Chapter 8

8.1

We are going to discuss two related concepts to regression analysis, namely the correlation analysis. With the partial- and semi partial-correlation. The multiple regression models involves the use of two or more predictors and one criterion variable; thus there are at minimum three variables involved in the analysis. So for the Pearson correlation we only have two variables at a time, so we need a solution for this. This is the partial- and semi partial- correlation.

First the partial correlation. The simplest situation is that we have three variables, which we label, X1, X2, and X3. So an example of a partial correlation is the correlation between X1 and X2, where X3 is held constant. Thus the partial correlation here represents the linear relationship between X1 and X2 independent of the linear influence of X3. This particular partial correlation is denoted by r12.3. We compute this as follows:

$$r_{12.3} = r_{12} - r_{13}r_{23}/\sqrt{(1-r_{13}^2)(1-r_{23}^2)} \qquad (80)$$

There can be some extreme results from this. One example is perfect collinearity, which is a serious problem. In this case, either r13, or r23 equal 1, then r12.3 cannot be calculated, as the denominator is equal to 0. In this situation the partial correlation is not defined.

We will now look at the concept of semipartial correlation (also called a part correlation). Again the simplest situation is that we have three variables, labelled X1, X2, and X3. Here an example of semipartial correlation would be the correlation between X1 and X2 where X3 is removed from X2 only. Thus the semi partial correlation here represents the linear relationship between X1 and X2 after that portion of X2 that can be linearly predicted from X3 has been removed from X2. This particular correlation is denoted by r1(2.3). We compute it:

$$r_{1(2.3)} = r_{12} - r_{13}r_{23}/\sqrt{(1-r_{23}^2)} \qquad (81)$$

8.2

In this section we will discuss the unstandardized and standardized multiple regression models, the coefficient of multiple determination, multiple correlation, tests of significance, and statistical assumptions.

The sample multiple linear regression model for predicting Y from m predictors X1,2,….m is (unstandardized multiple regression model)

$$Y_i = b_1X_{1i} + b_2X_{2i} + \ldots + b_mX_{mi} + \mathrm{a} + e_i \qquad (82)$$

Where

- Y is the criterion variable (dependent variable).
- Xk are the predictor (independent) variables where k = 1,…,m.
- bk is the sample partial slope of the regression line for Y as predicted by Xk.
- a is the sample intercept of the regression line for Y as predicted by the set of Xk.
- ei represents the residuals or errors of prediction.
- i represents an index for an individual or object. It can take values form 1,...,n.

The term partial slope is used because it represents the slope of Y for a particular Xk in which we have partialled out the influence of other Xk.

The sample prediction model is:

$$Y_i' = b_1X_{1i} + b_2X_{2i} + \dots + b_mX_{mi} + \text{a} \qquad (83)$$

Where Yi' is the predicted value of Y. Difference between prediction and the regression model is the same as in chapter 7. We compute the residuals as follows:

$$e_i = Y_i - Y'_i \qquad (84)$$

It is hard to determine the sample partial slopes and the intercept. To keep it simple, we use a two-predictor model for illustrative purposes. Mostly we rely on statistical software. For the two-predictor case, the sample partial slopes (b1 and b2) and the intercept (a) are:

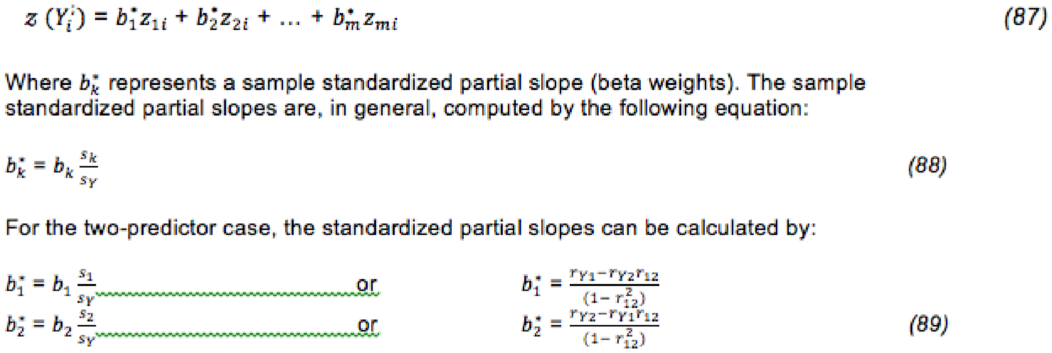
$$\begin{aligned} b_1 &= (r_{Y1} - r_{Y2}r_{12})s_Y/(1 - r_{12}^2)s_1 \\ b_2 &= (r_{Y2} - r_{Y1}r_{12})s_Y/(1 - r_{12}^2)s_2 \\ \text{a} &= \bar{Y} - b_1\bar{X}_1 - b_2\bar{X}_2 \end{aligned} \qquad (85)$$

An alternative method for computing the sample partial slopes that involves the use of a partial correlation is as follows:

$$\begin{aligned} b_1 &= r_{Y1.2}\, {}^{s_Y\sqrt{1 - r_{Y2}^2}}\Big/_{s_1\sqrt{1-r_{12}^2}} \\ b_2 &= r_{Y2.1}\, {}^{s_Y\sqrt{1 - r_{Y1}^2}}\Big/_{s_2\sqrt{1-r_{12}^2}} \end{aligned} \qquad (86)$$

In the multiple linear regression model, and almost in all general linear models (GLM), we use the least squares criterion. So we need to find a regression model, defined by a particular set of partial slopes and an intercept, which has the smallest sum of squares residuals.

We will now look at the standardized regression model. In this model the terms are expressed in standard z score units. The means and variances of the standardized variables are 0 and 1 respectively. The sample standardized linear prediction model becomes:



$$z\,(Y_i') = b_1^* z_{1i} + b_2^* z_{2i} + \ldots + b_m^* z_{mi} \qquad (87)$$

Where b_k^* represents a sample standardized partial slope (beta weights). The sample standardized partial slopes are, in general, computed by the following equation:

$$b_k^* = b_k \frac{s_k}{s_Y} \qquad (88)$$

For the two-predictor case, the standardized partial slopes can be calculated by:

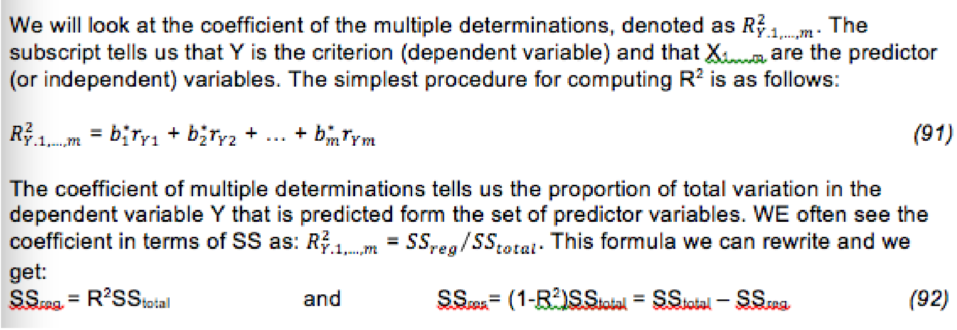
$$b_1^* = b_1 \frac{s_1}{s_Y} \quad \text{or} \quad b_1^* = \frac{r_{Y1} - r_{Y2} r_{12}}{(1 - r_{12}^2)}$$
$$b_2^* = b_2 \frac{s_2}{s_Y} \quad \text{or} \quad b_2^* = \frac{r_{Y2} - r_{Y1} r_{12}}{(1 - r_{12}^2)} \qquad (89)$$

So now what is the utility of the set of predictor variables? The easiest way to look at that involved the partitioning of the familiar total sum of squares in Y, which we denote as SStotal. In the multiple regression analysis, we can write it as follows:

$$SS_{total} = [n\sum Y_i^2 - (\sum Y_i)^2]/n \quad \text{or} \quad SS_{total} = (n\text{-}1)s_Y^2$$
$$SS_{total} = SS_{reg} + SS_{res}$$
$$\sum(Y_i - \bar{Y})^2 = \sum(Y_i' - \bar{Y})^2 + \sum(Y_i - Y_i')^2 \qquad (90)$$

Where

- SSreg is the regression sum of squares due to the prediction of Y from the Xk (also written as SSY').
- SSres is the sum of squares due to the residuals.



We will look at the coefficient of the multiple determinations, denoted as $R^2_{Y.1,\ldots,m}$. The subscript tells us that Y is the criterion (dependent variable) and that $X_{1,\ldots,m}$ are the predictor (or independent) variables. The simplest procedure for computing R^2 is as follows:

$$R^2_{Y.1,\ldots,m} = b_1^* r_{Y1} + b_2^* r_{Y2} + \ldots + b_m^* r_{Ym} \qquad (91)$$

The coefficient of multiple determinations tells us the proportion of total variation in the dependent variable Y that is predicted form the set of predictor variables. WE often see the coefficient in terms of SS as: $R^2_{Y.1,\ldots,m} = SS_{reg}/SS_{total}$. This formula we can rewrite and we get:

$$SS_{reg} = R^2 SS_{total} \quad \text{and} \quad SS_{res} = (1\text{-}R^2)SS_{total} = SS_{total} - SS_{reg} \qquad (92)$$

The coefficient is determined not just by the quality of the predictor variables included in the model, but also by the quality of the relevant predictor variables not included in the model, as well as by the amount of total variation in the dependent variable Y. This coefficient of determination can be used to determine effect size (Small effect: R2 = 0.10; medium effect: R2 = 0.30; large effect R2 = 0.50).

We should note that R2 is sensitive to sample size and to the number of predictor variables. As the samples size and/or the number of predictor variables increase, R2 will increase as well. R is a biased estimate of the population multiple correlation as to sampling error in the bivariate correlations and in the standard deviations of X and Y. In general R overestimates the population multiple correlation. Adjusted R2 is calculated as follows:

$$R^2_{adj} = 1 - (1- R^2)\left(\frac{n-1}{n-m-1}\right) \qquad (93)$$

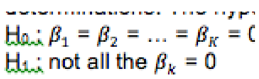
This adjusted value adjusts for sample size and the number of predictors. So now we can use it to compare models fitted to the same set of data with different numbers of predictors or with different samples of data. The difference between R2 and adjusted R is called shrinkage.

To make sure the power is enough you can use the power software. However you need to make sure that your ratio of n to m is large. Because this will minimize the bias and generalizations are likely to be better about the population values.

We will now look at the significance tests, we look at two methods used in the multiple regression model. The first is to test the significance of the overall regression model and second is test the significance of each individual partial slope:

Test of significance of overall regression model

You can frame this alternatively as the test of significance of the coefficient of multiple determinations. The hypotheses are as follows:



H0 : $\beta_1 = \beta_2 = \ldots = \beta_K =$ (

H1 : not all the $\beta_k = 0$

When H0 is rejected, then one or more of the individual regression coefficients is statistically significantly different from 0. The test is based on the following statistic:

$$F = \frac{R^2/m}{(1-R^2)/(n-m-1)} \qquad (94)$$

Where F indicates it is an F statistic. m is the number of predictors or independent variables. n is the sample size.

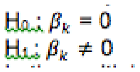
The F test statistic is compared to the F critical value, always a one-tailed test and at the alpha level, with the degrees of freedom being m and (n-m-1). Taken from the F table A.4. This test statistic can also be written as:

$$F = \frac{SS_{reg}/df_{reg}}{SS_{res}/df_{res}} = \frac{MS_{reg}}{MS_{res}} \qquad (95)$$

Where df(reg) = m, and df(res) = (n-m-1).

Test of significance of bk

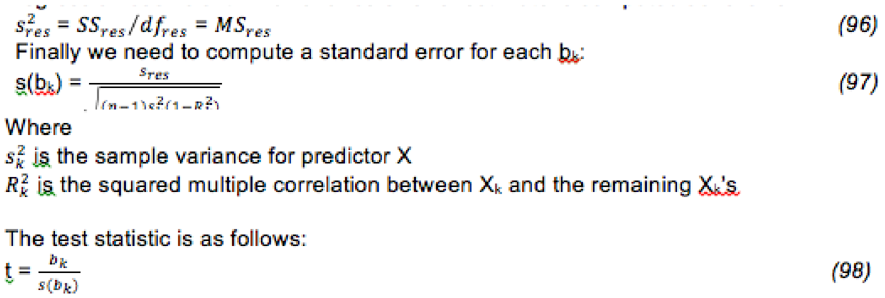
This tests whether all the individual unstandardized regression coefficients are statistically significant from 0. The hypotheses are:



H_0: $\beta_k = 0$

H_1: $\beta_k \neq 0$

In the multiple regressions, it is necessary to compute a standard error for each regression coefficient. The variance error of estimate is computed as follows:



$$s_{res}^2 = SS_{res}/df_{res} = MS_{res} \qquad (96)$$

Finally we need to compute a standard error for each b_k:

$$s(b_k) = \frac{s_{res}}{\sqrt{(n-1)s_k^2(1-R_k^2)}} \qquad (97)$$

Where

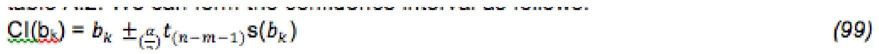
s_k^2 is the sample variance for predictor X

R_k^2 is the squared multiple correlation between X_k and the remaining X_k's

The test statistic is as follows:

$$t = \frac{b_k}{s(b_k)} \qquad (98)$$

This test statistic is compared to the critical value of t, a two-tailed test for a non-directional H1, at the level of alpha, and degrees of freedom of (n-m-1), taken from table A.2. We can form the confidence interval as follows:



$$CI(b_k) = b_k \pm_{(\frac{\alpha}{2})} t_{(n-m-1)} s(b_k) \qquad (99)$$

We will now discuss the assumptions of the multiple regression model. The assumptions are concerned with (a) independence, (b) homogeneity, (c) normality, (d) linearity, (e) fixed X, and (f) non-collinearity.

Independence

The simplest procedure to examine this assumption is a residual plot of e versus the predicted values of the dependent variable Y' and of e versus each independent variable Xk. If the assumption is satisfied, the residuals fall into a random display of points. Lack of independence affects the estimated standard errors of the model. For serious violations, one could consider generalized or weighted least squares as a method of estimation.

Homogeneity

In this assumption the conditional distributions have the same constant variance of all values of X. In the residual plot, the consistency of the variance of the conditional distributions may be examined. If the assumption is violated, estimates of the standard errors are larger, and the conditional distributions may also be non-normal.

Normality

The conditional distributions of the scores on Y, or the prediction errors are normal in shape. Violation of the normality assumption may be the result of outliers. You can use the frequency distributions, Q-Q plots, and skewness statistics to examine the normality assumption.

Linearity

There is a linear relationship between the observed scores of the dependent variable Y and the values of the independent variables, Xk. If satisfied, then the sample partial slopes and intercept are unbiased estimators of the population partial slopes and intercept. If a nonlinear relationship exists, it means that the expected increase in Y depends on the value of X. So the expected increase is not a constant value. Violation of the linearity assumption can be detected through residual plots. The residuals should be located within a band of (standard errors).

Fixed X

The independent variables, Xk are fixed variables rather than random variables. This results in the regression model being valid only for those particular values of Xk that where actually observed and used in the analysis. Generally we may not want to make predictions about individuals having combinations of Xk scores outside of the range of values used in developing the prediction model, this si defined as extrapolating. Also we may not be quite as concerned in making predictions about individuals having combinations of Xk scores within a range of values used in developing the prediction model, this is defined as interpolating. There is shown, that when all the other assumptions are met, regression analysis performs just as well when X is a random variable.

Noncollinearity

This assumption is unique for the multiple linear regression analysis. A violation of this assumption is known as collinearity where there is a very strong linear relationship between two or more of the predictors. This can be problematic in several aspects. First, it can lead to instability of the regression coefficients across samples, where the estimates will bounce around quite a bit in terms of magnitude and even occasionally result in changes in sign. This occurs because the standard errors are larger which makes it harder to achieve statistical significance. Second, it could occur that the overall regression is significant, but none of the individual predictors are significant. Collinearity will also restrict the utility and generalizability of the estimate regression model.

Collinearity may be indicated when there are large changes in estimated coefficients due to (a) a variable being added or deleted and/or (b) an observation being added or deleted.

We can detect violations of this assumption by conducting a series of special regression analyses, one for each X, where that predictor is predicted by all of the remaining X's. If any of the resultant values are close to 1 (greater than .9, rule of thumb), then there may be a collinearity problem. However large R2 value can also be due to small sample sizes.

Also if the number of predictors is greater than or equal to n, then perfect collinearity is a possibility (look at 8.1). Another method for detecting collinearity is to compute a variance inflation factor (VIF) for each predictor, which is equal to 1/(1-). The VIF is defined as the inflation that occurs for each regression coefficient above the ideal situation of uncorrelated predictors. Many suggest that the largest VIF should be less than 10 in order to satisfy this assumption.

There are several methods to deal with collinearity. Frist, one can remove one or more of the correlated predictors. Second, ridge regression techniques can be used. Third, principal component scores resulting from principal component analysis can be utilized rather than raw scores on each variable. Fourth, transformations of the variables can be used to remove or reduce the extent of the problem. The last solution is to use simple linear regression, as collinearity cannot exist with a single predictor.

Assumptions and violations of assumptions: Multiple Linear regression analysis

Assumption	Effect of assumption violation
Independence	Influences standard errors of the model
Homogeneity	Bias in variances of errors May inflate standard errors and thus increase likelihood of a Type II error May result in non-normal conditional distributions
Normality	Less precise slopes, intercept, and R2
Linearity	Bias in slope and intercept Expected change in Y is not a constant and depends on value of X
Fixed X values	Extrapolating beyond the range of X combinations: prediction errors larger, may also bias slopes and intercept Interpolating within the range of X combinations: smaller effect than earlier; if other assumptions met, negligible effect
Non-collinearity of X's	Regression coefficients can be quite unstable across samples (as standard errors are larger) R2 may be significant, yet none of the predictors are significant Restricted generalizability of the model.

8.3

The multiple predictor model which we have considered until now, can be viewed as simultaneous regression. That means, all of the predictors to be used are entered simultaneously, such that all of the regression parameters are estimated simultaneously. There are other methods of entering the independent variables where the predictor variables are entered systematically. This class of models is referred to as sequential regression (variable selection procedures). We will discuss different such procedures.

Backward elimination

Here the variables are eliminated from the model based on their minimal contribution to the prediction of the criterion variable. In the first stage of the analysis, all the potential predictors are included. In the second stage, the predictor is deleted from the model that makes the smallest contribution to the prediction of the dependent variable. Removing the variable with the smallest t or F statistic can do this. In subsequent stages, that predictor is deleted that makes the next smallest contribution. This analysis continues until each of the remaining predictors in the model is a significant predictor of Y. This can be determined by comparing the t or F statistics for each predictor to the critical value.

Forward selection

In this procedure, variables are added or selected into the model based on their maximal contribution to the prediction of the criterion variable. Initially none of the potential predictors are included in the model.

In the first stage, the predictor is added to the model that makes the largest contribution to the prediction (largest t or F statistic). In the stages following that, the predictor are selected that make the next largest contribution to the prediction of Y. This again continues until each of the selected predictors in the model is a significant predictor of the outcome Y (comparing t and F-statistics with critical values).

Stepwise selection

This procedure is a modification of the forward selection procedure with one important difference. Predictors that have been selected into the model can, at a later step, be deleted from the model. This situation can occur for a predictor when a significant contribution at an earlier step later becomes a non-significant contribution given the set of other predictors in the model. Initially in this model none of the potential predictors are included in the model. In the first step, the predictor is added to the model that makes the largest contribution to the explanation of the dependent variable (largest t or F statistic). In following stages, the predictor is selected that makes the next largest contribution. Also those predictors that have been entered at earlier stages are checked to see if their contribution remains significant. If not, this predictor is eliminated. This continues until each of the predictor remains significant predictor (compare the F or t statistic to critical value).

All possible subsets regression

Let us say that there are five potential predictors. In this procedure, all possible one-, two-, three-, and four-variable models are analysed. Thus, there will be 5 one-predictor models, 10 two-predictor models, 10 three-predictor models, and 5-four predictor models. The best k predictor model can be selected; this is the one that yields the largest R2.

The researcher is not advised to use this procedure, or for that matter, any of the other sequential regression procedures, when the number of potential predictor is large. The number of models will be equal to 2m.

Hierarchical regression

In this model the researcher specifies a priori sequence for the individual predictor variables. The analysis proceeds in a forward selection, backward elimination, or stepwise selection mode. This method is different from those previously discussed in that the researcher determines the order of entry from a careful consideration of the available theory and research. One type of hierarchical regression is known as set wise regression (block-wise, chunk-wise, or forced stepwise regression). Here the researcher specifies a priori sequence for sets or predictor variables. This method is the same as hierarchical in that the researcher determines the order of entry of the variables. The difference is that the set wise method uses sets of predictor variables at each stage rather than one individual predictor variable at a time.

There are some comments on the sequential regression procedures. First is that numerous statisticians have noted problems with the step wise methods which include (a) selecting noise rather than important predictors; (b) highly inflated R2 and adjusted R2 values; (c) CIs for partial slopes that are too narrow; (d) p values that are not trustworthy; (e) important predictors being barely edged out of the model, making it possible to miss the true model; and (f) potentially heavy capitalization on chance given the number of models analysed. Second theoretically based regression models have become the norm in many disciplines.

8.4

We will now look at how to deal with nonlinearity. We introduce several multiple regression models for when the criterion variables does not have a linear relationship with the predictor variables. First the polynomial regression models. In these models, the powers of the predictor variables are used. It is as follows:

$$Y = b_1X + b_2X^2 + \dots + b_mX^m + a + e \qquad (100)$$

If the model only consists of X taken to the first power, then this is a simple linear regression model (or first-degree polynomial). A second-degree polynomial includes X taken to the second power (quadratic model). A third-degree polynomial includes X taken to the third power (cubic model). It is important that when a higher-order polynomial is included (quadratic, cubic, or more) the first-order polynomial must also be included in the model.

8.5

Another type of model involves the use of an interaction term. These can be implemented in any type of regression model. A simple two-predictor interaction-type model is written as:

$$Y = b_1X_1 + b_2X_2 + b_3X_1X_2 + a + e \qquad (101)$$

Where X1X2 is the interaction of predictor variables 1 and 2. The definition of an interaction is the relationship between Y and X1 depends on the level of X2. In other words, X2 is the moderator variable. Note that if the predictors are very highly correlated, collinearity is likely.

8.6

Until now we have only looked at continuous predictors (independent variables that are interval or ratio in scale). However, it can also be that you want to have a categorical predictor. However these variables need to be recoded, so that they are on a scale of 0 and 1. This is called “dummy coding”. For example 0 is coded for females, and 1 is coded for males.

8.7

We will now discuss the steps to follow for conducting a multiple linear regression analyses in SPSS. We have data with one dependent variable, and two independent variables.

- Go to “analyse” and select “regression” and then select “linear”.
- Click dependent variable and move it into the “dependent” box. Click the independent variables in the “Independent(s)” box.
- From the “Linear regression” dialog box, clicking on “statistics” provides options to select. You need to select the following (a) estimates, (b) CIs, (c) model fit, (d) R squared change, (e) descriptive, (f) part and partial correlations, (g) collinearity diagnostics, (h) Durbin-Watson, and (i) case wise diagnostics. Click on “continue”.

- From the “Linear regression” dialog box clicking on “Plots” also gives options to select. You need to check the following: (a) histogram, (b) normal probability plot, (c) produce all partial plots. Click on “continue”.
- Now click in the “Linear regression” dialog box in “save”. Under the heading predicted values you need to check unstandardized. Under the heading Residuals, check the following (a) unstandardized and (b) studentized. Under the heading distances you need to check the following (a) Mahalanobis, (b) Cook’s, and (c) leverage values. Under the heading Influence Statistics, you need to check standardized DFBETA(s). Click on “continue” and click on “OK” to generate output.

The output is listed on page 395-399.

Some important interpretations from this output:

The adjusted R2 is interpreted as the percentage of variation in the dependent variable that is explained after adjusting for sample size and the number of predictors.

We will review the values that we have requested to be saved in our dataset:

- PRE_1 represents the unstandardized predicted values.
- RES_1 represents the unstandardized residuals, simply the difference between the observed and predicted values.
- SRE_1 represents the studentized residuals, a type of standardized residual that is more sensitive to outliers as compared to standardized residuals. These are computed as the unstandardized residuals divided by an estimate of the standard deviation with the case removed. Rule of thumb, studentized residuals with an absolute value greater than 3 are considered outliers.
- MAH_1 represents Mahalanobis distance values, which measure how far that particular case is from the average of the independent variable and thus can be helpful in detecting outliers. The squared Mahalanobis distances divided by the number of variables which are greater than 2.5 (small samples) or 3-4 (large samples) are suggestive of outliers.
- COO_1 Cook’s distance values and provides an indication of influence of individual cases. Rule of thumb, Cook’s values greater than 1 suggest that case is potentially problematic.
- LEV_1 represents leverage values, a measure of distance from the respective case to the average of the predictor.
- SDB0_1, SDB1_1, and SDB2_1 are standardized DFBETA values for the intercept and slopes, and are easier to interpret as compared to their unstandardized counterparts. Standardized DFBETA values greater than an absolute value of 2 suggest that the case may be exerting undue influence on the calculation of the parameters in the model.

To see if the assumptions are met we need to different things. For the independence assumptions we will plot the following (a) studentized residuals against unstandardized predicted values and (b) studentized residuals against each independent variable. If the assumption of independence is met the points should fall randomly within a band of -2.0 and +2.0.

We can use the same plots to test for homogeneity. Evidence for meeting the assumption is a plot where the spread of residuals appears fairly constant over the range of unstandardized predicted values and observed values of the independent variables.

We can also use these plots to review the assumption of linearity. There is linearity if you see a diagonal line.

For the normality assumption you can use the methods discussed before, such as the skewness and kurtosis statistic. Or you can use the boxplot, Q-Q plot, or the Shapiro-Wilk (S-W) test.

Multicollinearity refers to strong correlations between the independent variables. Detecting this can be done by reviewing VIF and tolerance statistics. Tolerance is calculated as (1- R2), and values close to 0 (0.10 or less) suggest potential Multicollinearity problems. Because a tolerance of 0.10 suggest that 90% of the variance in one of the independent variables can be explained by another independent variable. VIF is 1/tolerance. So values greater than 10 suggest potential Multicollinearity.

8.8

Again we will use G*power to measure the post hoc and priori power of the test. For the post hoc power analysis, we need to select the correct test family. This can be done by selecting “tests” then “correlation and regression”, and then “Linear multiple regression: fixed model, R2, deviation from zero”. This will automatically change the “test family” to the “F test”. The input parameters are: (1) effect size, (2) alpha level, (3) total sample size, and (4) number of predictors. We need to use the pop-out effect size calculator in G*power. Click on “calculate” to compute the effect size, then click on “calculate and transfer to main window” to transfer the calculated effect size to “input parameters”.

For the priori power, we can determine the total sample size needed for multiple linear regressions given the estimated size f2, alpha level, desired power, and number of predictors. We follow again small effect: r2=0.02, moderate effect: r2=0.15, and large effect: r2=0.35.