European Online Journal of Natural and Social Sciences 2015;
Vol.4, No.4 pp. 725-736
ISSN 1805-3602

Data Health Assurance in Social and Behavioral Sciences Research

Ch. Mahmood Anwar
Scholars Index, Jersey City, USA
Email: Mahmood.Anwar@scholarsindex.com; Tel: (302) 364-0795

Received for publication: 14 July 2015.
Accepted for publication: 20 October 2015.

Abstract

This illustrative study reincarnates the philosophical assumption of Methodology for science among other assumptions like Epistemology and Ontology in the context of social and behavioral sciences. Based on literature review the study divided the overlapping and perplexing Gaussian Linear Regression Model (GLRM) assumptions into two comprehensive groups. The study modeled straightforward diagnostics for GLRM assumptions violations by using the data collected from 150 postgraduate university students. Finally, the study provides the remedial directions to address possible problems created by GLRM assumptions violations.

Keywords: Gaussian Linear Regression Model, Statistics for Behavioral and Social Sciences, Research Methodology, statistical assumptions and diagnostics, Parametric tests.

Introduction

Unlike natural sciences, in which the scientific theories are rooted in hard facts of the world, social sciences raise few definitional issues because the social theories are built on personal opinions or tentative thoughts (Kangai, 2012) and people interpretation of world (phenomenological view) (Moustakas, 1994), while some are socially constructed realities (Searle, 1995). Generally, social sciences study human behavior, social groups and institutions; subdivided into multifarious areas like Anthropology, Commerce, Behavioural Science, Economics, Political Science, Education, Management, Psychology, Public Administration etc. Due to the complexity and unpredictability of *The Human Element*, social studies can only said to be scientific if all observations leading to theories are carried out carefully and impartially to attain objective and secure footing for science (see Chalmers, 1999). At this point, it can easily be observed in social knowledge context that among other philosophical assumptions for science like Epistemology and Ontology (Bryman, 2001), Methodology assumption (Cohen, Manion & Morrison, 2001) is more important whether using positivist or constructivist paradigms. No doubt, nowadays social scientists are conflicting with Epistemological assumption (Ghoshal, 2005), till present the Epistemological and Ontological assumptions hold the arena of social sciences strongly. The present precise introduction to scientific philosophy hence emerged the fact that beyond the vitality of Epistemological and Ontological assumptions the Methodological assumption plays an important role to label social studies knowledge as "science". This point satisfies the explanation of Chalmers (1999) for what constitutes science.

After establishing my position about true value of Methodology, I would rather move inside the assumption. In the context of positivist paradigm, methodology consists of research design, data collection, measurements and data analysis methods (Tharenou, Donohue & Cooper, 2007). Among these ingredients of methodology, data make soul for labelling the body of social knowledge as "social sciences". If health of the data is good than the social theories built on the data will reflect

good science and vice versa. Therefore, social scientists must investigate and report the health and unbiasedness of data to satisfy the data health assumptions and diagnostics (Gujarati, 2004).

I was motivated to write this article while surfing and studying the "Instructions for authors" given by top ranked management, behavioral and social science journals. I carefully examined the author resources of highest quality journals like Academy of Management Journal, Journal of Management, Journal of Organizational Behavior, Journal of Economic Literature etc, and found no specific detailed instructional statement endorsing or encouraging authors regarding reporting the health of data in terms of statistical assumptions and diagnostics. However, very few journals understand (e.g. American Psychological Association Journals, International Journal of Management, Economics and Social Sciences) the significance of data health and encourage the authors to report it in the submitted manuscripts. I personally feel that the instructions for authors by a particular journal are most influential masterpiece for authors submitting their manuscripts for publication. Therefore, this study is aimed at highlighting importance of investigating health of data (soul) leading inferences and estimations in social sciences (body). This article is very significant because researchers and students of social and behavioral sciences will now be able to understand Gaussian Linear Regression Model (GLRM) assumptions and their diagnostics evocatively under one umbrella.

Gaussian Linear Regression Model (GLRM) Assumptions

Cuthbertson, Hall and Taylor (1995) pen a million dollar truth that the application of statistics on data is not like a mechanical mechanism but it requires deep knowledge, intuition and adroitness. We know that ordinary least square (OLS) method is sufficient to approximate the population regression function (PRF) estimators, but in social sciences we are more interested to draw the inferences rather than just mathematical estimations, hence, need accurate values of model estimators. In addition, PRF also depends upon the disturbances which make the model more flimsy if underlying assumptions are violated. The accuracy can only be achieved by taking few assumptions into account. These assumptions are eleven in counting and known as Gaussian Linear Regression Model (GLRM) assumptions (Gujarati, 2004). To keep it simple, I would only review six assumptions because I think these assumptions are more important in social science research.

The range of diagnostic tests actually belongs to Gaussian Linear Regression Model (GLRM) assumptions which I will touch in methodology latter. A wide class of diagnostic tests have been reported in literature including examination of residuals, Durbin–Watson d-test, RESET test (Ramsey, 1969), Lagrange Multiplier test (Engle, 1982), discrimination and discerning (Harvey, 1990), J Test (Davidson and MacKinnon, 1981), the JA test, Cox test, Mizon–Richard test, the P test (Baltagi, 1998), outliers, leverage, and overly influential cases, recursive least squares, Chow's prediction failure test etc.

The breadth of the topic under consideration is as wide as many specialized books are required to cover it. However, following Peter Kennedy's keep it stochastically simple principle, I would discuss only those diagnostic statistics which are simple and easy to calculate or examine for students and researchers by using conventional statistical software like SPSS.

Now, the question arises that among these eleven assumptions how much one should pay attention to some while neglecting others? For instance, Gujarati (2004) weighted all assumptions equally but Tabachnick and Fidell's (2001) discussed linearity, normality, multicollinearity, homoscedasticity and outliers. Many text books amalgamate these assumptions which often confuse the readers and students. I feel that researchers should not understand these assumptions allegorically, but literally. In fact, there exist two types of assumptions; one is about model specification and disturbances, while other is about data. Linearity, homoscedasticity, independence,

model specs and Gaussianity assumptions belong to first type, however, singularity (multicollinearity) and model bias part of model specification assumption belongs to second type (see Wetherill, 1986). Now I will explain each assumption precisely and use to the point approach.

Linearity

This assumption states that the regression model should be linear in parameters. Most people think linearity as if the conditional expectation function (CEF) of regressand is a linear function of regressors i.e. a straight line. In fact, conditional expectation function should be a linear function of the model estimators. It simply means that the estimators must have a power of one and must not multiply or divide with each other. The parametric linearity of the regression function is essential because the regressand and regressors may be linear or non-linear but parameters should be linear to satisfy this assumption (Gujrati, 2014).

Homoscedasticity

According to this assumption the conditional variances of stochastic disturbances should be identical for conditional expectation function (i.e. equal variance). Linguistically, the meaning of homoscedasticity is equal spread (homogeneity of variance) and the word was derived from Greek word. The opposite situation is known as heteroscedasticity in which disturbances are not identical for conditional expectation function. Homoscedasticity simply means that the variation around the regression line should be identical across the values of regressor. The probability of heteroscedasticity is greater on cross sectional data (Gujrati, 2014).

Independence

This assumption states that all regressors should be independent form each other. In simple words, the correlations among the disturbances of two or more regressors must be zero (no autocorrelation). For cross sectional data, the chance of dependence among regressors is less with the random sampling and increase with convenience sampling. Non-random sampled data may sometimes indicate spatial dependence among regressors, but the problem of autocorrelation is more serious in time series data, especially when the time interval between data collection points is short.

Nowadays, many researchers use the term autocorrelation and serial correlation synonymously. But, in fact autocorrelation is the lagged correlation of a series of data with itself, whereas, lagged correlation between different data series is known as serial correlation (Tintner, 1965).

Singularity

There should be no perfect linear relationship among the regressors according to this assumption. This concept was first introduced by Ragnar Frisch in 1934, which simply means the perfect linear relation among few or all regressors in the regression model. We know that in real life nothing is perfect, so, nowadays researchers are using it as multicollinearity. However, in the case of perfect multicollinearity (singularity), model estimators would be sitting on the fence having infinite disturbances. While, with less than perfect multicollinearity, the estimators can be determined, having large disturbances leading to inaccurate and imprecise estimators. In the near to multicollinearity or with small number of observations, the OLS estimators are still BLUE but have large variances and co-variances leading to imprecise estimations and wrong statistical inferences. Goldberger introduced the term micronumerosity for effects of sample size on estimation. Montgomery and Peck (1982) indicated many sources of multicollinearity like data collections errors, model specifications, over determined model, model constraints and regressors sharing common trend in time series data.

Gaussianity

The GLRM also assume that the disturbances of each regressor should follow Gaussian (i.e. normal) distribution. To understand this, it is essential to inform readers that the theoretical

justification of Gaussianity is rooted in the famous central limit theorem (CLT) (Fischer, 2011). According to the theorem, the sum of large number of independent and identically distributed (*i.i.d*) random variables leads to Gaussian distribution as the variables increase indefinitely. This hints that the dependent variable is actually influenced by the disturbances from the number of independent variables in the regression model. In addition, because the linear function of Gaussian random variable is itself Gaussian, hence, the probability distributions of model estimators can easily be derived.

The assumption of Gaussianity is not important if the objective is only estimation, because the OLS estimators are BLUE even if disturbances are non-Gaussian. But mostly, the objective of researchers in social sciences is testing hypotheses and making inferences with small or medium sized sample, in this case the assumption of Gaussianity becomes critical. Tharenou, Donohue & Cooper (2007) explained that multivariate Gaussianity is more difficult to test; researchers should ensure univariate Gaussianity which reduces the chances of multivariate Gaussianity.

Model Specification and Bias

Diagnostic tests are the sub-procedures to check the assumption of selecting correctly specified regression model for analysis whilst violation of the assumption leads to model specific errors (under or over fitting) or bias (Gujarati, 2004). The presence of these errors in the regression model can be looked with the help of regression fishing.

We know that conditional expectation function (CEF) is parametric in nature and needs transformation into statistic called stochastic sample regression function (SRF) which estimates the CEF. SRF informs that differences between the actual and estimated values of any dependent variable are important and termed as residuals. The residuals can have positive or negative values. Now the question of interest is that how these residuals influence regression model? To have an answer, I will proceed to next section.

Outliers and Overly Influential Cases

Outliers and Overly Influential Cases are nothing but data points in regression model. To understand them precisely, recall few basic concepts of regression estimations. SRF can only be estimated precisely if the sum of residuals is as small as possible, but in reality, some residuals receive equal weights while others receive unequal weights. Thus, Gauss a well known scientist proposed Ordinary Least Square (OLS) method to resolve this problem because least square criteria assigns more weight to underestimated residuals (Gujarati, 2004). Hence, outliers can easily be understood as cases from different population or the case having greater effect than the majority of other sample cases. Both outliers and overly influential cases distort the regression line and reduce generalizability of the regression model

There are two types of outliers, simple and multivariate. Simple outliers reflect cases having extreme value with respect to one variable, whereas, multivariate outliers are the cases with excessive values with respect to several variables (Garson, 2012).

Methodology

Due to the illustrative nature of this study, the main focus was not on special constructs or any specialized sampling technique. Self-administered 7-point Likert type questionnaires were distributed among 185 postgraduate university level students while comprehensively briefing them about the objectives of the study. The questionnaire comprises of three sample constructs i.e. proactive personality (IND1), creative-self efficacy (IND2) and knowledge sharing behavior (DEP) along with questions related to basic demographics of sample. Out of 185 questionnaires 161 were received back (response rate 87.02%). After analyzing the questionnaires for wrong entries and uncompleted ones, 150 questionnaires (93.10% of total received questionnaires, with acceptable

Cronbach's alpha values) were finally selected for testing of GLRM diagnostics. Sample comprises of 76.66% male respondents 23.33% female respondents.

Linearity Diagnostics

There are many tests available to confirm linearity of regression model like graphical methods, RESET test (Ramsey, 1969), Eta test, ANOVA linearity test, linear to non-linear comparison and curve fitting test etc. However, the purpose of this study is to provide easy to understand diagnostics to students and researchers (fingertips approach) although many difficult tests are also available. To test the linearity assumption, I applied mean procedures test (MPT) first and then crosschecked the results by running non-linear association (*Eta*) test. To do this, run MPT test in SPSS, enter dependent variable and independent variables into the model and select test of linearity. The following result is produced:

Table 1: Mean Procedure Test For Linearity

Variable Pairs	Condition	*SS*	*f-ratio*	*Sig*
DEP*IND 1	Combined	189.232	4.306	.001
	Linearity	158.087	154.669	.001
	Non-Linearity	31.145	0.726	.868
DEP*IND 2	Combined	145.517	2.259	.001
	Linearity	127.205	108.298	.001
	Non-Linearity	18.312	1.114	.356

The mean procedure test (ANOVA) for linearity splits down the paired groups between their linear and non-linear components making more easy for researcher to diagnose presence of non-linearity. It can be seen that for first pair of variables, the *f-ratio* for linearity component is significant at .001, whereas, the *f-ratio* for non-linearity component is insignificant (p< .86). This simply means that the groups are linear and data meet the assumption of linearity significantly. For the second pair of variables, the *f-ratio* for linearity component is significant at .001, whereas, the *f-ratio* for non-linearity component is insignificant (p< .35). This again means that the groups are linear and data meet the assumption of linearity significantly. If the groups have non-linearity, the level of significance for non-linear component would be less than .05 with the significant (p<.05) or insignificant linear component (p>.05). Hence, this test informs researchers about total linearity, total non-linearity and partial linearity.

To crosscheck the results, non-linear association (NAT) test was also run. For this, run mean comparison (ANOVA) test by entering the model variables and selecting the ANOVA table and *Eta*. The following results were produced. For the first pair of variables, the correlation coefficient (*R*) is .69 where as coefficient of non-linear association (*Eta*) is .74. The difference between correlation coefficient and non-linear coefficient is .04. The rule of thumb is that a model is perfectly linear if non-linear coefficient is equal to correlation coefficient. As the daily life is not ideal, there is nothing perfect hence we concluded that our model is linear.

Similarly for the first pair of variables, the correlation coefficient (*R*) is .77 where as coefficient of non-linear association (*Eta*) is .85. The difference between correlation coefficient and non-linear coefficient is .07, hence we concluded that our model satisfactorily fulfils the assumption of linearity. Researchers should note that the difference of non-linear coefficient and correlation coefficient will determine the extent of non-linearity in the model.

Table 2: Non-Linear Association Test

Measures of Association				
Variables	***R***	***R Squared***	***Eta***	***Eta Squared***
DEP * IND 1	.697	.486	.746	.556
DEP * IND 2	.777	.604	.850	.723

Singularity Diagnostics

Singularity (multicollinearity) was tested by calculating Variance Inflation Factor (VIF), Tolerance (TOL) and Condition Indices (CI) as suggested by Tharenou, Donohue and Cooper (2007) and Garson (2012). Kline (2005) suggest that for a multiple regression model, value of Variance Inflation Factor greater than 10 and tolerance less than .10 may indicate multicollinearity. However, many statisticians follow more strict rules and do not allow VIF greater than 4 and TOL value less than .25. At the same time the values of Condition Index above 15 indicate possible multicollinearity and above 30 indicates serious multicollinearity (singularity) problem (Garson, 2012; Gugrati, 2004). The VIF values calculated for this study were lower than 4, TOL greater than .25 and CI less than 15, thus multicollinearity does not appear to be a problem for this study model. Hence, no evidence of multicollinearity was found in the model as the results are reported in Table 3.

Table 3: Collinearity Diagnostics

	Collinearity Statistics		
Variables	**Tolerance**	**VIF**	**CI**
IND 1	.487	2.053	6.600
IND 2	.390	2.564	9.800

Independence Diagnostics

Independence of the residuals was first tested by estimating the popular Durbin-Watson scores for detection of independence of variables (Gujrati, 2004) and then crosschecked by calculating Intra-class Correlation Coefficient (ICC) as suggested by Garson (2012). The Durbin-Watson test uses studentized disturbances to calculate test coefficient. It is suggested by the author that the score far away from 2 and near to 0 indicates the problem with independence. However, the score near 2 indicates that the model satisfy the independence assumption. Many strict statisticians recommend that DW coefficient should lie between 1.5 to 2.5 for independent observations. A DW score of 1.47 was calculated for our regression model which showed the variables are independent and have no significant evidence of autocorrelation.

Table 4: Test of Independence

Model Summary[b]					
Model	**R**	**R Square**	**Adjusted R Square**	**Std. Error of the Estimate**	**Durbin-Watson**
1	.803[a]	.644	638	911	1.47
a. Predictors: (Constant), IND 1, IND 2					
b. Dependent Variable: DEP					

The results were cross validated by calculation ICC value for our model. ICC method constructs a null linear model to estimate reliability coefficient. In this study, I used a two way random effect model with consistency type mix at .05 significance level to determine IC coefficient against the value zero (0). It can be seen that for this study model the average scores of independent variables are highly reliable, generating the ICC value of .83 (interval .75 to .88 with 95% confidence). The significant ICC value shows significant data independence.

Table 5: ICC Test of Independence

Intraclass Correlation Coefficient							
Type	**Intraclass Correlation[a]**	**95% Confidence Interval**		***f-ratio* with True Value 0**			
		Lower Bound	**Upper Bound**	**Value**	***df1***	***df2***	***Sig***
Single Measures	.713[b]	.609	.792	5.961	114	114	.001
Average Measures	.832[c]	.757	.884	5.961	114	114	.001
a. Type C intraclass correlation coefficients using a consistency definition-the between-measure variance is excluded from the denominator variance.							
b. The estimator is the same, whether the interaction effect is present or not.							
c. This estimate is computed assuming the interaction effect is absent, because it is not estimable otherwise.							

Hetroscedasticity Diagnostics

There are many diagnostic techniques available to detect hetroscedasticity. For instance, one can use graphical methods, Goldfeld-Quandt test, weighted least square regression, Glejser model, Park test, Breusch-Pagan Godfrey test, White's test or Szroeter test to detect hetroscadesticity of regressors. There are no hard and fast rules to detect this problem and all available tests are useful for situations dependent upon sample size and type of hetroscedasticity. I have similar believe like Wilkinson and TFSI (1999) that instead of calculating complicated statistics, one should use simple graphical methods to check this assumption quickly. For large to medium sample size, the best way is to plot the squared disturbances in the regression model and see the patterns of residuals. However, I have plotted disturbances of both independent variables (IND 1 and IND 2).

These residual plots can easily be drawn in SPSS. For instance, from the plot of our first study variable (see Figure 1), it can be seen that many overlapping residual patterns exist in the variable with irregular shapes. To view this plot more precisely, a linear line has been drawn with SPSS automated feature. As the result, a sinusoidal shape along the linear line (automatically drawn by SPSS) clearly reflects that our first variable (IND 1) is suffering from hetroscadesticity.

The plot for second variable (IND 2) is shown below (see Figure 2). It can be seen that few hetroscadastic patterns are there but linear line is clear and not showing any sinusoidal distribution. Hence, it can be inferred that our second variable is not affected by hetroscadesticity.

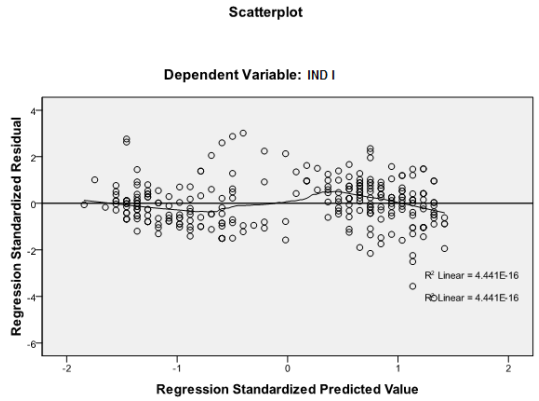

Figure 1: Residual Plot for IND 1

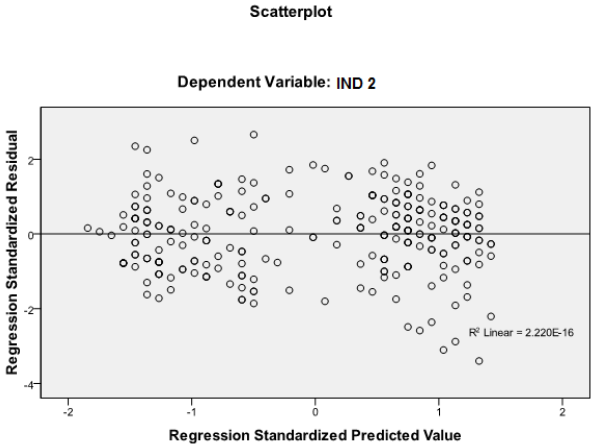

Figure 2: Residual Plot for IND 2

Gaussianity Diagnostics

The assumption of Gaussianity was tested with Skewness and Kurtosis tests as suggested by Tabachnick and Fidell (2001). Ayentimi et al. (2013) explained skewness as a measure of asymmetry of the distribution. A distribution can be Gaussian i.e normal, positively or negatively skewed. In the same study, they defined kurtosis as the peakedness or flatness of a distribution. Non-normal distributions can be Leptokurtic and Platykurtic. Kendall and Stuart (1958) indicated that to satisfy the assumption of Gaussianity the absolute values of skewness should not approach 2 and kurtosis should not be greater than 5. Whereas, strict statisticians say that kurtosis range should be +3 to -3 (Garson, 2012). The result obtained for skewnes and kurtosis for our model are shown below.

Table 6: Gaussianity Diagnostics

Gaussianity Statistics

Variables	Skewness	Kurtosis
DEP	0.145	-1.677
IND 1	0.161	-1.295
IND2	0.095	-1.414

The Gaussianity histograms for the study variables are also shown in Figure 3 and Figure 4, showing satisfactory result for the assumption of Gaussianity.

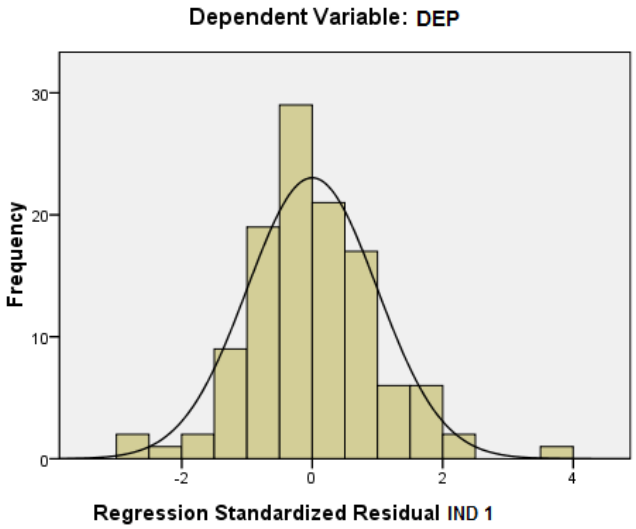

Figure 3: Gaussianity Histogram for IND 1

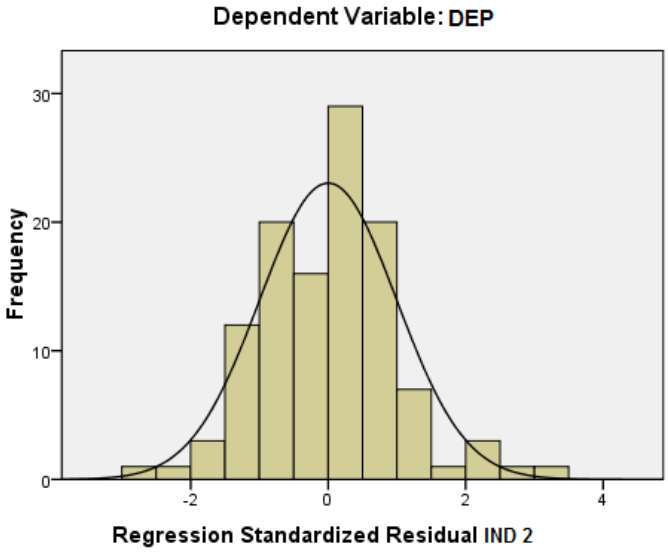

Figure 4: Gaussianity Histogram for IND 2

Diagnostics for Outliers and Overly Influential Cases

Simple outliers are those located at plus minus three standard deviations from the mean or more. In most research studies, the main problem is the detection of multivariate outliers. Multivariate outliers can be detected by graphical examination of disturbances and four statistical methods i.e. Cook's distance, Standardized residuals, Leverage and Mahalanobis distance. If the value of standardized residuals is greater than equals to 3, Cook's distance greater than 1 and cases with highest Mahalanobis D-square value, the outliers should be removed (Garson, 2012). Overly influential cases should be removed of the value of leverage is > .5, however, the safe range is ≤ .2.

Table 7: Outliers and Overly Influential Cases Detection

Diagnostic Statistics	Value
Standardized Residuals	$\leq \pm 2$
Leverage	.067
Cook's Distance	.159
Mahalanobis Distance	.038 − 7.687

It can be seen in Table 7 that our study model diagnostic statistics values lie in safe range as suggested by statisticians like Tabachnick and Fidell (2001) and Garson (2012). It should be also worth mentioned that Orr, Sackett and Dubois (1991) study concluded that researchers should prefer visual examination of outliers and overly influential cases than numerical examination.

Discussion

A good study or teaching aid should make several contributions. First, this illustrative study established the true value and importance of Methodology assumption (among other philosophical assumptions for science) responsible to label social knowledge as "science". Second, the study established the fact that in social sciences the most important issue is the diagnosis of data health which should be done by researchers painstakingly. The meticulous analyses of data will determine whether the data will provide true inferences in testing hypotheses. Third, the study clarified the confusion of researchers and students by introducing two groups of GLRM assumptions. After establishing this position, fourth, the study provided simple visual and numerical diagnostic techniques to detect possible data health problems. Fifth, the study endorsed that researcher should try to build their skills up to the level so that they may analyze the violations to the assumptions graphically as the use of distributional tests and stats shape indexes are not preferred substitute of graphical methods for analysis of disturbances. The summary based tests, shape indexes and increase in sample size create issues in detecting distributional irregularities in disturbances. Therefore, according to many researchers the graphical methods should be the first priority to analyze the GLRM assumptions violations (Wilkinson & TFSI, 1999).

Now, question arises, what to do if data are found to be suffered from GLRM assumptions violations? Although this question is out of the scope of this study but I will provide helpful directions to the researchers to rectify the violations to the assumptions.

In the view of Blanchard (1967) "do nothing school of thought", multicollinearity is essentially due to micronumerosity, and in social science researchers have no control over data available for empirical analysis. Therefore, researchers should follow a do nothing approach. Although, we cannot estimate one or more OLS estimators with quality precision but estimable function can be estimated efficiently (Conlisk, 1971). On the contrary, statisticians suggest to drop variables and specification biases, transformation of variables, new data collection, reducing collinearity in polynomial regressions, orthogonal polynomials (Draper & Smith, 1981), principal components and ridge regression (Chatterjee & Bertram Price,1977; Vinod, 1978) methods to deal with multicollinearity.

Although heteroscedasticity does not annihilate consistency and unbiasedness of OLS estimators but it can affect the precision of hypothesis testing and make it ambiguous. Gujati (2004) suggests, if constant or homoscedastic variance of residuals is known than weighted least square method is useful to get BLUE estimators (heteroscedasticity correction). But these constants are rarely know hence the other method to correct heteroscedasticity is to measure White's Heteroscedasticity-Consistent Variances and Standard Errors.

Just like the case of multicollinearity, transformation of variables are also recommended for heteroscedasticity, outlier removal, non-normality and non-linearity of data (Tabachnick & Fidell, 2001). The ultimate objective of transformation (i.e. log transformation, square root transformation and inverse transformation) is to normalize your data. But Tabachnick and Fidell (2001) suggested avoiding transformations and using them in only extreme cases. This is also worth mention to inform readers that many researchers like Bollinger & Chandra (2005) and (Garson, 2012) prefer winsorizing of data instead of dropping outliers directly.

Conclusion

This illustrative study motivates researchers to understand the true value of GLRM assumptions and provides "fingertips approach" to data health diagnostics. Social scientists and researchers must test collected data for its health before testing, building or extending social science theories. It is better to realize soul-body relationship example in case of social sciences as body has no value without soul.

References

Ayentimi, D.T., Mensah, A.E. & Naa-Idar, F. (2013). Stock Market Efficiency of Ghana Stock Exchange: An Objective Analysis, Int. J. Manag. Econom. Soc. Sci. 2, 54-75.

Baltagi, B.H. (1998). Econometrics: 209-222, New York: Springer.

Blanchard, O. J. (1967). Comment. J. Bus. Econ. Stat., 5, 449–451.

Bollinger, C. R. & Chandra, A. (2005). Letrogentic specification error: A cautionary tale of cleaning data. J. Lab. Econ., 23(2): 235-257.

Bryman. (2001). Social Research Methods, Oxford University Press: Oxford.

Chalmers, A.F. (1999). What is This Thing We Called Science?, McGraw-Hill, Maidenhead.

Chatterjee, S. & Price, B. (1977). Regression Analysis by Example, John Wiley & Sons, New York.

Cohen, L., Manion, L. & Morrison, K. (2001). Research Methods in Education, Routledge, New York.

Conlisk, J. (1971). When Collinearity is Desirable, West. Econ. J. 9.

Cuthbertson, K., Stephen, G. (1995). Taylor, Applied Econometric Techniques, The University of Michigan Press, pp. 130.

Draper, N., & Smith, H. (1981). Applied Regression Analysis , John Wiley & Sons, New York, pp. 266–274.

Davidson, R. & MacKinnon, J.G. (1981). Several Tests for Model Specification in the Presence of Alternative Hypotheses, Econometrica, 49, 781-793.

Engle, R.F. (1982). A general approach to Lagrangian multiplier model diagnostics," J. Econometrics, 20, 83-104.

Fischer, H. (2011). A history of the Central Limit Theorem: From Classical to Modern Probability Theory, Springer Science+Business Media, New York.

Garson, D.G. (2012). Testing Statistical Assumptions, Garson & Statistical Associates Publishing, Asheboro.

Ghoshal, S. (2005). Bad Management Theories are Destroying Good Management Practices, Acad. Manag. Learn. Edu. 4, 75-91

Gujarati, D.N. (2004). Basic Econometrics, McGraw-Hill Book Co.

Harvey, A.C. (1990). The Econometric Analysis of Time Series, MIT Press, Cambridge, Mass.

Kangai, C., Bukalia, R.M. and Mapuranga, B. (2011). Content Analysis of Research Implications for Quality Assessment. Turk. Online. J. Dist. Edu.12, 1.

Kendall, M.G. & Stuart, A. (1958). The Advanced Theory of Statistics, Hafner, New York.

Moustakas, C. (1994). Phenomenological Research Methods. Thousand Oaks: Sage California.
Montgomery, D.C. & Peck, L.A. (1982). Introduction to Linear Regression Analyss, John Willy & Sons: New York.
Orr, J.M., Sackett, P.R. & Dubois, C.L.Z. (1991). Outlier Detection and Treatment in I/O Psychology, Pers. Psychol., 44, 474-486.
Ramsey, J.B. (1969). Tests for Specification Errors in Classical Linear Least Squares Regression Analysis, J. R. Stat. Soc., series B, 31, 350-371.
Searle, J.R. (1995). The Construction of Social Reality, Simon & Schuster, New York.
Tabachnick, L.S. & Fidell, L.S. (2001). Using multivariate statistics, Allyn and Bacon, New York.
Tharenou, P., Donohue, R. & Cooper, B. (2007). Management Research Methods, Cambridge University Press, New York .
Tintner, G. (1965). Econometrics, John Wiley & Sons, New York.
Vinod, H. D. (1978). A Survey of Ridge Regression and Related Techniques for Improvements over Ordinary Least Squares, Rev. Econ. Stat, 60, 121-131.
Wetherill, G.B. (1986). Regression Analysis with Applications, Chapman and Hall, New York, pp. 14–15.
Wilkinson, L. & The Task Force on Statistical Inference (TFSI) (1999). Statistical methods in psychology journals: Guidelines and explanation. Amer. Psychol. 54, 594-604.