The Importance of Assumptions in Multiple Regression and How to Test Them

Ronelle M. Krieger

University of Calgary

Introduction

Multiple regression analysis is a statistical tool used to predict a dependent variable from multiple independent variables (Harlow, 2005; Stevens, 2009). The independent variables (also called predictor variables) are usually not under experimental control and the variations observed in them are to be accepted for what they are. The focus of multiple regression is to investigate which, if any, of these predictor variables can significantly predict the dependent variable. Multiple regression holds increase utility within the social sciences as it allows for more comprehensive analysis of constructs related to human behaviour (Stevens, 2009). However, it is critical to recognize that multiple regression is inherently a correlation technique and cannot explain the causalities that may underlie the relationship between the variables being observed (Steven, 2009).

Most statistical tests rely on certain assumptions about the variables used within an analysis to ensure that the analysis is as accurate and true as possible, and therefore valid (Osborne & Waters, 2002; Stevens, 2009). Assumptions are critical in statistics because if the underlying assumptions are not valid, then the process is unreliable, unpredictable, and out of the researcher's control (Stevens, 2009). This could lead the researcher to draw conclusions that are not valid or scientifically unsupported by the data. Researchers are encouraged to examine the data of an analysis to ensure the values are plausible and reasonable. The assumptions of multiple regression include the assumptions of linearity, normality, independence, and homoscedasticty, which will be discussed separately in the proceeding sections.

Assumption of Linearity

Relationships between variables are considered linear when they are consistent and directly proportional to each other (Stevens, 2009; Tabachnick & Fidell, 2006). It is imperative to examine analysis for nonlinearity as there are many instances in the social sciences where nonlinear relationships occur (Kivilu, 2003; Steven, 2009). Violations of this assumption may result in the estimates obtained from the analysis, such as R^2, regression coefficients, standard errors, and statistical significance, being biased; therefore, not portraying the accurate or true population values (Osborne & Waters, 2002; Tabachnick & Fidell, 2006). According to Hoxx (1995), the results from the analysis will underestimate the true relationship between the independent variables (predictor variables) and dependent variable if the relationship is not linear. This underestimation of the results could lead to two areas of concern; first, an increase risk of Type 11 error could occur for that predictor variable, and second, an increase risk of Type 1 error (which is an overestimation) for the other predictor variable(s) that share variance with that predictor variable could occur (Hoxx, 1995; Osborne & Water, 2002).

Testing the Linearity Assumptions

The linearity assumption can be tested through the visual examination of residual plots (Kivilu, 2003; Osborne & Waters, 2002; Stevens, 2009). A residual scatterplot is a figure that depicts one axis for the standardized residuals (r_i) and the other axis for the predicted values (y^i) (Stevens, 2009). If the linearity assumption is met, the standardized residuals will scatter randomly around a horizontal line which represents the standardized residuals equaling zero ($r_i=0$) (Stevens, 2009; Tabachnick & Fidell, 2006). Figure 1.1 depicts an example of a residual plot portraying a clustering of residuals along the horizontal line in a rectangular shape, therefore, a linear relationship is present (Osborne & Waters, 2002). When linearity is violated,

the residual plot portrays a c-curved or u-curved shape of distribution around the horizontal line. Figure 1.2 depicts an example of residual plot demonstrates a curvilinear relationship between the standardized residuals and standardized predicted values, demonstrating that the residuals are no longer random and are either mostly above or below the zero line at differing predicted values, causing the shape to be curved instead of rectangular (Osborne & Waters, 2002).

Figure 1.1 Linear Relationship

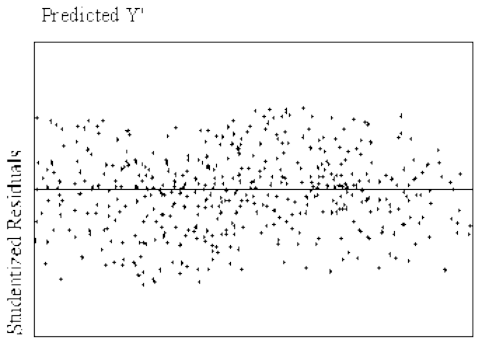
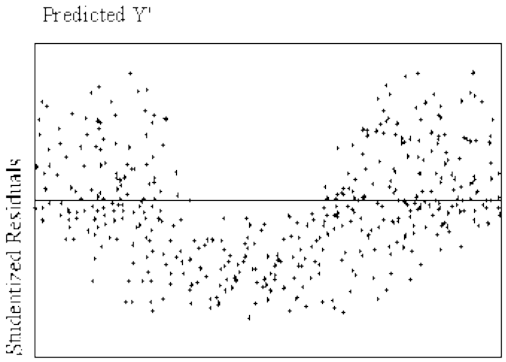


Figure 1.2 Curvilinear Relationship



(Osborne & Waters, 2002)

Residuals plots can be viewed during an initial screening run of the analysis or after the analysis has been conducted (Hoxx, 1995). However, by investigating a residual plot in the early stages of an analysis, the sooner detection of possible nonlinear relationships can occur; therefore, allowing the researcher to manage data information and time allotment for analysis more effectively and accurately.

Residuals plots can be created and obtained through the completion of multiple regression analysis in SPSS by selecting Analyze from the drop down menu, followed by Regression, and then select Linear. Upon completing this section, the Linear Regression window should appear. Ensure to place the appropriate variables into the correct variable box; the dependent variable into the Dependent variable rectangular box, and the predictor variables into

the Independent(s) variable(s) rectangular box. The next step is to select the 'Plots' button found on the upper right hand side of the window. Select the ZPredicted and place it into the Y: rectangular box, and select the ZResidual for the X: rectangular box to plot the predicted score against the residual score to determine if they are linearly related. Upon completing this task, click on the Continue button located on the bottom left hand side of the window, which should return you back to the Linear Regression Window. To run the residual plot, select the OK button located on the bottom left hand side of the Linear Regression Window.

Assumption of Independence

Multiple regression assumes that the errors, which are the residuals between the actual score and the estimated score obtained through the regression equation, are independent and there is no serial correlation (Stevens, 2009). Having no serial correlation between the residuals implies that the size of the residual for one variable has no impact on the size of the residual for another variable. Therefore, the independence assumption requires that the variables and residuals are independent and the subjects are responding independently of each other (Stevens, 2009). The independence assumption is a significant assumption that should be investigated prior to any interpretation of multiple regression analysis, as violation of this assumption could hold critical implications (Stevens, 2009). Even a slight violation of the independence assumption should be taken seriously, as it can greatly increase the risk of Type 1 error, resulting in the risk of falsely rejecting the null hypothesis several times greater than the level of error assumed for the test (Stevens, 2009).

Testing the Independence Assumption

The Durbin-Watson is a statistic test which can be used to test for the occurrence of serial correlation between residuals. The value of Durbin-Watson statistics ranges between 0 and 4, however, the residuals are considered not correlated if the Durbin-Watson statistic is between 1.5 and 2.5. Figure 2 depicts a Model Summary table which includes a Durbin-Watson statistic. The Durbin-Watson statistic is 1.951, indicating that the residuals are uncorrelated; therefore, the independence assumption is met for this analysis.

Figure 2 Durbin-Watson statistic

Model	R^2	Adjusted R^2	Std. error	Durbin-Watson
	0.532	0.532	1.615	1.951

Notes: [a]Predictors: (Constant), *HM2*; [b]Dependent variable: *ACC*

(Durbin-Watson statistic obtained through Google Image clipart)

When completing multiple regression analysis using SPSS, select Analyze from the drop down menu, followed by Regression, and then select Linear. The Linear Regression window should appear allowing the insertion of the dependent and predictor variables being investigated in the analysis. The next step is to select the 'Statistics' button found on the upper right hand side of the window. Select Durbin-Watson, which is located in the Residuals section in the bottom right hand side of the window. Upon completing this task, click on the Continue button, which should return you back to the Linear Regression Window. To run the Durbin-Watson statistic, select the OK button located on the bottom left hand side of the window. The Durbin-Watson coefficient can be found in the Model Summary table in the multiple regression output.

Assumption of Normality

Screening for normality is an important early step when conducting multiple regression, as residuals are normally distributed is assumed (Stevens, 2009; Tabachnick & Fidell, 2006). Non-normal distributions that are positively or negatively skewed, contain large kurtosis, or have extreme outliers can distort the obtained significance levels of the analysis, resulting in the standard errors becoming biased (Osborne & Waters, 2002). Though multiple regression is generally considered to be quite robust to violations of normality, a small sample size can actually increase the seriousness of non-normality of a distribution (Osborne & Waters, 2002). Outliers may have stronger influence on normal distribution when the sample size is small, whereas standard errors for both skewness and kurtosis decrease with larger samples, as there will most likely be only minor deviations from normality (Tabachnick & Fidell, 2006).

Testing the Normality Assumptions

Graphical methods, such as histograms and normality plots, can be conducted to provide a visual inspection of the normal distribution of a data set prior to further interpretation of the regression analysis (Tabachnick & Fidell, 2006). Histograms can provide important information about the shape of a distribution. If most of the scores are gathered around the middle of the continuum and a gradual, symmetric decrease of frequency on either side of the centre score occurs, it is considered a normal distribution. However, it the scores are not symmetric and are spread out away from the majority it is considered skewed. If the 'tail' (a small number of the distribution) is spread out to the right, it is considered positively skewed, and if the 'tail' is spread out to the left, it is considered negatively skewed. Kurtosis is the shape of any or lack of peaks within a distribution (Tabachnick & Fidell, 2006). Though no distribution can be considered 'perfect', a distribution is regarded as normal when the values of both skewness and

kurtosis are zero; however a suggested acceptable range for both is between -2 and +2. Figures 3.1through 3.4 depict examples of histograms which are normal, contain outliers, skewed, and kurtotic.

Figure 3.1 Normal Distribution

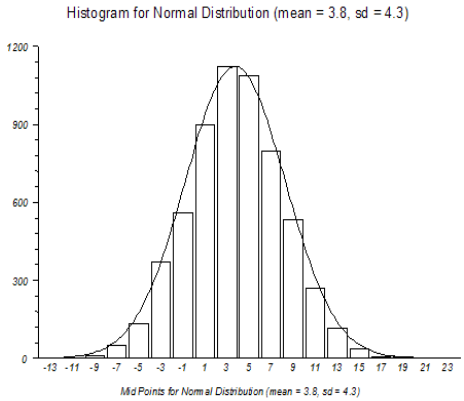


Figure 3.2 Distribution Containing Outlier

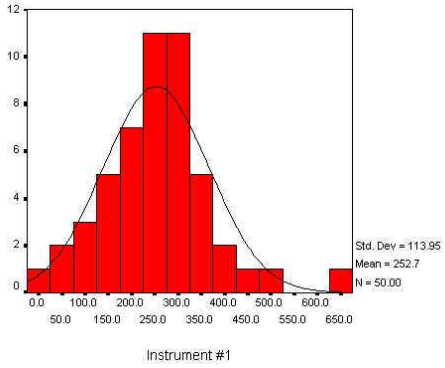


Figure 3.3 Positively Skewed Distribution

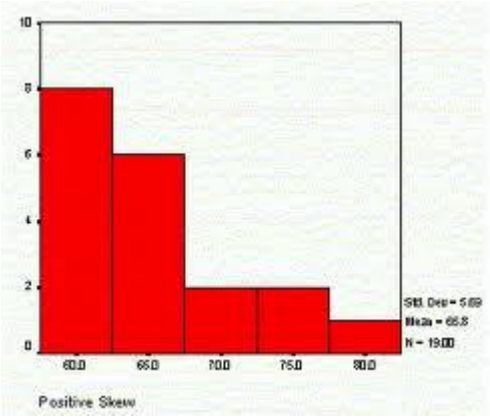
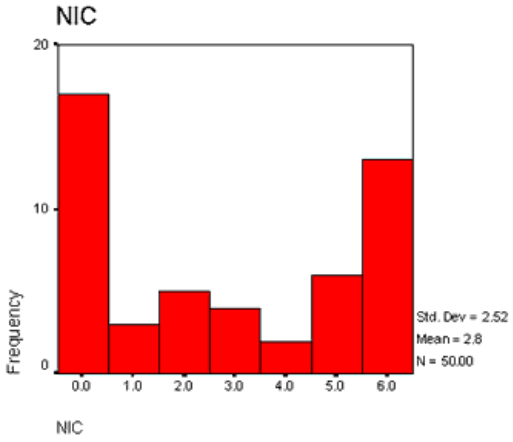


Figure 3.4 Kurtotic Distribution (Bimodal)



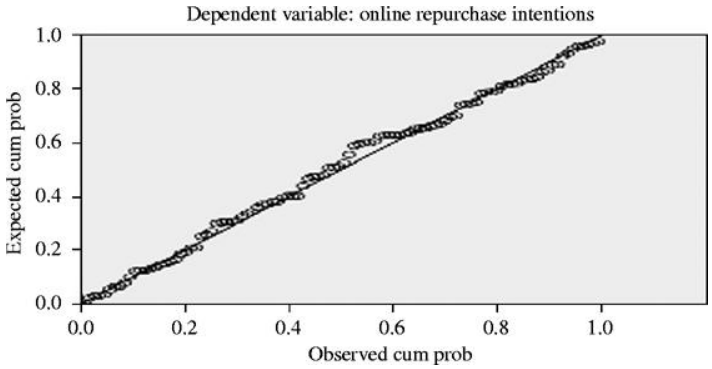
(All histograms were obtained through Google Image clipart)

When using SPSS, histograms can be obtained through multiple regression analysis by selecting Analyze from the drop down menu, followed by Regression, and then select Linear, upon which the Linear Regression window should then appear. Ensure to place the appropriate

variables into the correct variable box; the dependent variable into the Dependent variable rectangular box, and the predictor variables into the Independent(s) variable(s) rectangular box, though be aware that a histogram for the dependent variable will be the only one conducted. The next step is to select the 'Plots' button found on the upper right hand side of the window. Within the Linear Regression: Plots window, select Histograms, which is located in the Standardized Residual Plots section in the bottom right hand side of the window. Upon completing this task, click on the Continue button located on the bottom left hand side of the window, which should return you back to the Linear Regression Window. To run the histogram for the dependent variable, select the OK button located on the bottom left hand side of the window.

The normality assumption can also be tested through the visual examination of normal probability plots (P-P plots) of the standardized residuals. In a P-P plot, the normal distribution is depicted by a random scatter of plots around a 45 degree line. Figure 3.5 portrays an example of a Normal P-P plot of the standardized residuals (Tabachnick & Fidell, 2006).

Figure 3.5 Normal P-P plot



(P-P plot obtained through Google Image clipart)

When using SPSS, P-P plots can be obtained through multiple regression analysis by selecting Analyze from the drop down menu, followed by Regression, and then select Linear, upon which the Linear Regression window should then appear. Ensure to place the appropriate variables into the correct variable box; the dependent variable into the Dependent variable rectangular box, and the predictor variables into the Independent(s) variable(s) rectangular box, though be aware that a P-P plot for the dependent variable will be the only one conducted. The next step is to select the 'Plots' button found on the upper right hand side of the window. Within the Linear Regression: Plots window, select Normal Probability Plot, which is located in the Standardized Residual Plots section in the bottom right hand side of the window. Upon completing this task, click on the Continue button located on the bottom left hand side of the window, which should return you back to the Linear Regression Window. To run the P-Plot for the dependent variable, select the OK button located on the bottom left hand side of the window.

Assumption of Homoscedasticity

The assumption of homoscedacity indicates that the variance of errors is equal and constant across all levels of the variables (Osborne & Waters, 2002; Stevens, 2009). Homoscedasticity is related to the assumption of normality because when the assumption of normality is met, the relationship between the variables is homoscedastic (Tabachnick & Fidell, 2006). Heteroscedasticity occurs when the variance of errors differs at different values of the independent variables (Osborne & Waters, 2002). Slight heteroscedasticity has little effect on significance tests; however when heteroscedasticity is marked it can lead to serious distortions of findings and seriously weaken the analysis thus increasing the possibility of a Type 1 error for small sample size (Osborne & Waters, 2002).

Testing the Homoscedasticity Assumption

The homoscedasticity assumption can be tested through the visual examination of the same residual plots of the standardized residuals and predicted values depicted in the assumption of linearity section of this paper (Osborne & Waters, 2002). When the homoscedasticity assumption has been met, the residuals will present as being randomly scattered around the horizontal line depicting $r_i=0$. Figure 4.1 portrays an example of a residual plot demonstrating a relative equal clustering of residuals along the horizontal line in a rectangular shape, therefore, the homoscedasticity assumption seems to have been met. There are many forms heteroscedacity can take, two of which are bow-tie and fan shape (Osborne & Waters, 2002). Figure 4.2 depicts an example of a residual plot which demonstrates a bow-tie shape of the distribution, indicating that the variables are heteroscedastic.

Figure 4.1 Homoscedasticity

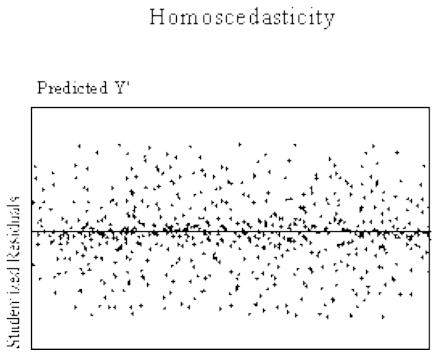
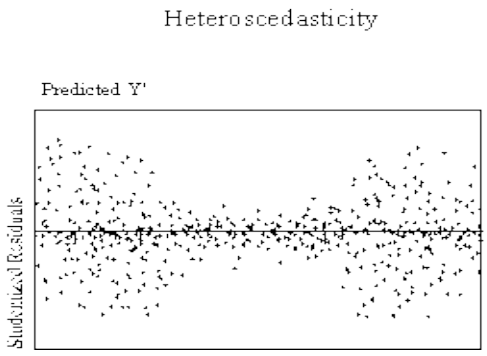


Figure 4.2 Heteroscedacity



(Osborne & Waters, 2002)

A description on how to create and obtain residual plots in SPSS was provided in the Testing of Linearity Assumption section of this paper.

Conclusion

When completing multiple regression analysis, it is essential to rest the regression model ensure that the assumptions of multiple regression have been satisfied (Stevens, 2009). When the assumptions are violated, the significance may be over or under estimated, increasing the risk of committing a Type I or Type II error (Osborne & Waters, 2002). The assumptions of multiple regression can be tested through a visual examination of histograms of the standardized residuals, residual plots of standardized residuals and predicted values, and by the Durbin Watson statistic, all of which may be obtained through multiple regression analysis using SPSS, as well as other methods which were not discussed in this paper. Though Kerlinger and Lee (2000) discussed that there is some controversy amongst statisticians and researchers regarding the actual implications of violations of assumptions, however, most statisticians and researchers have indicated that such analysis findings would be questionable. Therefore, not only is it essential to understand the reasoning of testing and meeting the assumptions of multiple regression analysis, but to also ensure to include that the assumptions had been investigated when reporting the research findings so others will have confidence in the validity of the findings being reported on.

Web-Based Resources

The following links provide comprehensive information in the area of multiple regression, the assumptions of multiple regression, and multiple regression analysis using SPSS.

Statnotes from University of North Texas

Multiple Regression : http://faculty.chass.ncsu.edu/garson/PA765/regress.htm

Testing Assumptions : http://faculty.chass.ncsu.edu/garson/PA765/assumpt.htm

Multiple Regression Using SPSS

The Multiple Linear Regression Analysis in SPSS: http://www.statisticssolutions.com/resources/directory-of-statistical-analyses/the-multiple-linear-regression-analysis-in-spss

Multiple Regression Using SPSS: http://www.statisticshell.com/multireg.pdf

Multiple Linear Regression in SPSS: http://www.unt.edu/rss/class/Jon/SPSS_SC/Module9/M9_Regression/SPSS_M9_Regression2.htm

Multiple Regression Using SPSS/PASW: http://www.youtube.com/watch?v=4EFXic4sGdE

References

Harlow, L.L. (2005). What is multivariate thinking. *The essence of multivariate thinking* (pp.3-27). Mahwah, N.J.: Lawrence Erlbaum Assoc., Inc. Retrieved from https://blackboard.ucalgary.ca/webapps/portal/frameset.jsp?tab_id=_2_1&url=%2fwebapps%2fblackboard%2fexecute%2flauncher%3ftype%3dCourse%26id%3d_86998_1%26url%3d

Hoxx, J.J. (1995). *Applied multilevel analysis*. Amsterdam: TT-Publikaties. Retrieved June 1, 2011, from http://igitur-archive.library.uu.nl/fss/2007-1114-201211/hox_95_applied%20multilevel%20analysis.pdf

Kerlinger, F.N. and Lee, H.B, (2000). *Foundations of behavioral research* (4th Ed) Harcourt

Kivilu, M. (2003). Understanding the structure of data when planning for analysis: Application of hierarchical linear models. *South African Journal of Education,* 23(4), 249-253. Retrieved June 1, 2011, from http://www.ajol.info/index.php/saje/article/viewFile/24942/20628

Osborne and Waters (2002). Four assumptions of multiple regression that researcher should always test. *Practical Assessment, Research & Evaluation, 8*(2). Retrieved June 1, 2011, from http://www.worldcat.org/title/practical-assessment-research-evaluation-pare/oclc/222890640 ?title=&detail=&page=frame&url=http%3A%2F%2Fbibpurl.oclc.org%2Fweb%2F4752%26checksum%3D4d3bcab9508caf22d2a62261417aa86a&linktype=digitalObject

SPSS Inc. (2010). SPSS Base 19 for Windows User's Guide. SPSS Inc., Chicago.

Stevens, J.P. (2009). *Applied multivariate statistics for the social sciences* (5th ed.). New York: Routledge.

Tabachnick, B. & Fidell, L. (2006). Cleaning up your act screening data prior to analysis. Using multivariate analysis (5th ed.). Needham Heights, MA: Allyn &

Bacon. Retrieved from https://blackboard.ucalgary.ca/webapps/portal/frameset.jsp?tab_id=_2_1&url=%2fwebapps%2fblackboard%2fexecute%2flauncher%3ftype%3dCourse%26id%3d_86998_1%26url%3d