Advances
2022; 3(3): 49-59
http://www.sciencepublishinggroup.com/j/advances
doi: 10.11648/j.advances.20220303.12

Diagnostic Tests for Econometric Problems in Multiple Regression Analysis

Abeer Mohamed Abd El Razek Youssef

Faculty of Graduate Studies for Statistical Research, Cairo University, Giza, Egypt

Email address:
abeer_mahamed_a@yahoo.com

To cite this article:
Abeer Mohamed Abd El Razek Youssef. Diagnostic Tests for Econometric Problems in Multiple Regression Analysis. *Advances*. Vol. 3, No. 3, 2022, pp. 49-59. doi: 10.11648/j.advances.20220303.12

Received: June 19, 2022; **Accepted**: July 19, 2022; **Published**: July 29, 2022

Abstract: Most econometric models suffer from the problems of autocorrelation, multicollinearity, and heteroscedasticity. This paper presents a brief on these problems, their causes, how can be detected, tested, and minimized. The OLS method is based on several assumptions, and if these assumptions are fulfilled, we obtain unbiased, consistent, and efficient estimates (less variance compared to other methods). We discuss these problems as follows: First: the problem of multicollinearity Second: The problem of autocorrelation Third: Variation Heteroscedasticity. This article presents inference for many commonly used estimators - Box Plot on Normal Distribution, skewness, kurtosis, and Assumptions for Multiple Regression, that are asymptotically normally distributed. The Section Inference focuses on multicollinearity and hypothesis tests based on correlation matrix estimates measures a goodness of fit that are determine if a data set is well-modeled, heteroskedastic and, if relevant, Autocorrelation test. The Section Model Tests and Diagnostics summarizes tests of model adequacy and model diagnostics. The Section of practical application presents diagnostic tests that are used to judge the quality of the model, whether it is efficiency, convenience, fitness and flawless. The validity its ability to measure sensitivity and specificity. where it is essential indicators of test accuracy and allow to determine the appropriateness of the diagnostic tool.

Keywords: Multicollinearity, Autocorrelation, Heteroscedasticity

1. Introduction

The hypothesis of the model will focus on three points: 1- Linear regression model 2- The hypotheses of the model must be fulfilled. So, if you realize this, how do you discover this, the effect of not being fulfilled, will it be on the results of the model? 3- Methods for selecting independent variables: Linear regression model: independent variables, whether one variable in the simple regression. Multiple regression model: More than one independent variable in a multiple regression. I chose to finish the variables to be included with me in the model so that I can finally see the results that reach me the efficiency of the model so that it is of high quality, and I can rely on it.

2. Material and Methods

The most famous of the test methods here are three. They are considered the most common: 1- In order to analyze my data before I work or apply regression, I must do a processing of raw data, meaning are there errors in data entry (Data cleaning and preprocessing). 2- Missing value. Are there missing values or not, such as if you do hypothesis tests that need Hypothesis tests, do them in the Regression «An ova, F - Test, Total Model». for the model, see it for the full model; I look at the significant or not, or the T-Test for individual variables, so I see whether each variable of a unit is considered significant or not. [1].

2.1. Measuring Goodness of Fit

If the data in my data has missing values, it will affect the power of the test, meaning the result that he does not receive from the test will remain few and the power in it will be very less, both from F-Test and T-Test. Therefore, I must make an estimation for the missing data. Outliers: Whether in the dependent variable or the independent variables, I begin to treat them, and they are far from the rest of the data. [2].

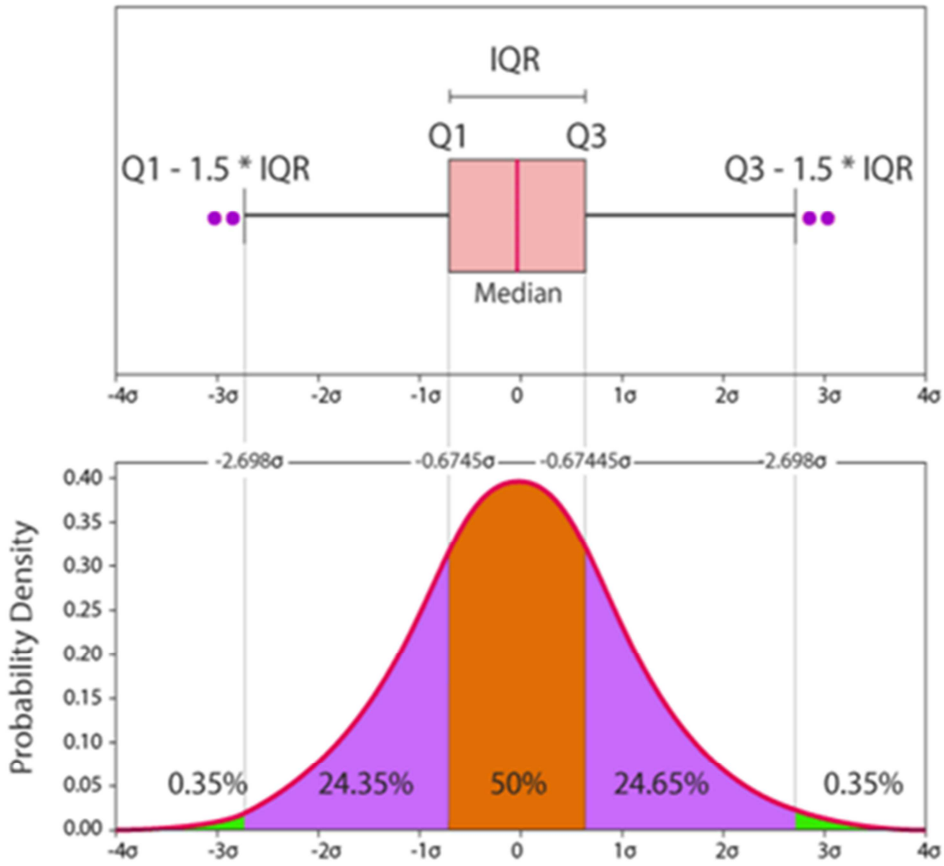
For example, if I talk about income, a person earns one million pounds and another person earns 300 pounds and 600 pounds, so this large number is called an abnormal value. Since it has a specific method of treatment such as robust and reweighting, because the abnormal values will change all the

results to me, so the results will not reflect the whole or the whole community or they will not express the sample, but they're expressing themselves. 3- Unusual Distribution If I had a model that assumed a specific distribution, we would suppose the distribution would be as valid as the normal distribution in the case of regression needing the condition of normality. [3].

Violation	Consequence	Detection method
Non-normality of errors	*F*- and *t*-tests unreliable	-Normal Q-Q plots -Formal tests of normality
Non-linearity	$s_{\theta i}$ unreliable	-Partial residual plots -Scatter plots of Y on X_i
Heteroscedasticity	$s_{\theta i}$ unreliable	-Residual plots -White's test
Autocorrelation	$s_{\theta i}$ unreliable	-Residual plots -Durbin-Watson test
Multicolinearity	$s_{\theta i}$ unreliable *t*-test non-significant	-Variance inflation factor (VIF) -Tolerance
Measurement error s	Errors in X_i bias $\bar{\theta}_i$ towards 0 Errors in Y inflate $s_{\theta i}$	-Small *t*-values; $P > 0.05$ -Inflated $s_{\theta i}$ values
Outliers	$s_{\theta i}$ unreliable	-Studentized residuals

Source: lecture note Dr. Mahmoud A. Abdel-Fattah, AS611_Lecture01, 2022

Figure 1. *Clarity and Cleaning Data.*



Son, N. H. (2006). Data cleaning and Data preprocessing. Mimuw University.

Figure 2. *Box Plot on Normal Distribution.*

Positively skewed:	If the distance from the average to the maximum is greater than the distance from the average to the minimum, the scatterplot is positively skewed.
Negatively skewed:	If the distance from the average to the lowest is greater than the distance from the medium to the maximum, the scatterplot of chart deviates negatively.
Symmetrical:	The graph diagram, is said to be the same if the broker is at an equal distance from the maximum and minimum values.

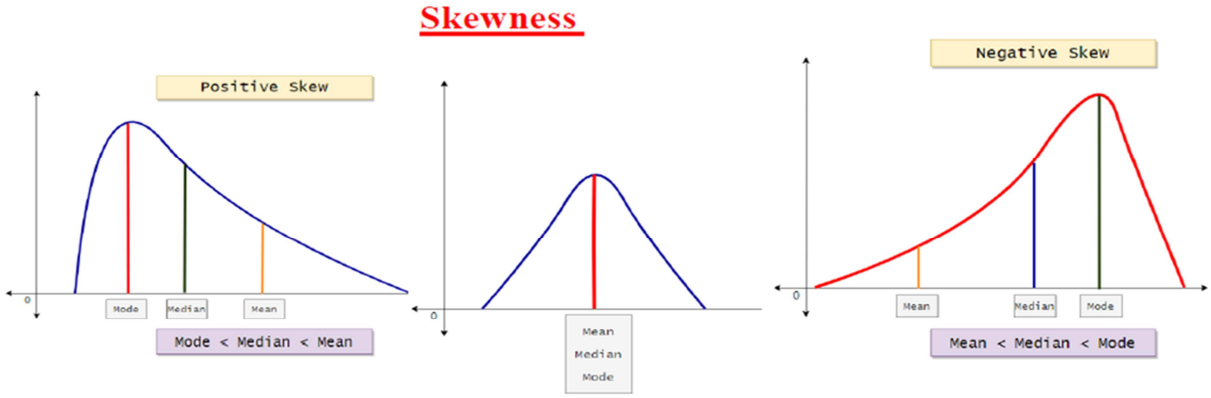


Figure 3. *Types of skewness.*

2.2. Correlation

It is used to find relationships between independent and dependent variables.

Changes in variability How does the variation factor in the 5-number summaries? 4- Parts of Box Plots The distribution of the plot chart will explain how tightly the data collection is collected, how data is skewed or kurtosis, as well as about data consistency and efficient. [4].

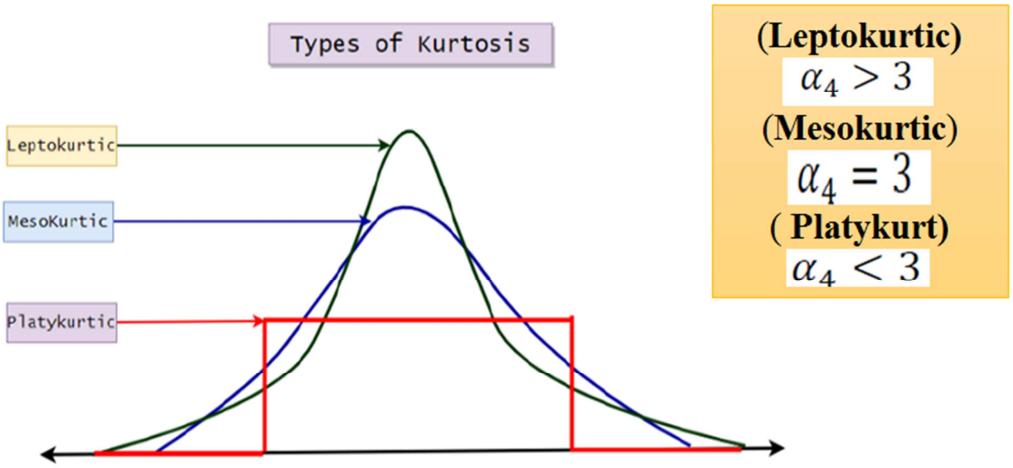


Figure 4. *Types of kurtoses.*

3. The Purpose of Regression Analysis

Its purpose is that I have a dependent variable that I want to explain its behavior, and this is known through its relationship with other variables (how independent variables affect the dependent variable).

$$Y = f(x_1, x_2, \ldots\ldots, x_k)$$

3.1. The Results Obtained in the Model Help to

Predicting the dependent variable in the future, such as forecasting the weather, such as marine and air navigation, or the state of flight, is important in insurance in which forecasting is used. [5].

Or the (machine learning regression) or predicting whether a person has a certain disease or not or with a full probability, through the overlap between computer science and statistics? Predicting the dependent variable depending on the values of the independent variables. [6].

Meaning that whenever the x changes by one unit, will the y change by how many? Like if you were talking about the relationship between income and saving or income and consumption.

Whenever income increases by one unit, consumption will change in the direction of increase or decrease by an amount of any?

$$Y_i = b_0 + b_1 x_i + e_i$$

In other words, I have other variables that affect the behavior of the dependent variable y, but I don't include them in the model.

And the existence of random error continues in the model because no matter how many variables entered the model, other variables are sure to explain the dependent variable. [7].

Table 1. *Data Cleaning and preprocessing.*

Changes in variability	
Scatter plots:	It is used in 1- detecting outliers. 2- To reveal the relationship is linear or not? 3- To reveal the direction and strength of the relationship, positive or negative.
Histogram	1- It is used to detect the normality, whether it has abnormal values or not? 2- I see how the Rang of the variability factor.
Box Plots	1- By revealing the normality 2- Detecting bad outliers 3- Looking at the boundaries of the first spring and the second spring.

The multiple linear regression equation is as follows:

$$\hat{Y} = B_0 + B_1 X_1 + B_2 X_2 + \ldots\ldots + B_P X_P$$

Where:

1) yi is the dependent or predicted variable
2) β0 is the y-intercept, i.e., the value of y when both xi and x2 are 0.
3) β1 and β2 are the regression coefficients representing the change in y relative to a one-unit change in xi1 and xi2, respectively.
4) βp is the slope coefficient for each independent variable.
5) ϵ is the model's random error (residual) term.

3.2. Assumptions for Multiple Regression

The assumptions for multiple linear regression are largely the same as those for simple linear regression models . However there are a few new issues to think about and it is worth redundancy our assumptions for using multiple explanatory variables.

Table 2. *Econometric Problems.*

Linear relationship:	The model is a roughly linear one. This is slightly different from simple linear regression as we have multiple explanatory variables. This time we want the outcome variable to have a roughly linear relationship with each of the explanatory variables, taking into account the other explanatory variables in the model.
Homoscedasticity:	this means that the variance of the residuals should be the same at each level of the explanatory variable/s. This can be tested for each separate explanatory variable, though it is more common just to check that the variance of the residuals is constant at all levels of the predicted outcome from the full model (i.e. the model including all the explanatory variables). This means that residuals should be uncorrelated.
Independent errors:	As with simple regression, the assumptions are the most important issues to consider but there are also other potential problems you should look out for: Normally distributed residuals: The residuals should be normally distributed.
Outliers/influential cases:	As with simple linear regression, it is important to look out for cases which may have a disproportionate influence over your regression model.
Variance in all predictors:	It is important that your explanatory variables ,Explanatory variables may be continuous, ordinal or nominal but each must have at least a small range of values even if there are only two categorical possibilities.
Multicollinearity:	Multicollinearity exists when two or more of the explanatory variables are highly correlated. This is a problem as it can be hard to disentangle which of them best explains any shared variance with the outcome. It also suggests that the two variables may represent the same underlying factor.

3.3. Assumptions of Multiple Linear Regression

Multiple linear regression is based on the following assumptions:

1. Linear relationships, outliers/influential cases: This set of assumptions can be examined to a satisfactory extent simply by drawing scattered charts of the relationship between each illustrative variable and the result variable. It's important to check that each scatter chart shows a linear relationship between variables (you might add a slope line to help you do so). Alternatively, you can only check the dispersion chart for the actual result variable versus the expected result.

Now that you're somewhat comfortable with the regression and the residual term, you might want to think about the difference between extreme values and slightly influential situations. Look at the two scatterplots below (Figures 5 & 6):

Note how the two problematic data points affect the slope line in different ways. A simple deviation affects the line to a much lesser extent but will have a very large remaining (distance to the slope line). SPSS can help you identify extreme values by identifying situations with particularly large residues. The dramatic anomaly changes the slope line but may be difficult to determine as the remaining one is small - smaller than most other points that are more representative! A situation of this extreme is very rare! In addition to examining the dispersion chart, you can also use impact statistics to identify points that may inappropriately affect the form. We'll talk about these statistics and how to interpret them in our example.

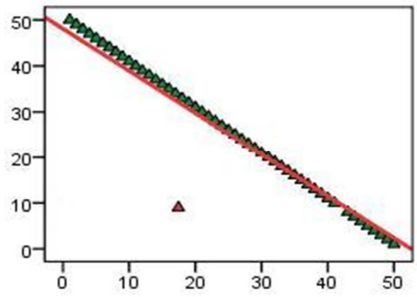


Figure 5. *Scatterplot showing a simple outlier.*

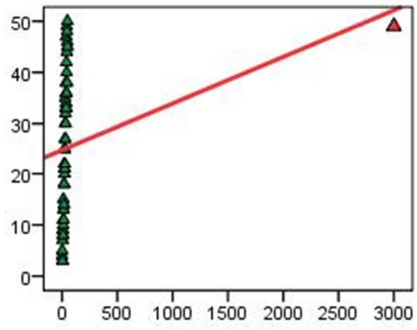


Figure 6. *Scatterplot showing an outlier that is an influential case.*

Bai, J. (1997). Estimation of a change point in multiple regression models. Review of Economics and Statistics, 79 (4), 551-563.

2. Variation in all explanatory variables: It's easy to verify this variable - just create a graph for each variable to make sure there's a range of values or that the data is divided into multiple categories. This assumption is rarely violated if you create good metrics for the variables, you care about. [8].
3. Multicollinearity: The simplest way to ascertain whether your explanatory variables are highly correlated with each other is to examine a correlation matrix. If correlations are above .80 then you may have a problem. A more precise approach is to use the collinearity statistics that SPSS can provide. The Variance inflation factor (VIF) and tolerance statistic can tell you whether or not a given explanatory variable has a strong relationship with the other explanatory variables. Again, we'll show you how to obtain these statistics when we run through the example!
4. Homoscedasticity: We can verify that the residuals `values do not systematically differ with the values expected by drawing the residuals values against the values predicted by the regression model. Let's get into this a little bit more deeply than we did before. We are looking for any evidence that residues differ in a clear pattern. Let's look at the examples below (Figure 7).

This scatterplot is an example of what a dispersal scheme might look like if the assumption of homoscedasticity is not met (this can be described as heteroscedasticity). Data points appear to be moving towards the negative end of the x-axis, indicating a greater variation in residual values at predicted values higher than lower expected values. This is a problem because it indicates that our model is more accurate when estimating lower values than higher values! In cases where the assumption of homoscedasticity is not met, it may be possible to convert the score scale. [9].

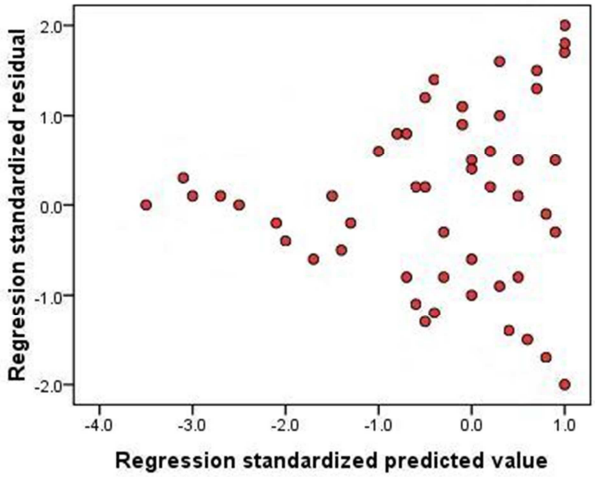


Figure 7. *Scatterplot showing heteroscedasticity - assumption violated.*

In Figure 8 the data points seem randomly distributed with an even spread of residuals at all predicted values.

5. Independent errors: As we have already stated, this assumption is difficult to test, but fortunately it applies only to data in which repeated measures have been taken at several point in time. It should be noted that if there is a high degree of assembly, multiple regression may be appropriate. The use of a complex SPSS sample unit, mixed unit, or separate multi-level modeling packages such as MLWin may be the only solution.

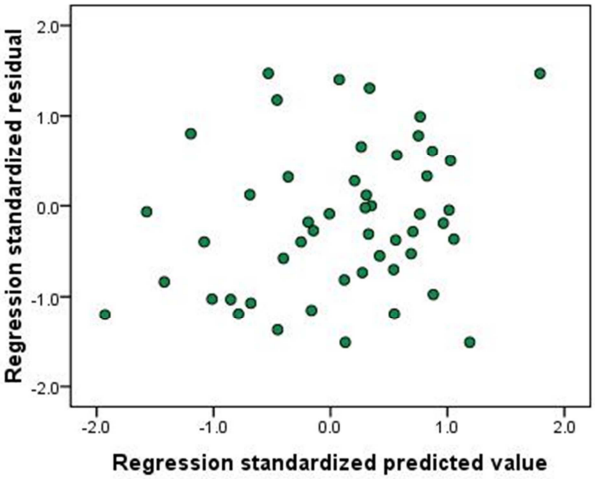
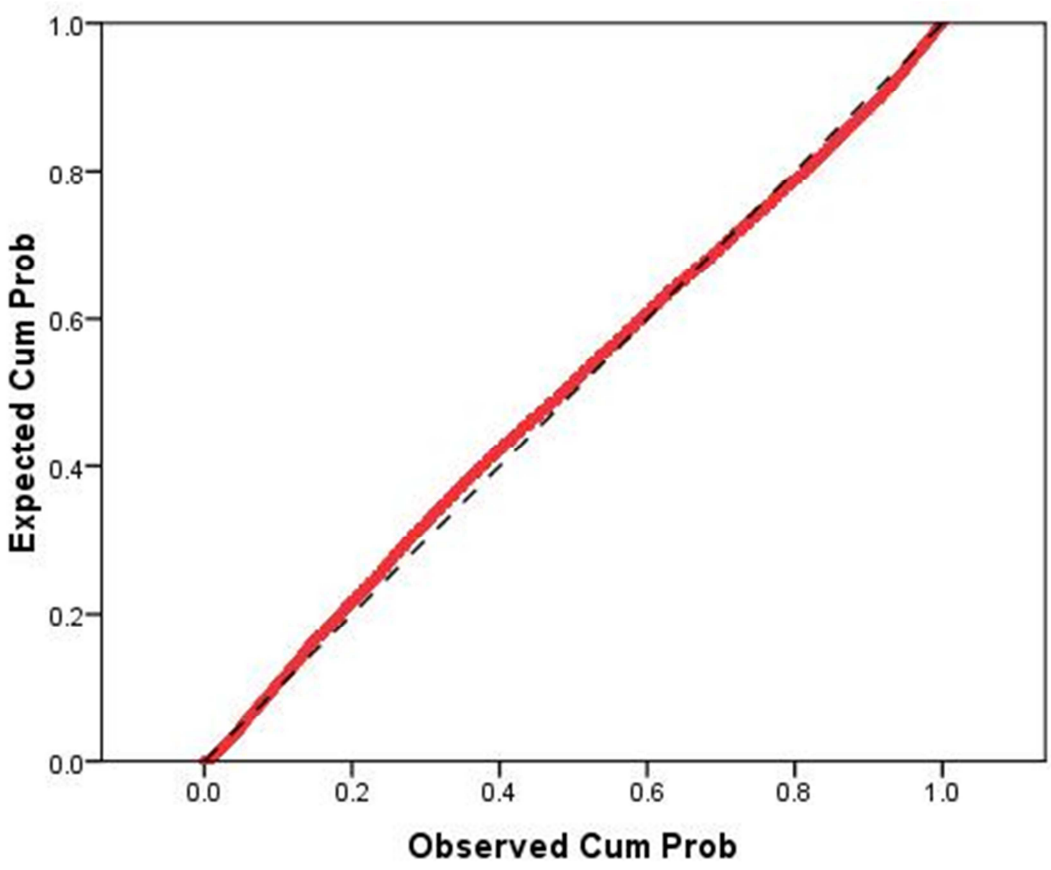


Figure 8. *Scatterplot showing homoscedasticity - assumption met.*

Poole, M. A., & O'Farrell, P. N. (1971). The assumptions of the linear regression model. Transactions of the Institute of British Geographers, 145-158.

6. Normally distributed residuals: A histogram of the residuals (errors) in our model can be used to check that they are normally distributed. However it is often hard to tell if the distribution is normal from just a histogram so additionally you should use a P-P plot as shown below (Figure 9):



Osborne, J. W., & Waters, E. (2002). Multiple Regression Assumptions. ERIC Digest.

Figure 9. *P-P plot of standardised regression residual.*

As you can see, the expected and noticeable cumulative possibilities, although not quite identical, are similar. This indicates that the residues are distributed almost naturally. In this example, the assumption is not violated.

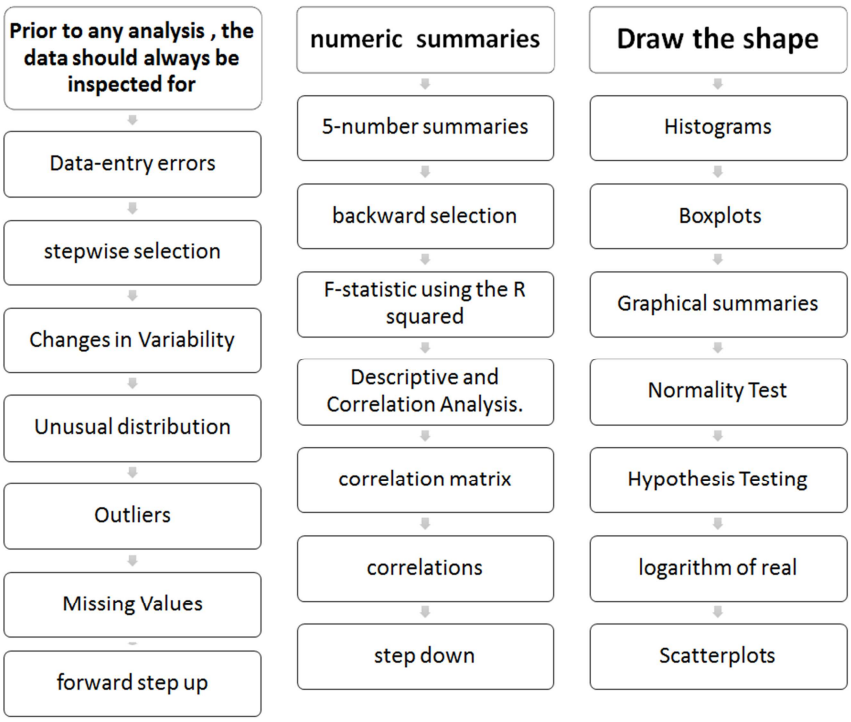
3.4. Detecting Problems

The following table lists the names of the most common assessment issues, a brief definition of each, its consequences, the typical tools used to detect them, and generally accepted ways to solve each problem. [10].

Table 3. *Detecting Problems in Multiple Regression Analysis.*

Problem	Definition	Consequences	Detection	Solution
High multicollinearity	Two or more independent variables in a regression model exhibit a close linear relationship.	Large standard errors and insignificant *t*-statistics Coefficient estimates sensitive to minor changes in model specification Nonsensical coefficient signs and magnitudes	Pairwise correlation coefficients Variance inflation factor (VIF)	1. Collect additional data. 2. Re-specify the model. 3. Drop redundant variables.
Heteroskedasticity	The variance of the error term changes in response to a change in the value of the independent variables.	Inefficient coefficient estimates Biased standard errors Unreliable hypothesis tests.	Park test Goldfeld-Quandt test Breusch-Pagan test White test.	1. Weighted least squares (WLS) 2. Robust standard errors.
Autocorrelation	An identifiable relationship (positive or negative) exists between the values of the error in one period and the values of the error in another period.	Inefficient coefficient estimates Biased standard errors Unreliable hypothesis tests.	Geary or runs test Durbin-Watson test Breusch-Godfrey test.	1. Cochrane-Orcutt transformation 2. Prais-Winsten transformation 3. Newey-West robust standard errors.

Source: lecture note Dr. Mahmoud A. Abdel-Fattah, AS611_Lecture01, 2022



Source: lecture note Dr. Mahmoud A. Abdel-Fattah, AS611_Lecture01, 2022

Figure 10. *Data cleaning and preprocessing.*

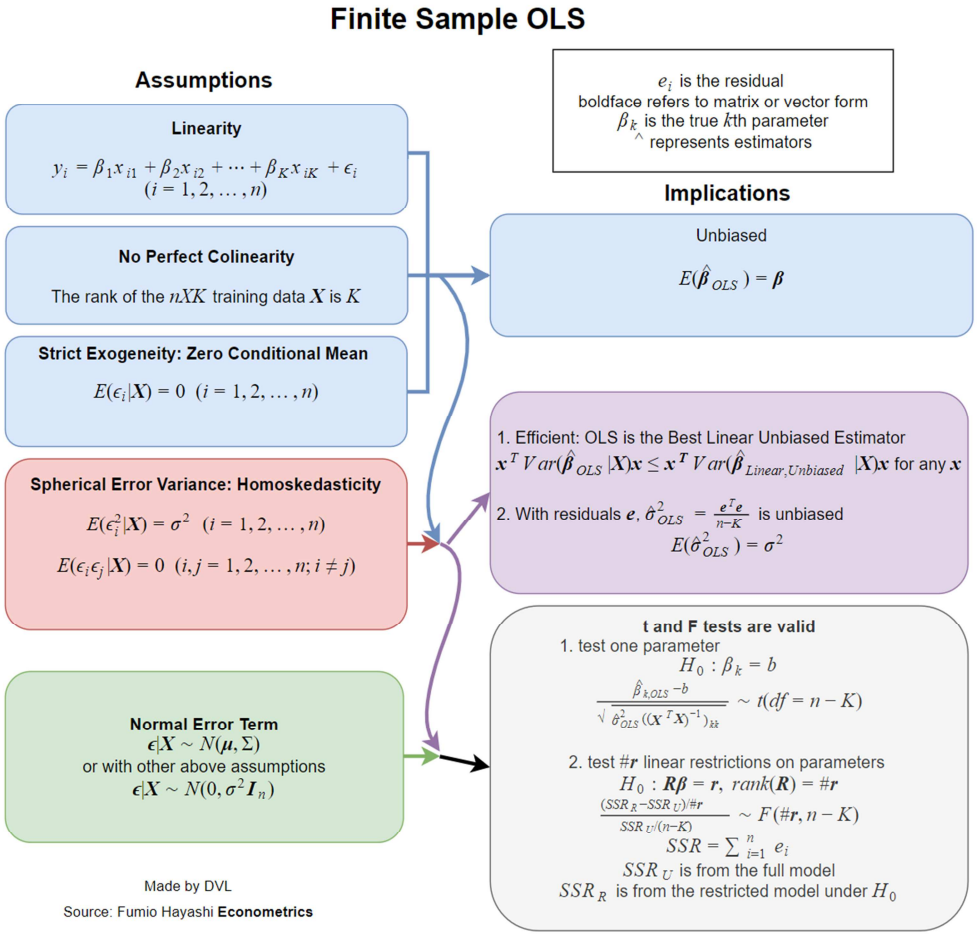
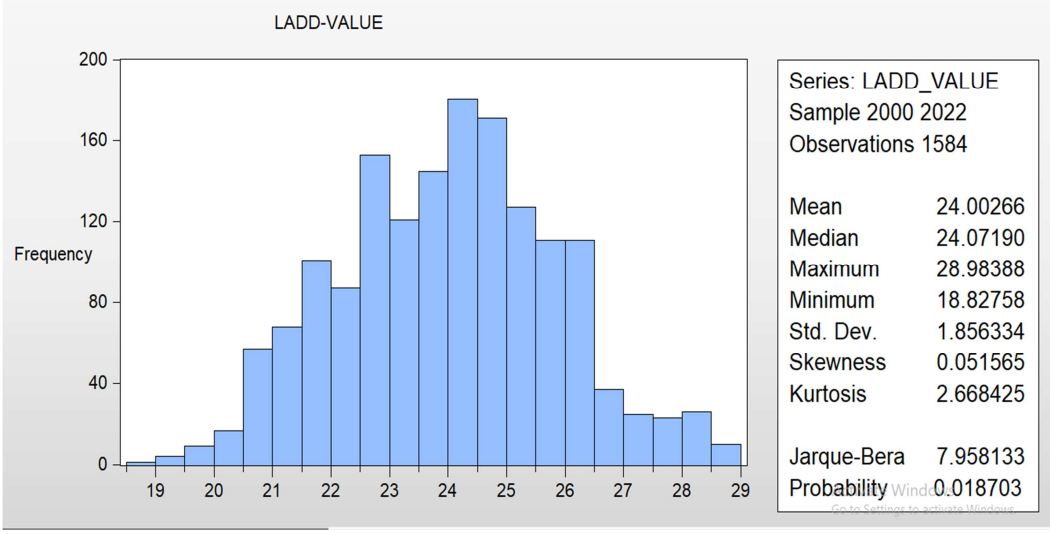


Figure 11. *Finite sample ols.*

4. Practical Application for Diagnostic Tests: I Present the Empirical Results in This Section

4.1. Normality Test for the Dependent Variable

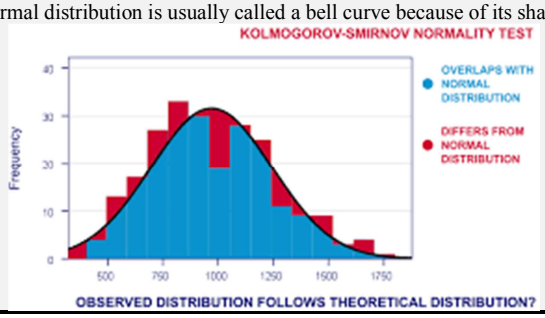
Normality tests are used to determine if a data set is well-modeled by a normal distribution and measures a goodness of fit of a normal model to the data.



Source: Prepared by the researcher based on the statistical program EViews 10th Edition

Figure 12. *Test for normality.*

Table 4. *Properties of Normal distributions.*

A normal distribution of data is one in which most data points are relatively similar, occurring within a small range of values, while there are fewer outliers on the higher and lower ends of the data range.	When the data is normally distributed, plotting it on the graph results in a bell-shaped, symmetrical image. In such a distribution of data, the mean, median, and mode are the same value and correspond to the peak of the curve.	The normal distribution is usually called a bell curve because of its shape. 

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3/4), 591-611.

From Figure 12, data of Add Value are given. Normality of the above data was assessed. Result showed that data were not normally distributed as skewness (0.0515) and kurtosis (2.668) individually were within ±1. Jarque-Bera test (P = 0. 0187) were statistically significant, that is, data were considered unnormal distributed.

Although both methods indicated that data were not normally distributed. As SD of the Add Value was less than half mean value (1.85 <24.006), data were considered unnormally distributed.

4.2. Correlation Matrix

A correlation matrix is a table showing correlation coefficients between sets of variables. Each random variable (Xi) in the table is correlated with each of the other values in the table (Xj). This allows you to see which pairs have the highest correlation.

Table 5. *Correlation matrix.*

Covariance Analysis: Ordinary Date: 10/25/21 Time: 14:13 Sample: 2000 2021 Included observations: 1563 Balanced sample (listwise missing value deletion) Covariance

Correlation	LHIGH_TECH	LADD_VALUE	LNANOTECHNOLOGY
LHIGH_TECH	0.308772		
	1.000000		
LADD_VALUE	0.685226	3.398561	
	0.668909	1.000000	
LNANOTECHNOLOGY	0.748596	3.299685	4.999181
	0.602531	0.800526	1.000000

Source: Prepared by the researcher based on the statistical program EViews 10th Edition.

Table 5 shows the relationship between the independent variables and shows the correlation MATRIX between the variable LNANOTECHNOLOGY and LHIGH_TECH (0.602531). This means that there is a strong positive relationship and it is called Multicollinearity problem.

4.3. Boxplot Test for the Variables

Boxplot using the inter-quartile range (IQR) to judge outliers in a dataset. Outliers are elements that exist outside of a pattern.

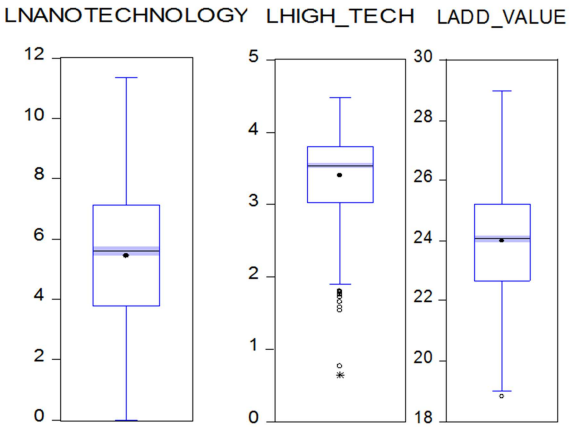


Figure 13. *Boxplot Test.*

Source: Prepared by the researcher based on the statistical program EViews 10th Edition.

Form figure 13 LHIGH_TECH, the boxplot shows that the median in the sample data is approximately 3.7, The minimum value is about 1.9, and the maximum value is about 4.8, in graph a dot plot to represent the outliers. An outlier is an observation that lies an abnormal distance from other values in a random sample from a population.

Form figure LNANOTECHNOLOGY, the boxplot shows that the median in the sample data is approximately 5.8, The minimum value is about 0.00, and the maximum value is about 11.9.

Form figure LADD_VALUE, the boxplot shows that the median in the sample data is approximately 24.00, The minimum value is about 19.00, and the maximum value is about 29.00.

4.4. Draw the Estimated Equation

LADD_VALUE = 22.1876727936 + 0.274374917648*LNANOTECHNOLOGY + 0.0979094342961*LHIGH_TECH

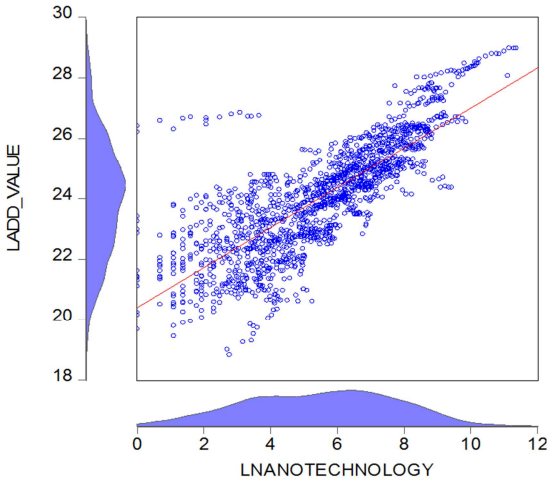


Figure 14. *Simple regression.*

Source: Prepared by the researcher based on the statistical program EViews 10th Edition

As appears in figure 14 Scatter diagram between independent variable and it is placed at the point corresponding to the measurement of the LNANOTECHNOLOGY (horizontal axis) and the LADD_VALUE (vertical axis). shows increasing positive of relation among variables.

LADD_VALUE = C(1) + C(2)* NANOTECHNOLOGY + C(3)* HIGH_TECH

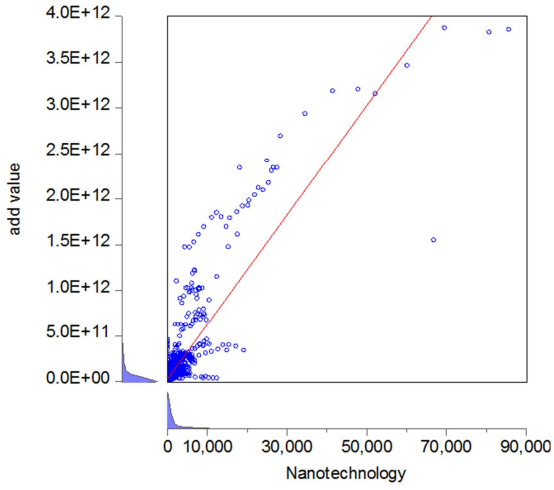


Figure 15. *Simple regression.*

Source: Prepared by the researcher based on the statistical program EViews 10th Edition.

As appears of observations collected in figure 15 is linear positively related, the pattern shows the covered area by the dots center's on a straight line. In this case the type of a straight line can adequately describe the general trend of the dots.

4.5. Autocorrelation Test

An identifiable relationship (positive or negative) exists between the values of the error in one period and the values of the error in another period.

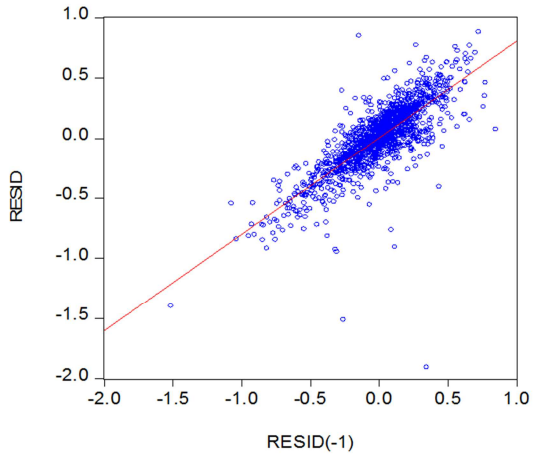


Figure 16. *Autocorrelation test.*

Source: Prepared by the researcher based on the statistical program EViews 10th Edition.

In figure 16 the diagram shows that there is a positive relationship in the form of a straight line. This means that there is an autocorrelation of the term of the random error, meaning that the term of the random error in any time period is related with the term of the random error in another time period. As it shown in equation:

$$u_t = \rho u_{t-1} + v_t \quad v_t \sim N(0, \sigma_u^2)$$

$$u_t \sim N(0, \sigma_u^2) \quad for\ all\ t$$
$$E(u_t, u_s) = 0 \quad for\ all\ t$$

4.6. Heteroscedasticity Test

1) The error term of our regression model is homoscedastic if the variance of the conditional distribution of ui given Xi,

Var(ui|Xi=x), is constant for all observations in our sample:

$$Var(ui|Xi=x)=\sigma 2\ \forall\ i=1,..., n.$$

2) If instead there is dependence of the conditional variance of ui on Xi, the error term is said to be heteroskedastic. We then write:

$$Var(ui|Xi=x)=\sigma 2i\ \forall\ i=1,..., n.$$

3) Homoskedasticity is a special case of heteroskedasticity.

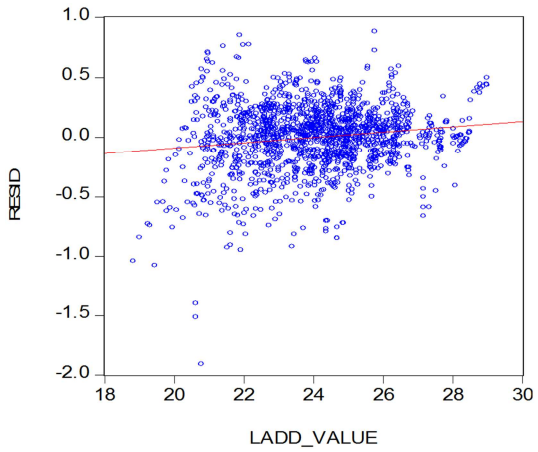


Figure 17. *Heteroscedasticity test.*

In figure 17 the diagram shows that there is a relation between residual and add value in the form of a straight line. meaning that there are heterogeneous disturbances in a linear regression model.

5. Conclusion

From the above we conclude the following:

1. The model suffers from the problem of heterogeneity of variance, and this leads to that the predictions in the variable Y depending on the estimators Bˆ 's (the coefficients of the independent variables) from the original data will have large variances, and this means that the prediction will be inefficient and the reason for this is that the variance The predictions will include the U variance as well as the parameters variance.
2. The model suffers from a problem of autocorrelation, which means that Cov(uj,ui) ≠ 0, and therefore the standard errors σ 2 are rather large, which means that the accuracy in the model is low and therefore the confidence intervals and the model's significance will be unacceptable and unreliable in and inefficient.
3. The model suffers from the problem of linear interference. This means that the estimators' values are very large and biased, as well as the variances of these estimators and the covariances are very large, so the properties of estimators are not BLUE.

Appendix of Study

The data was collected by the researcher from the data of the World Bank and referred to the link No 11 in the references.

Acknowledgements

I thank my father and mother Susan for what she endured during difficult times during the days of exams, and in recognition and gratitude to her and loyalty and compliance with her. I dedicate this humble effort. I wish god almighty to heal her and preserve her health and well-being, and to make

my good work in the balance of her good deeds.

I thank Dr. Hossam Elden M. Abdelkader, Associate professor Economic Dep., Faculty of Administrative Sciences, Ain Shams University, and at Faculty of– King Salman International University (KSIU). Egypt, I thank him for teaching the monetary policy course, benefiting from his knowledge, and teaching us advanced econometrics lectures in the preparatory year for the doctorate program, He is a distinguished young doctor, genius, and a role model for the youth. Dr. Hossam is like a sherbet that is watered with roses in sugar.

References

[1] Tae, K. H., Roh, Y., Oh, Y. H., Kim, H., & Whang, S. E. (2019, June). Data cleaning for accurate, fair, and robust models: A big data-AI integration approach. In Proceedings of the 3rd International Workshop on Data Management for End-to-End Machine Learning (pp. 1-4).

[2] Gharatkar, S., Ingle, A., Naik, T., & Save, A. (2017, March). Review preprocessing using data cleaning and stemming technique. In 2017 international conference on innovations in information, embedded and communication systems (iciiecs) (pp. 1-4). IEEE.

[3] Mahmoud Abd El Fattah (April, 2022), Faculty of Graduate Studies for Statistical Research and Econometrics of Statistical Studies and Research - Statistics Department, Cairo University.

[4] Hu, Y., & Plonsky, L. (2021). Statistical assumptions in L2 research: A systematic review. Second Language Research, 37 (1), 171-184.

[5] Meuleman, B., Loosveldt, G., & Emonds, V. (2015). Regression analysis: Assumptions and diagnostics. The SAGE handbook of regression analysis and causal inference, 83-110.

[6] Parke, C. S. (2013). Module 7: evaluating model assumptions for multiple regression analysis. Essential first steps to data analysis: Scenario-based examples using SPSS, 147-178.

[7] Garson, G. D. (2012). Testing statistical assumptions. Asheboro, NC: Statistical Associates Publishing.

[8] Osborne, J. W., & Waters, E. (2002). Four assumptions of multiple regression that researchers should always test. Practical assessment, research, and evaluation, 8 (1), 2.

[9] Ezekiel, M. (1925). The assumptions implied in the multiple regression equation. Journal of the American Statistical Association, 20 (151), 405-408.

[10] Roberto Pedace, (2016), Typical Problems Estimating Econometric Models, From The Book: Econometrics For Dummies.

[11] https://docs.google.com/spreadsheets/d/1rFz1kqV56NvQfw0vwHk4ZNfMzJL4IhUW/edit?usp=sharing&ouid=106936511433736093821&rtpof=true&sd=true.