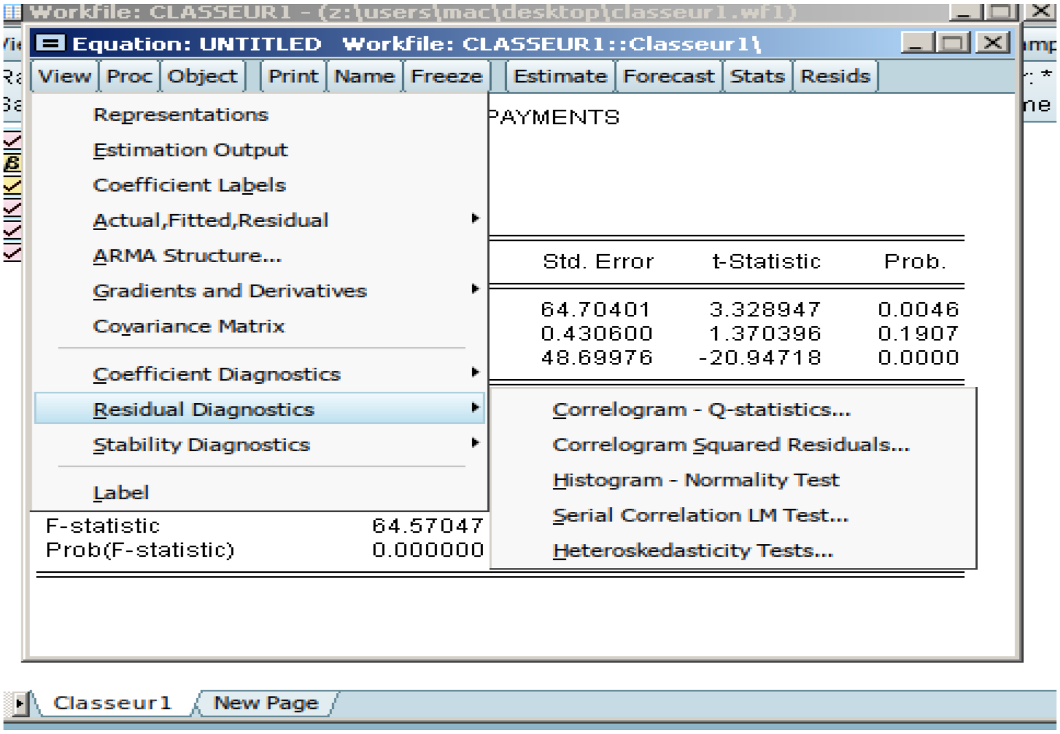
```
Range:  2000 2017  --  18 obs                Filter: *
Sample: 2000 2017  --  18 obs                Order: Name
balance_of_payments
c
resid
usa_exports
usa_fdi
year
```


Equation Estimation
Specification
Options
Equation specification
Dependent variable followed by list of regressors including ARMA and PDL terms, OR an explicit equation like Y=c(1)+c(2)*X.
Estimation settings
OK
Cancel

```
balance_of_payments usa_exports usa_fdi c
Method: LS - Least Squares (NLS and ARMA)
Sample: 2000 2017
```

2. Go to the " View" menu in EViews.

3. Navigate to " Residual Diagnostic" and click on " Histogram - Normality Test "

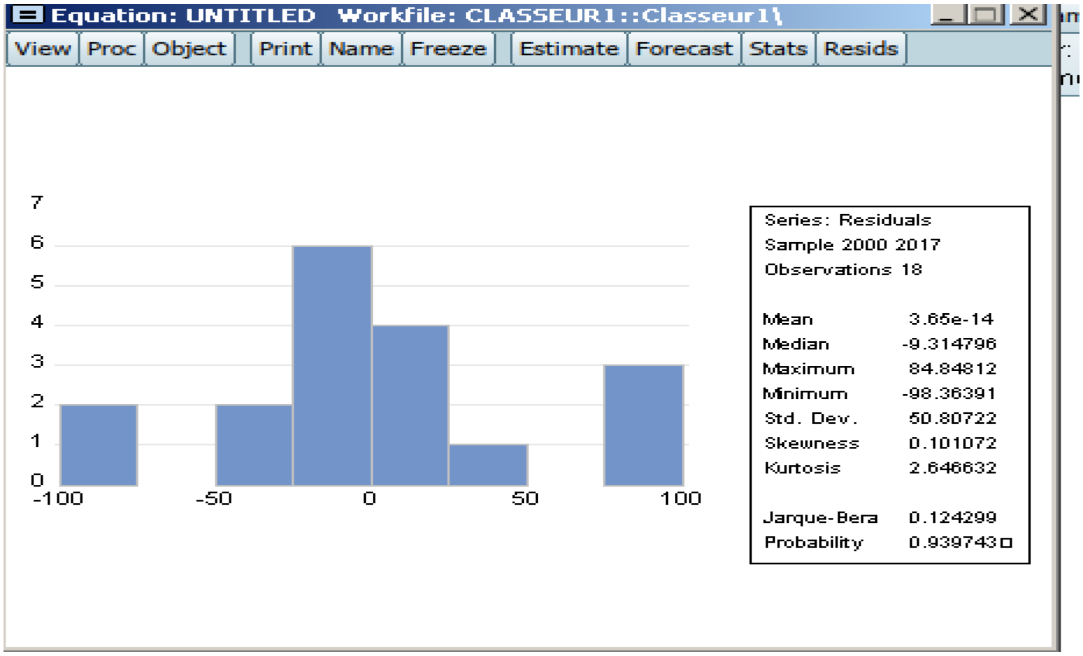

Workfile: CLASSEUR1 - (z:\users\mac\desktop\classeur1.wf1)
Equation: UNTITLED Workfile: CLASSEUR1::Classeur1\
View
Proc
Object
Print
Name
Freeze
Estimate
Forecast
Stats
Resids
Representations
Estimation Output
Coefficient Labels
Actual,Fitted,Residual
ARMA Structure...
Gradients and Derivatives
Covariance Matrix
Coefficient Diagnostics
Residual Diagnostics
Stability Diagnostics
Label
Correlogram - Q-statistics...
Correlogram Squared Residuals...
Histogram - Normality Test
Serial Correlation LM Test...
Heteroskedasticity Tests...
Classeur1
New Page

```
PAYMENTS

Std. Error    t-Statistic    Prob.

64.70401       3.328947      0.0046
0.430600       1.370396      0.1907
48.69976     -20.94718       0.0000

F-statistic            64.57047
Prob(F-statistic)      0.000000
```

4. EViews will compute the Jarque-Bera test statistic and its associated p-value.

5. The results will typically include the Jarque-Bera statistic, and p-value.

Figure 18: Histogram of residuals BoP- export, FDI



```
Series: Residuals
Sample 2000 2017
Observations 18

Mean         3.65e-14
Median      -9.314796
Maximum      84.84812
Minimum     -98.36391
Std. Dev.    50.80722
Skewness     0.101072
Kurtosis     2.646632

Jarque-Bera  0.124299
Probability  0.939743
```

Source: EViews output based on the data from our example

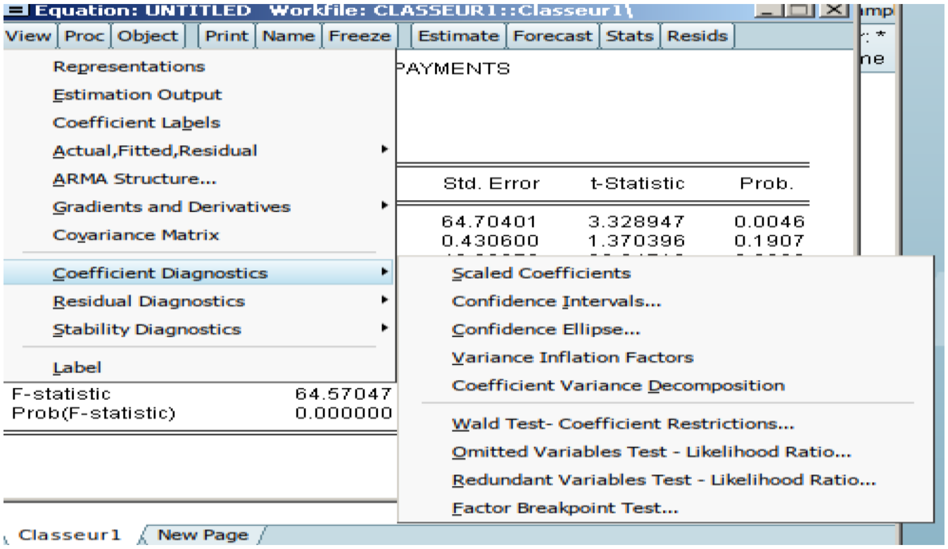
According to our results the probaility of the Jarque-Bera is 0.9397, which is greater than the 0.05 (the chosen significance level) this mean we fail to reject the null hypothesis suggesting that the residuals are normally distributed

Check for Multicollinearity: We can employ diverse approaches to assess multicollinearity, one of which involves computing variance inflation factors (VIFs). The null hypothesis and alternative hypothesis for VIF can be stated as follows:

Null Hypothesis (H0): There is no multicollinearity among the independent variables in the regression model.

Alternative Hypothesis (H1): There is multicollinearity among the independent variables in the regression model.

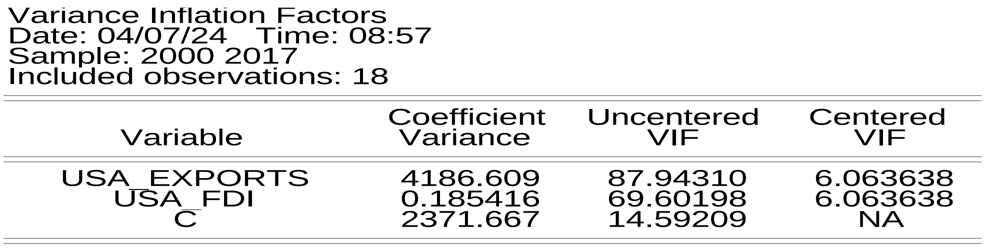
- After estimating the regression model, go to the "View" menu, select "Coefficient Diagnostics...", and then choose "Variance Inflation Factors".



```
PAYMENTS

Std. Error     t-Statistic     Prob.

64.70401       3.328947        0.0046
0.430600       1.370396        0.1907

F-statistic            64.57047
Prob(F-statistic)      0.000000
```

- The results will appear as in table 21

Table 21: variance inflation factor results



Variance Inflation Factors
Date: 04/07/24 Time: 08:57
Sample: 2000 2017
Included observations: 18

Variable	Coefficient Variance	Uncentered VIF	Centered VIF
USA_EXPORTS	4186.609	87.94310	6.063638
USA_FDI	0.185416	69.60198	6.063638
C	2371.667	14.59209	NA

Source: EViews output

The VIF values exceeding 5 indicate significant multicollinearity in the data. Table 21 displays all VIF (centered VIF) values surpassing this threshold.

ii. **Check for Homoskedasticity:** To check for homoskedasticity, we can use several diagnostic techniques, and the most commonly used is the Breusch-Pagan Test

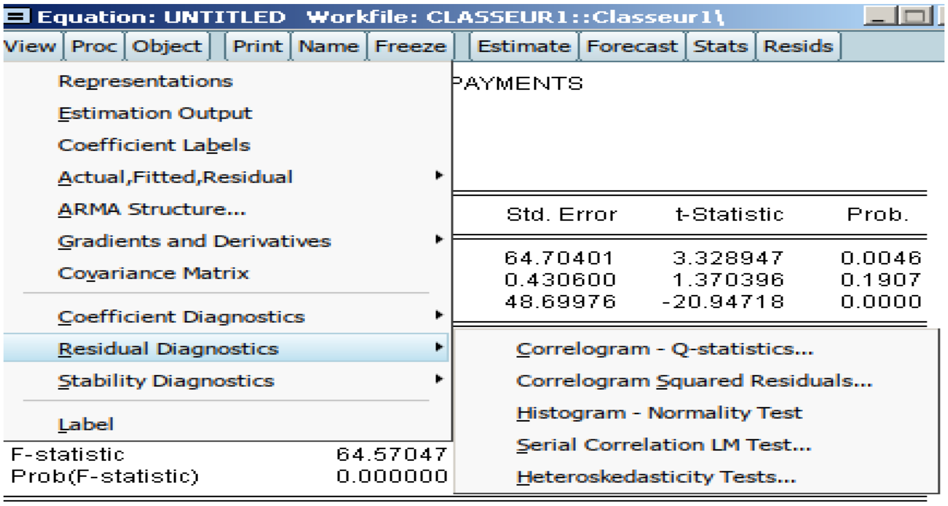
Breusch-Pagan test Hypothesis

The null hypothesis (H_0) of the Breusch-Pagan test: the variance of the errors is constant (homoskedasticity).

The alternative hypothesis (H_a): the variance of the errors is not constant (heteroskedasticity).

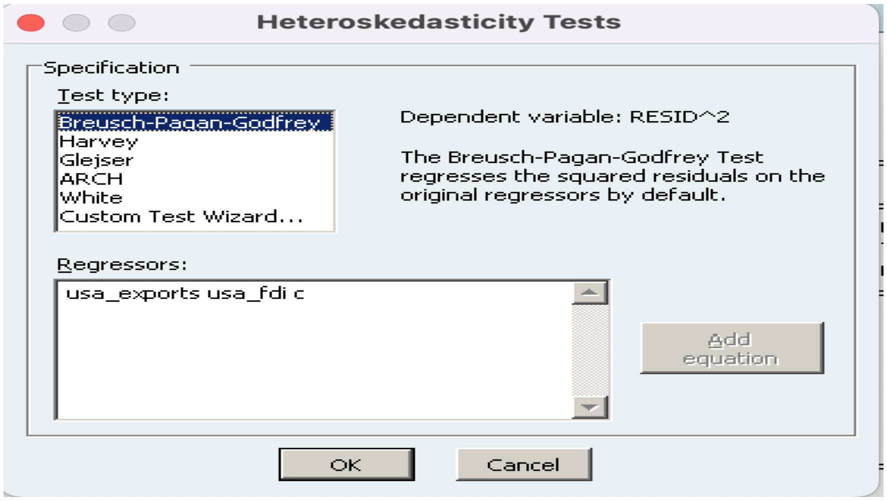
a. After running the regression, we will get the result of the OLS estimation;

b. Go to " View" menu in EViews; Navigate to " Residual Diagnostics" and click on " Heteroskedasticity Test "



```
PAYMENTS

Std. Error     t-Statistic     Prob.

64.70401       3.328947        0.0046
0.430600       1.370396        0.1907
48.69976      -20.94718        0.0000

F-statistic              64.57047
Prob(F-statistic)        0.000000
```

c. In the Options dialog box, go to the "Test types" select **Breusch-Pagan Test.** Click "OK" to run the regression.



Test type: Breusch-Pagan-Godfrey

Dependent variable: RESID^2

The Breusch-Pagan-Godfrey Test regresses the squared residuals on the original regressors by default.

Regressors: usa_exports usa_fdi c

d. Look for the section of the output that corresponds to the Breusch-Pagan test. The test results will include the f- statistic, its associated p-value, and other relevant information.

Table 22: heteroskedasticity test results

Heteroskedasticity Test: Breusch-Pagan-Godfrey
Null hypothesis: Homoskedasticity

F-statistic	5.587368	Prob. F(2,15)	0.0154
Obs*R-squared	7.684710	Prob. Chi-Square(2)	0.0214
Scaled explained SS	4.393711	Prob. Chi-Square(2)	0.1112

Source: EViews output

Based on the results in the table above (Table 22), the regression of squared residuals on the independent variables yields a significant R-squared value. This suggests that the assumption of homoskedasticity is not fulfilled. In addition, the p-value related to the F-statistic is less than the significance level. Therefore, we reject the null hypothesis of homoskedasticity, indicating evidence of a heteroskedasticity issue.

iii. **Check for autocorrelation: The Durbin-Watson test** is a statistical test used to detect the presence of autocorrelation in the residuals of a regression analysis.

v. **Durbin Watson test Hypothesis**: the null and alternative hypotheses for the Durbin-Watson test are as follows:

- **Null Hypothesis (*H*0):** There is no autocorrelation in the residuals.
- ***Alternative Hypothesis (Ha):*** There is autocorrelation in the residuals**.**

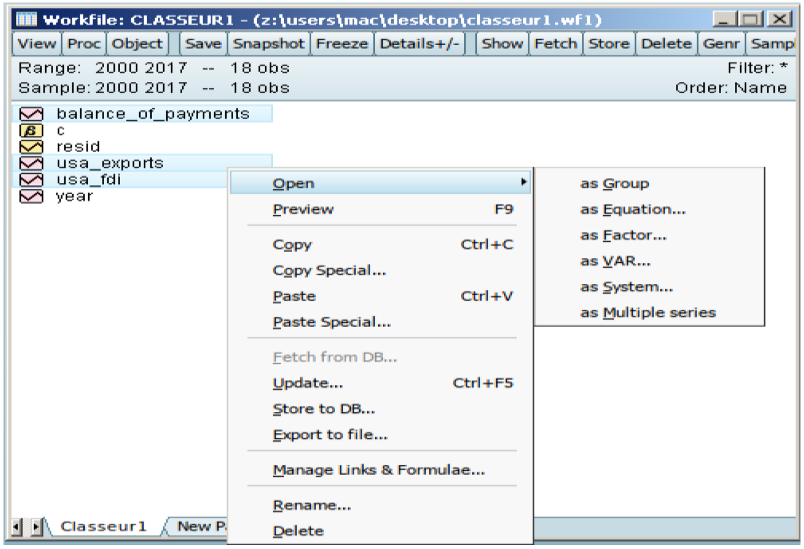
The Durban Watson statistic, dd, is provided in Table 22 of the OLS estimation results

(d=0.324779)

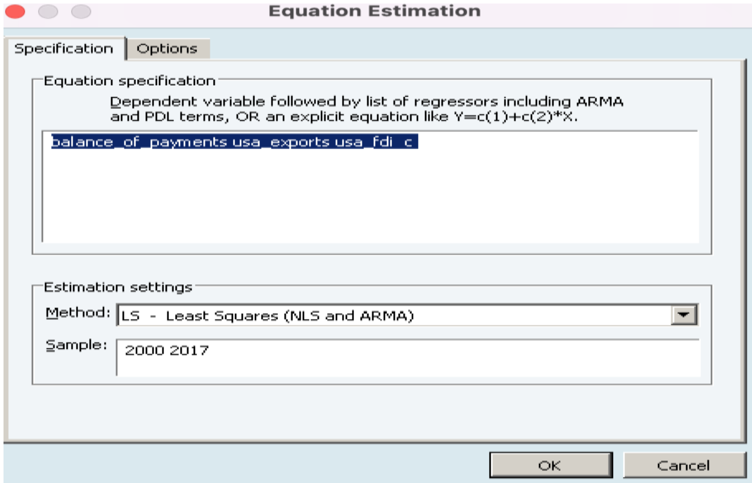
The OLS estimation is performed in EViews following these steps

Upload the data in EViews > Select the dependent variable & independent variables

Go to View > select Open > Open as equation > estimation equation dialog box will appear > Click OK



```
Workfile: CLASSEUR1 - (z:\users\mac\desktop\classeur1.wf1)
Range:  2000 2017  --  18 obs                                Filter: *
Sample: 2000 2017  --  18 obs                            Order: Name
```



Equation specification

Dependent variable followed by list of regressors including ARMA and PDL terms, OR an explicit equation like Y=c(1)+c(2)*X.

```
balance_of_payments usa_exports usa_fdi c
```

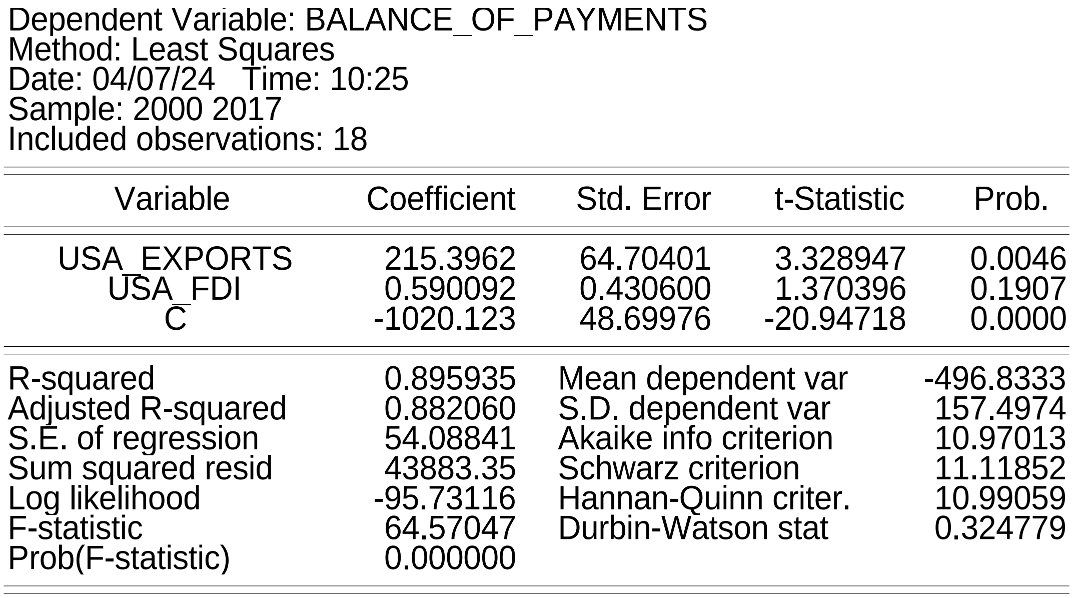
Estimation settings

Method: LS - Least Squares (NLS and ARMA)

Sample: 2000 2017

After performing the analysis, the EViews will display the result as in the table 22

Table 22: OLS estimation result



```
Dependent Variable: BALANCE_OF_PAYMENTS
Method: Least Squares
Date: 04/07/24   Time: 10:25
Sample: 2000 2017
Included observations: 18
==============================================================================
        Variable          Coefficient    Std. Error    t-Statistic    Prob.
==============================================================================
      USA_EXPORTS          215.3962      64.70401      3.328947      0.0046
        USA_FDI            0.590092      0.430600      1.370396      0.1907
           C              -1020.123      48.69976     -20.94718      0.0000
==============================================================================
R-squared                  0.895935    Mean dependent var          -496.8333
Adjusted R-squared         0.882060    S.D. dependent var           157.4974
S.E. of regression         54.08841    Akaike info criterion        10.97013
Sum squared resid          43883.35    Schwarz criterion            11.11852
Log likelihood            -95.73116    Hannan-Quinn criter.         10.99059
F-statistic                64.57047    Durbin-Watson stat           0.324779
Prob(F-statistic)          0.000000
==============================================================================
```

Source: EViews output

The Durbin Watson test, requires the use of tables which you can find in the appendix (see appendix A, Durban Watson significance table A4), since we have two regressor (Export, FDI) means k=2 and the sample size is n=18, so the critical value at 5% level of significance are: ***d_U* =1.259**, ***d_L* =0.805**

- According to our results the calculated Durbin-Watson statistic ***d*** is below the ***d_L***, so we reject the null hypothesis, then there is evidence that the data is positively autocorrelated (autocorrelation in the residual).

Due to the presence of autocorrelation, heteroskedasticity, and multicollinearity, the coefficients provided by the OLS may not be statistically reliable or interpretable. To obtain more reliable coefficients, it is necessary to address these issues through techniques such as transforming variables to mitigate multicollinearity or employing time-series methods to handle autocorrelation.

6. If the assumptions are not met, how you can correct the issues?

When the assumptions of linear regression are not met, several techniques can be employed to address the issues. Here are some common approaches:

1. **Autocorrelation:**
    - Use Autoregressive Integrated Moving Average (ARIMA) models or other time-series techniques to handle autocorrelation in the residuals (Shumway & Stoffer , 2011).
    - Employ generalized least squares (GLS) regression, which allows for correlated errors (Wooldridge, 2012).
2. **Heteroskedasticity:**
    - Utilize robust standard errors or White's heteroskedasticity-consistent standard errors to correct the standard errors of the coefficients (Wooldridge, 2012).
    - Perform weighted least squares (WLS), where weights are applied to observations to account for varying levels of error variance (Greene, 2017).
3. **Multicollinearity:**
    - Remove highly correlated independent variables from the model (Judge , Hill, Griffiths, & Lütkepohl, 1982).
    - Use dimensionality reduction techniques such as principal component analysis (PCA) or ridge regression to address multicollinearity (Johnson & Wichern, 2007).
    - Gather more data to increase the sample size, which can sometimes mitigate multicollinearity (Gareth , Witten, Hastie, & Tibshirani, 2017).

4. **Nonlinearity:**

    - Apply transformations to the independent or dependent variables, such as logarithmic, exponential, or polynomial transformations, to capture nonlinear relationships (Gareth , Witten, Hastie, & Tibshirani, 2017).

    - Consider using nonlinear regression techniques like polynomial regression or spline regression (Douglas & Watts , 1988).

Practice Questions

Exercise 1:

Consider that you examined the effect of US interest rates (*r_t*) and US Dollar exchange rates (*er_t*) on oil prices (*OP_t*), and the results you obtained are as follow:

Variable	Coefficient	Std. Error	t-Statistic
Const	98.67	16.45	……….
r_t	- 1.875	0. 793	………..
er_t	-3.651	0.9653	………….
R-Squared …………		F-Stat ………….. prob (f-stat) = 0.00001	
Adjusted R-Squared ………..		SSE ……….	
Observations 63		SSR 950	
		SST 1150	

1. Calculate the omitted values from the table.
2. Write the regression equation that showing the relationship between variables under the study.
3. Determine the study hypotheses regarding the relationship between the independent variables and the dependent variable
4. Test the hypothesis at 1% level of significance. Give the interpretation of the relationship between oil prices and the US interest rates and US Dollar exchange rates.
5. Did the two independent variables explain well the variation of the dependent variable?

Part three: Stationarity Autocorrelation and Partial Autocorrelation

Chapter 1: Stationarity and Non-stationarity

In time series analysis, stationarity is a fundamental concept that refers to the statistical properties of a time series data set remaining constant over time. A stationary time series is one where the mean, variance, and autocovariance structure do not change over time. Understanding stationarity is crucial for modeling and forecasting time series data accurately.

1-1- Concepts of stationarity and non-stationarity

Understanding the concepts of stationarity and non-stationarity is essential for analyzing and modeling time series data accurately. Stationarity simplifies the modeling process by providing stable statistical properties, while non-stationarity presents challenges and requires appropriate techniques for analysis and forecasting.

By recognizing the characteristics of stationarity and identifying deviations from stationarity in real-world data, analysts can select appropriate modeling techniques and make reliable predictions from time series data.

i. Stationarity

Stationarity refers to the statistical properties of a time series data set remaining constant over time. A stationary time series exhibits the following characteristics (Shumway & Stoffer , 2011):

a. **Constant Mean**: The mean of the time series remains constant over time. T This implies that there is no systematic trend in the data. Mathematically, it can be expressed as $E(Y_t) = \mu$, where Y_t represents the time series at time *t*, and μ is a constant.

b. **Constant Variance:** The variance of the time series remains constant over time. This implies that the magnitude of fluctuations around the mean remains consistent. Mathematically, it can be expressed as $Var\ (Y_t) = \sigma^2$ where σ^2

c. **Constant Autocovariance**: The covariance between any two observations in the time series only depends on the time lag between them and not on the specific points in time. This implies that the relationship between observations does not change over time. Mathematically, it can be expressed as $Cov\ (Y_t, Y_{t+k}) = \gamma(k)$ where $\gamma(k)$ is a function of the lag k and remains constant for all time points.

ii. **Non-stationarity:** Non-stationarity occurs when one or more of the characteristics of stationarity are violated. A non-stationary time series may exhibit trends, seasonality, changing variances, or changing autocorrelations over time. Some common forms of non-stationarity include (Shumway & Stoffer , 2011):

a. **Trend:** A systematic increase or decrease in the mean of the time series over time, indicating a long-term pattern or trend. Trends can be linear, quadratic, exponential, or more complex.

b. **Seasonality:** Periodic patterns or fluctuations in the time series at fixed intervals, such as daily, weekly, monthly, or yearly cycles. Seasonality often arises from external factors like weather, holidays, or business cycles.

c. **Heteroscedasticity:** Variation in the variance of the time series over time, where the magnitude of fluctuations around the mean varies. Heteroscedasticity can lead to unequal spread of data points and challenges in modeling.

d. **Autocorrelation:** Changes in the relationship between observations over time, where the autocovariance between observations depends on the specific points in time rather than just the time lag between them. Autocorrelation indicates that past observations influence future observations.

1-2- Tests for stationarity (Unit root test)

Unit root tests are statistical tests used to determine whether a time series is stationary or non-stationary, particularly in the context of identifying the presence of a unit root in the data. A unit root implies that a time series has a root that equals one, indicating that the series is non-stationary.

i. Unit Root Test:

Unit root tests are commonly employed to investigate the stationarity of time series data. The most widely used unit root test is the Augmented Dickey-Fuller (ADF) test. Other popular unit root tests include the Phillips-Perron (PP) test and the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test (Brooks, 2008).

a. **Augmented Dickey-Fuller (ADF) Test:** The ADF test is based on the Dickey-Fuller test and is augmented to allow for higher-order autoregressive terms in the model.

   The null hypothesis of the ADF test is that the time series contains a unit root, indicating non-stationarity. The alternative hypothesis is that the time series is stationary.

The Augmented Dickey-Fuller (ADF) test is a statistical test commonly used to determine whether a time series is stationary or non-stationary. It is an extension of the Dickey-Fuller test, which was developed to test for the presence of a unit root in a univariate time series (Brooks, 2008).

1) Key Components of the ADF Test:

- **Null Hypothesis (H0):** The null hypothesis of the ADF test is that the time series contains a unit root, indicating non-stationarity. In other words, if the null hypothesis cannot be rejected, it suggests that the time series is non-stationary.
- **Alternative Hypothesis (H1):** The alternative hypothesis of the ADF test is that the time series is stationary. Rejecting the null hypothesis in favor of the alternative hypothesis indicates that the time series is stationary.
- **Test Statistic:** The ADF test statistic is based on the regression of the differenced time series on lagged values of the time series itself. The test statistic is compared to critical values from the Dickey-Fuller distribution to determine whether to reject the null hypothesis.
- **Lags:** The choice of lag length in the ADF test is crucial and can affect the test's power. Commonly used lag selection criteria include Akaike Information Criterion (AIC), Schwarz Bayesian Criterion (SBC), and Hannan-Quinn Criterion (HQ).

2) Steps for Conducting the ADF Test:

Estimating the regression model for the Augmented Dickey-Fuller (ADF) test involves regressing the differenced time series on lagged values of the time series itself. The purpose of this regression is to capture the relationship between the current value of the differenced series and its lagged values, which helps in testing for the presence of a unit root.

Here are the steps to estimate the regression model for the ADF test (Brooks, 2008):

1. **Difference the Time Series:** Before estimating the regression model, you need to difference the original time series to make it stationary. This involves taking the first difference of the time series, which is calculated as $\Delta Y_t = Y_t - Y_{t-1}$, where Y_t represents the original time series and ΔY_t represents the differenced series.
2. **Select the Number of Lags:** Determine the appropriate number of lags to include in the regression model. The number of lags is crucial for capturing the autocorrelation

structure of the time series and ensuring the efficiency of the ADF test. Commonly used lag selection criteria include Akaike Information Criterion (AIC), Schwarz Bayesian Criterion (SBC), and Hannan-Quinn Criterion (HQ).

3. **Estimate the Regression Model:** Once you have determined the number of lags, regress the differenced time series on lagged values of itself. The regression equation takes the form: $\Delta Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots\cdots\cdots + \beta_k Y_{t-k} + \varepsilon_t$ Where:
   - ΔY_t is the differenced time series at time t.
   - α is the intercept term.
   - β_1, β_2 ,…, β_k are the coefficients of the lagged differences.
   - ε_t is the error term.

3) **Perform the ADF Test in EViews:** Performing an Augmented Dickey-Fuller (ADF) test using EViews software involves several steps. Here's a general guide on how to do it: (EViews, 2020)**:**

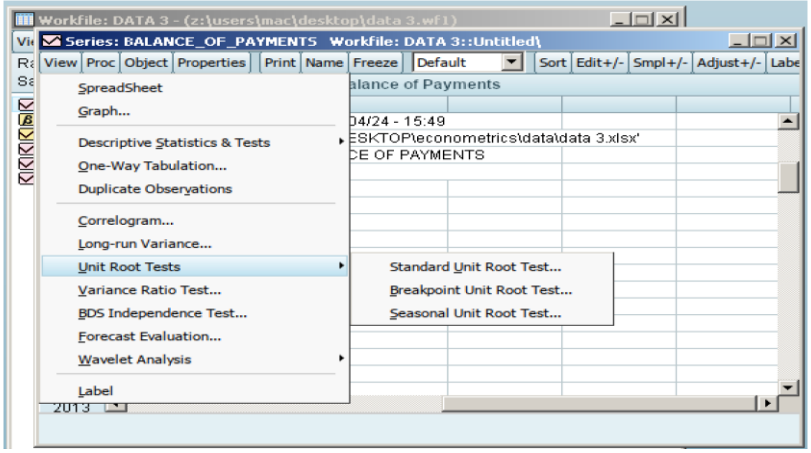
❖ **Import the Data**: First, you need to import your time series data into EViews. You can do this by opening EViews and importing it from a file (e.g., Excel, CSV).

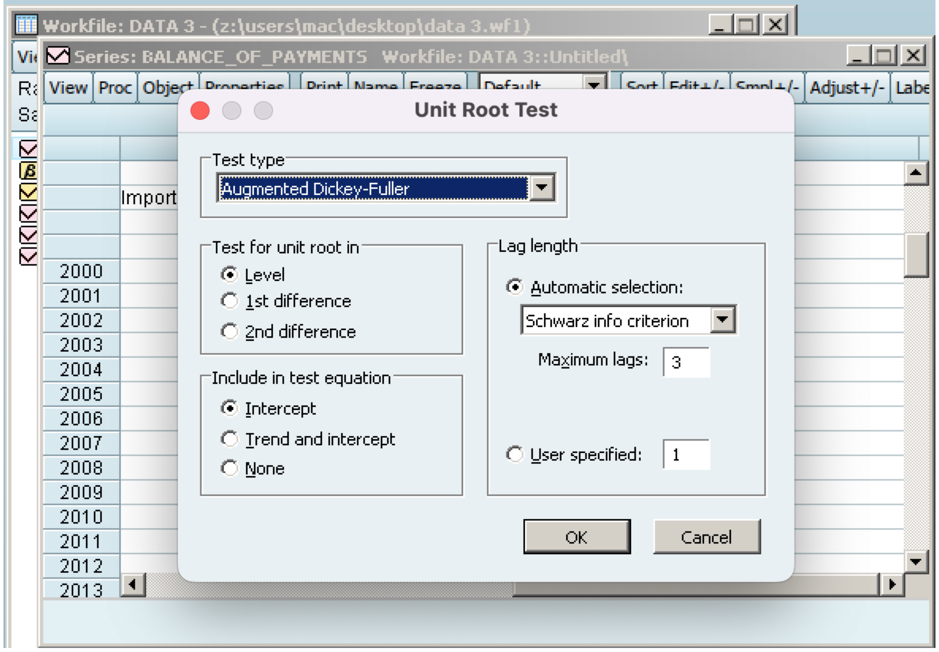
❖ **Open the Series**: Once your data is imported, open the series you want to perform the ADF test on. You can do this by double-clicking on the series name in the workfile window.

❖ **Run the ADF Test**: With the series window go to “View” menu at the top of the EViews window choose Unit root test. Then, under "Unit Root Tests", select "Augmented Dickey-Fuller Test". This will open a dialog box where you can specify the lag order, trend specification, and any other options for the ADF test.

❖ **Set Test Options**: In the dialog box, you can specify options such as the lag length to include in the test, whether to include a trend or intercept in the test equation, and the method for selecting the lag length (e.g., AIC, BIC).

Figure 19 : ADF test


Workfile: DATA 3 - (z:\users\mac\desktop\data 3.wf1)
Series: BALANCE_OF_PAYMENTS Workfile: DATA 3::Untitled\
View | Proc | Object | Properties | Print | Name | Freeze | Default | Sort | Edit+/- | Smpl+/- | Adjust+/- | Labe
SpreadSheet
Graph...
Descriptive Statistics & Tests
One-Way Tabulation...
Duplicate Observations
Correlogram...
Long-run Variance...
Unit Root Tests
Variance Ratio Test...
BDS Independence Test...
Forecast Evaluation...
Wavelet Analysis
Label
Standard Unit Root Test...
Breakpoint Unit Root Test...
Seasonal Unit Root Test...
alance of Payments
04/24 - 15:49
ESKTOP\econometrics\data\data 3.xlsx'
CE OF PAYMENTS
2013


Workfile: DATA 3 - (z:\users\mac\desktop\data 3.wf1)
Series: BALANCE_OF_PAYMENTS Workfile: DATA 3::Untitled\
View | Proc | Object | Properties | Print | Name | Freeze | Default | Sort | Edit+/- | Smpl+/- | Adjust+/- | Labe
Import
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
OK
Cancel

Unit Root Test

Test type: Augmented Dickey-Fuller

Test for unit root in
- ● Level
- ○ 1st difference
- ○ 2nd difference

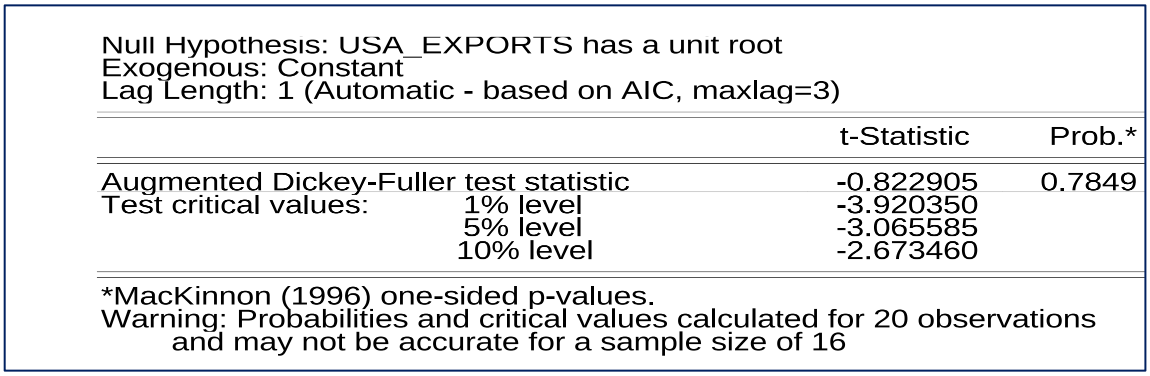
Include in test equation
- ● Intercept
- ○ Trend and intercept
- ○ None

Lag length
- ● Automatic selection: Schwarz info criterion
  - Maximum lags: 3
- ○ User specified: 1

Source : EViews 12 output

- **Run the Test**: After specifying the test options, click "OK" to run the ADF test. EViews will then perform the test and display the results in a new window.
- **Interpret the Results**: Once the test is complete, you can interpret the results to determine whether the series is stationary or not. Look at the test statistic and compare it to critical values from the ADF table. If the test statistic is less than the critical value, you can reject the null hypothesis of a unit root and conclude that the series is stationary.

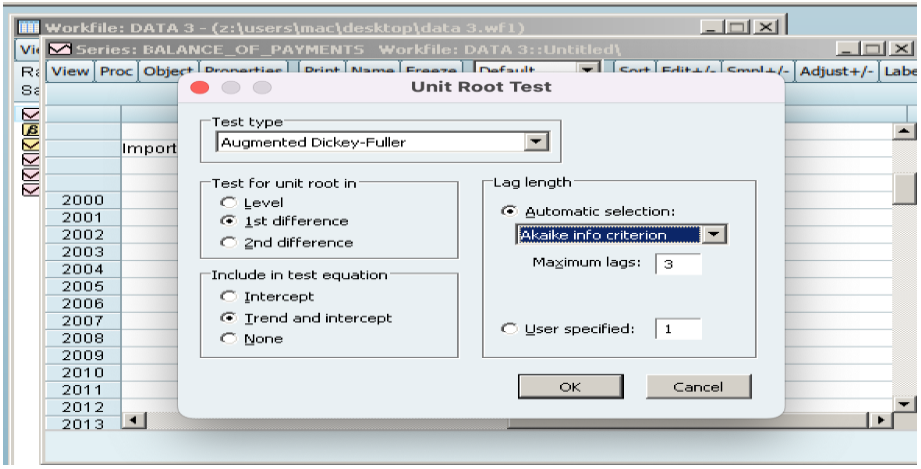
Table 20: ADF test result at level



```
Null Hypothesis: USA_EXPORTS has a unit root
Exogenous: Constant
Lag Length: 1 (Automatic - based on AIC, maxlag=3)

                                               t-Statistic     Prob.*

Augmented Dickey-Fuller test statistic         -0.822905      0.7849
Test critical values:        1% level          -3.920350
                             5% level          -3.065585
                            10% level          -2.673460

*MacKinnon (1996) one-sided p-values.
Warning: Probabilities and critical values calculated for 20 observations
      and may not be accurate for a sample size of 16
```

Source: EViews 12 output

According to the table above the absolute value of the computed t-statistic is less than the critical value, so, accept the null hypothesis and conclude that the time series is not stationary (or has Unit root). In this case we go to the first difference



Workfile: DATA 3 - (z:\users\mac\desktop\data 3.wf1)

Series: BALANCE_OF_PAYMENTS Workfile: DATA 3::Untitled\

Unit Root Test

Test type: Augmented Dickey-Fuller

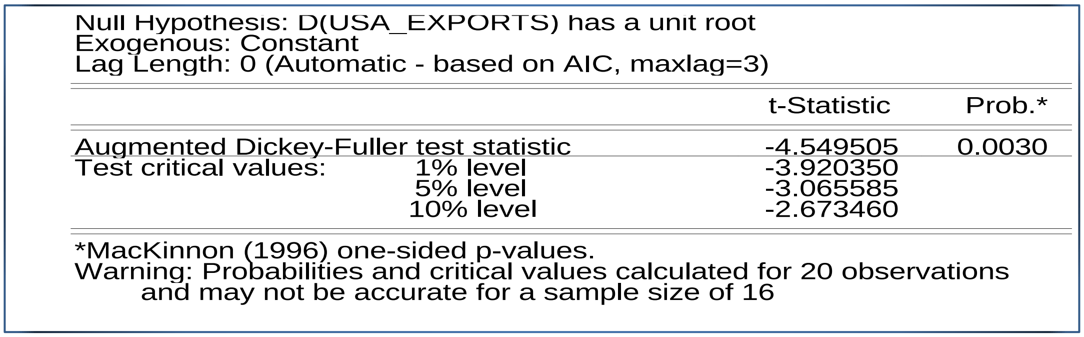
Test for unit root in
○ Level
● 1st difference
○ 2nd difference

Include in test equation
○ Intercept
● Trend and intercept
○ None

Lag length
● Automatic selection: Akaike info criterion
Maximum lags: 3
○ User specified: 1

Source: EViews 12 output

Table 21 : ADF test result at first difference



```
Null Hypothesis: D(USA_EXPORTS) has a unit root
Exogenous: Constant
Lag Length: 0 (Automatic - based on AIC, maxlag=3)

                                               t-Statistic     Prob.*

Augmented Dickey-Fuller test statistic         -4.549505      0.0030
Test critical values:        1% level          -3.920350
                             5% level          -3.065585
                            10% level          -2.673460

*MacKinnon (1996) one-sided p-values.
Warning: Probabilities and critical values calculated for 20 observations
      and may not be accurate for a sample size of 16
```

Source: EViews 12 output

According to the table the absolute value of the computed t-statistic is Greater than the critical value, so, reject the null hypothesis and conclude that the time series is stationary.

We can also use the p-value to interpret the result, in the table 20 the p-value related to the ADF statistic is greater than the chosen significance level 1%, 5% & 10% so we fail to reject the null hypothesis of Unit root; this suggests that the data exhibits non-stationarity. Conversely in the table 21, the p-value of ADF test is (0.0030 = 0.30%) is less than the selected significance level 1%, so we reject the null hypothesis and we conclude that the data is stationary at first difference at 1% level of significance. Specifically, the USA-Export data is integrated I(1).

b. **Phillips-Perron (PP) Test**: The PP test is similar to the ADF test but is based on different assumptions about the underlying process generating the data. Like the ADF test, the PP test is used to determine whether a time series contains a unit root. The PP test is another commonly used statistical test for unit root detection in time series data. Like the Augmented Dickey-Fuller (ADF) test, the PP test is used to determine whether a time series contains a unit root, indicating non-stationarity (Wooldridge, 2012).

1) **Key Components of the PP Test** (Greene, 2017)**:**

- **Null Hypothesis (H0):** The null hypothesis of the PP test is that the time series contains a unit root, indicating non-stationarity. In other words, if the null hypothesis cannot be rejected, it suggests that the time series is non-stationary.
- **Alternative Hypothesis (H1):** The alternative hypothesis of the PP test is that the time series is stationary. Rejecting the null hypothesis in favor of the alternative hypothesis indicates that the time series is stationary.
- **Test Statistic:** The PP test statistic is based on the regression of the differenced time series on lagged values of the time series itself, similar to the ADF test. However, the PP test uses a different estimation procedure and correction for serial correlation compared to the ADF test.
- **Lags:** As with the ADF test, the choice of lag length in the PP test is important and can affect the test's power. The lag length can be determined using lag selection criteria such as Akaike Information Criterion (AIC), Schwarz Bayesian Criterion (SBC), or other criteria.

2) **Differences between ADF and PP Tests** (Greene, 2017)**:**

While both the ADF and PP tests are used to test for the presence of a unit root in time series data, there are some differences between them:

1. **Estimation Procedure:** The ADF test uses an augmented regression model that includes lagged differences of the time series, while the PP test uses a different estimation procedure that involves computing the test statistic directly from the residuals of the regression.
2. **Serial Correlation Correction:** The PP test employs a different correction for serial correlation in the test statistic compared to the ADF test. This difference in correction procedures can lead to slightly different results between the two tests.
3. **Robustness:** The ADF test is often considered more robust to heteroscedasticity and autocorrelation in the residuals compared to the PP test. However, the PP test is still widely used and provides an alternative approach to unit root testing.

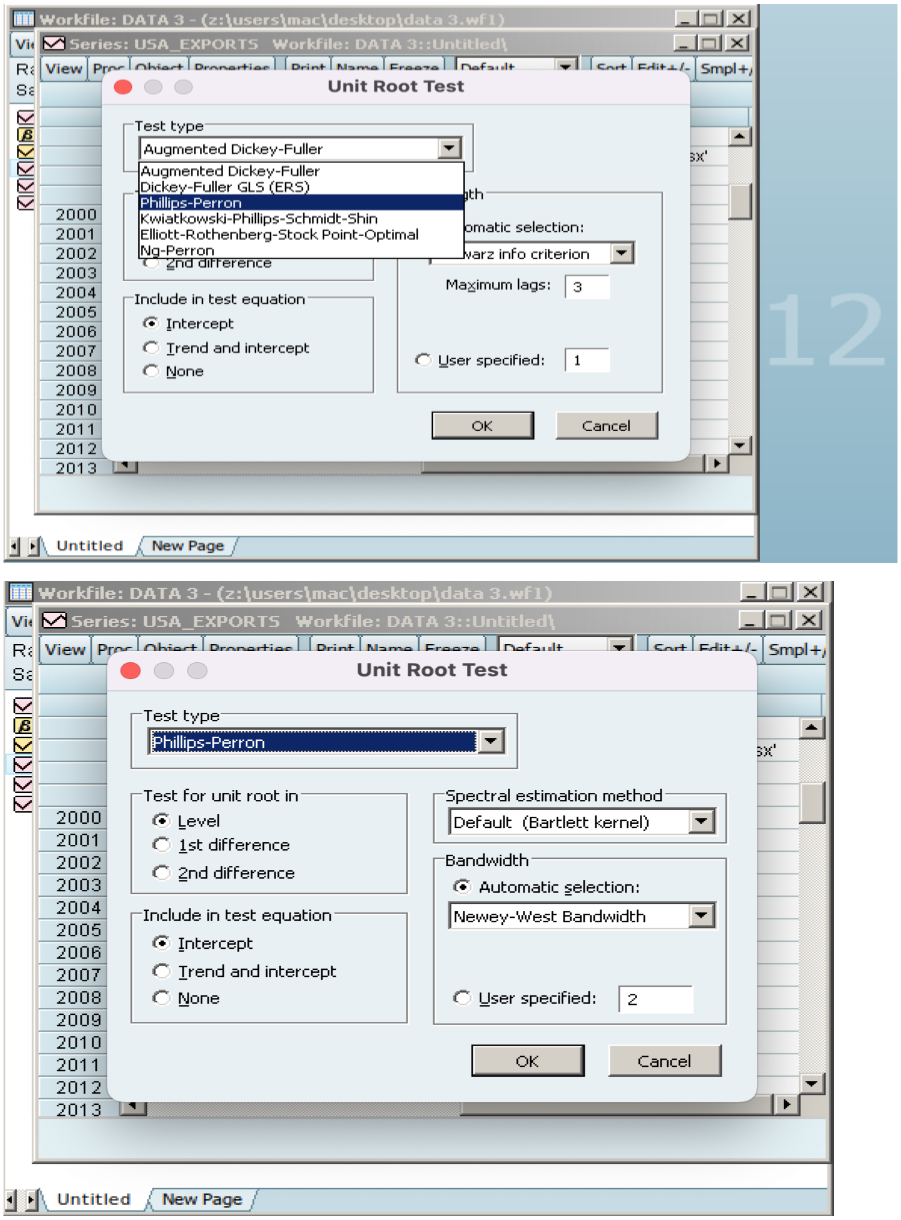
3) Perform the Phillips-Perron (PP) Test in EViews Software:

The Phillips-Perron (PP) test also formulates an autoregressive regression model, but it differs from the ADF test in terms of the variables included in the regression. The PP test uses the original level of the time series variable, rather than the differenced series, in the regression model. The general form of the regression model used in the PP test is simpler compared to the augmented regression model used in the ADF test (EViews, 2020).

Performing an Augmented Dickey-Fuller (ADF) test using EViews software involves several steps. Here's a general guide on how to do it: (EViews, 2020)**:**

- **Import the Data**: First, you need to import your time series data into EViews. You can do this by opening EViews and importing it from a file (e.g., Excel, CSV).
- **Open the Series**: Once your data is imported, open the series you want to perform the ADF test on. You can do this by double-clicking on the series name in the workfile window.
- **Run the ADF Test**: With the series window go to “View” menu at the top of the EViews window choose Unit root test. Then, under "Unit Root Tests", select "Phillips-Perron Test". This will open a dialog box where you can specify the lag order, trend specification, and any other options for the ADF test.
- **Set Test Options**: In the dialog box, you can specify options such as the lag length to include in the test, whether to include a trend or intercept in the test equation, and the method for selecting the lag length (e.g., AIC, BIC).

Figure 20 : Phillips Perron test



Unit Root Test

Test type: Augmented Dickey-Fuller

- Augmented Dickey-Fuller
- Dickey-Fuller GLS (ERS)
- Phillips-Perron
- Kwiatkowski-Phillips-Schmidt-Shin
- Elliott-Rothenberg-Stock Point-Optimal
- Ng-Perron

○ 2nd difference

Include in test equation

● Intercept
○ Trend and intercept
○ None

…th

…omatic selection: …warz info criterion

Maximum lags: 3

○ User specified: 1

OK Cancel

Unit Root Test

Test type: Phillips-Perron

Test for unit root in

● Level
○ 1st difference
○ 2nd difference

Include in test equation

● Intercept
○ Trend and intercept
○ None

Spectral estimation method: Default (Bartlett kernel)

Bandwidth

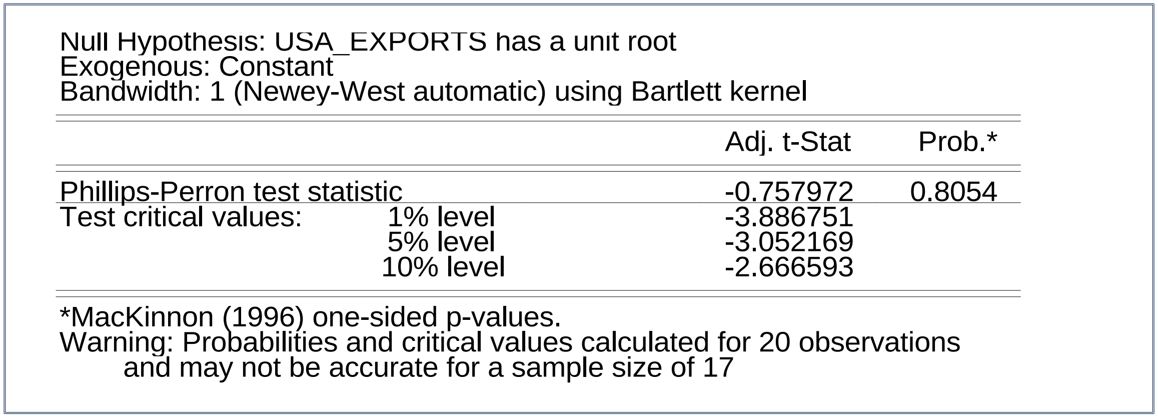
● Automatic selection: Newey-West Bandwidth
○ User specified: 2

OK Cancel

Source : EViews12 Output

- **Run the Test**: After specifying the test options, click "OK" to run the PP test. EViews will then perform the test and display the results in a new window.
- **Interpret the Results**: Once the test is complete, you can interpret the results to determine whether the series is stationary or not. Look at the test statistic and compare it to critical values from the PP table. If the test statistic is less than the critical value, you can reject the null hypothesis of a unit root and conclude that the series is stationary.

Table 22: Phillips-Perron Test Results at Level

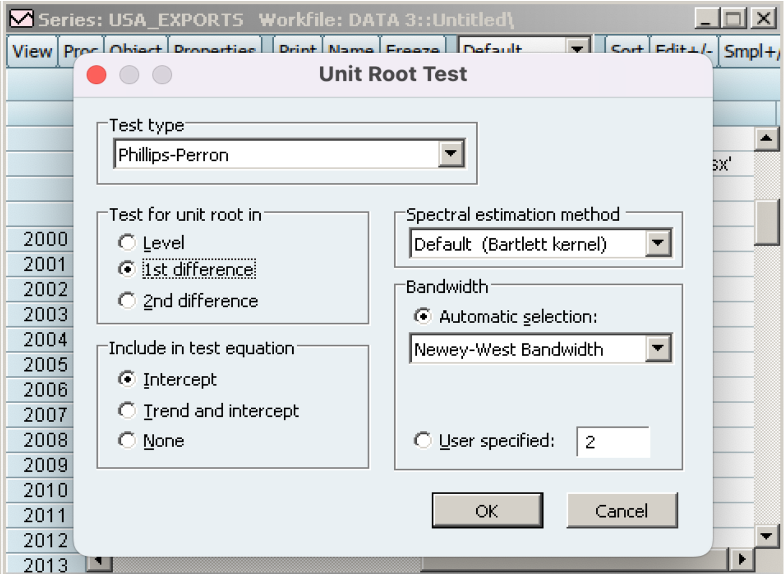


```
Null Hypothesis: USA_EXPORTS has a unit root
Exogenous: Constant
Bandwidth: 1 (Newey-West automatic) using Bartlett kernel

                                                  Adj. t-Stat     Prob.*

Phillips-Perron test statistic                    -0.757972      0.8054
Test critical values:          1% level           -3.886751
                               5% level           -3.052169
                               10% level          -2.666593

*MacKinnon (1996) one-sided p-values.
Warning: Probabilities and critical values calculated for 20 observations
      and may not be accurate for a sample size of 17
```

Source : EViews 12 output

According to the table above the absolute value of the computed t-statistic is less than the critical value, so, accept the null hypothesis and conclude that the time series is not stationary (or has Unit root). In this case we go to the first difference

Figure 21: Phillips-Perron test at first difference



Unit Root Test

Test type: Phillips-Perron

Test for unit root in
○ Level
● 1st difference
○ 2nd difference

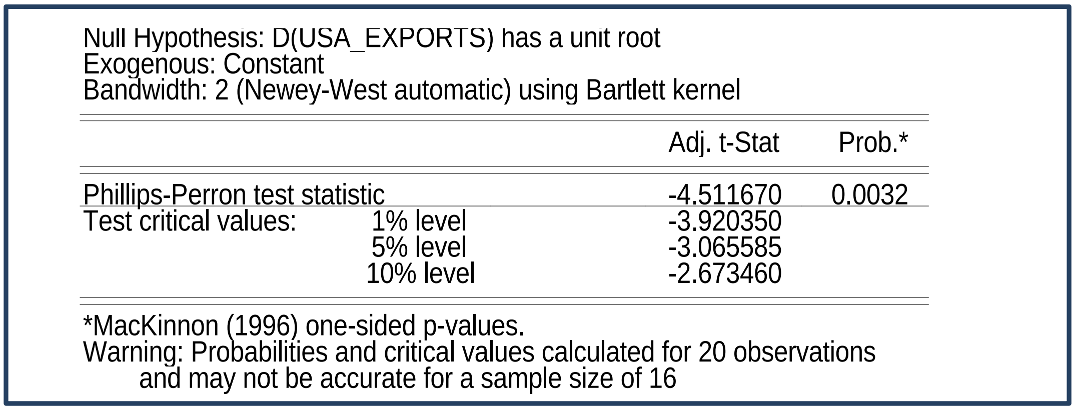
Include in test equation
● Intercept
○ Trend and intercept
○ None

Spectral estimation method: Default (Bartlett kernel)

Bandwidth
● Automatic selection: Newey-West Bandwidth
○ User specified: 2

Source: EViews 12 Output

Table 23: Phillips-Perron Test Results at First Difference



```
Null Hypothesis: D(USA_EXPORTS) has a unit root
Exogenous: Constant
Bandwidth: 2 (Newey-West automatic) using Bartlett kernel

                                                Adj. t-Stat     Prob.*

Phillips-Perron test statistic                  -4.511670       0.0032
Test critical values:        1% level           -3.920350
                             5% level           -3.065585
                             10% level          -2.673460

*MacKinnon (1996) one-sided p-values.
Warning: Probabilities and critical values calculated for 20 observations
        and may not be accurate for a sample size of 16
```

Source: EViews 12 Output

According to the table the absolute value of the computed t-statistic is Greater than the critical value, so, reject the null hypothesis and conclude that the time series is stationary at first difference at 1% level of significance.

We can also use the p-value to interpret the result, in the table 20 the p-value related to the ADF statistic is greater than the chosen significance level 1%, 5% & 10% so we fail to reject the null hypothesis of Unit root; this suggests that the data exhibits non-stationarity. Conversely in the table 21, the p-value of ADF test is (0.0030 = 0.30%) is less than the selected significance level 1% , so we reject the null hypothesis and we conclude that the data is stationary at first difference at 1% level of significance. Specifically, the USA-Export data is integrated I(1).

c. **Kwiatkowski-Phillips-Schmidt-Shin (KPSS) Test**: is another test used to test for stationarity around a deterministic trend. The null hypothesis of the KPSS test is that the time series is stationary, while the alternative hypothesis is that the time series is non-stationary (Kwiatkowski, Phillips, P. C. B. , Schmidt, & Shin, 1992).

1) Kwiatkowski-Phillips-Schmidt-Shin (KPSS) Test

The KPSS test is a statistical test used to determine the stationarity of a time series, specifically focusing on the presence of a deterministic trend. Unlike the Augmented Dickey-Fuller (ADF) and Phillips-Perron (PP) tests, which test for the presence of a unit root, the KPSS test examines whether the time series is stationary around a deterministic trend (Kwiatkowski, Phillips, P. C. B. , Schmidt, & Shin, 1992).

2) **Key Components of the KPSS Test:** based on the definition above we can summarize the component of KPSS test as follow:

- **Null Hypothesis (H0):** The null hypothesis of the KPSS test is that the time series is stationary around a deterministic trend. In other words, if the null hypothesis cannot be rejected, it suggests that the time series is stationary with respect to a trend.
- **Alternative Hypothesis (H1):** The alternative hypothesis of the KPSS test is that the time series is non-stationary. Rejecting the null hypothesis in favor of the alternative hypothesis indicates that the time series is non-stationary with respect to a trend.
- **Test Statistic:** The KPSS test statistic is computed based on the cumulative sum of squared deviations of the series from a deterministic trend. The test statistic is compared to critical values from the KPSS distribution to determine statistical significance.
- **Trend Specification:** The KPSS test allows for different specifications of the deterministic trend, including a linear trend, a quadratic trend, or a cubic trend. The choice of trend specification depends on the underlying characteristics of the time series (Kwiatkowski, Phillips, P. C. B. , Schmidt, & Shin, 1992).

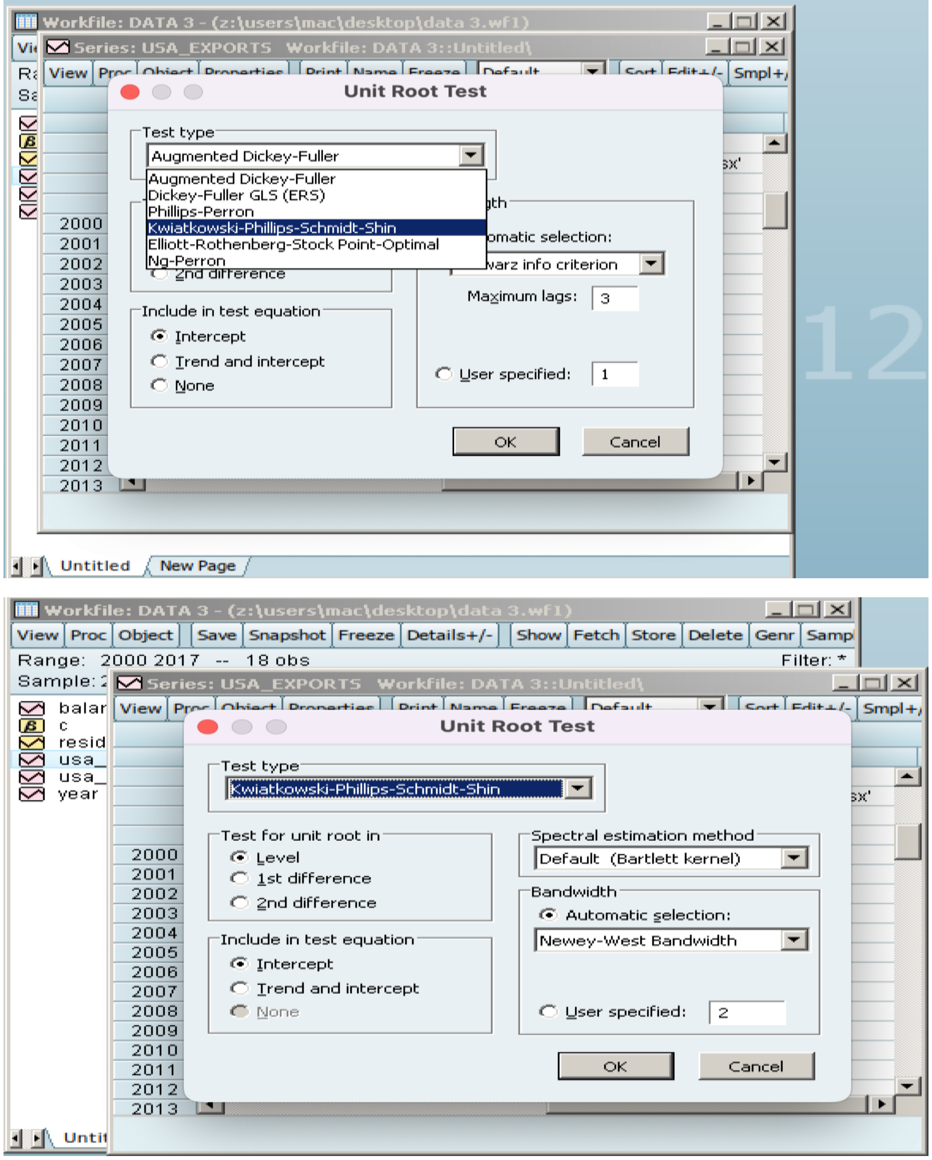
3) **Perform the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) Test in EViews Software:** Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test begins by formulating a regression model that includes a deterministic trend component. The trend component can be specified as a linear trend, quadratic trend, or cubic trend, depending on the chosen specification. (EViews, 2020).

Performing Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test using EViews software involves several steps. Here's a general guide on how to do it: (EViews, 2020)**:**

- **Import the Data**: First, you need to import your time series data into EViews. You can do this by opening EViews and importing it from a file (e.g., Excel, CSV).
- **Open the Series**: Once your data is imported, open the series you want to perform the ADF test on. You can do this by double-clicking on the series name in the workfile window.

- **Run the KPSS Test**: With the series window go to “View” menu at the top of the EViews window choose Unit root test. Then, under "Unit Root Tests", select " Kwiatkowski-Phillips-Schmidt-Shin (KPSS) ". This will open a dialog box where you can specify the lag order, trend specification, and any other options for the KPSS test.
- **Set Test Options**: In the dialog box, you can specify options such as the lag length to include in the test, whether to include a trend or intercept in the test equation, and the method for selecting the lag length (e.g., AIC, BIC).

Figure 23 : Phillips Perron test


Workfile: DATA 3 - (z:\users\mac\desktop\data 3.wf1)
Series: USA_EXPORTS Workfile: DATA 3::Untitled\
View Proc Object Properties Print Name Freeze Default Sort Edit+/- Smpl+/
Untitled New Page
View Proc Object Save Snapshot Freeze Details+/- Show Fetch Store Delete Genr Samp
OK
Cancel

Unit Root Test

Test type: Augmented Dickey-Fuller
- Augmented Dickey-Fuller
- Dickey-Fuller GLS (ERS)
- Phillips-Perron
- Kwiatkowski-Phillips-Schmidt-Shin
- Elliott-Rothenberg-Stock Point-Optimal
- Ng-Perron

Include in test equation: ● Intercept ○ Trend and intercept ○ None

Maximum lags: 3

○ User specified: 1

Range: 2000 2017 -- 18 obs Filter: *

Unit Root Test

Test type: Kwiatkowski-Phillips-Schmidt-Shin

Test for unit root in: ● Level ○ 1st difference ○ 2nd difference

Include in test equation: ● Intercept ○ Trend and intercept ○ None

Spectral estimation method: Default (Bartlett kernel)

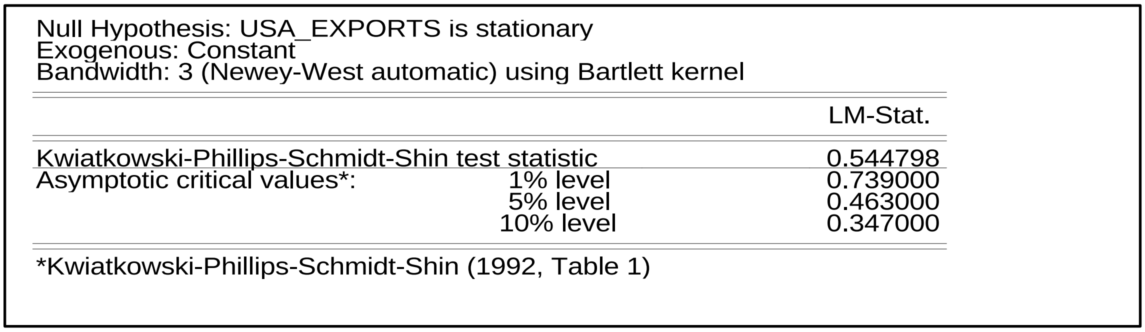
Bandwidth: ● Automatic selection: Newey-West Bandwidth

○ User specified: 2

Source: EViews 12 Output

- **Run the Test**: After specifying the test options, click "OK" to run the KPSS test. EViews will then perform the test and display the results in a new window.
- **Interpret the Results**: Once the test is complete, you can interpret the results to determine whether the series is stationary or not. Look at the test statistic and compare it to critical values from the PP table. If the test statistic is less than the critical value, you can reject the null hypothesis of a unit root and conclude that the series is stationary.

Table 24: Kwiatkowski-Phillips-Schmidt-Shin (KPSS)



```
Null Hypothesis: USA_EXPORTS is stationary
Exogenous: Constant
Bandwidth: 3 (Newey-West automatic) using Bartlett kernel

                                                              LM-Stat.

Kwiatkowski-Phillips-Schmidt-Shin test statistic              0.544798
Asymptotic critical values*:                  1% level        0.739000
                                              5% level        0.463000
                                             10% level        0.347000

*Kwiatkowski-Phillips-Schmidt-Shin (1992, Table 1)
```

Source: EViews 12 output

According to the table above the absolute value of the computed t-statistic is less than the critical value, so, accept the null hypothesis and conclude that the time series is stationary.

1-3- Techniques for achieving stationarity (differencing, detrending, etc.)

Achieving stationarity is essential in time series analysis because it allows for more reliable and accurate modeling of data. These techniques can be used alone or in combination depending on the characteristics of the time series data and the specific objectives of the analysis. Here are some techniques for achieving stationarity:

i. **Differencing**: This involves taking the difference between consecutive observations. It's a common technique to remove trend and seasonality from a time series. If the original series has a trend, differencing can help make it stationary. First-order differencing involves subtracting each observation from the one preceding it. Higher-order differencing involves repeating this process (Shumway & Stoffer, 2017).

ii. **Detrending**: Detrending involves removing the underlying trend from the time series data. This can be achieved through various methods such as fitting a regression line to the data and subtracting it, or using techniques like moving averages or LOESS (Locally Estimated

Scatterplot Smoothing) to smooth out the trend and then subtracting it from the original series (Montgomery , Jennings, & Kulahci, 2015).

iii. **Seasonal Adjustment**: If there's a seasonal component in the data, seasonal adjustment techniques such as seasonal decomposition can be used to remove it. Seasonal decomposition involves separating the time series into seasonal, trend, and residual components using methods like seasonal decomposition of time series (STL) or classical decomposition methods like X-11 or SEATS (Shumway & Stoffer, 2017).

iv. **Transformation**: Transforming the data using mathematical functions like logarithms, square roots, or Box-Cox transformations can stabilize the variance of the time series and make it more stationary.

v. **Rolling Statistics**: Computing rolling statistics like rolling mean and rolling standard deviation can help identify and remove trends and seasonality by subtracting them from the original series (Shumway & Stoffer, 2017).

vi. **Weighted Moving Averages**: Instead of giving equal weight to all observations, weighted moving averages assign different weights to different observations, giving more importance to recent observations. This can help in removing noise and identifying underlying trends (Montgomery , Jennings, & Kulahci, 2015).

vii. **Deseasonalization**: This involves removing the seasonal component from the time series data. This can be done by subtracting the seasonal component obtained from seasonal decomposition techniques like STL or by using seasonal adjustment methods (Montgomery , Jennings, & Kulahci, 2015).

viii. **Exponential Smoothing**: Exponential smoothing techniques like single exponential smoothing, double exponential smoothing (Holt's method), or triple exponential smoothing (Holt-Winters method) can be used to remove trends and seasonality from the data, making it stationary (Montgomery , Jennings, & Kulahci, 2015).

Chapter 2: Autocorrelation and Partial Autocorrelation

Autocorrelation and partial autocorrelation are statistical concepts commonly used in time series analysis to understand the relationship between observations at different time points. while autocorrelation measures the overall correlation between a time series and its past values, partial autocorrelation focuses on the specific and direct correlation between two time points, adjusting for the influence of other intermediate time points. Both autocorrelation and partial autocorrelation are essential tools in time series analysis for understanding the temporal dependencies within a dataset and for building accurate predictive models.

2-1- Definition and interpretation of autocorrelation (ACF)

Understanding autocorrelation is crucial in various fields such as economics, finance, meteorology, and signal processing. It helps in identifying trends, seasonality, and potential forecasting models for the time series data.

i. **Definition:** Autocorrelation (ACF), also known as serial correlation, is a statistical concept that measures the degree to which a time series is correlated with itself over successive time intervals. In simpler terms, it assesses the correlation between observations of a variable at different time points within the same series (Shumway & Stoffer, 2017).

The autocorrelation function (ACF) is a plot or mathematical function that illustrates the correlation coefficients at different lags. The lag represents the number of time intervals between each pair of observations being compared. A positive autocorrelation indicates that values at adjacent time points are positively correlated, while a negative autocorrelation implies an inverse relationship between adjacent values (Shumway & Stoffer, 2017).

ii. **Interpretation**:

The interpretation of autocorrelation depends on the context of the data being analyzed. In the case of time series analysis, autocorrelation can provide insights into the underlying structure and patterns within the data (Shumway & Stoffer, 2017). For example:

❖ **Positive Autocorrelation**: Indicates a pattern where high values tend to follow high values and low values follow low values. This suggests persistence or momentum in the series.

- ❖ **Negative Autocorrelation**: Suggests an alternating pattern where high values are followed by low values and vice versa. This might indicate a mean-reverting behavior.
- ❖ **Zero Autocorrelation**: Implies no discernible pattern or correlation between adjacent observations.

iii. **How to detect the autocorrelation in time series dataset:** Detecting autocorrelation involves analyzing a time series dataset to identify patterns of correlation between observations at different time points. Here's a step-by-step guide on how to detect autocorrelation:

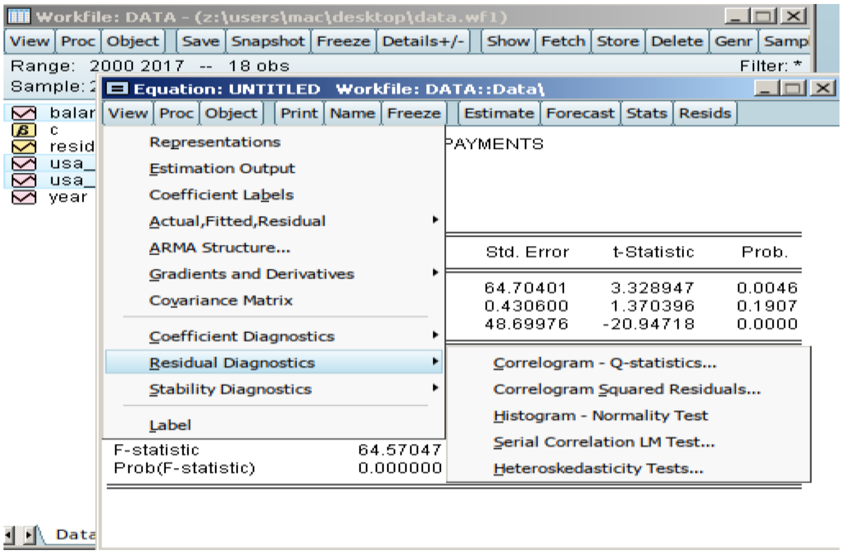
- ❖ **Visual Inspection**: Begin by visualizing your time series data. Plot the data over time to observe any apparent trends, cycles, or irregularities. Look for patterns where high or low values seem to repeat at regular intervals. This initial visualization can provide insights into potential autocorrelation (Shumway & Stoffer, 2017).
- ❖ **Autocorrelation Function (ACF)**: Conduct statistical tests to assess the significance of autocorrelation ; by using statistical software packages like EViews and Stata….. Common tests include the Durbin-Watson test, Ljung-Box test, and Breusch-Godfrey test, . These tests evaluate whether autocorrelation exists in the data and help determine whether the observed correlations are statistically significant (EViews, 2020).

- **Example of detecting autocorrelation using** Breusch-Godfrey test in EViews software: here's the steps to perform the Breusch-Godfrey test for autocorrelation in EViews:

1. **Run the Regression Model**: Start by estimating the regression model using your time series data. This could be a simple linear regression model (OLS) as **shown in the part two** or a more complex model depending on your analysis.

2. Navigate to the "View" menu in EViews and select "Residual Diagnostics" > "Serial Correlation LM Test (Breusch-Godfrey)".

Figure 24: Steps of performing Breusch-Godfrey test



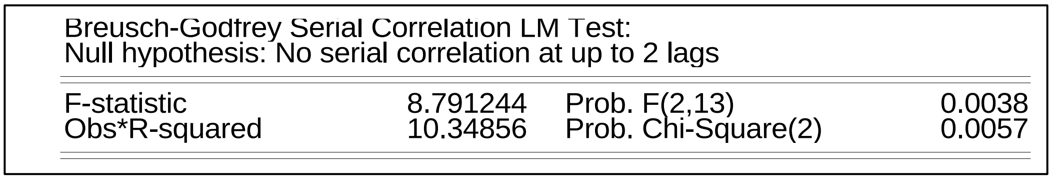
```
PAYMENTS

Std. Error    t-Statistic    Prob.

64.70401       3.328947      0.0046
0.430600       1.370396      0.1907
48.69976     -20.94718       0.0000

F-statistic              64.57047
Prob(F-statistic)        0.000000
```

Source: EViews 12 output

3. **Set Lag Length**: In the Breusch-Godfrey test dialog box, you will need to specify the maximum lag length for the test. This determines the number of lagged residuals included in the test. You can choose a lag length based on the characteristics of your data and the requirements of your analysis.

4. **Interpret Results**: After running the test, EViews will provide output displaying the test statistic, p-value, and other relevant information.

Table 25: Breusch-Godfrey test Results



```
Breusch-Godfrey Serial Correlation LM Test:
Null hypothesis: No serial correlation at up to 2 lags

F-statistic          8.791244    Prob. F(2,13)            0.0038
Obs*R-squared        10.34856    Prob. Chi-Square(2)      0.0057
```

Source : EViews Output

The p-value indicates the significance of the autocorrelation in the residuals. If the p-value is less than the chosen significance level (e.g., 0.05), we reject the null hypothesis of no autocorrelation and conclude that autocorrelation is present in the residuals ; conversely, if the p-value is greater than the chosen significance level (e.g., 0.05) we fail to reject the null hypothesis.

According to our results (**Table 25**) the p-value related to the F-statistic and Obs*R-squared is less than 5%, so we reject the null hypothesis which confirms the presence of Autocorrelation

❖ **Model Residual Analysis**: If you're working with a predictive model, analyze the residuals (the differences between observed and predicted values). Autocorrelation in the residuals indicates that the model may not be capturing all the underlying patterns in the data. Plotting the residuals against lagged values can reveal autocorrelation patterns (Cowpertwait & Metcalfe, 2009).

iv. **Correcting autocorrelation in time series dataset:** Here are several strategies to address autocorrelation:

❖ **Differencing**: One of the most common methods for mitigating autocorrelation is differencing. This involves computing the difference between consecutive observations in the time series. By removing the trend or seasonality, differencing can help reduce or eliminate autocorrelation. You can apply first-order differencing (subtracting each observation from its previous value) or higher-order differencing if needed.

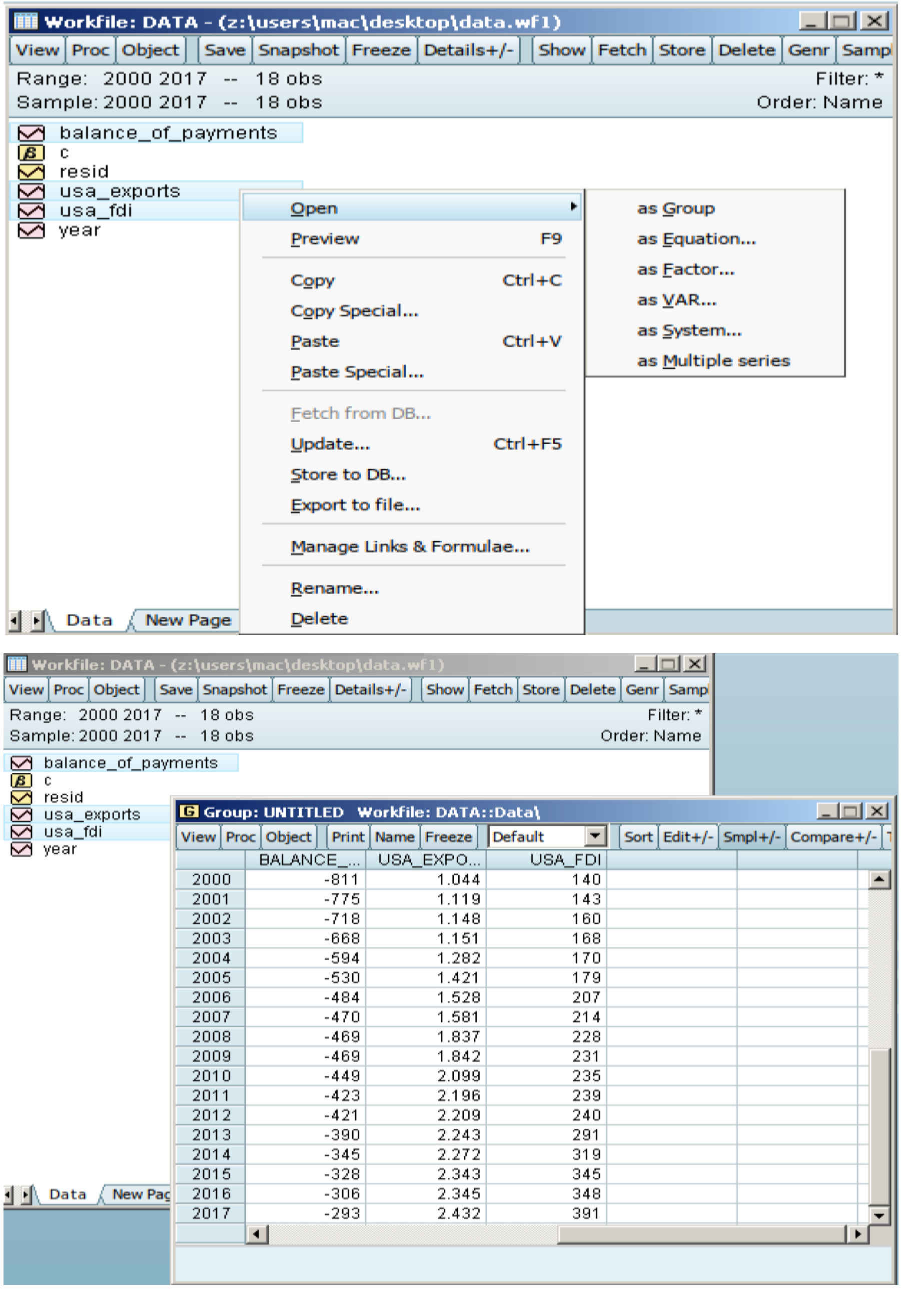
- **In Excel**, you can perform differencing to mitigate autocorrelation by computing the differences between consecutive observations in a time series data. Here's how you can do it (Moore, McCabe, & Craig's, 2009):

1) **Prepare Data**: Organize the time series data in an Excel spreadsheet. Ensure that the data is arranged in sequential order with each observation in a separate row or column.
2) **Create a New Column for Differenced Data**: Insert a new column next to the original data. This column will store the differenced values.
3) **Compute First-Order Differencing**: if your original data is in column A, and your new column starts from B2 ; so, in the first row of the new column you would enter
 the formula "=A2-A1" in cell B2.
4) **Fill Down the Formula**: Once the difference for the first observation have been computed Select the cell containing the formula (B2), and then double-click on the small square in the bottom-right corner of the cell, or drag it down to fill the formula for all observations.

- **In EViews**: In EViews, differencing can be performed easily using the built-in features for data transformation. Here's how you can do it:

- **Open Your Time Series Data**: Start by importing your data file into Eviews >> Select the data >> go to the open choose open as group >> than the EViews will show this results (figure 25)

Figure 25: How to open Dataset as group



Workfile: DATA - (z:\users\mac\desktop\data.wf1)
View Proc Object Save Snapshot Freeze Details+/- Show Fetch Store Delete Genr Samp
Open
Preview F9
Copy Ctrl+C
Copy Special...
Paste Ctrl+V
Paste Special...
Fetch from DB...
Update... Ctrl+F5
Store to DB...
Export to file...
Manage Links & Formulae...
Rename...
Delete
as Group
as Equation...
as Factor...
as VAR...
as System...
as Multiple series
Data
New Page
Group: UNTITLED Workfile: DATA::Data\
View Proc Object Print Name Freeze Default Sort Edit+/- Smpl+/- Compare+/-

Range: 2000 2017 -- 18 obs Filter: *
Sample: 2000 2017 -- 18 obs Order: Name

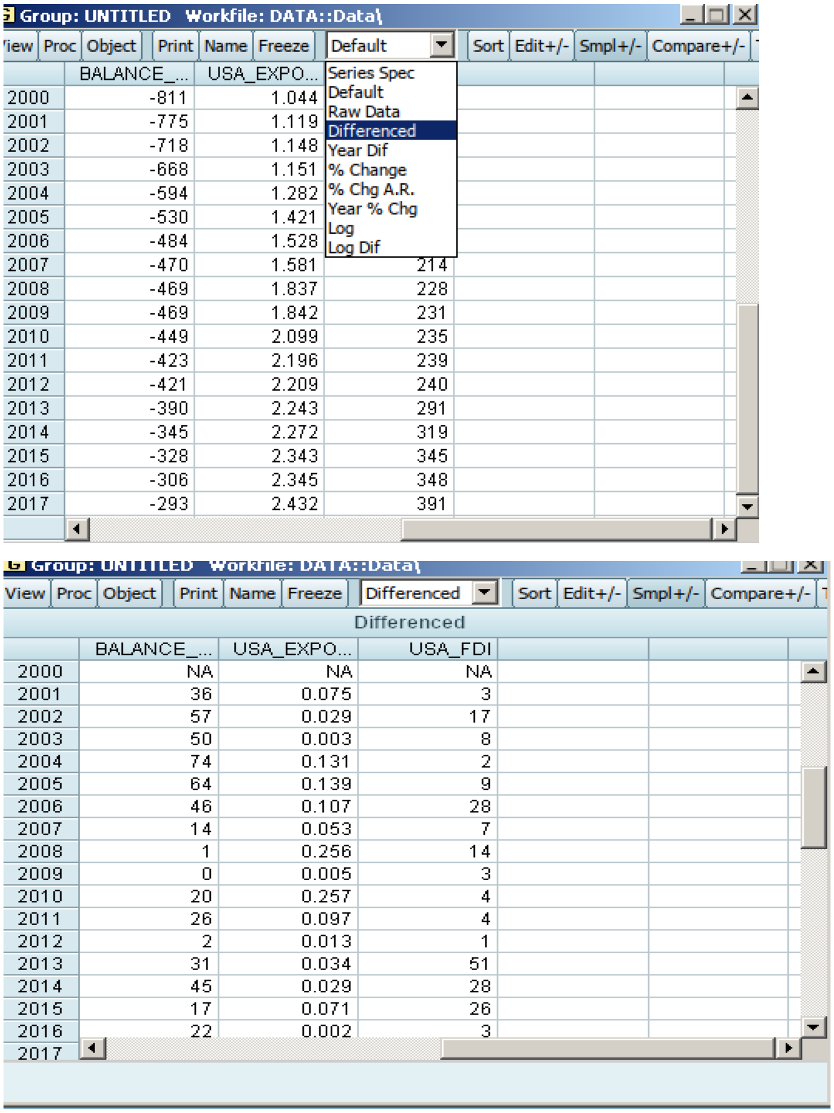
balance_of_payments
c
resid
usa_exports
usa_fdi
year

	BALANCE_...	USA_EXPO...	USA_FDI
2000	-811	1.044	140
2001	-775	1.119	143
2002	-718	1.148	160
2003	-668	1.151	168
2004	-594	1.282	170
2005	-530	1.421	179
2006	-484	1.528	207
2007	-470	1.581	214
2008	-469	1.837	228
2009	-469	1.842	231
2010	-449	2.099	235
2011	-423	2.196	239
2012	-421	2.209	240
2013	-390	2.243	291
2014	-345	2.272	319
2015	-328	2.343	345
2016	-306	2.345	348
2017	-293	2.432	391

Source: EViews output

When the workfile is open, go to 'Default', select the series you want to differenced and specify the order of differencing (e.g., first-order differencing). After selecting the appropriate options, EViews will automatically create a new series containing the differenced values (as shown in the figure 26)

Figure 26: Differenced data



	BALANCE_...	USA_EXPO...	
2000	-811	1.044	
2001	-775	1.119	
2002	-718	1.148	
2003	-668	1.151	
2004	-594	1.282	
2005	-530	1.421	
2006	-484	1.528	
2007	-470	1.581	214
2008	-469	1.837	228
2009	-469	1.842	231
2010	-449	2.099	235
2011	-423	2.196	239
2012	-421	2.209	240
2013	-390	2.243	291
2014	-345	2.272	319
2015	-328	2.343	345
2016	-306	2.345	348
2017	-293	2.432	391

Differenced

	BALANCE_...	USA_EXPO...	USA_FDI
2000	NA	NA	NA
2001	36	0.075	3
2002	57	0.029	17
2003	50	0.003	8
2004	74	0.131	2
2005	64	0.139	9
2006	46	0.107	28
2007	14	0.053	7
2008	1	0.256	14
2009	0	0.005	3
2010	20	0.257	4
2011	26	0.097	4
2012	2	0.013	1
2013	31	0.034	51
2014	45	0.029	28
2015	17	0.071	26
2016	22	0.002	3
2017			

Source: EViews Output

- **Detrending**: If the time series exhibits a clear trend, removing it can reduce autocorrelation. This can be achieved by fitting a trend line (e.g., linear regression) to the data and subtracting it from the original series. Once detrended, the residuals may exhibit less autocorrelation (Cowpertwait & Metcalfe, 2009).

  When detrending data, it's essential to consider the characteristics of the time series and choose the most appropriate method based on factors such as the presence of seasonality, the complexity of the trend, and the desired level of detail in the detrended series. Additionally, it's important to assess the effectiveness of detrending by examining the detrended series and evaluating whether it adequately captures the underlying patterns in the data.

- **Seasonal Adjustment**: If the time series exhibits seasonal patterns, removing the seasonal component can help reduce autocorrelation. Seasonal adjustment techniques, such as seasonal decomposition or seasonal differencing, can be applied to isolate and remove the seasonal effects from the data (Shumway & Stoffer, 2017).

Seasonal adjustment involves removing the seasonal component from a time series dataset to better analyze the underlying trend and irregular components. There are several methods for seasonal adjustment, including seasonal decomposition and seasonal differencing. Here's how you can perform seasonal adjustment using these methods:

1) **Seasonal Decomposition** (Shumway & Stoffer, 2017):

   - **Classical Decomposition**: This method decomposes the time series into seasonal, trend, and irregular components.

   - Compute the seasonal indices by averaging the observations for each season (e.g., each month for monthly data).
   - Divide the original time series by the seasonal indices to obtain the seasonally adjusted series.

     Example if we have monthly sales data for two years, assume the sales of January are 100 (in first year) & 210 (in second year),

     the seasonal indices: For January: (100 + 210) / 2 = 155

     then the seasonal series related to January would be 100/155= 0.645 for the first year and 210/155=1.354 for the second year

     Here's the formula for computing the seasonally adjusted series using seasonal indices:

     Let S_i denote the seasonal index for month i, and let Y_i denote the original sales value for month i. Then, the seasonally adjusted sales value for month i (SA_i) can be calculated as: $SA_i = \frac{S_i}{Y_i}$

Where:

- SA_i = Seasonally adjusted sales value for month *i*
- Y_i = Original sales value for month *i*
- S_i = Seasonal index for month *i*

  - **X-11 Method**: This method is more sophisticated and accounts for factors such as moving holidays and trading day effects.

  - Use specialized software or statistical packages (e.g., X-12-ARIMA, Census X-13ARIMA-SEATS) to perform the seasonal decomposition.
  - Obtain the seasonally adjusted series from the decomposition results

2. **Seasonal Differencing** (Hyndman & Athanasopoulos, 2018):

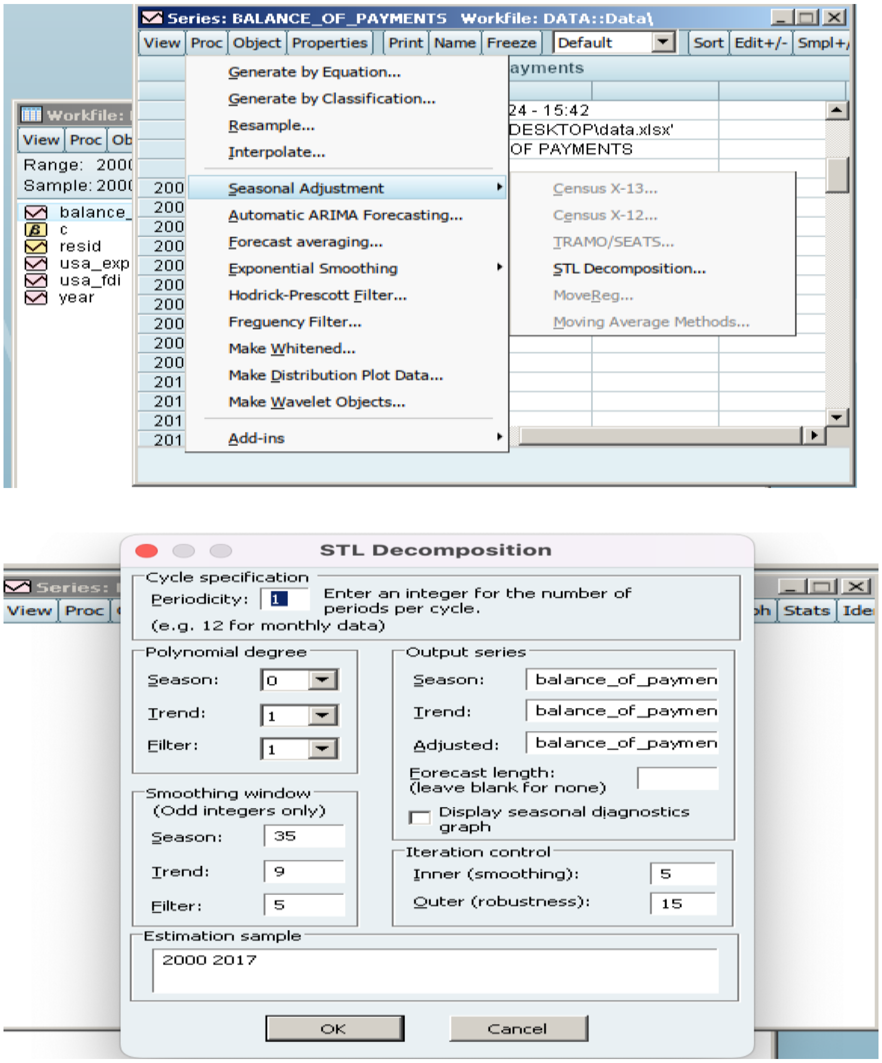
   - Take the difference between the original series and the seasonal lagged series (e.g., subtracting each observation from the corresponding observation from the same season in the previous year).

   - This method removes the seasonal pattern by subtracting the corresponding values from the previous seasonal cycle.

Here's a step-by-step guide to performing seasonal adjustment using seasonal using software like EViews (EViews, 2020):

- **Open Your Time Series Data:** Start by opening your time series data in EViews. You can do this by importing your data file into Eviews and select you dataset file double click on it (or go “View” menu choose Open selected).
- **After opening the dataset, go to** “Proc” menu choose Seasonal Adjustment then “STL Decomposition” then SLT decomposition window will open and enable you choose the option appropriate click ok then a new seasonally adjusted series will appear.

Figure 27: Seasonal adjustment in steps E-views



STL Decomposition

Cycle specification
Periodicity: 1 Enter an integer for the number of periods per cycle. (e.g. 12 for monthly data)

Polynomial degree
Season: 0
Trend: 1
Filter: 1

Output series
Season: balance_of_paymen
Trend: balance_of_paymen
Adjusted: balance_of_paymen
Forecast length: (leave blank for none)
□ Display seasonal diagnostics graph

Smoothing window (Odd integers only)
Season: 35
Trend: 9
Filter: 5

Iteration control
Inner (smoothing): 5
Outer (robustness): 15

Estimation sample
2000 2017

Source: EViews output

- **Autoregressive Integrated Moving Average (ARIMA) Modeling**: ARIMA models are specifically designed to handle autocorrelation in time series data. By incorporating autoregressive (AR), differencing (I), and moving average (MA) components, ARIMA models can capture and model autocorrelation patterns effectively. You can use techniques such as model diagnostics and parameter optimization to select the appropriate ARIMA model for your data (Hyndman & Athanasopoulos, 2018).
- **Adding Lagged Variables**: Including lagged values of the dependent variable as predictors in a regression model can account for autocorrelation. By capturing the effects

of past observations on the current value, lagged variables help model the autocorrelation structure in the data (Stock & Watson, 2020).

- **Exponential Smoothing**: Exponential smoothing methods, such as exponential moving averages (EMA), can be used to smooth the time series data and reduce autocorrelation. These methods assign exponentially decreasing weights to past observations, giving more weight to recent data while still considering historical values (Hyndman & Athanasopoulos, 2018).
- **Generalized Least Squares (GLS)**: GLS is a regression technique that accounts for autocorrelation by estimating a covariance matrix that incorporates the autocorrelation structure. GLS adjusts the standard errors of the coefficients to reflect the autocorrelation present in the data, providing more efficient and unbiased parameter estimates (Greene, 2017).

2-2- Definition and interpretation of partial autocorrelation (PACF)

i. **Definition:** The partial autocorrelation is a statistical concept used in time series analysis. It quantifies the direct relationship between observations at different time points, while controlling for the influence of other intermediate observations (Box , Jenkins, & Renis, 2015).

ii. Interpretation of PACF

Partial Autocorrelation (PACF): While autocorrelation measures the direct correlation between an observation and its lagged versions, partial autocorrelation measures the correlation between an observation and its lagged versions, controlling for the influence of other observations between them. Essentially, PACF gives you the correlation between two variables, with the influence of other variables removed. PACF values range between -1 and 1, where (Box , Jenkins,, & Reinsel, Time Series Analysis: Forecasting and Control, 2015):

- 1 indicates a perfect positive relationship.
- -1 indicates a perfect negative relationship.
- 0 indicates no relationship.

❖ In practice, partial autocorrelation coefficients are often displayed as a PACF plot, which shows the partial autocorrelation coefficients for different lags.

❖ In a PACF plot, significant spikes or bars indicate potential lags that are correlated with the current observation after controlling for other lags.

iii. **Compute Autocorrelation Function (PACF)**: Before calculating PACF, you need to compute the autocorrelation function (ACF) for the time series data. This can be done using statistical software packages like Stata, EViews….., or other similar tools. Once you have the ACF values, you can calculate the PACF using mathematical formulas or built-in functions in statistical software.

Detecting partial autocorrelation in time series data is crucial for identifying the relationship between observations separated by various lags, while controlling for the relationships at shorter lags. Several statistical tests and methods can be used to detect and measure partial autocorrelation. PACF plots are typically the starting point, providing a visual representation, while statistical tests like the Box-Ljung and Breusch-Godfrey tests offer formal hypothesis testing to confirm the presence of significant partial autocorrelation (Brockwell & Davis, Introduction to Time Series and Forecasting., 2016).

a. **Plot PACF:** After calculating the PACF values, it's common to plot them to visually inspect the partial autocorrelation structure of the time series. You can create a PACF plot where the x-axis represents the lagged time points and the y-axis represents the partial autocorrelation coefficients. Significant spikes or bars in the PACF plot indicate significant partial autocorrelation at those lagged time points (Shumway & Stoffer, 2017).

b. **Box-Ljung Test**: This test assesses whether a group of autocorrelations of a time series are different from zero. Apply the test to the residuals of a fitted model to check if any remaining partial autocorrelation is statistically significant, which suggests model inadequacy (Ljung & Box, 1978).

Objective: The Box-Ljung test evaluates whether there are significant autocorrelations in a time series up to a specified lag. If the test rejects the null hypothesis, it indicates that the time series has significant autocorrelations (Ljung & Box, 1978).

Test Statistic:

- The Box-Ljung test statistic Q is computed as (Ljung & Box, 1978):

$$Q = n(n+2)\sum_{k=1}^{m}\frac{\hat{\rho}_k^2}{n-k}$$

where:

- n is the sample size.
- $\hat{\rho}_k^2$ is the sample autocorrelation at lag k.
- m is the number of lags being tested.

This statistic approximately follows a chi-square distribution with m degrees of freedom under the null hypothesis that there is no autocorrelation up to lag mm.

Interpretation:

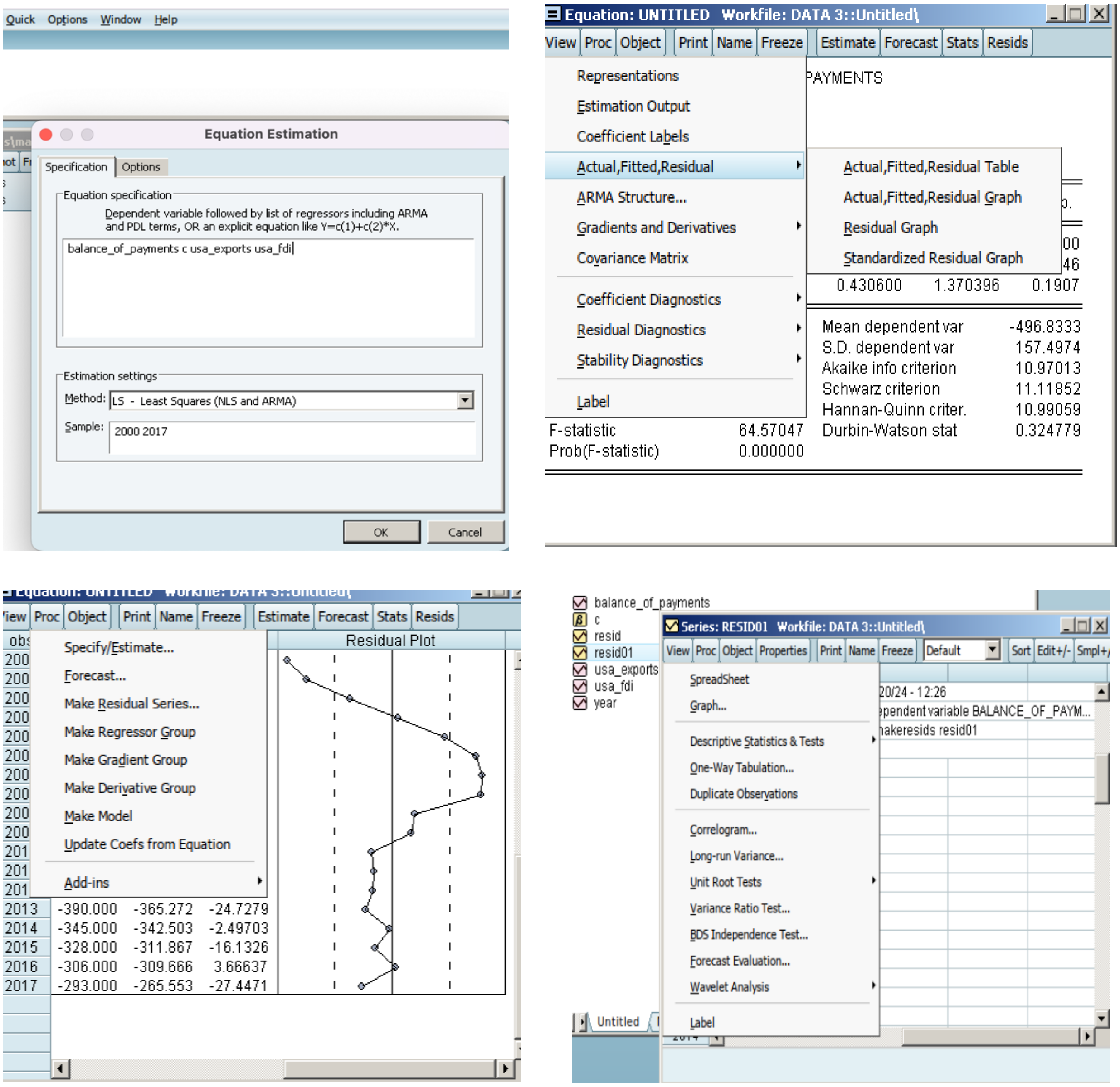
If the computed *Q* statistic is larger than the critical value from the chi-squared distribution with *m* degrees of freedom, the null hypothesis is rejected, indicating that the series has significant autocorrelations up to lag *m*.

Steps to Perform the Box-Ljung Test in EViews

EViews allows to perform various statistical tests, including the Box-Ljung test. Here's how to apply it (EViews, 2020):

- **Prepare Your Data**: Load the time series data into Eviews >> Ensure that the series is stationary. If not, apply differencing or other transformations.
- **Fit an Appropriate Model**: Go to **Quick** > **Estimate Equation** (Enter the appropriate ARIMA model specification or any other model you are fitting. Estimate the model)
- **Generate Residuals**: After estimating the model, go to **View** > **Actual, Fitted, Residuals**. Save the residuals by clicking on **Proc** > **Make Residual Series** and give a name to the residual series (e.g., **resid**).
- **Perform the Box-Ljung Test**: Select the residual series in the workfile. Go to **View** > **Correlogram**. In the Correlogram dialog box, select the **Box-Ljung Q-Statistic** option. Specify the number of lags you want to test (commonly, 20 lags are used). Click **OK**.

Figure 28: Steps of performing Box-Ljung Test



balance_of_payments c usa_exports usa_fdi

Method: LS - Least Squares (NLS and ARMA)

Sample: 2000 2017

```
PAYMENTS

                              0.430600      1.370396      0.1907

                              Mean dependent var      -496.8333
                              S.D. dependent var       157.4974
                              Akaike info criterion    10.97013
                              Schwarz criterion        11.11852
                              Hannan-Quinn criter.     10.99059
F-statistic         64.57047  Durbin-Watson stat       0.324779
Prob(F-statistic)   0.000000
```

```
obs
2013   -390.000   -365.272   -24.7279
2014   -345.000   -342.503   -2.49703
2015   -328.000   -311.867   -16.1326
2016   -306.000   -309.666    3.66637
2017   -293.000   -265.553   -27.4471
```

```
☑ balance_of_payments
β c
☑ resid
☑ resid01
☑ usa_exports
☑ usa_fdi
☑ year
```

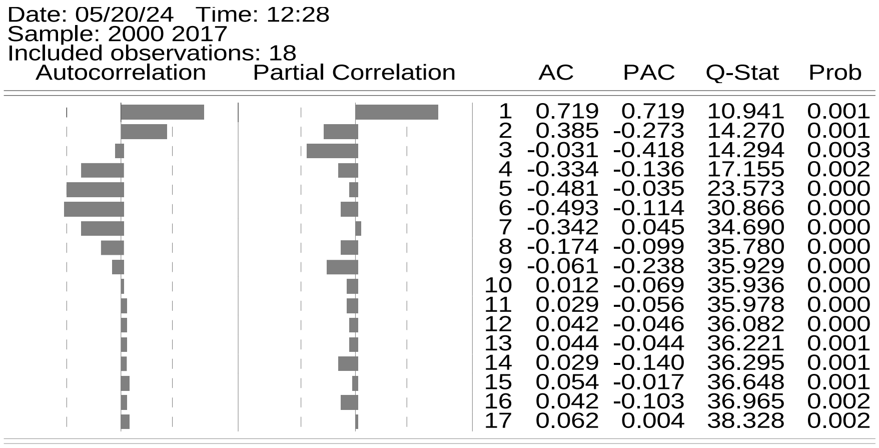
```
20/24 - 12:26
ependent variable BALANCE_OF_PAYM...
nakeresids resid01
```

Source : EViews output

❖ **Interpret the Results**:

- EViews will display a correlogram with autocorrelations, partial autocorrelations, and the Box-Ljung Q-statistics.

Table 26: Box-Ljung Test results



```
Date: 05/20/24   Time: 12:28
Sample: 2000 2017
Included observations: 18
 Autocorrelation   Partial Correlation        AC     PAC    Q-Stat   Prob

                                         1   0.719   0.719  10.941  0.001
                                         2   0.385  -0.273  14.270  0.001
                                         3  -0.031  -0.418  14.294  0.003
                                         4  -0.334  -0.136  17.155  0.002
                                         5  -0.481  -0.035  23.573  0.000
                                         6  -0.493  -0.114  30.866  0.000
                                         7  -0.342   0.045  34.690  0.000
                                         8  -0.174  -0.099  35.780  0.000
                                         9  -0.061  -0.238  35.929  0.000
                                        10   0.012  -0.069  35.936  0.000
                                        11   0.029  -0.056  35.978  0.000
                                        12   0.042  -0.046  36.082  0.000
                                        13   0.044  -0.044  36.221  0.001
                                        14   0.029  -0.140  36.295  0.001
                                        15   0.054  -0.017  36.648  0.001
                                        16   0.042  -0.103  36.965  0.002
                                        17   0.062   0.004  38.328  0.002
```

Source: EViews output

Check the Q-statistics and their corresponding p-values :

- If the p-value is less than your chosen significance level (e.g., 0.05), it indicates significant autocorrelation at that lag ; suggesting model inadequacy.

  According to the results in the above table all the p-values are less than the chosen significance level 5% or 0.05, this indicates significant autocorrelation and partial autocorrelation.

c. **Breusch-Godfrey Test**: Used to test for higher-order serial correlation in the residuals from a regression model. Apply this test to check for partial autocorrelation at various lags, beyond just the first lag.This test uses the following hypotheses:

***H_0* (null hypothesis):** There is no autocorrelation at any order less than or equal to *p*.
***H_a* (alternative hypothesis):** There exists autocorrelation at some order less than or equal to *p*.
The test statistic follows a Chi-Square distribution with *p* degrees of freedom.
If the p-value that corresponds to this test statistic is less than a certain significance level (e.g. 0.05) then we can reject the null hypothesis and conclude that

autocorrelation exists among the residuals at some order less than or equal to *p*. (Bobbitt, 2021).

- **Test Procedure**:

- Estimate the original regression model and obtain the residuals.
- Conduct an auxiliary regression where the residuals are regressed on the original independent variables and their lagged values up to the *p*-th order. This auxiliary regression captures how much of the residual's variation can be explained by past values (lags), similar to examining partial autocorrelations.
- Compute the LM (Lagrange Multiplier) statistic $n \times R^2$, where n is the number of observations and R^2 is from the auxiliary regression. This statistic follows a chi-squared distribution.
- By regressing the residuals on their own lags and the original regressors, the BG test effectively examines if past values (lagged residuals) add significant explanatory power, which is what partial autocorrelation measures. If the test finds significant serial correlation, it implies that the residuals are partially correlated at those lags.

Steps to Perform the Breusch-Godfrey Test in EViews

Here's how to conduct the Breusch-Godfrey test in EViews and interpret it in the context of partial autocorrelation (EViews, 2020):

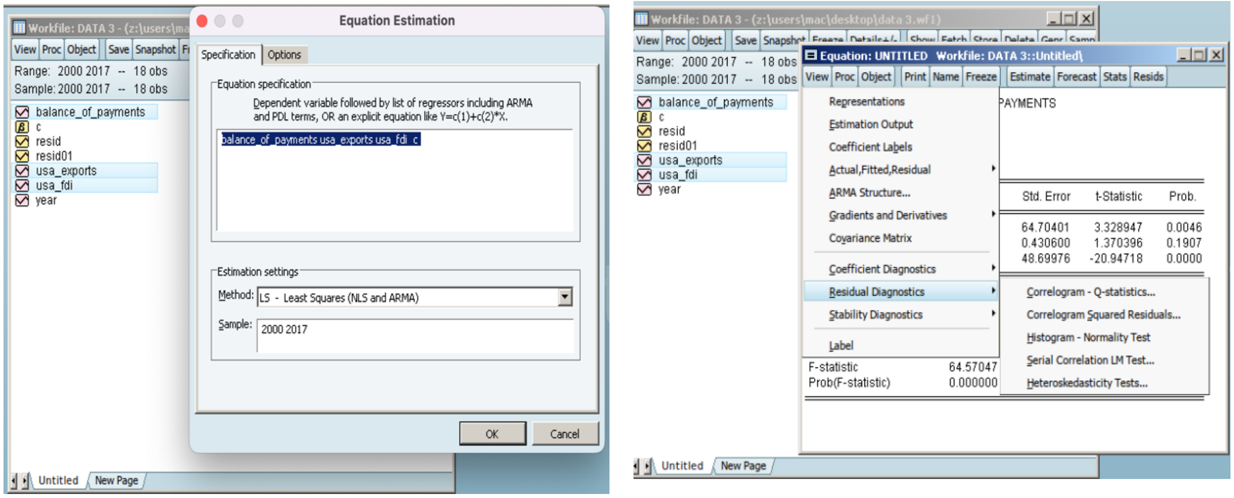
Estimate Your Regression Model:

- Load the dataset into EViews.
- Go to **Quick** > **Estimate Equation**.
- Specify your regression model (e.g., **Y c X1 X2**) and click **OK**.

Perform the Breusch-Godfrey Test:

- After estimating the model, go to **View** > **Residual Diagnostics** > **Serial Correlation LM Test**.
- In the Serial Correlation LM Test dialog box, specify the number of lags to test for (e.g., 2 for up to second-order serial correlation).
- Click **OK**.

Figure 29: Steps of performing the Breusch-Godfrey Test in EViews



Workfile: DATA 3 - (z:\users\mac\desktop\data 3.wf1)

Range: 2000 2017 -- 18 obs
Sample: 2000 2017 -- 18 obs

Equation Estimation

Equation specification: Dependent variable followed by list of regressors including ARMA and PDL terms, OR an explicit equation like Y=c(1)+c(2)*X.

balance_of_payments usa_exports usa_fdi c

Estimation settings

Method: LS - Least Squares (NLS and ARMA)

Sample: 2000 2017

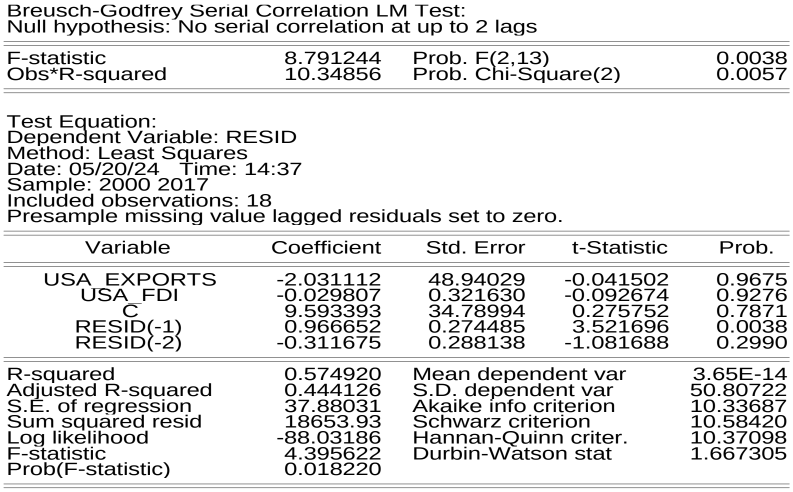
Equation: UNTITLED Workfile: DATA 3::Untitled\

	Std. Error	t-Statistic	Prob.
	64.70401	3.328947	0.0046
	0.430600	1.370396	0.1907
	48.69976	-20.94718	0.0000

F-statistic 64.57047
Prob(F-statistic) 0.000000

Source: EViews output

Table 27 : Breusch- Godfrey serial correlation Lm test



```
Breusch-Godfrey Serial Correlation LM Test:
Null hypothesis: No serial correlation at up to 2 lags

F-statistic           8.791244    Prob. F(2,13)              0.0038
Obs*R-squared         10.34856    Prob. Chi-Square(2)        0.0057

Test Equation:
Dependent Variable: RESID
Method: Least Squares
Date: 05/20/24   Time: 14:37
Sample: 2000 2017
Included observations: 18
Presample missing value lagged residuals set to zero.

      Variable         Coefficient    Std. Error    t-Statistic     Prob.

    USA_EXPORTS         -2.031112     48.94029     -0.041502      0.9675
      USA_FDI           -0.029807     0.321630     -0.092674      0.9276
         C               9.593393     34.78994      0.275752      0.7871
      RESID(-1)          0.966652     0.274485      3.521696      0.0038
      RESID(-2)         -0.311675     0.288138     -1.081688      0.2990

R-squared              0.574920    Mean dependent var       3.65E-14
Adjusted R-squared     0.444126    S.D. dependent var       50.80722
S.E. of regression     37.88031    Akaike info criterion    10.33687
Sum squared resid      18653.93    Schwarz criterion        10.58420
Log likelihood        -88.03186    Hannan-Quinn criter.     10.37098
F-statistic            4.395622    Durbin-Watson stat       1.667305
Prob(F-statistic)      0.018220
```

Source: EViews output

View and Interpret Results:

- EViews will display the test results, including the **Obs*R-squared** value and the corresponding p-value.
- If the p-value is less than your chosen significance level (e.g., 0.05), it indicates significant serial correlation up to the specified lag, suggesting partial autocorrelation at those lags.

According to the result of the table 30, the p-value related to Obs*R-squared and F-statistic are less than the chosen significance level 5%, this confirms the existence of serial correlation up to lag 2, suggesting partial autocorrelation at 2 lags.

d. **Yule-Walker Equations**: A set of equations used to estimate the parameters of an autoregressive model. These parameters are directly related to partial autocorrelations. Solve the Yule-Walker equations to obtain estimates of partial autocorrelations for a given autoregressive process (Brockwell & Davis, Introduction to Time Series and Forecasting., 2016).

Practice question

Exercise 1 :

1. **What does it mean for a time series to be stationary?**

a. The mean, variance, and autocorrelation structure of the series do not change over time.
b. The series has no trend or seasonal component.
c. The series values are bounded within a certain range.
d. The series is non-linear.

2. **Which of the following tests is commonly used to check for stationarity in a time series?**

a. Durbin-Watson test
b. Augmented Dickey-Fuller (ADF) test
c. Breusch-Pagan test
d. Jarque-Bera test

3. **Autocorrelation measures the correlation between:**

a. A time series and an independent series.
b. Two different time series at the same time period.
c. A time series and its lagged values.
d. A time series and its first differences.

4. **What does the autocorrelation function (ACF) plot help identify in a time series?**

a. The level of seasonality.
b. The degree of non-stationarity.
c. The lagged correlations of the time series.
d. The linearity of the series.

5. **The partial autocorrelation function (PACF) measures:**

a. The correlation between a time series and its lagged values after removing the effects of intermediate lags.
b. The overall trend in the time series.
c. The autocorrelation at different time lags.
d. The non-linear dependencies in the time series.

6. **If the autocorrelation function (ACF) of a time series decreases exponentially, what does this indicate about the series?**

a. The series is white noise.
b. The series has a unit root.
c. The series is likely an autoregressive process.

d. The series is likely a moving average process.

7. **Which of the following is true about a stationary time series?**

a. It can be transformed into a non-stationary series by differencing.
b. It often displays a clear trend over time.
c. It has constant mean and variance over time.
d. It is always periodic.

8. **In a partial autocorrelation function (PACF) plot of a pure autoregressive process, what is typically observed?**

a. Significant spikes at multiple lags.
b. Significant spikes at the first lag and tapering off afterwards.
c. Significant spikes at the first few lags and none afterwards.
d. No significant spikes at any lag.

9. **What characteristic is indicative of a non-stationary time series?**

a. ACF that drops to zero quickly.
b. Mean and variance that change over time.
c. ACF that shows no significant patterns.
d. PACF with significant spikes only at lag one.

10. **Why is it important to transform a non-stationary time series into a stationary one for modeling?**

a. To ensure that the time series can be used for linear regression.
b. To ensure that the properties of the series are consistent over time, allowing for reliable modeling and forecasting.
c. To reduce the computational complexity of the model.
d. To eliminate the need for differencing.

11. Which of the following statements about the Augmented Dickey-Fuller (ADF) test is true?
a) It is used to test for the presence of cointegration.
b) It is used to test for stationarity in a time series.
c) It is used to test for the presence of unit root.
d) Both b & c.

12. What is the null hypothesis in an Augmented Dickey-Fuller (ADF) test?
a) The time series is stationary.
b) The time series has a unit root.
c) The time series has no unit root test.
d) The time series is cointegrated.

13. How do you interpret the p-value obtained from the ADF test?

a) If the p-value is less than the significance level (e.g., 0.05), we reject the null hypothesis of non-stationarity.
b) If the p-value is greater than the significance level, we reject the null hypothesis of stationarity.
c) The p-value indicates the strength of the stationarity in the time series.
d) The p-value represents the test statistic divided by the number of observations.

14. What do the critical values in the ADF test indicate?

a) They represent the upper and lower bounds for the test statistic beyond which we reject the null hypothesis.
b) They represent the significance level at which the null hypothesis is rejected.
c) They measure the autocorrelation in the time series data.
d) They quantify the strength of the stationarity in the time series.

15. Which of the following statements about unit root tests is true?

a) A unit root indicates that a time series is stationary.
b) A unit root indicates that a time series is non-stationary.
c) Unit root tests are used to test for the presence of trends in a time series.
d) A significant p-value in a unit root test implies stationarity.

16. Which test is commonly used to detect the presence of a unit root in a time series?

A) Engle-Granger test
B) Augmented Dickey-Fuller (ADF) test
C) Johansen test
D) Granger causality test

Exercise 2:

Suppose you have conducted an Augmented Dickey-Fuller (ADF) test to assess the stationarity of a time series data. After running the test, you obtain the following results table:

Lag length	**3**		
Obs. 18		**t-statistic**	**Prob.***
Augmented dikey-Fuller test statistic		-0.742941	0.8096
Test critical values	1% level	-3.886751	
	5% level	-3.052169	
	10% level	-2.666593	

Based on the given results, answer the following questions:

1. What is the test statistic obtained from the ADF test?
2. What is the p-value associated with the test?
3. How many lags were used in the test?
4. How many observations were included in the analysis?
5. What are the critical values at the 1%, 5%, and 10% significance levels?

Exercise 3: Assume we have monthly sales data for 24 months:

Month	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Sales	100	110	105	120	130	125	135	140	145	150	155	160
Month	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**
Sales	165	170	175	180	185	190	195	200	205	210	215	220
Month	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**	**36**
Sales	225	230	235	240	245	250	255	260	265	270	275	280

- Perform an Augmented Dickey-Fuller (ADF) test on this time series data to determine if there's a unit root present (Define the null hypothesis (H0) and alternative hypothesis (H1). Calculate the ADF test statistic) .
- Compare the ADF test statistic with the critical value to make a decision. Interpret the result.

Answers

Exercise 1

question	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Answer	*a*	*b*	*c*	*c*	*a*	*c*	*c*	*c*	*b*	*b*
question	*11*	*12*	*13*	*14*	*15*	*16*				
Answer	*d*	*b*	*c*	*b*	*d*	*b*				

Exercise 2

1. The test statistic obtained from the ADF test is -0.742941.
2. The p-value associated with the test is 0.8096.
3. The number of lags used in the test is 3.
4. The number of observations included in the analysis is 18.
5. The critical values at the 1%, 5%, and 10% significance levels are:
    - 1% level: -3.886751
    - 5% level: -3.052169
    - 10% level: -2.666593

References

i. Books

1. Angrist , J., & Pischke, J. (2008). *Mostly Harmless Econometrics: An Empiricist's Companion A.* Princeton University Press.
2. Bobbitt, Z. (2021, April 16). *How to Perform a Breusch-Godfrey Test in R.* Retrieved May 2024, from https://www.statology.org/breusch-godfrey-test-in-r/
3. Box , G., Jenkins,, G., & Reinsel, G. (2015). *Time Series Analysis: Forecasting and Control.* Wiley.
4. Breusch, T., & Pagan, A. (1979). A Simple Test for Heteroscedasticity and Random Coefficient Variation. *Econometrica, 5*(47), 1287-1294.
5. Brockwell, P., & Davis, R. (2016). *Introduction to Time Series and Forecasting.* Springer.
6. Brooks, C. (2008). *Introductory Econometrics for Finance.* New York: Cambridge University Press.
7. Chatfield, C. (2003). *The Analysis of Time Series: An Introduction.* New York: Chapman & Hall/CRC.
8. Chatterjee., S., & Hadi, A. (2012). *Regression Analysis by Example.* Wiley.
9. Cowpertwait, P., & Metcalfe, A. (2009). *Introductory Time Series with R.* New York: Springer.
10. Doe , J. (2024). Closed-Form Solution for Simple Linear Regression. (Jane Smith Publication).
11. Douglas , M., & Watts , D. (1988). *Nonlinear Regression Analysis and Its Applications.* Wiley.
12. Draper , N., & Smith, H. (1998). *Applied Regression Analysis and Other Multivariable Methods.* Wiley.
13. EViews. (2020). *Eviews 12 User's Guide II.* IHS Markit.
14. Field, A. (2005). *Discovering Statistics Using IBM SPSS Statistics.* Los Ageles: Sage edge
15. Field, A. (2009). *Discovering statistics using SPSS.* Sage Publications.
16. Gareth , J., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An Introduction to Statistical Learning: with Applications in R.* Springer.
17. Griliches, Z. (1986). *Handbook of Econometrics.* Amsterdam: North-Holland.
18. Greene, W. (2017). *Econometric Analysis.* Pearson.
19. Gonick, L., & Smith, W. (1993). *The Cartoon Guide to Statistics.* HarperPerennial.
20. Gujarati , D., & Porter, D. (2009). *Gujarati, D. N., & Porter, D. C. Basic Econometrics.*
21. McGraw-Hill Education.
22. Hair, J., Black, W., Babin, B., & Anderso, R. (2010). *Multivariate Data Analysis.* Upper Saddle River, New Jersy: Prentice Hall.
23. Hamilton, J. (1994). *Time Series Analysis.* Princeton, NJ: Princeton University Press.
24. Hardin , J., & Hilbe, J. (2012). *Generalized Estimating Equations.* Chapman and Hall/CRC.
25. Hyndman , R., & Athanasopoulos, G. (2018). *Forecasting: Principles and Practice.* OTexts .
26. Johnson , R., & Wichern, D. (2007). *Applied Multivariate Statistical Analysis.* Pearson.
27. Judge , G., Hill, R., Griffiths, W., & Lütkepohl, H. (1982). *Introduction to the Theory and Practice of Econometrics.* New York: Wiley.
28. Kaplan, R., & Saccuzzo, D. (2013). *Mean, median, and mode In "Comprehensive and psychiatry".* . Fort Worth, TX: Holt, Rinehart, and Winston.
29. Keller , G., & Warrack, B. (2005). *Statistics for Management and Economics.* Thomson/South-Western.

30. Kutner, M., Nachtsheim, C., Neter, J., & LI, W. (2005). *Applied Linear Statistical Models.* MCGraw-Hill.
31. Montgomery , D., Jennings, C., & Kulahci, M. (2015). *Introduction to Time Series Analysis and Forecasting.* John Wiley & Sons.
32. Montgomery, D., Peck, E., & Vining, G. (2012). *Introduction to Linear Regression Analysis.* Wiley.
33. Moore, McCabe, & Craig's. (2009). *EXCEL MANUAL Introduction to the Practice of Statistics.* (W. F. Company, Ed.) Austin: Betsy Greenberg University of Texas.
34. Moore, D., & McCabe, G. (2017). *Introduction to the practice of statistics.* W. H. Freeman.
35. Pindyck, R., & Rubinfeld, D. (2011). *Microeconomics.* Pearson.
36. SANFORD, W. (2014). *Applied Linear Regression.* School of Statistics University of Minnesota, Minneapolis, MN: Wiley.
37. Shumway, R., & Stoffer , D. (2011). *Time Series Analysis and Its Applications With R Examples* . Panselvania: Springer.
38. Shumway, R., & Stoffer, D. (2017). *Time Series Analysis and Its Applications: With R Examples.* Springer.
39. Stock , J., & Watson, M. (2020). *Introduction to Econometrics.* Pearson.
40. Varmuza, K., & Filzmoser, P. (2009). *Introduction to Multivariate Statistical Analysis in Chemometrics.* CRC Press.
41. Weisberg , S. (2014). *Applied Linear Regression.* Wiley.
42. Wooldridge, J. (2012). *Introductory Econometrics A Modern Approach* (Vol. Fifth Edition). Michigan: Cengage Learning.

ii. Article from Journal

1. Ljung, G., & Box, G. (1978). On a Measure of Lack of Fit in Time Series Models. *Biometrika, 2*(65), 297-303.
2. Kwiatkowski, D., Phillips, P. C. B. , P., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? . *Journal of Econometrics, (1-3)*(54), 159-178.

iii. Electronic Sources

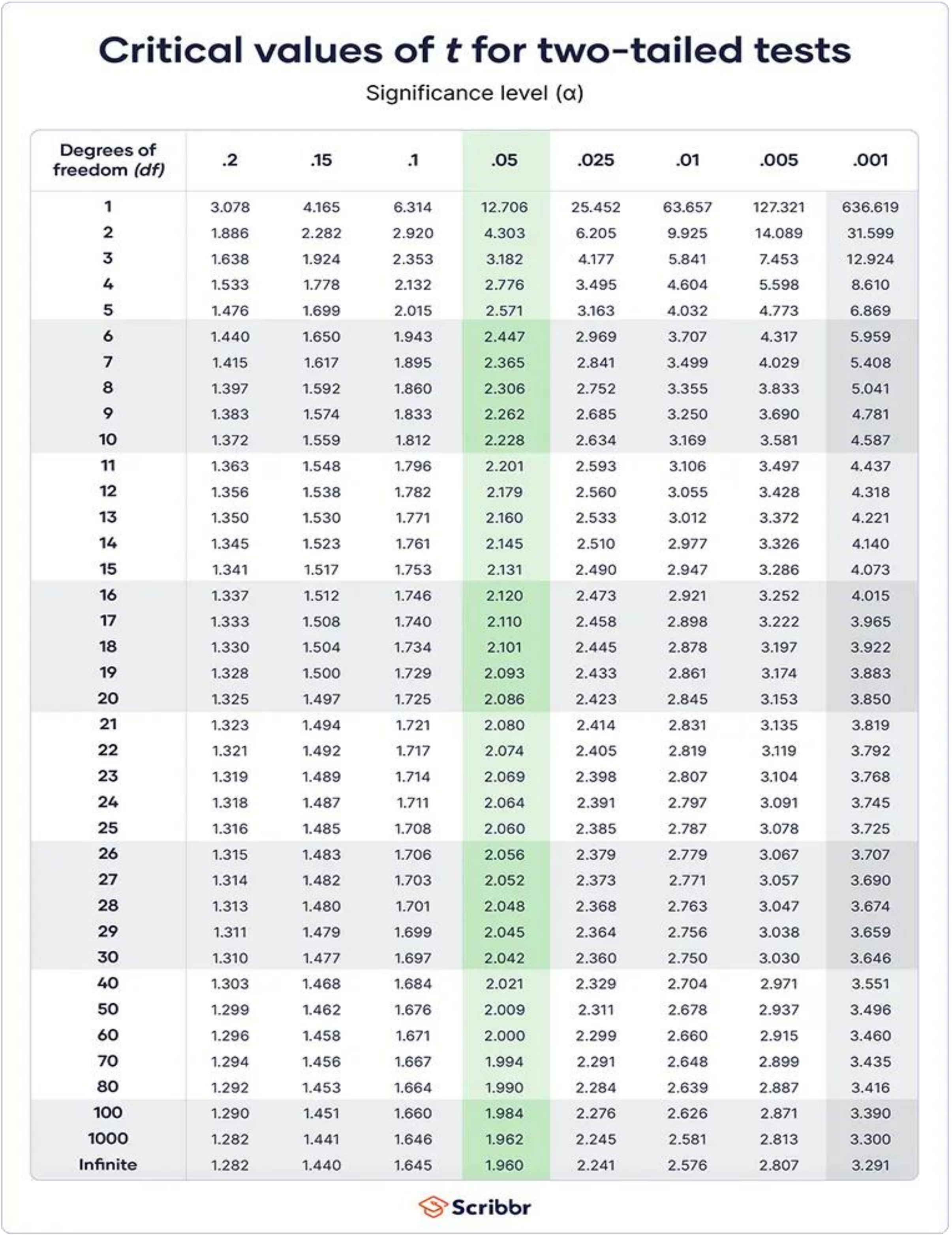
Shaun Turney. (2022, April 29). *Student's t Table (Free Download) | Guide & Examples*. Retrieved March 2024, from Scribbr: https://www.scribbr.com/statistics/students-t-table/

iv.Arabic references

خالد محمد السواعي. (2011). *ا Eviews و القياس الاقتصادي .* عمان، الاردن: دائرة المكتبة الوطنية.

Appendices

Table A.1: the *t* Distribution table

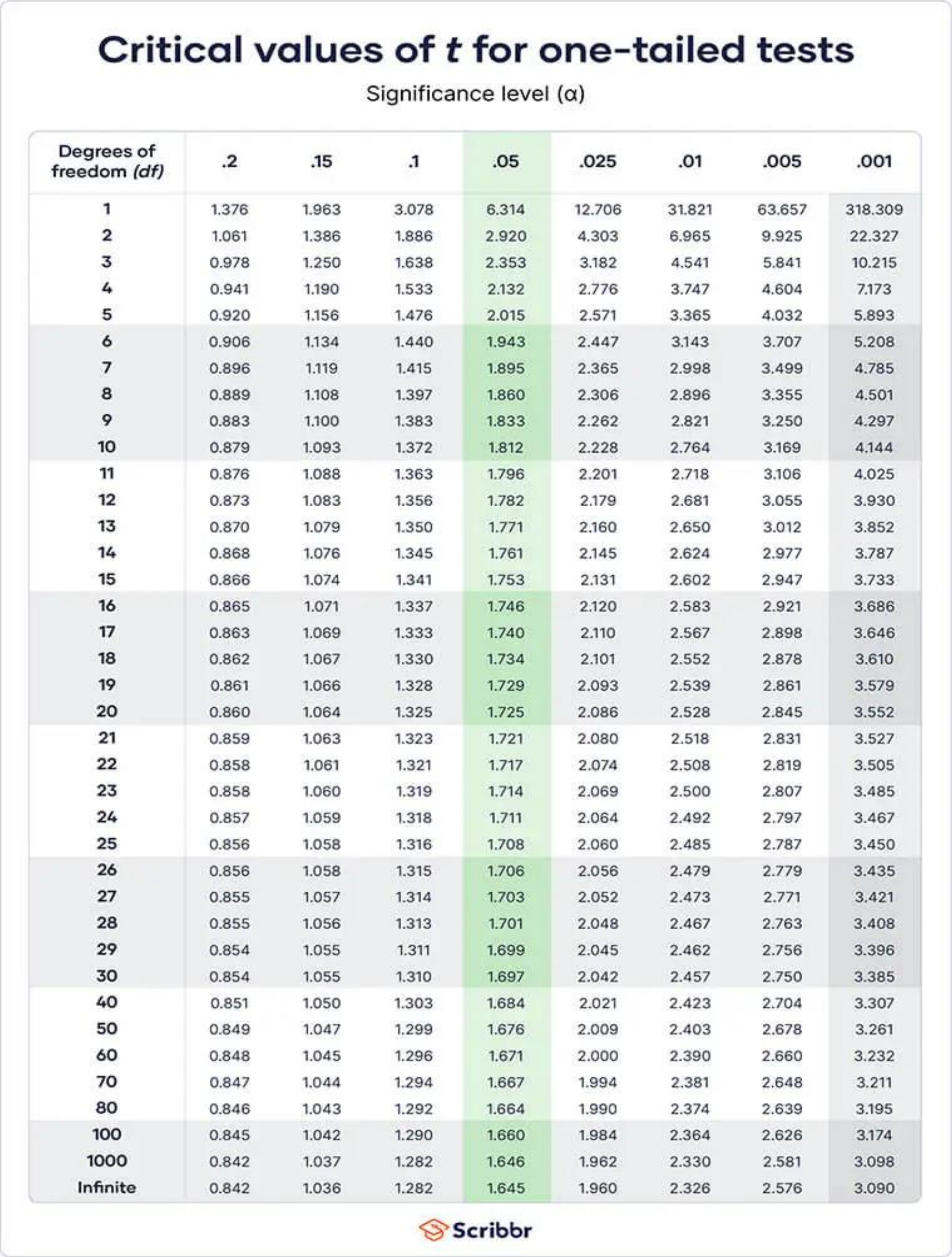


Critical values of *t* for two-tailed tests

Significance level (α)

Degrees of freedom *(df)*	.2	.15	.1	.05	.025	.01	.005	.001
1	3.078	4.165	6.314	12.706	25.452	63.657	127.321	636.619
2	1.886	2.282	2.920	4.303	6.205	9.925	14.089	31.599
3	1.638	1.924	2.353	3.182	4.177	5.841	7.453	12.924
4	1.533	1.778	2.132	2.776	3.495	4.604	5.598	8.610
5	1.476	1.699	2.015	2.571	3.163	4.032	4.773	6.869
6	1.440	1.650	1.943	2.447	2.969	3.707	4.317	5.959
7	1.415	1.617	1.895	2.365	2.841	3.499	4.029	5.408
8	1.397	1.592	1.860	2.306	2.752	3.355	3.833	5.041
9	1.383	1.574	1.833	2.262	2.685	3.250	3.690	4.781
10	1.372	1.559	1.812	2.228	2.634	3.169	3.581	4.587
11	1.363	1.548	1.796	2.201	2.593	3.106	3.497	4.437
12	1.356	1.538	1.782	2.179	2.560	3.055	3.428	4.318
13	1.350	1.530	1.771	2.160	2.533	3.012	3.372	4.221
14	1.345	1.523	1.761	2.145	2.510	2.977	3.326	4.140
15	1.341	1.517	1.753	2.131	2.490	2.947	3.286	4.073
16	1.337	1.512	1.746	2.120	2.473	2.921	3.252	4.015
17	1.333	1.508	1.740	2.110	2.458	2.898	3.222	3.965
18	1.330	1.504	1.734	2.101	2.445	2.878	3.197	3.922
19	1.328	1.500	1.729	2.093	2.433	2.861	3.174	3.883
20	1.325	1.497	1.725	2.086	2.423	2.845	3.153	3.850
21	1.323	1.494	1.721	2.080	2.414	2.831	3.135	3.819
22	1.321	1.492	1.717	2.074	2.405	2.819	3.119	3.792
23	1.319	1.489	1.714	2.069	2.398	2.807	3.104	3.768
24	1.318	1.487	1.711	2.064	2.391	2.797	3.091	3.745
25	1.316	1.485	1.708	2.060	2.385	2.787	3.078	3.725
26	1.315	1.483	1.706	2.056	2.379	2.779	3.067	3.707
27	1.314	1.482	1.703	2.052	2.373	2.771	3.057	3.690
28	1.313	1.480	1.701	2.048	2.368	2.763	3.047	3.674
29	1.311	1.479	1.699	2.045	2.364	2.756	3.038	3.659
30	1.310	1.477	1.697	2.042	2.360	2.750	3.030	3.646
40	1.303	1.468	1.684	2.021	2.329	2.704	2.971	3.551
50	1.299	1.462	1.676	2.009	2.311	2.678	2.937	3.496
60	1.296	1.458	1.671	2.000	2.299	2.660	2.915	3.460
70	1.294	1.456	1.667	1.994	2.291	2.648	2.899	3.435
80	1.292	1.453	1.664	1.990	2.284	2.639	2.887	3.416
100	1.290	1.451	1.660	1.984	2.276	2.626	2.871	3.390
1000	1.282	1.441	1.646	1.962	2.245	2.581	2.813	3.300
Infinite	1.282	1.440	1.645	1.960	2.241	2.576	2.807	3.291

Scribbr



Critical values of *t* for one-tailed tests

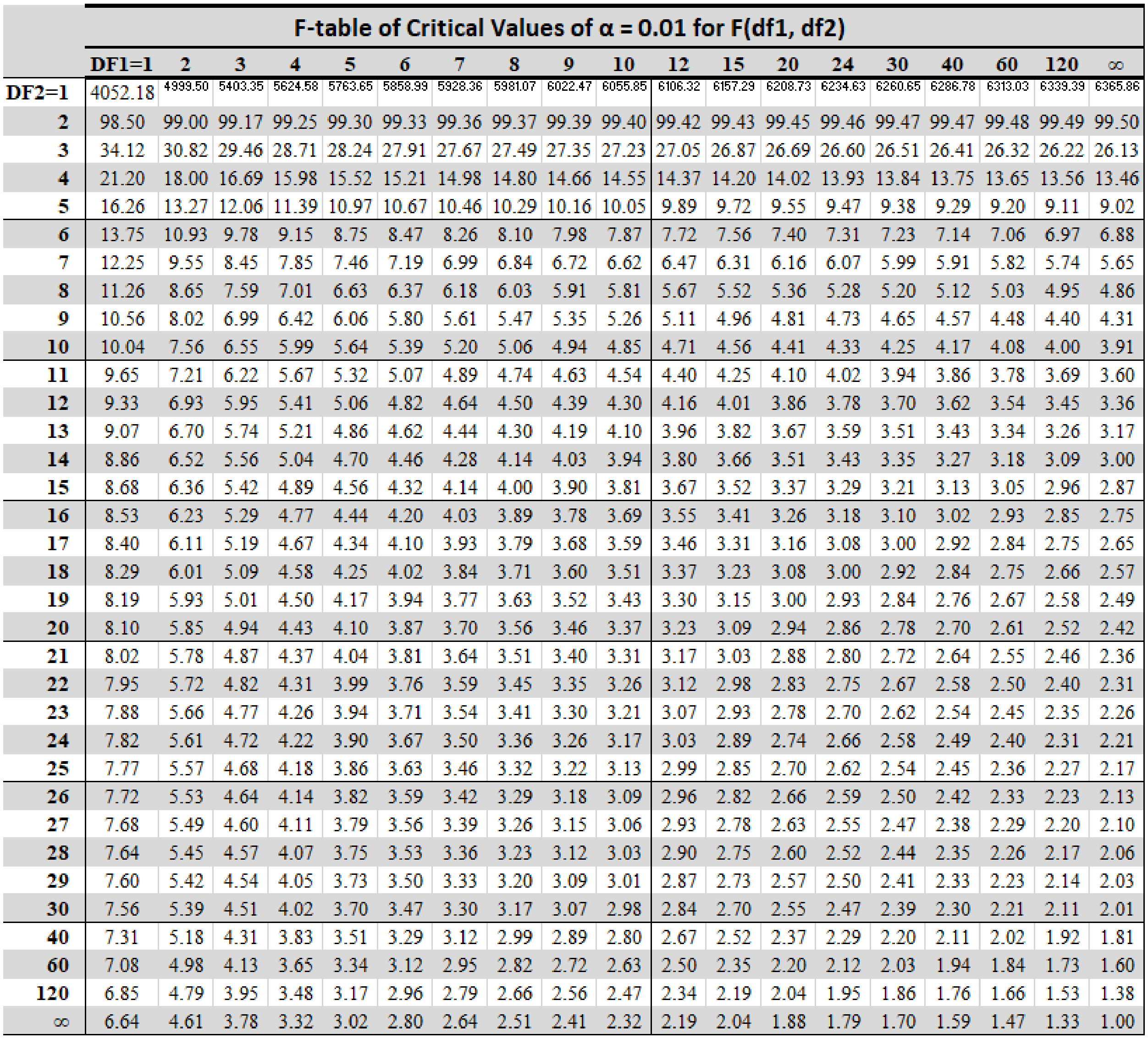
Significance level (α)

Degrees of freedom *(df)*	.2	.15	.1	.05	.025	.01	.005	.001
1	1.376	1.963	3.078	6.314	12.706	31.821	63.657	318.309
2	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327
3	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215
4	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173
5	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893
6	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208
7	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785
8	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501
9	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297
10	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144
11	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025
12	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930
13	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852
14	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787
15	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733
16	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686
17	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646
18	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610
19	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579
20	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552
21	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527
22	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505
23	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485
24	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467
25	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450
26	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435
27	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421
28	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408
29	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396
30	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385
40	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307
50	0.849	1.047	1.299	1.676	2.009	2.403	2.678	3.261
60	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232
70	0.847	1.044	1.294	1.667	1.994	2.381	2.648	3.211
80	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195
100	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174
1000	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098
Infinite	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090

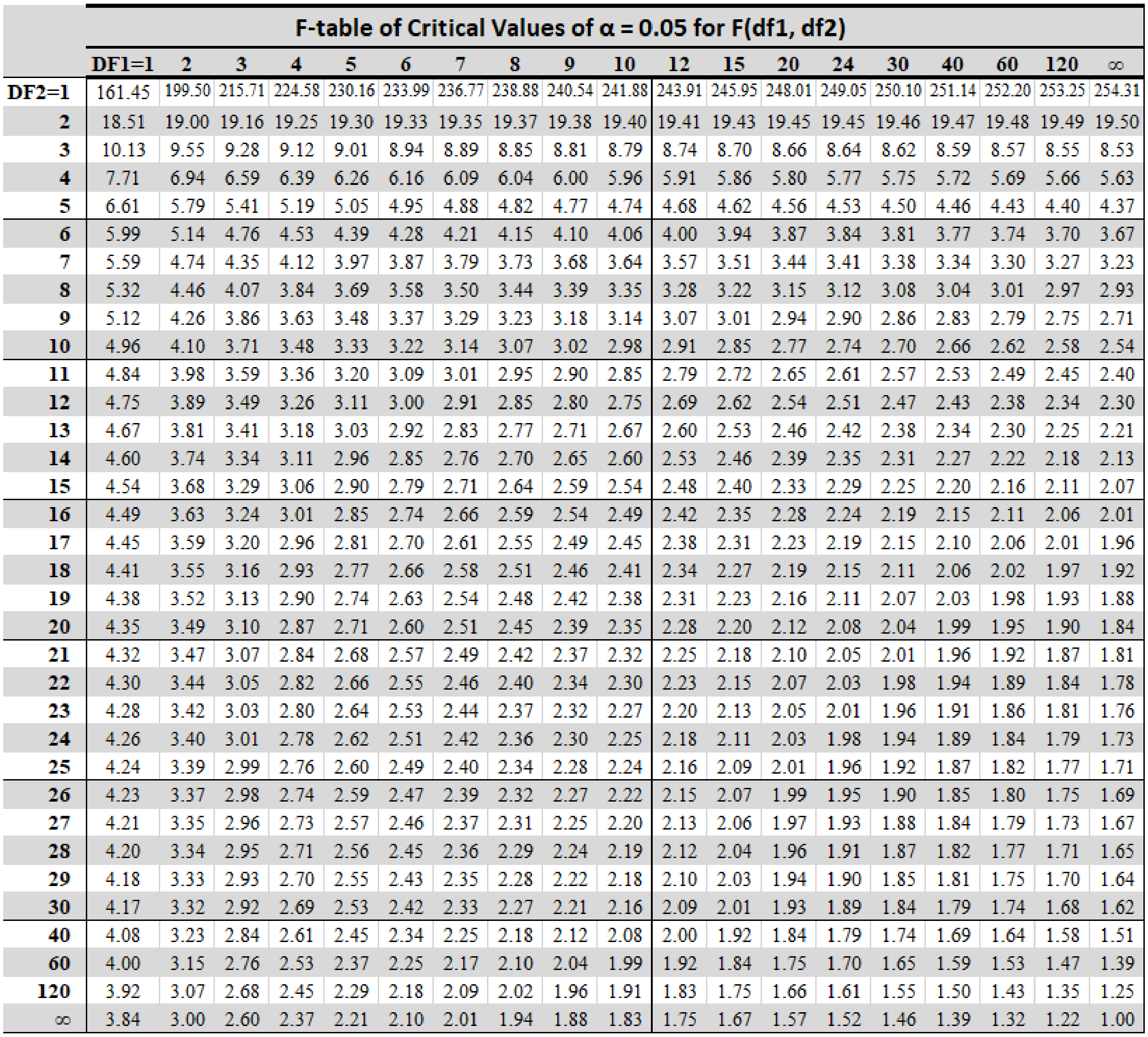
Scribbr

Source: https://www.scribbr.com/statistics/students-t-table/

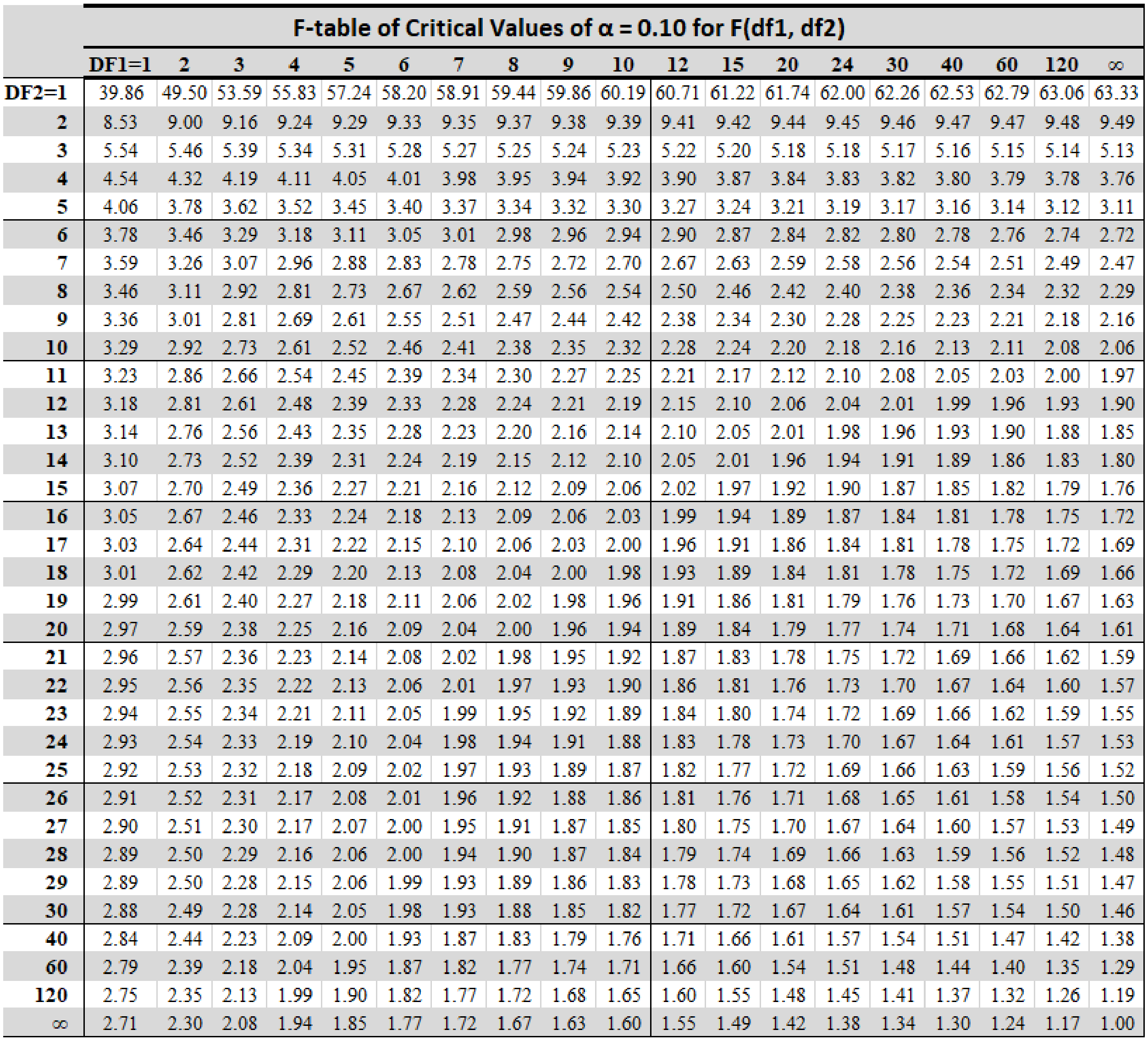
Table A.2 : Upper Percentage Points of the *F* Distribution



F-table of Critical Values of α = 0.01 for F(df1, df2)																			
	DF1=1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**	**15**	**20**	**24**	**30**	**40**	**60**	**120**	**∞**
DF2=1	4052.18	4999.50	5403.35	5624.58	5763.65	5858.99	5928.36	5981.07	6022.47	6055.85	6106.32	6157.29	6208.73	6234.63	6260.65	6286.78	6313.03	6339.39	6365.86
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.42	99.43	99.45	99.46	99.47	99.47	99.48	99.49	99.50
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.05	26.87	26.69	26.60	26.51	26.41	26.32	26.22	26.13
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37	14.20	14.02	13.93	13.84	13.75	13.65	13.56	13.46
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
6	13.75	10.93	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.67	3.59	3.51	3.43	3.34	3.26	3.17
14	8.86	6.52	5.56	5.04	4.70	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.90	3.81	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.85	2.75
17	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.84	2.75	2.65
18	8.29	6.01	5.09	4.58	4.25	4.02	3.84	3.71	3.60	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
19	8.19	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.93	2.84	2.76	2.67	2.58	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.70	2.61	2.52	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
23	7.88	5.66	4.77	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
25	7.77	5.57	4.68	4.18	3.86	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.62	2.54	2.45	2.36	2.27	2.17
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.96	2.82	2.66	2.59	2.50	2.42	2.33	2.23	2.13
27	7.68	5.49	4.60	4.11	3.79	3.56	3.39	3.26	3.15	3.06	2.93	2.78	2.63	2.55	2.47	2.38	2.29	2.20	2.10
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	2.06
29	7.60	5.42	4.54	4.05	3.73	3.50	3.33	3.20	3.09	3.01	2.87	2.73	2.57	2.50	2.41	2.33	2.23	2.14	2.03
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.67	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.81
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.04	1.95	1.86	1.76	1.66	1.53	1.38
∞	6.64	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.19	2.04	1.88	1.79	1.70	1.59	1.47	1.33	1.00



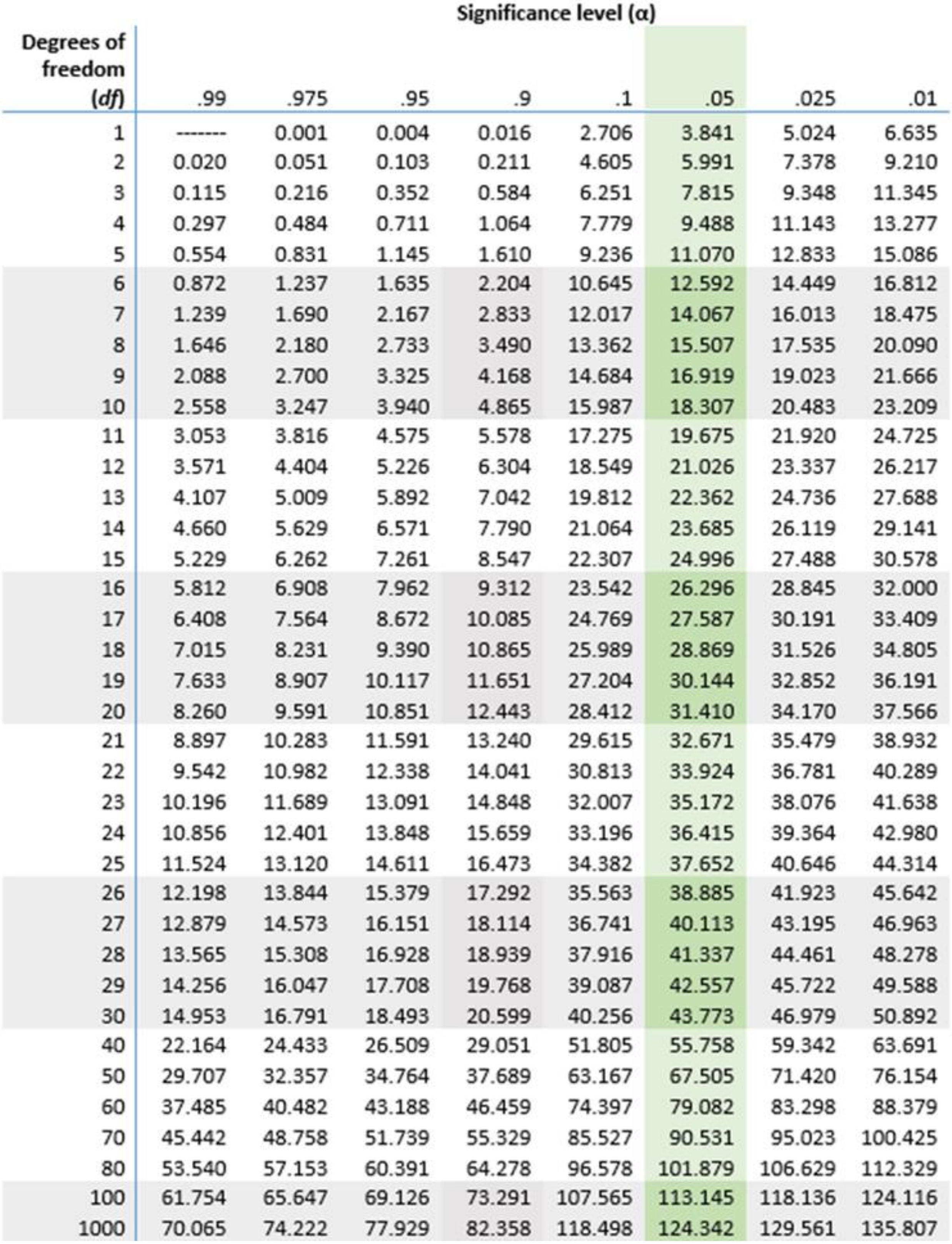
	F-table of Critical Values of α = 0.05 for F(df1, df2)																		
	DF1=1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**	**15**	**20**	**24**	**30**	**40**	**60**	**120**	**∞**
DF2=1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	243.91	245.95	248.01	249.05	250.10	251.14	252.20	253.25	254.31
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00



	F-table of Critical Values of α = 0.10 for F(df1, df2)																		
	DF1=1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**	**15**	**20**	**24**	**30**	**40**	**60**	**120**	**∞**
DF2=1	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86	60.19	60.71	61.22	61.74	62.00	62.26	62.53	62.79	63.06	63.33
2	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9.41	9.42	9.44	9.45	9.46	9.47	9.47	9.48	9.49
3	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.22	5.20	5.18	5.18	5.17	5.16	5.15	5.14	5.13
4	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	3.90	3.87	3.84	3.83	3.82	3.80	3.79	3.78	3.76
5	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3.27	3.24	3.21	3.19	3.17	3.16	3.14	3.12	3.11
6	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.90	2.87	2.84	2.82	2.80	2.78	2.76	2.74	2.72
7	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70	2.67	2.63	2.59	2.58	2.56	2.54	2.51	2.49	2.47
8	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54	2.50	2.46	2.42	2.40	2.38	2.36	2.34	2.32	2.29
9	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42	2.38	2.34	2.30	2.28	2.25	2.23	2.21	2.18	2.16
10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32	2.28	2.24	2.20	2.18	2.16	2.13	2.11	2.08	2.06
11	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25	2.21	2.17	2.12	2.10	2.08	2.05	2.03	2.00	1.97
12	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19	2.15	2.10	2.06	2.04	2.01	1.99	1.96	1.93	1.90
13	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14	2.10	2.05	2.01	1.98	1.96	1.93	1.90	1.88	1.85
14	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10	2.05	2.01	1.96	1.94	1.91	1.89	1.86	1.83	1.80
15	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06	2.02	1.97	1.92	1.90	1.87	1.85	1.82	1.79	1.76
16	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03	1.99	1.94	1.89	1.87	1.84	1.81	1.78	1.75	1.72
17	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00	1.96	1.91	1.86	1.84	1.81	1.78	1.75	1.72	1.69
18	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98	1.93	1.89	1.84	1.81	1.78	1.75	1.72	1.69	1.66
19	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96	1.91	1.86	1.81	1.79	1.76	1.73	1.70	1.67	1.63
20	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94	1.89	1.84	1.79	1.77	1.74	1.71	1.68	1.64	1.61
21	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95	1.92	1.87	1.83	1.78	1.75	1.72	1.69	1.66	1.62	1.59
22	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90	1.86	1.81	1.76	1.73	1.70	1.67	1.64	1.60	1.57
23	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92	1.89	1.84	1.80	1.74	1.72	1.69	1.66	1.62	1.59	1.55
24	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88	1.83	1.78	1.73	1.70	1.67	1.64	1.61	1.57	1.53
25	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89	1.87	1.82	1.77	1.72	1.69	1.66	1.63	1.59	1.56	1.52
26	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86	1.81	1.76	1.71	1.68	1.65	1.61	1.58	1.54	1.50
27	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87	1.85	1.80	1.75	1.70	1.67	1.64	1.60	1.57	1.53	1.49
28	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87	1.84	1.79	1.74	1.69	1.66	1.63	1.59	1.56	1.52	1.48
29	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.86	1.83	1.78	1.73	1.68	1.65	1.62	1.58	1.55	1.51	1.47
30	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82	1.77	1.72	1.67	1.64	1.61	1.57	1.54	1.50	1.46
40	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76	1.71	1.66	1.61	1.57	1.54	1.51	1.47	1.42	1.38
60	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71	1.66	1.60	1.54	1.51	1.48	1.44	1.40	1.35	1.29
120	2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.68	1.65	1.60	1.55	1.48	1.45	1.41	1.37	1.32	1.26	1.19
∞	2.71	2.30	2.08	1.94	1.85	1.77	1.72	1.67	1.63	1.60	1.55	1.49	1.42	1.38	1.34	1.30	1.24	1.17	1.00

Source: https://statisticsbyjim.com/hypothesis-testing/f-table/

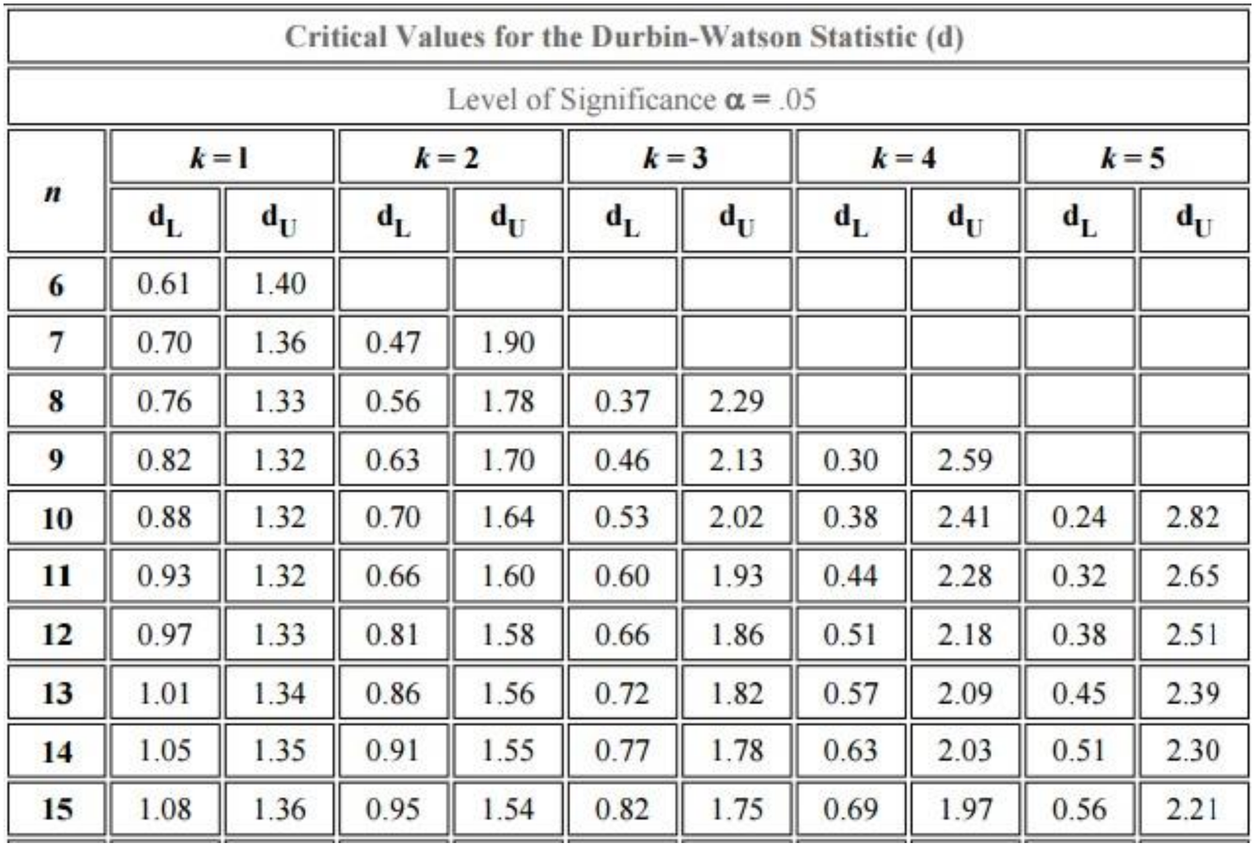
Table A.3: Chi-square χ^2 distribution table (right-tail probabilities)



Degrees of freedom (*df*)	Significance level (α)							
	.99	.975	.95	.9	.1	.05	.025	.01
1	-------	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086
6	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475
8	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666
10	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209
11	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725
12	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217
13	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688
14	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141
15	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578
16	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000
17	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409
18	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805
19	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191
20	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566
21	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932
22	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289
23	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638
24	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980
25	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314
26	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642
27	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963
28	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278
29	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588
30	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892
40	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691
50	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154
60	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379
70	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425
80	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329
100	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116
1000	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807

Source: https://www.scribbr.com/statistics/chi-square-distribution-table/

Table A.4 : Durbin–Watson *d* Statistic: Significance Points of d_L and d_U at 0.05 Level of Significance



Critical Values for the Durbin-Watson Statistic (d)

Level of Significance α = .05

n	*k* = 1		*k* = 2		*k* = 3		*k* = 4		*k* = 5	
	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
6	0.61	1.40								
7	0.70	1.36	0.47	1.90						
8	0.76	1.33	0.56	1.78	0.37	2.29				
9	0.82	1.32	0.63	1.70	0.46	2.13	0.30	2.59		
10	0.88	1.32	0.70	1.64	0.53	2.02	0.38	2.41	0.24	2.82
11	0.93	1.32	0.66	1.60	0.60	1.93	0.44	2.28	0.32	2.65
12	0.97	1.33	0.81	1.58	0.66	1.86	0.51	2.18	0.38	2.51
13	1.01	1.34	0.86	1.56	0.72	1.82	0.57	2.09	0.45	2.39
14	1.05	1.35	0.91	1.55	0.77	1.78	0.63	2.03	0.51	2.30
15	1.08	1.36	0.95	1.54	0.82	1.75	0.69	1.97	0.56	2.21

Source : https://www.statisticshowto.com/durbin-watson-test-coefficient/

For more table about Durbin–Watson *d* Statistic: Significance Points of d_L and d_U at different Level of Significance check this link: https://real-statistics.com/statistics-tables/durbin-watson-table/