# Assumptions of Multiple Linear Regression

linear relationship between the outcome variable and the independent variables. This does not restrict you to use a nonlinear transformation of the independent variables. Scatter plots can show whether there is a linear or curvilinear relationship. This assumption may be checked by looking at a histogram or a Q-Q-Plot.  Normality can also be checked with a goodness of fit test (e.g., the Kolmogorov-Smirnov test), though this test must be conducted on the residuals themselves.

Multivariate Normality–Multiple regression assumes that the residuals are normally distributed.

No Multicollinearity—Multiple regression assumes that the independent variables are not highly correlated with each other. Multicollinearity may be checked multiple ways:

1) Correlation matrix – When computing a matrix of Pearson’s bivariate correlations among all independent variables, the magnitude of the correlation coefficients should be less than .80.

2) Variance Inflation Factor (VIF) – The VIFs of the linear regression indicate the degree that the variances in the regression estimates are increased due to multicollinearity. VIF values higher than 10 indicate that multicollinearity is a problem.

If multicollinearity is found in the data, one possible solution is to center the data.  To center the data, subtract the mean score from each observation for each independent variable. However, the simplest solution is to identify the variables causing multicollinearity issues (i.e., through correlations or VIF values) and removing those variables from the regression.

Homoscedasticity–This assumption states that the variance of error terms are similar across the values of the independent variables.  A plot of standardized residuals versus predicted values can show whether points are equally distributed across all values of the independent variables. A scatterplot of residuals versus predicted values is good way to check for homoscedasticity.  There should be no clear pattern in the distribution; if there is a cone-shaped pattern (as shown below), the data is heteroscedastic. If the data are heteroscedastic, a non-linear data transformation or addition of a quadratic term might fix the problem.