linear regression model (OLS technique)

Prepared by: Prof. Yasmina GUECHARI

1. Linear Regression Model

- Linear regression is a statistical method used to model the relationship between dependent variable and one or more independent variables.
- The relationship is assumed to be linear, meaning the change in the dependent variable is proportional to the change in the independent variable(s).

2. Estimation of the Slope and Intercept

- Simple linear regression is the regression with one independent variable, the model equation is:
- $Y = \beta_0 + \beta_1 X + \varepsilon$
- Where, β_0 is the intercept, β_1 is the slope, ε is the error term.
- OLS estimates the coefficients (β_0, β_1) by minimizing the sum of the squared differences between the observed Y_i and predicted $\hat{Y}_i$values.

3. Interpretation of Coefficients

- The intercept β_0 represents the value of the dependent variable when all independent variables are zero. The slope β_1 represents the change in the dependent variable for a one-unit change in the independent variable, holding other variables constant.

5. Multiple Linear regression

- Multiple linear regression is a model that include multiple independent variables:
- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots\cdots + \beta_n X_n + \varepsilon$
- OLS technique estimates the coefficients for each independent variable.

6. Interpretation of Coefficients in Multiple Regression

- Each coefficient β_i represents the change in the dependent variable for a one-unit change in the corresponding independent variable, holding all other variables constant.
- **Example** the value of β_1 represent the change in the dependent variable for one unit change in X_1 holding all the other independent variable ($X_2, X_3, \ldots . X_n$) constant;

4. Assumptions of Linear Regression

1. Errors are normally distributed. (how to check it???)
2. Errors have constant variance (homosKedasticity). (how to check it?)
3. Errors are uncorrelated (no autocorrelation). (how to check it???)

How to check Normality Assumption

- **The Jarque-Bera test** is a statistical test used to assess whether a sample comes from a normally distributed population.
- **Here's how the Jarque-Bera test works:**
- First, calculate the skewness, a skewness value of 0 indicates perfect symmetry.
- A kurtosis value of 3 is often considered as the benchmark for normality (since the kurtosis of a normal distribution is 3)

How to check Normality Assumption

- **Calculate the Test Statistic**: The Jarque-Bera test statistic is calculated as
- $JB = \frac{n}{6}(S^2 + \frac{1}{4}(K-3)^2)$
- Where:
- *n* is the sample size.
- *S* is the sample skewness.
- *K* is the sample kurtosis.
- The null hypothesis of the Jarque-Bera test is that the errors is normally distributed.
- The alternative hypothesis is the errors is not normally distributed.

How to check Normality Assumption

- After calculating the Jarque-Bera test statistic (JB) for your dataset using the formula above
- Use a chi-squared distribution table to find the critical value corresponding to your chosen significance level (e.g., 0.05) and 2 degrees of freedom.

❖If the **JB calculated exceeds the critical value**, suggesting that the data deviates significantly from normality (Data is 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)

❖In our example the JB calculate is 0.576, JB critical is 5.991, $JB_{cal} < JB_{crit}$;

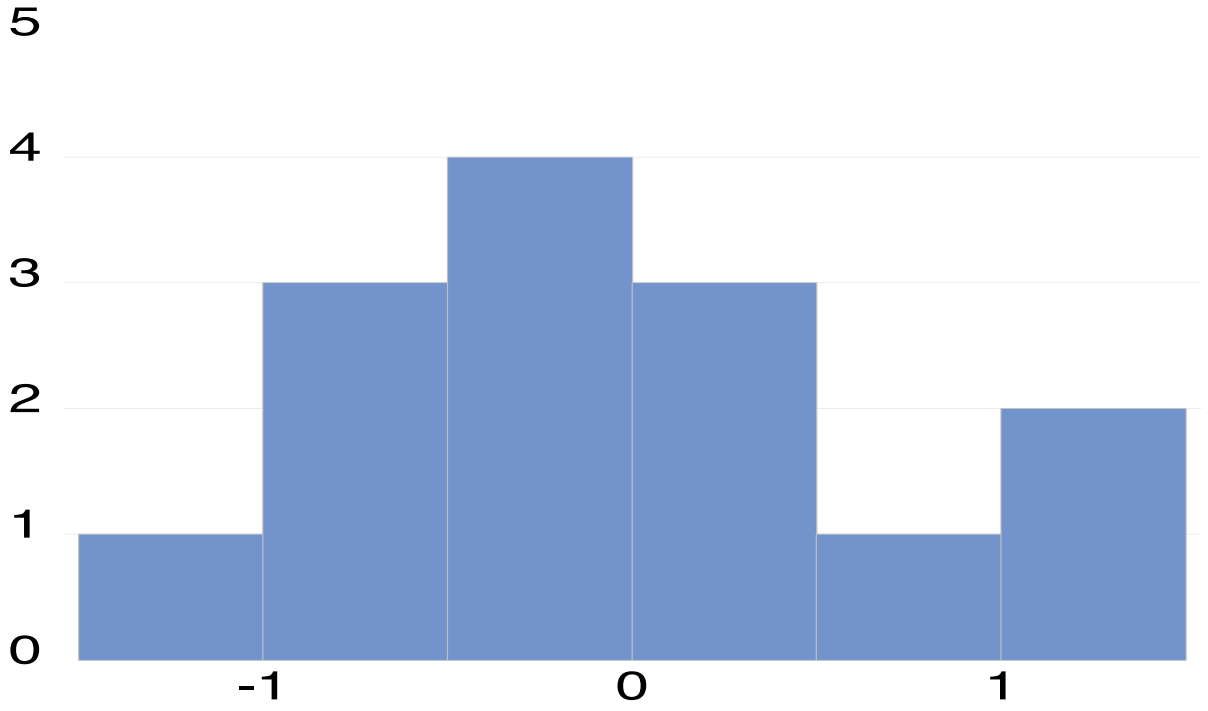
❖Accept the null hypothesis means the data is normally distributed

How to test the Normality Assumption using Eviews software

- To run a Jarque-Bera test in **EViews,** you can follow these steps:

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

Eviews output, Normality test



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

How to test Normality Assumption using Eviews software

4. Interpret the results based on the p-value:
   i. If the **p-value is less than the chosen significance level (e.g., 0.05),** you can reject the null hypothesis of normality, suggesting that the data is not normally distributed.
   ii. If the **p-value is greater than the significance level**, you fail to reject the null hypothesis, indicating that the data is normally distributed.

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

How to check the for Homoskedasticity Assumption

- **Homoskedasticity** refers to the assumption in regression analysis that the variance of the residuals (errors) is constant across all levels of the independent variables. To check for homoscedasticity, you can use several diagnostic techniques:

1. **Breusch-Pagan Test:** is a formal statistical test to check for homoscedasticity.

How to check the Homoskedasticity Assumption

- To perform the Breusch-Pagan test for homoscedasticity in a regression analysis, you can follow these steps:

i. First, estimate your regression model using ordinary least squares (OLS) or any other regression technique.

ii. Obtain Residuals

iii. Square the Residuals

iv. Run an Auxiliary Regression: Run an auxiliary regression where the squared residuals are regressed on the independent variables used in your original regression model.

How to check the Homoskedasticity Assumption

- The null hypothesis of the Breusch-Pagan test is that the variance of the errors is constant (homoscedasticity).
- The alternative hypothesis is hat the variance of the errors is not constant (heteroscedasticity).
- You can test this hypothesis by examining 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.

How to check the Homoskedasticity Assumption

- 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 homoscedasticity.

How to test the Homoskedasticity Assumption using Eviews software

1. Upload your data in Eviews software
2. Go to »View » menu at the top of the Eviews windows,
3. select « open data as equation » click OK
4. You will get the result of the OLS estimation;
5. Go to " View" menu in EViews.
6. Navigate to " Residual Diagnostic" and click on " Heteroskedasticity Test "
7. In the Options dialog box, go to the "Test types" select **Breusch-Pagan Test**

How to test the Homoskedasticity Assumption using Eviews software

8. Run the Regression: Once you've specified your regression model and test options, click "OK" to run the regression.
9. 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.
10. **Interpret Results:**
    i. If the p-value is less than your chosen significance level (e.g., 0.05), you may conclude that there is evidence of heteroskedasticity, indicating that the variance of the error term is not constant across observations.
    ii. If the p-value is greater than the significance level, you may fail to reject the null hypothesis of homoskedasticity, there is evidence of heteroskedasticity.

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		

How to Correct the heteroscedasticity issue?

- If heteroscedasticity is detected, consider using techniques such as:

i. Weighted least squares regression (WLS),

ii. Transforming the dependent variable, by applying a mathematical transformation (such as Logarithm) to the variable itself. The goal is to stabilize the variance of the residuals and make it more constant across different levels of the independent variable.

10. How to Correct the heteroscedasticity issue?

- **Logarithmic Transformation:** Taking the natural logarithm of the dependent variable (*Y*) is a common transformation. Transformed Y=ln(Y).
- **Square Root Transformation:**The square root transformation is another option to stabilize the variance.
- Transformed $Y = \sqrt{y}$
- It's essential to address heteroscedasticity to obtain valid statistical inferences from your regression model.

How to check the autocorrelation assumption

- **Autocorrelation,** also known as serial correlation, is crucial in regression analysis, especially when dealing with time series data.
- Autocorrelation occurs when the error terms in a regression model are correlated with each other over time, violating the assumption of independence.
- Here's how you can test for autocorrelation:
- **Durbin-Watson Test**: This test is a widely used for autocorrelation in the residuals of a regression model.

11. How to Check the Autocorrelation assumption?

- The null and alternative hypotheses for the Durbin-Watson test are as follows:
- Null Hypothesis (*H*0) Of Durbin Watson test:**There is no first-order autocorrelation in the residuals.**
- **Alternative Hypothesis (**H1**):There is first-order autocorrelation in the residuals.**

11. How to Check the Autocorrelation Assumption?

- The Durbin-Watson test statistic is computed using the following formula:
- $d = \frac{\sum_{t=2}^{n}(\varepsilon_t - \varepsilon_{t-1})^2}{\sum_{t=1}^{n} \varepsilon_t{}^2}$

ε_t is the residual at time t,*n* is the number of observations.

In this formula, the numerator represents the sum of squared differences between consecutive residuals, and the denominator represents the sum of squared residuals. The test statistic *d* ranges from 0 to 4.

11. How to Check the Autocorrelation?

- The test statistic, denoted as *d*, ranges between 0 and 4.

a. Value of *d* close to 2 (around 2 ± 0.2) indicates no first-order autocorrelation (null hypothesis not rejected).

b. Values significantly less than 2 suggest positive autocorrelation,

c. Values significantly greater than 2 suggest negative autocorrelation.

- $d \approx 2$: No evidence of first-order autocorrelation.
- $d < 2$: Evidence of positive autocorrelation.
- $d > 2$: Evidence of negative autocorrelation.

How to test the autocorrelation assumption using Eviews Software

- The Durbin-Watson test statistic will be displayed, along with its corresponding p-value.
- If the test statistic is close to 2 (around 2 ± 0.2), it suggests no significant autocorrelation.
- A value significantly below 2 indicates positive autocorrelation,
- Avalue significantly above 2 indicates negative autocorrelation.

How to test the autocorrelation assumption using Eviews Software

- **Perform serial correaltion Test in e-views**: After estimating the model and obtaining the residuals, you can perform the serial correlation LM test using the following 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.
f. The result will appear as follow:

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		

i. If the p-value is less than your chosen significance level (e.g., 0.05), you may conclude that there is evidence of autocorrelation, indicating that the the error terms are serially correlated.
ii. If the p-value is greater than the significance level, you may fail to reject the null hypothesis of no serial correlation, means the error are not autocorrelated.

How to correct the Autocorrelation issue?

- If autocorrelation is detected, consider these technique to correct it:

i. Incorporating **lagged values** into your model;

ii. Differencing data involves taking the first difference of the dependent variable or the independent variables. This is often done in time series analysis. Differencing can help remove the trend and make the data more stationary, reducing autocorrelation.

7. How to check for Multicollinearity?

- Multicollinearity refers to the presence of high correlations among independent variables in a regression model.
- It can cause issues in the estimation of regression coefficients and their interpretation.
- Here the most important methods to test for multicollinearity:
- Variance Inflation Factor (VIF): 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

How to check for Multicolinearity?

- For each independent variable, calculate its VIF using the formula: $VIF = \frac{1}{1-R^2}$ where R^2 is the coefficient of determination from regressing one independent variable against all the other independent variables.
- 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.

How to check for Multicolinearity using Eviews Software?

1. **Variance Inflation Factors (VIF)**:
   i. Run a regression model in EViews.
   ii. After estimating the regression model, go to the "View" menu, select "Coefficient Diagnostics...", and then choose "Variance Inflation Factors".
   iii. In the VIF dialog box, select the variables you want to include in the VIF calculation.
   iv. Click "OK" to generate the VIF values. VIF values greater than 10 or 5 may indicate multicollinearity.

Variance Inflation Factors
Date: 02/20/24 Time: 15:40
Sample: 2010 2023
Included observations: 14

Variable	Coefficient Variance	Uncentered VIF	Centered VIF
INFLATION_RATE__	0.268587	33.39856	1.012294
GDP_GROWTH_RAT	0.293126	50.96839	1.012294
C	3.287718	74.47698	NA

The VIF (centred VIF) are all less than 5, means there is no multicolinearity issue in the data

How to Correct the multicollinearity issue?

- We can remove the multicollinearity by using the following techniques:

i. Identify pairs of independent variables with high correlation and consider removing one of them.

ii. If possible, combine highly correlated variables into a single variable. For example, if you have variables measuring similar concepts in different units, consider creating an index or using principal component analysis.

How to Correct the multicollinearity issue?

iii. Transform variables, such as taking the logarithm or square root, to reduce the impact of extreme values and potentially alleviate multicollinearity.

iv. Collect more data to increase the sample size. A larger sample size can help stabilize estimates and reduce the impact of multicollinearity.

v. Centring variables: Centering involves subtracting the mean of a variable from all observations. Centering can sometimes help reduce multicollinearity, especially if variables have different scales.

How to Correct the multicollinearity issue?

vi. Use Stepwise Regression: Perform stepwise regression to iteratively add or remove variables based on statistical criteria. This can help select a subset of variables that minimizes multicollinearity.

8. What it OLS & GLS

- **Ordinary Least Squares (OLS):** It's a method used in regression analysis to estimate the parameters of a linear regression model (simple or multiple linear regression).
- In OLS, the goal is to find the line that minimizes the sum of the squared differences between the observed values and the values predicted by the model.
- Here's how OLS works:

1. **Model Specification**: Define a linear relationship between the dependent variable (Y) and one or more independent variables (X).

8. OLS

2. **Estimation of Coefficients**: Use the observed data to estimate the coefficients $\beta_0, \beta_1, \ldots, \beta_k$. The estimates are chosen to minimize the sum of the squared differences between the observed values of Y and the values predicted by the model; hence the name "least squares."

- **Statistical Inference**: Once the coefficients are estimated, statistical tests can be performed to assess the significance of the coefficients, the overall fit of the model, and other properties such as multicollinearity, heteroscedasticity, and autocorrelation.

8. OLS

- OLS is widely used because it's computationally simple, has closed-form solutions, and is well-understood statistically. However, it assumes that the errors are normally distributed, have constant variance (homoscedasticity), and are independent of each other (no autocorrelation). When these assumptions are violated, alternative estimation techniques may be more appropriate.

9. Performing Ordinary Least Squares (OLS) regression in EViews

- Performing Ordinary Least Squares (OLS) regression in EViews is straightforward. Here's a step-by-step guide:

1. **Load Data**: Open EViews and load your dataset containing the variables you want to include in the regression analysis.
2. **Specify Regression Equation**: Go to "Quick/Estimate Equation" or "Object/New Object/Equation" to open the equation specification window.
3. **Enter Equation**: In the equation specification window, enter your regression equation.

9. Performing Ordinary Least Squares (OLS) regression in EViews

- You can either type the equation directly or select variables from the list.
- **Estimate Equation**: Once you've entered the equation, click on the "Estimate" button to estimate the coefficients using OLS.
- **View Results**: After estimation, EViews will display the results of the regression analysis, including the estimated coefficients, standard errors, t-statistics, p-values, R^2, adjusted R^2 , and other diagnostic statistics.
- **Interpret Results**: Analyze the results to understand the relationship between the dependent and independent variables, as well as the significance and direction of the coefficients.

9. Performing Ordinary Least Squares (OLS) regression in EViews

- When the assumptions of Ordinary Least Squares (OLS) regression are violated, alternative estimation techniques may be more appropriate. Here are some of the techniques commonly used in such cases:

1. **Weighted Least Squares (WLS)**: WLS is used when the errors have heteroscedasticity.
2. **Generalized Least Squares model(GLM)**: GLS and GLM is a more general extension of OLS that allows for correlation among the errors and heteroscedasticity. GLS can be used when errors are both heteroscedastic and correlated.

9. Performing Ordinary Least Squares (OLS) regression in EViews

3. **Time-Series Analysis Techniques**: For time-series data, alternative techniques such as autoregressive integrated moving average (ARIMA) models, vector autoregression (VAR), or state-space models may be more appropriate, especially when dealing with autocorrelation and non-stationarity.
4. **Panel Data Analysis**: For panel data, techniques such as fixed effects models, random effects models, or dynamic panel data models can account for individual or time effects and address issues like heteroscedasticity and autocorrelation.

9. Performing Ordinary Least Squares (OLS) regression in EViews

- These techniques provide robust alternatives to OLS regression when its assumptions are violated, allowing researchers to obtain more reliable estimates and draw valid conclusions from their data.
- The choice of method depends on the specific nature of the data and the violations of OLS assumptions present in the model.