Testing Assumptions in Linear Regression Models: A Case Study

CHUDA PRASAD DHAKAL

Tribhuwan University, Institute of Agriculture and Animal Sciences, Rampur Campus, Chitwan

ABSTRACT : Assumptions testing techniques in linear regression are not simply abundant, but are also severely confusing in many occasions. To shed light on such ambiguous scenario *through a real life example*, in this paper, two popular assumptions testing tools: regression diagnostic and statistical test of hypothesis are mentioned in detail. Accordingly the degree of clarity in the subject matter is met by testing every single assumption through both of these mentioned approaches. In this, the statistical test of hypothesis is the objective complement to its subjective counterpartregression diagnostic for testing assumptions. As such this paper will apparently help suffice assumption testing in linear regression analysis.

Key words: Assumption testing; Bruesh-Pagon, heteroscedasticity, regression diagnostics, Shaprio-Wilk.

Linear regression focuses on the assumptions of errors and researchers should be aware for checking these assumptions. A range of literature can be found in testing assumptions in linear regression model. For instance, in a few of them [(Poole&O'Farrell, 1970); (Osborne & Waters, 2002); (Assumption of linear regression, 2016); and, (Andrew, 2013)] in one or the other way all have discussed linear regression assumption testing. However, what these literatures consider to the assumptions and what their coverage are not the same in all respect. Some have only considered the assumption which are not robust to violation and some have even considered the reliability of the data taken as of the assumption of linear regression. But primarily, since when (Osborne & Waters, 2002) wrote the famous article entitled "Four Assumptions of Multiple Regression That Researchers Should Always Test" even to contrary to this article, most often while said assumption testing in linear regression; these are considered to be: Linearity, independence, constant variance and normality which has been dealt thoroughly in this paper. The central theme of the study in this case is, to demonstrate the testing of these assumptions via the popular approaches, regression diagnostic and the statistical test of significance.

Checking of these assumptions help avoid Type I and II errors.When any of these assumptions are violated, results and or the inferences yielded by a regression model would be either inefficient or badly misleading. However, as mentioned earlier it should be kept in mind that, not all assumptions are of the same in degree to robustness in violation.

Nau (2014) has been specific for what happens when an individual assumption is violated and has given the remedial in such cases. Violation of linearity causes problem when the fitted regression line needs extrapolation for prediction. The remedy to this is an appropriate nonlinear transformation to the dependent and/or independent variables. In case independence assumption is violated one may end up with a mis-specified model. Presence of heteroscedasticity make it difficult to gauge the true standard deviation of the forecast errors and this may also have the effect of giving too much weight to a small subset of the data. As a remedy for this, a log transformation applied to the dependent variable may be appropriate. Sometimes violation of independence and homoscedasticity could be the byproduct of violation of linearity assumption and/or also due to the violation of independence assumption in the case of heteroscedasticity. And hence these could be fixed as byproduct fixings. Finally, violation of normality create problems for determining whether model coefficients are significantly different from zero and for calculating confidence intervals for forecasts but this is less important if the goal is only to establish a correct model equation. Non-normality often arises either because (a) the distributions of the dependent and/or independent variables are themselves significantly non-normal, and/or (b) the linearity assumption is violated. In such cases, a nonlinear transformation of variables might cure both problems.

As mentioned earlier, two basic approaches to test the assumptions in linear regression are: regression

diagnostic and statistical tests of hypothesis. Regression diagnostic is the post model fit visual inspection of graphs and figures where as statistical tests of hypothesis are mathematical approaches. However, by what means should the assumptions be tested are neverdebate less. Below are some literatures to support this point.

Modern alternatives to residual plots are to plot the absolute value of the residuals and fit LOWESS curves through them. One approach for checking univariate normality is to examine histograms or stem-and-leaf plots for each variable. However, this is imprecise and only provides an indication. Many books present formal tests for residuals – I find these not particularly useful, and prefer the simple residual plots. However, one useful diagnostic is the Durbin-Watson test for autocorrelation. (Multiple Linear Regression:Chapter 21, n.d.).

Also, what raises the another confusion is, "While many statisticians argue that the independent and dependent variables don't need to be normally distributed, they all appear to agree that the error term must be normally distributed (Multiple Regression: Diagnostics and Solutions, n.d.)." Similar to this there are bountiful of literatures that one can investigate discussing how assumptions in linear regression analysis are tested or should be tested.Somehow while testing assumptions if a parametric statistical technique is not met non-parametric techniques are used.

To suffice the clarity on the mentioned issue of assumption testing,this paper demonstrates how assumptions are to be tested so that there would barely be any questions regarding the validity of a fitted regression model.

MATERIALS AND METHODS

Time series (1961-2013) data for the response variable PRODN [Annual Rice Production in Nepal (t)], and the predictors: ARHA [(Harvested Area [(Ha)], RURLPOPLN [Rural Population (000)] and FRMPR [Price at Harvest, NRS/t] were obtained from (IRRI, 2013). Assumptions were to be tested for the multiple regression model investigated from this data.

Finding answers to research questions involves developing theories and testing hypothesis (Sweet and Martin, 2012). Most often the context is, researchers will have a small portion of potential observation represented in the study, to know that relationship observed in such small sample also exists in the much larger population. Hypothesis testing (considers various statistical tests to be used properly as per the nature of the research question and the type of the data the study uses) is a tool used by the researchers, which help them understand, if there is enough evidence to reasonably believe a hypothesis to be true. Two logically opposing statements, called null (H0), and the alternative (Ha) hypothesis are used for this. The null hypothesis says that the relationship is due to chance and the alternative hypothesis says that the relationship is real. Researcher's with the help of probabilities, makes inference on which hypothesis is more strongly supported.

In the context of this study, for regression diagnostics we have followed the guidelines given by Pardoe (2014)and the various statistical tests of hypothesis we have employed with regression diagnostics are; Lack of fit test: for testing linearity, Durbin-Watson test for test of independence, Bruesh- Pagon test: for heteroscedasticity and Saprio-Wilk test for normality. Software used were Minitab vs. 16, SPSS vs. 20, and STATA vs.12. All regression diagnostics were conducted in Minitab, whereas for the tests of statistical hypothesis the software have been chosen as per the ease in analysis and also as the availability of the facility in the software for that specific assumption testing.

For linearity and the constant variance we createdplot of standardized residual vs. fitted values (figure 1) (a scatterplot with the residuals on the vertical axis and the fitted values, on the horizontal axis).

Accompanied was Lack of fit test in Minitab. The procedure follows:
Stat > Regression >Regression…>Options…>Lack of Fit Tests>OK.

For testing homoscedasticity as the diagnostic tool, the same plot (plot of residuals vs. fitted values) (Figure 1) was used.

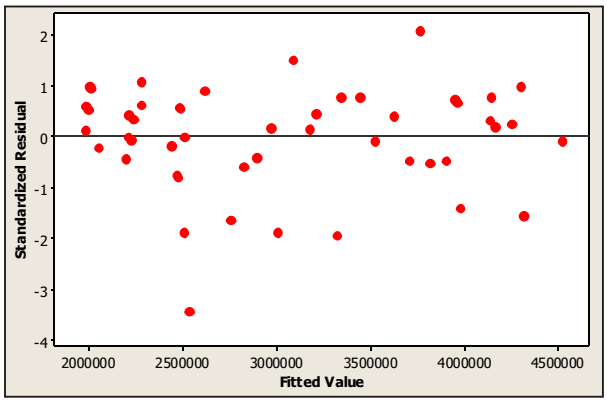

Fig. 1: Plot of standardized residual vs. fitted value

"Homoscedasticity and linearity: the two assumptions of random errors $\in^1$ have constant variation, and the random errors $\in^1$ have zero mean, can be checked at the same time. To do this, we use a residual plot. A residual plot is a scatter plot of the standardized residuals against the fitted values (Larsen, n.d.)."

Accompanied was Bruesh-Pagon test of constant variance. The procedure in Stata (Hamrick, 2012) follows: The command hettest ARHA FRMPR followed by enter yields the output of the test.

For the condition of independence of the errors i.e. serial/auto correlation assumption, we createdplot of standardized residual vs. order (Figure 2) (a scatterplot with the residuals on the vertical axis and the time sequence on the horizontal axis). Accompanied statistical test of hypothesis was: Durbin-Watson test in Minitab.
Stat > Regression >Regression…>Options…>Durbin-Watson statistics>OK

Finally, for the test of normality out of several alternatives the best one (Pardoe,2014) we created normal probability plot of the standardized residuals. Other alternatives for this (test of normality) could have been series of scatterplots with the residuals, on the vertical axis and each of the predictors in the model on the horizontal axes or to create a histogram and or a box plot of the residuals

Accompanied was Shaprio-Wilk test in SPSS:
Analyze> Descriptive Statistics> Explore…>Plots…
>Normality plots with tests>Continue>Ok

RESULTS AND DISCUSSION

We tested individual assumption with two approaches each. In regression diagnostic we assess the scatter plots visually and draw inferences, accordingly whereas in the statistical tests of hypothesis about any assumption tested either it was to reject a null hypothesis set up or to fail to reject the null hypothesis and accordingly the inferences were drawn.

Figure1 is the plot of the residuals versus fitted values. In this plot residuals are roughly centered and symmetric around the horizontal axis. However we can see some of the residuals fastened a bit apart from the constant variance range, these are the outliers. Ignoring some points as the possible outliers, the (vertical) average of the residuals remains close to 0 as we scan the plot from left to right. This affirms the plot demonstrates linearity assumption.

But, if the variation in the residuals had shown some asymmetrical pattern, assumption of linearity and constant variation would have come into question. And the condition was to be called nonlinearity.

Lack of it test was conducted to check linearity assumption. This test assesses the fit of the model with, H_0: The fit with the predictors is linear, and H_1: The fit is not linear. Model does not accurately fit the data, if the p-value computed by the test is less than the selected alpha-level (Minitab). The test showed, overall lack of fit test was significant $P=0.010 < 0.05$, at which led to reject the null hypothesis 'the fit is linear". It shows, the residuals no longer follow a linear trend and if the model is used as it is, this will not give reliable results. Meaning that true trend of the data is yet to captured by the model through further treatment and or transformation of data or so. For this at the first hand, outliers which we have observed in the diagnostic plots should be dealt properly to check if it solved the problem of nonlinearity.

The same (Figure 1) was used for testing the assumption of homoscedasticity. This time in the figure we check whether or not the (vertical) spread of the residuals was approximately constant as we scan the plot from left to right. And again putting some unusual residuals aside, visual assessment of the residuals showed it to be true. This confirmed the equality of variance i.e. the homoscedasticity condition was satisfied.

But, had the variation in the residuals shown some asymmetrical pattern and assumption of linearity and constant variation would have come into question and a heteroscedastic condition would have been to be dealt with.

For homoscedasticity assumption *Breusch-Pagan / Cook-Weisberg test* was conducted. This test was not significant ($X^2 = 0.45, >p = 0.05$). This meant, we failed to reject the null hypothesis of homoscedasticity i.e. the assumption of homogeneity of error variance was satisfied.

(Figure 2) is the plot of the residuals against time i.e., in the order that the data was collected. This is also called the order plot of the residuals. The plot when examined visually showed there was no systematic non-random pattern. This confirmed the assumption of independence of errors. The

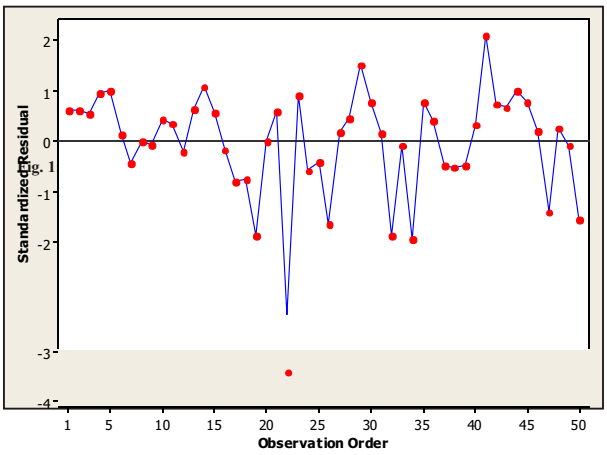

Fig. 2: Plot of standardized residual vs. order

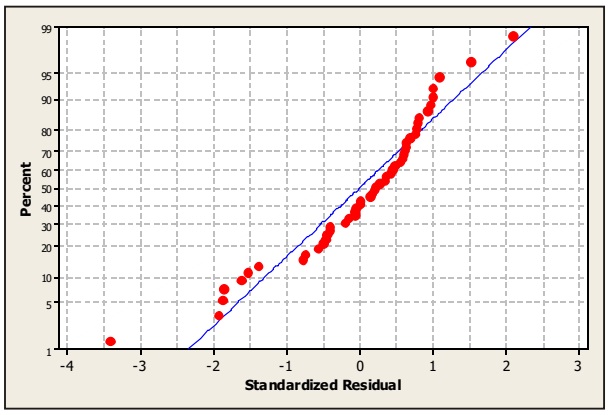

Fig. 3: Normal probability plot of standardized residual

errors therefore were not correlated. If the errors were correlated there would have been some error pattern and this may have suggested the need for some other type of the model to capture the true trend of the data.

Durbin-Watson Test was conducted for testing the assumption of autocorrelation. Assumptions of Linear Regression (2016)has described "Durbin-Watson Statistics, ' D' assumes values between 0 and 4, and values around 2 indicate no autocorrelation. As a rule of thumb the values of 1.5 < D < 2.5 show that there is no auto-correlation in the multiple linear regression data." The test result (1.5<DW=1.74< 2.5) implied failure to reject the null hypothesis: "residuals are not linearly auto correlated" and leads us to conclude that there was no evidence of autocorrelation.

Figure 3 is the normal probability plot of standardized residuals. Normality assumption is satisfied if normal probability plot reveals a straight line. That is, if the residuals uniformly encircle the line without deviating much from it, then this is reasonable to assume that the observed sample comes from a normal distribution. However, in (figure 3), though close, the normal probability plot is not convincingly revealing a straight line. We can clearly see some residuals deviate quite a bit far from the straight line. This shows the extent of departure from normality. So the normality assumption might not be satisfied for this data. Despite failure in normality assumption is commonly dealt with transforming the data into a new set which possibly could satisfy the assumption, it should be kept in mind that the effect might be due to one or more of the other assumptions are broken.

Shaprio-Wilk Test for normality conducted was found significant (SW(50) = 0.924, p < 0.05). This means the null hypothesis "the residuals are normally distributed" was rejected; which led to conclude that there was problem in the normality of the residuals. Hence, remedy is to be sought to solve this problem. Outliers cause a model not to be normal, therefore possible reason should be some of the residuals those which were observed in the diagnostic plots above.

CONCLUSION

Assumption testing is essential and it should not be undermined for having reliable linear regression model. Four basic assumptions to test are linearity, heteroscedasticity, independence and normality. Many a times ambiguity prevails on which methods to employ. However, the reality is assumption testing could be made sufficiently apparent to validate the model as is explained in this article. Therefore, this article could have a significant practical implication on mimicking the procedure in similar situation. In the line, one recommendation we make is because regression diagnostic is the visual assessment of assumption testing, always it may not be sufficient to rely on. For this reason, researchers are to cross check such visual testing methods by their objective counterparts called the hypothesis testing methods for testing assumptions.

However again, the challenging limitation is there never will be a case when the assumptions are perfectly satisfied. And in the one, the researcher will have to use his insight and or take help from other sources to make a relatively wiser decision. The subjectivity and the possible biasedness therefore, always prevails.

ACKNOWLEDGEMENTS

The writers of all papers and the books which are cited are all indebted. In addition thanks also go to all writers of any books, papers, and notes etc. which were helpful in any means in the course of preparing this research article.

REFERENCES

Andrew. (2013). What are the Key Assumptions of Linear Regression? *Statistical Modelling, Casual Inference and Social Science.*Published on internet. http://andrewgelman.com/ 2013/08/ 04/19470/. Accessed on 5 August 2016.

Assumptions of Linear Regression.(2016). *Statistics Solutions: Advancement Through Clarity*. Published on internet. http://www. statistics solutions.com/assumptions-of-linear-regression/. Accessed on 5 August 2016.

Hamrick, J. (2012). Executing the Breusch-Pagan Test in Stata. *You Tube*. Published on internet. http://www.youtube.com/watch?v=ZXxiv8MYQ IY.Accessed on 3 February 2014.

IRRI. (2013). Published on internet.http://ricestat.irri. org:8080/wrs2/entrypoint.htm. Accessed on 27 December 2013.

Larsen, P.V. (n.d.). Master of Applied Statistics. ST111: Regression and analysis of variance. *ST111: Data.* Published on internet. http://statmaster.sdu.dk/ maskel/docs/sample/ST111/data/data.ps. Accessed on4 February 2014.

Minitab (Version 16). (2010). Statistical Software. State College, PA: Minitab, Inc.

Multiple linear regression: Chapter 21. (n.d.). *Statistics and Acturial Science*: *Simon Faser University.* Published on internet. www.stat.sfu.ca/.../ Stat.../ JMP-part022.pdf. Accessed on 3 February 2014.

Multiple regression: Diagnostics and Solutions. (n.d.). Published on internet. http://www.yeatts. us/ 6200. Multivariate%2520Stats/Lectures-Tests/ Test%25202/Week-11-diagnostics-s...Accessed on 4 February 2014.

Nau, R. (2014). Regression diagnostics: testing the assumptions of linear regression. *Forecasting Home Page.* Published on internet. http://people. duke.edu/~rnau/testing.htm. Accessed on 26 December 2014.

Osborne, J. W. andWaters, E. (2002). Four Assumptions of Multiple Regression that Researchers Should Always Test. *Practical Assessment, Research & Evaluation*, 8(2). Published on internet. http://PAREonline.net/getvn.asp?v=8&n=2. Accessed on 29 December 2014.

Pardoe, I. (2014). Lesson 4: MLR model assumptions. *PennSate, Stat 501, Regression Methods.* Published on internet. https://onlinecourses. science.psu.edu/stat501/node/316. Accessed on 26 December 2014.

Pardoe, I. (2014). Lesson 4: SLR model assumptions. *Penn Sate, Stat 501, Regression Methods.* Published on internet. https://onlinecourses. science.psu.edu/stat501/node/275. Accessed on 26 December 2014.

Poole, M. A., and O'Farrell, P.N. (1970). The assumptions of the linear regression model. Published on internet. http://kharazmi-statistics.ir/Uploads/ Public/MY%20article/The%20assumptions%20 of%20the%20linear.pdf. Accessed on 5 August 2016.

SPSS (Version 20). (2011). IBM SPSS Statistics for Windows. IBM Corp, Armonk, NY.

Stata (Version 12). (2011). Stata Statistical Software. College Station TX: StataCorp LP.

Sweet, S. and Martin, G. K. (2012). Data Analysis with SPSS: A First Course in Applied Statistics (4th ed.). Allyn and Baco, Boston USA.

Received: February 26, 2016
Accepted: September 5, 2016