Linear Models: Looking for Bias

The following sections have been adapted from Field (2013) Chapter 8. These sections have been edited down considerably and I suggest (especially if you're confused) that you read this Chapter in its entirety. You will also need to read this chapter to help you interpret the output. If you're having problems there is plenty of support available: you can (1) email or see your seminar tutor (2) post a message on the course bulletin board or (3) drop into my office hour.

More on Bias

Outliers

We have seen that outliers can bias a model: they bias estimates of the regression parameters. we know that an outlier, by its nature, is very different from all of the other scores. Therefore, if we were to work out the differences between the data values that were collected, and the values predicted by the model, we could detect an outlier by looking for large differences. The differences between the values of the outcome predicted by the model and the values of the outcome observed in the sample are called *residuals*. If a model is a poor fit of the sample data then the residuals will be large. Also, if any cases stand out as having a large residual, then they could be outliers.

The *normal* or **unstandardized residuals** described above are measured in the same units as the outcome variable and so are difficult to interpret across different models. All we can do is to look for residuals that stand out as being particularly large: we cannot define a universal cut-off point for what constitutes a large residual. To overcome this problem, we use **standardized residuals**, which are the residuals converted to *z*-scores, which means they are converted into standard deviation units (i.e., they are distributed around a mean of 0 with a standard deviation of 1). By converting residuals into *z*-scores (standardized residuals) we can compare residuals from different models and use what we know about the properties of *z*-scores to devise universal guidelines for what constitutes an acceptable (or unacceptable) value. For example, in a normally distributed sample, 95% of *z*-scores should lie between −1.96 and +1.96, 99% should lie between −2.58 and +2.58, and 99.9% (i.e., nearly all of them) should lie between −3.29 and +3.29. Some general rules for standardized residuals are derived from these facts: (1) standardized residuals with an absolute value greater than 3.29 (we can use 3 as an approximation) are cause for concern because in an average sample a value this high is unlikely to occur; (2) if more than 1% of our sample cases have standardized residuals with an absolute value greater than 2.58 (we usually just say 2.5) there is evidence that the level of error within our model is unacceptable (the model is a fairly poor fit of the sample data); and (3) if more than 5% of cases have standardized residuals with an absolute value greater than 1.96 (we can use 2 for convenience) then there is also evidence that the model is a poor representation of the actual data.

Influential Cases

As well as testing for outliers by looking at the error in the model, it is also possible to look at whether certain cases exert undue influence over the parameters of the model. So, if we were to delete a certain case, would we obtain different regression coefficients? This type of analysis can help to determine whether the regression model is stable across the sample, or whether it is biased by a few influential cases. There are numerous ways to look for influential cases, all described in scintillating detail in Field (2013). We'll just look at 1 of them, **Cook's distance**, which quantifies the effect of a single case on the model as a whole. Cook and Weisberg (1982) have suggested that values greater than 1 may be cause for concern.

Generalization

Remember from your lecture on bias that linear models assume:

- Linearity and additivity: the relationship you're trying to model is, in fact, linear and with several predictors, they combine additively.
- Normality: For *b* estimates to be optimal the residuals should be normally distributed. For *p*-values and confidence intervals to be accurate, the sampling distribution of *bs* should be normal.
- Homoscedasticity: necessary for *b* estimates to be optimal and significance tests and CIs of the parameters to be accurate.

However, there are some other assumptions that are important if we want to generalize the model we fit beyond out sample. The most important is:

- **Independent errors**: For any two observations the residual terms should be uncorrelated (i.e., independent). This eventuality is sometimes described as a lack of **autocorrelation**. If we violate the assumption of independence then our confidence intervals and significance tests will be invalid. This assumption can be tested with the **Durbin-Watson test** (1951). The test statistic can vary between 0 and 4 with a value of 2 meaning that the residuals are uncorrelated. A value greater than 2 indicates a negative correlation between adjacent residuals, whereas a value below 2 indicates a positive correlation. The size of the Durbin-Watson statistic depends upon the number of predictors in the model and the number of observations. As a very conservative rule of thumb, values less than 1 or greater than 3 are definitely cause for concern; however, values closer to 2 may still be problematic depending on your sample and model.

There are some other considerations that we have not yet discussed (see Berry, 1993):

- *Predictors are uncorrelated with 'external variables'*: *External variables* are variables that haven't been included in the regression model which influence the outcome variable.
- *Variable types*: All predictor variables must be quantitative or categorical (with 2 categories), and the outcome variable must be quantitative, continuous and unbounded.
- *No perfect* **multicollinearity**: If your model has more than one predictor then there should be no perfect linear relationship between two or more of the predictors. So, the predictor variables should not correlate too.
- *Non-zero variance*: The predictors should have some variation in value (i.e., they do not have variances of 0). This is self-evident really.

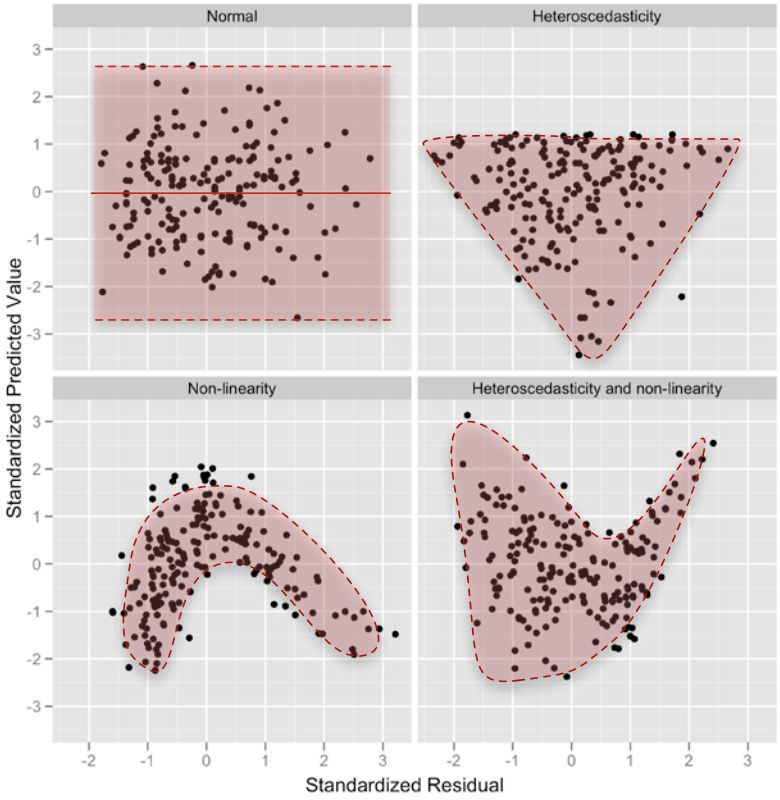


Figure 1: Plots of standardized residuals against predicted (fitted) values

The four most important conditions are linearity and additivity, normality, homoscedasticity, and independent errors. These can be tested graphically using a plot of standardized residuals (*zresid*) against standardized predicted values (*zpred*). Figure 1 shows several examples of the plot of standardized residuals against standardized predicted values. The top left panel shows a situation in which the assumptions of linearity, independent errors and homoscedasticity have been met. Independent errors are shown by a random pattern of dots. The top right panel shows a similar plot for

a data set that violates the assumption of homoscedasticity. Note that the points form a funnel: they become more spread out across the graph. This funnel shape is typical of heteroscedasticity and indicates increasing variance across the residuals. The bottom left panel shows a plot of some data in which there is a non-linear relationship between the outcome and the predictor: there is a clear curve in the residuals. Finally, the bottom right panel illustrates data that not only have a non-linear relationship, but also show heteroscedasticity. Note first the curved trend in the residuals, and then also note that at one end of the plot the points are very close together whereas at the other end they are widely dispersed. When these assumptions have been violated you will not see these exact patterns, but hopefully these plots will help you to understand the general anomalies you should look out for.

Methods of Regression

Last week we looked at a situation where we forced predictors into the model. However, there are other options. We can select predictors in several ways:

- In **hierarchical regression** predictors are selected based on past work and the researcher decides in which order to enter the predictors into the model. As a general rule, known predictors (from other research) should be entered into the model first in order of their importance in predicting the outcome. After known predictors have been entered, the experimenter can add any new predictors into the model. New predictors can be entered either all in one go, in a stepwise manner, or hierarchically (such that the new predictor suspected to be the most important is entered first).
- **Forced entry** (or *Enter* as it is known in SPSS) is a method in which all predictors are forced into the model simultaneously. Like hierarchical, this method relies on good theoretical reasons for including the chosen predictors, but unlike hierarchical the experimenter makes no decision about the order in which variables are entered.
- Stepwise methods are generally frowned upon by statisticians. In **stepwise regressions** decisions about the order in which predictors are entered into the model are based on a purely mathematical criterion. In the *forward* method, an initial model is defined that contains only the constant (b_0). The computer then searches for the predictor (out of the ones available) that best predicts the outcome variable—it does this by selecting the predictor that has the highest simple correlation with the outcome. If this predictor significantly improves the ability of the model to predict the outcome, then this predictor is retained in the model and the computer searches for a second predictor. The criterion used for selecting this second predictor is that it is the variable that has the largest semi-partial correlation with the outcome. In plain English, imagine that the first predictor can explain 40% of the variation in the outcome variable; then there is still 60% left unexplained. The computer searches for the predictor that can explain the biggest part of the remaining 60% (it is not interested in the 40% that is already explained). As such, this semi-partial correlation gives a measure of how much 'new variance' in the outcome can be explained by each remaining predictor. The predictor that accounts for the most new variance is added to the model and, if it makes a significant contribution to the predictive power of the model, it is retained and another predictor is considered.

Many writers argue that stepwise methods take the important methodological decisions out of the hands of the researcher. What's more, the models derived by stepwise methods often take advantage of random sampling variation and so decisions about which variables should be included will be based upon slight differences in their semi-partial correlation. However, these slight statistical differences may contrast dramatically with the theoretical importance of a predictor to the model. There is also the danger of over-fitting (having too many variables in the model that essentially make little contribution to predicting the outcome) and under-fitting (leaving out important predictors) the model. However, when little theory exists stepwise methods might be the only practical option.

The Example

We'll look at data collected from several questionnaires relating to clinical psychology, and we will use these measures to predict social anxiety using multiple regression. Anxiety disorders take on different shapes and forms, and each disorder is believed to be distinct and have unique causes. We can summarise the disorders and some popular theories as follows:

- **Social Anxiety**: Social anxiety disorder is a marked and persistent fear of 1 or more social or performance situations in which the person is exposed to unfamiliar people or possible scrutiny by others. This anxiety leads to avoidance of these situations. People with social phobia are believed to feel elevated feelings of shame.
- **Obsessive Compulsive Disorder (OCD)**: OCD is characterised by the everyday intrusion into conscious thinking of intense, repetitive, personally abhorrent, absurd and alien thoughts (Obsessions), leading to the endless repetition of specific acts or to the rehearsal of bizarre and irrational mental and behavioural rituals (compulsions).

Social anxiety and obsessive compulsive disorder are seen as distinct disorders having different causes. However, there are some similarities.

- They both involve some kind of attentional bias: attention to bodily sensation in social anxiety and attention to things that could have negative consequences in OCD.
- They both involve repetitive thinking styles: social phobics ruminate about social encounters after the event (known as post-event processing), and people with OCD have recurring intrusive thoughts and images.
- They both involve safety behaviours (i.e. trying to avoid the thing that makes you anxious).

This might lead us to think that, rather than being different disorders, they are manifestations of the same core processes. One way to research this possibility would be to see whether social anxiety can be predicted from measures of *other* anxiety disorders. If social anxiety disorder and OCD are distinct we should expect that measures of OCD will not predict social anxiety. However, if there are core processes underlying all anxiety disorders, then measures of OCD should predict social anxiety.

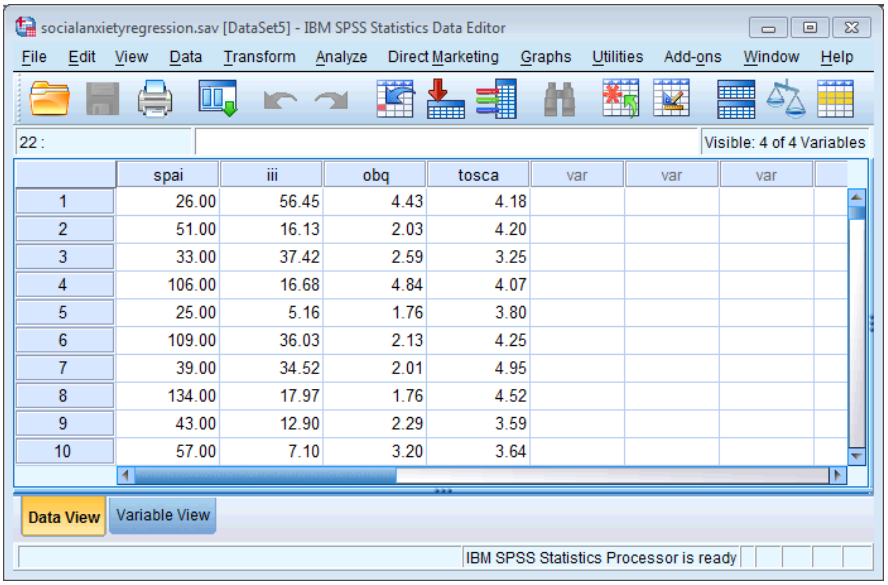


Figure 2: Data layout for multiple regression

The data are in the file **SocialAnxietyRegression.sav** which can be downloaded from Study Direct. This file contains four variables:

- The Social Phobia and Anxiety Inventory (**SPAI**), which measures levels of social anxiety.
- Interpretation of Intrusions Inventory (**III**), which measures the degree to which a person experiences intrusive thoughts like those found in OCD.
- Obsessive Beliefs Questionnaire (**OBQ**), which measures the degree to which people experience obsessive beliefs like those found in OCD.
- The Test of Self-Conscious Affect (**TOSCA**), which measures shame.

Each of 134 people was administered all four questionnaires. You should note that each questionnaire has its own column and each row represents a different person (see Figure 2).

What analysis will we do?

We are going to do a multiple regression analysis. Specifically, we're going to do a hierarchical multiple regression analysis. All this means is that we enter variables into the regression model in an order determined by past research and expectations. So, for your analysis, we will enter variables in so-called 'blocks':

- **Block 1**: the first block will contain any predictors that we expect to predict social anxiety. These variables should be entered using *forced entry*. In this example we have only one variable that we expect, theoretically, to predict social anxiety and that is shame (measured by the TOSCA).
- **Block 2**: the second block will contain our exploratory predictor variables (the one's we don't necessarily expect to predict social anxiety). This bock should contain the measures of OCD (OBQ and III) because these variables shouldn't predict social anxiety if social anxiety is indeed distinct from OCD. These variables should be entered using a *stepwise* method because we are 'exploring them' (think back to your lecture).

Doing Multiple Regression on SPSS

Specifying the First Block in Hierarchical Regression

Theory indicates that shame is a significant predictor of social phobia, and so this variable should be included in the model first. The exploratory variables (**obq** and **iii**) should, therefore, be entered into the model after shame. This method is called hierarchical (the researcher decides in which order to enter variables into the model based on past research). To do a hierarchical regression in SPSS we enter the variables in blocks (each block representing one step in the hierarchy). To get to the main regression dialog box select Analyze Regression Linear... . The main dialog box is shown in Figure 3.

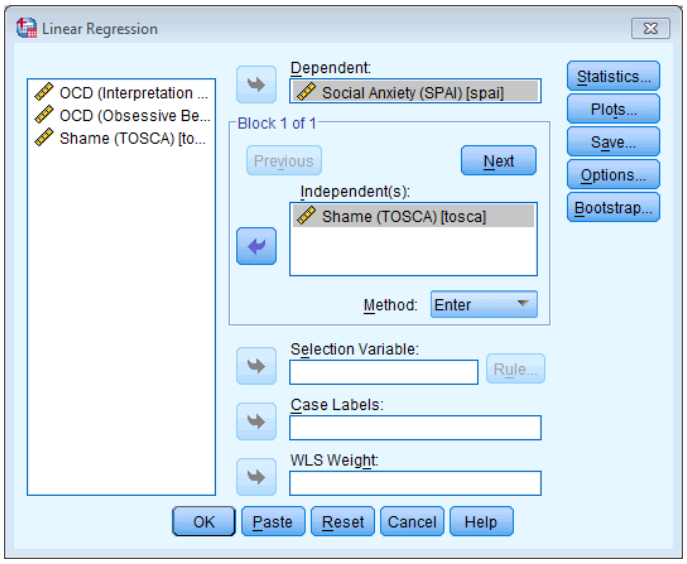
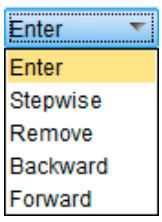


Figure 3: Main dialog box for block 1 of the multiple regression

The main dialog box is fairly self-explanatory in that there is a space to specify the dependent variable (outcome), and a space to place one or more independent variables (predictor variables). As usual, the variables in the data editor are listed on the left-hand side of the box. Highlight the outcome variable (SPAI scores) in this list by clicking on it and then transfer it to the box labelled *Dependent* by clicking on or dragging it across. We also need to specify the predictor variable for the first block. We decided that shame should be entered into the model first (because theory indicates that it is an important predictor), so, highlight this variable in the list and transfer it to the box labelled *Independent(s)* by clicking on or dragging it across. Underneath the *Independent(s)* box, there is a drop-down menu for specifying the *Method* of regression. You can select a different method of variable entry for each block by clicking on Enter, next to where it says *Method*. The default option is forced entry, and this is the option we want, but if you were carrying



out more exploratory work, you might decide to use one of the stepwise methods (forward, backward, stepwise or remove).

Specifying the Second Block in Hierarchical Regression

Having specified the first block in the hierarchy, we move onto to the second. To tell the computer that you want to specify a new block of predictors you must click on Next. This process clears the *Independent(s)* box so that you can enter the new predictors (you should also note that above this box it now reads *Block 2 of 2* indicating that you are in the second block of the two that you have so far specified). We decided that the second block would contain both of the new predictors and so you should click on **obq** and **iii** in the variables list and transfer them, one by one, to the *Independent(s)* box by clicking on . The dialog box should now look like Figure 4. To move between blocks use the Previous and Next buttons (so, for example, to move back to block 1, click on Previous).

It is possible to select different methods of variable entry for different blocks in a hierarchy. So, although we specified forced entry for the first block, we could now specify a stepwise method for the second. Given that we have no previous research regarding the effects of **obq** and **iii** on SPAI scores, we might be justified in requesting a stepwise method for this block (see your lecture notes and my textbook). For this analysis select a stepwise method for this second block.

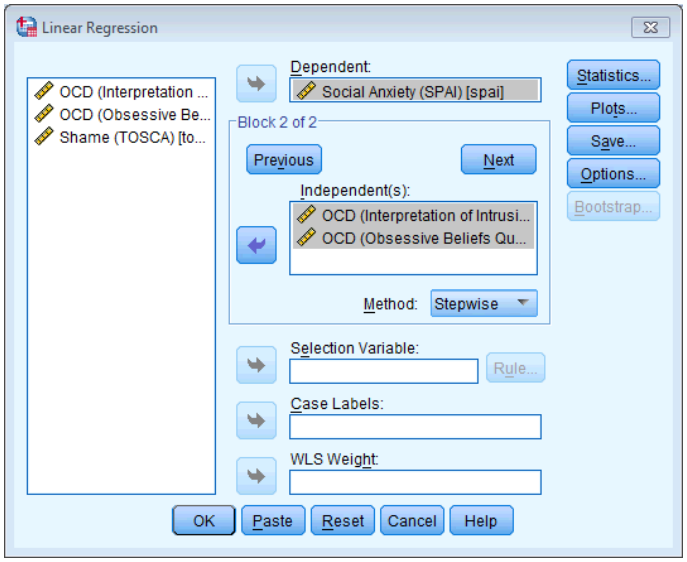


Figure 4: Main dialog box for block 2 of the multiple regression

Statistics

In the main *regression* dialog box click on Statistics... to open a dialog box for selecting various important options relating to the model (Figure 5). Most of these options relate to the parameters of the model; however, there are procedures available for checking the assumptions of no multicollinearity (*Collinearity diagnostics*) and independence of errors (*Durbin-Watson*). When you have selected the statistics you require (I recommend all but the covariance matrix as a general rule) click on Continue to return to the main dialog box.

- → **Estimates**: This option is selected by default because it gives us the estimated coefficients of the regression model (i.e. the estimated *b*-values).
- → **Confidence intervals**: This option produces confidence intervals for each of the unstandardized regression coefficients.
- → **Model fit**: This option is vital and is selected by default. It provides not only a statistical test of the model's ability to predict the outcome variable (the *F*-test), but also the value of *R* (or multiple *R*), the corresponding R^2, and the adjusted R^2.
- → **R squared change**: This option displays the change in R^2 resulting from the inclusion of a new predictor (or block of predictors). This measure is a useful way to assess the unique contribution of new predictors (or blocks) to explaining variance in the outcome.

- → **Descriptives**: If selected, this option displays a table of the mean, standard deviation and number of observations of all of the variables included in the analysis. A correlation matrix is also displayed showing the correlation between all of the variables and the one-tailed probability for each correlation coefficient. This correlation matrix can be used to establish whether there is multicollinearity.
- → **Part and partial correlations**: This option produces the zero-order correlation (the Pearson correlation) between each predictor and the outcome variable. It also produces the partial correlation between each predictor and the outcome, controlling for all other predictors in the model.
- → **Collinearity diagnostics**: This option is for obtaining collinearity statistics such as the VIF, tolerance, eigenvalues of the scaled, uncentred cross-products matrix, condition indexes and variance proportions (see Field, 2013, and your lecture notes).
- → **Durbin-Watson**: This option produces the Durbin-Watson test statistic, which tests for correlations between errors.
- → **Casewise diagnostics**: This option lists the observed value of the outcome, the predicted value of the outcome, the difference between these values (the residual) and this difference standardized. Furthermore, it will list these values either for all cases, or just for cases for which the standardized residual is greater than 3 (when the ± sign is ignored). This criterion value of 3 can be changed, and I recommend changing it to 2 for reasons that will become apparent.

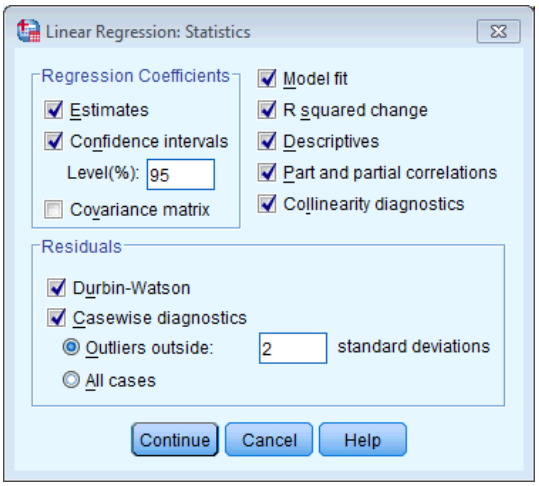


Figure 5: *Statistics* dialog box for regression analysis

Regression Plots

Once you are back in the main dialog box, click on Plots... to activate the regression *plots* dialog box shown in Figure 6. This dialog box provides the means to specify a number of graphs, which can help to establish the validity of some regression assumptions. Most of these plots involve various *residual* values. On the left-hand side of the dialog box is a list of several variables:

- **DEPENDNT** (the outcome variable).
- ***ZPRED** (the standardized predicted values of the dependent variable based on the model). These values are standardized forms of the values predicted by the model.
- ***ZRESID** (the standardized residuals, or errors). These values are the standardized differences between the observed data and the values that the model predicts).
- ***DRESID** (the deleted residuals).
- ***ADJPRED** (the adjusted predicted values).
- ***SRESID** (the Studentized residual).
- ***SDRESID** (the Studentized deleted residual).This value is the deleted residual divided by its standard error.

The variables listed in this dialog box all come under the general heading of residuals, and are discussed in detail in my book (sorry for all of the self-referencing, but I'm trying to condense a 60 page chapter into a manageable handout!). For a basic analysis it is worth plotting *ZRESID (*Y*-axis) against *ZPRED (*X*-axis), because this plot is useful to determine whether the assumptions of random errors and homoscedasticity have been met (see ealier). To create these plots

select a variable from the list, and transfer it to the space labelled either *X* or *Y* (which refer to the axes) by clicking . When you have selected two variables for the first plot (as is the case in Figure 6) you can specify a new plot by clicking on Next. This process clears the spaces in which variables are specified. If you click on Next and would like to return to the plot that you last specified, then simply click on Previous.

You can also select the tick-box labelled *Produce all partial plots* which will produce scatterplots of the residuals of the outcome variable and each of the predictors when both variables are regressed separately on the remaining predictors. Any obvious outliers on a partial plot represent cases that might have undue influence on a predictor's regression coefficient. Also, non-linear relationships between a predictor and the outcome variable are much more detectable using these plots. Finally, they are a useful way of detecting collinearity. There are several options for plots of the standardized residuals. First, you can select a histogram of the standardized residuals (this is extremely useful for checking the assumption of normality of errors). Second, you can ask for a normal probability plot, which also provides information about whether the residuals in the model are normally distributed. When you have selected the options you require, click on Continue to take you back to the main *regression* dialog box.

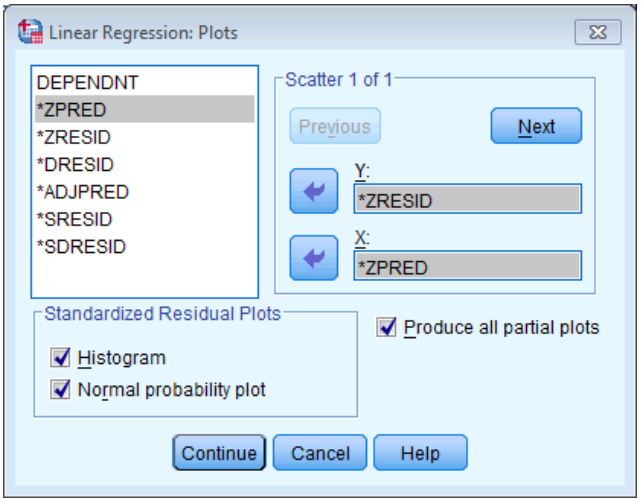


Figure 6: Linear regression: plots dialog box

Saving Regression Diagnostics

In this week's lecture we met two types of regression diagnostics: those that help us assess how well our model fits our sample and those that help us detect cases that have a large influence on the model generated. In SPSS we can choose to save these diagnostic variables in the data editor (so, SPSS will calculate them and then create new columns in the data editor in which the values are placed).

Click on Save... in the main *regression* dialog box to activate the *save* new variables dialog box (see Figure 7). Once this dialog box is active, it is a simple matter to tick the boxes next to the required statistics. Most of the available options are explained in Field (2013) and Figure 7 shows, what I consider to be, a bare minimum set of diagnostic statistics. Standardized versions of these diagnostics are generally easier to interpret and so I suggest selecting them in preference to the unstandardized versions. Once the regression has been run, SPSS creates a column in your data editor for each statistic requested and it has a standard set of variable names to describe each one (**zpr_1**: standardized predicted value; **zre_1**: standardized residual; **coo_1**: Cook's distance). After the name, there will be a number that refers to the analysis that has been run. So, for the first regression run on a data set the variable names will be followed by a 1, if you carry out a second regression it will create a new set of variables with names followed by a 2 and so on. When you have selected the diagnostics you require (by clicking in the appropriate boxes), click on Continue to return to the main *regression* dialog box.

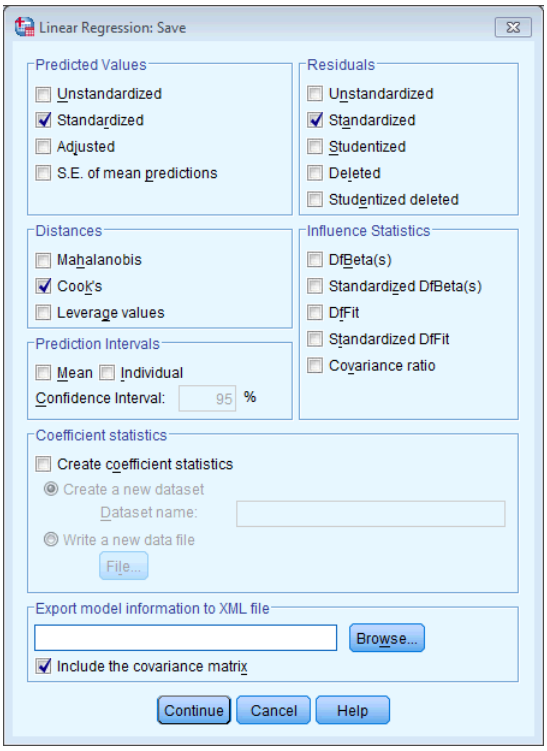


Figure 7: Dialog box for regression diagnostics

Bootstrapping

We can get bootstrapped confidence intervals for the regression coefficients by clicking Bootstrap... (see last week's handout). However, this function doesn't work when we have used the Save... option to save residuals, so we can't use it now. However, once you have run the analysis and inspected the residuals and influential cases, you might want to re-run the analysis selecting the bootstrap option (and remembering to deselect all of the options for saving variables).

A Brief Guide to Interpretation

Model Summary

The model summary (Output 1) contains two models. Model 1 refers to the first stage in the hierarchy when only TOSCA is used as a predictor. Model 2 refers to the final model (TOSCA, and OBQ and III if they end up being included).

→ In the column labelled R are the values of the multiple correlation coefficient between the predictors and the outcome. When only TOSCA is used as a predictor, this is the simple correlation between SPAI and TOSCA (0.34).

→ The next column gives us a value of R^2, which is a measure of how much of the variability in the outcome is accounted for by the predictors. For the first model its value is 0.116, which means that TOSCA accounts for 11.6% of the variation in social anxiety. However, for the final model (model 2), this value increases to 0.157 or 15.7% of the variance in SPAI. Therefore, whatever variables enter the model in block 2 account for an extra (15.7-11.6) 4.1% of the variance in SPAI scores (this is also the value in the column labelled R-square change but expressed as a percentage).

→ The adjusted R^2 gives us some idea of how well our model generalizes and ideally we would like its value to be the same, or very close to, the value of R^2. In this example the difference for the final model is a fair bit (0.157 – 0.143 = 0.014 or 1.4%). This shrinkage means that if the model were derived from the population rather than a sample it would account for approximately 1.4% less variance in the outcome.

→ Finally, if you requested the Durbin-Watson statistic it will be found in the last column. This statistic informs us about whether the assumption of independent errors is tenable. The closer to 2 that the value is, the better, and for these data the value is 2.084, which is so close to 2 that the assumption has almost certainly been met.

Model Summary[c]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	.340[a]	.116	.109	28.38137	.116	16.515	1	126	.000	
2	.396[b]	.157	.143	27.82969	.041	6.045	1	125	.015	2.084

a. Predictors: (Constant), Shame (TOSCA)

b. Predictors: (Constant), Shame (TOSCA), OCD (Obsessive Beliefs Questionnaire)

c. Dependent Variable: Social Anxiety (SPAI)

Output 1

ANOVA Table

Output 2contains an analysis of variance (ANOVA) that tests whether the model is significantly better at predicting the outcome than using the mean as a 'best guess'. This table is again split into two sections: one for each model. If the improvement due to fitting the regression model is much greater than the inaccuracy within the model then the value of *F* will be greater than 1 and SPSS calculates the exact probability of obtaining the value of *F* at least this big if there were no effect. For the initial model the *F*-ratio is 16.52 ($p < .001$), and for the second model the value of *F* is 11.61, which is also highly significant ($p < .001$). We can interpret these results as meaning that the final model significantly improves our ability to predict the outcome variable.

ANOVA[c]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	13302.700	1	13302.700	16.515	.000[a]
	Residual	101493.3	126	805.502		
	Total	114796.0	127			
2	Regression	17984.538	2	8992.269	11.611	.000[b]
	Residual	96811.431	125	774.491		
	Total	114796.0	127			

a. Predictors: (Constant), Shame (TOSCA)

b. Predictors: (Constant), Shame (TOSCA), OCD (Obsessive Beliefs Questionnaire)

c. Dependent Variable: Social Anxiety (SPAI)

Output 2

Model Parameters

The next part of the output is concerned with the parameters of the model. The first step in our hierarchy included TOSCA and although these parameters are interesting up to a point, we're more interested in the final model because this includes all predictors that make a significant contribution to predicting social anxiety. So, we'll look only at the lower half of the table (Model 2).

In multiple regression the model takes the form of an equation that contains a coefficient (*b*) for each predictor. The first part of the table gives us estimates for these *b* values and these values indicate the individual contribution of each predictor to the model.

The *b* values tell us about the relationship between social anxiety and each predictor. If the value is positive we can tell that there is a positive relationship between the predictor and the outcome whereas a negative coefficient represents a negative relationship. For these data both predictors have positive *b* values indicating positive relationships. So, as shame (TOSCA) increases, social anxiety increases and as obsessive beliefs increase so does social anxiety; The *b* values also tell us to what degree each predictor affects the outcome if the effects of all other predictors are held constant.

Each of these beta values has an associated standard error indicating to what extent these values would vary across different samples, and these standard errors are used to determine whether or not the *b* value differs significantly from zero (using the *t*-statistic). Therefore, if the *t*-test associated with a *b* value is significant (if the value in the column labelled Sig. is less than 0.05) then that predictor is making a significant contribution to the model. For this model,

Shame (TOSCA), $t(125) = 3.16$, $p = .002$, and obsessive beliefs, $t(125) = 2.46$, $p = .015$, are significant predictors of social anxiety. From the magnitude of the *t*-statistics we can see that the Shame (TOSCA) had slightly more impact than obsessive beliefs. This conclusion is also borne out by the standardized beta values, which are measured in standard deviation units and so are directly comparable: the standardized beta values for Shame (TOSCA) is 0.273, and for obsessive beliefs is 0.213. This tells us that shame has slightly more impact in the model.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B		Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	-54.368	28.618		-1.900	.060	-111.002	2.267					
	Shame (TOSCA)	27.448	6.754	.340	4.064	.000	14.081	40.814	.340	.340	.340	1.000	1.000
2	(Constant)	-51.493	28.086		-1.833	.069	-107.079	4.094					
	Shame (TOSCA)	22.047	6.978	.273	3.160	.002	8.237	35.856	.340	.272	.260	.901	1.110
	OCD (Obsessive Beliefs Questionnaire)	6.920	2.815	.213	2.459	.015	1.350	12.491	.299	.215	.202	.901	1.110

a. Dependent Variable: Social Anxiety (SPAI)

Output 3

Excluded Variables

At each stage of a regression analysis SPSS provides a summary of any variables that have not yet been entered into the model. In a hierarchical model, this summary has details of the variables that have been specified to be entered in subsequent steps, and in stepwise regression this table contains summaries of the variables that SPSS is considering entering into the model. The summary gives an estimate of each predictor's *b* value if it was entered into the equation at this point and calculates a *t*-test for this value. In a stepwise regression, SPSS should enter the predictor with the highest *t*-statistic and will continue entering predictors until there are none left with t-statistics that have significance values less than 0.05. Therefore, the final model might not include all of the variables you asked SPSS to enter.

In this case it tells us that if the interpretation of intrusions (III) is entered into the model it would not have a significant impact on the model's ability to predict social anxiety, $t = -0.049$, $p = .961$. In fact the significance of this variable is almost 1 indicating it would have virtually no impact whatsoever (note also that its beta value is extremely close to zero!).

Excluded Variables[c]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics		
						Tolerance	VIF	Minimum Tolerance
1	OCD (Interpretation of Intrusions Inventory)	.132[a]	1.515	.132	.134	.917	1.091	.917
	OCD (Obsessive Beliefs Questionnaire)	.213[a]	2.459	.015	.215	.901	1.110	.901
2	OCD (Interpretation of Intrusions Inventory)	-.005[b]	-.049	.961	-.004	.541	1.849	.531

a. Predictors in the Model: (Constant), Shame (TOSCA)

b. Predictors in the Model: (Constant), Shame (TOSCA), OCD (Obsessive Beliefs Questionnaire)

c. Dependent Variable: Social Anxiety (SPAI)

Output 4

Checking for Bias

SPSS produces a summary table of the residual statistics and these should be examined for extreme cases. Output 5 shows any cases that have a standardized residual less than −2 or greater than 2 (remember that we changed the default criterion from 3 to 2). In an ordinary sample we would expect 95% of cases to have standardized residuals within about ±2. We have a sample of 134, therefore it is reasonable to expect about 7 cases (5% approx..) to have standardized residuals outside of these limits. From Output 5 we can see that we have 7 cases (5%) that are outside of the limits: therefore, our sample is basically what we would expect. In addition, 99% of cases should lie within ±2.5 and so we would expect only 1% of cases to lie outside of these limits. From the cases listed here, it is clear that two cases (1.5%)

lie outside of the limits (cases 8 and 45). Therefore, our sample appears to conform roughly to what we would expect for a fairly accurate model. There are also no standardized residuals greater than 3, which is good news.

We should also scan the data editor to see if any cases have Cook's distance (**COO_1**) greater than 1. [You could also use SPSS to find the maximum value of Cook's distance by using the descriptive statistics command]. You should find that all of Cook's distances are below 1, which means that no cases are having an undue influence.

Casewise Diagnostics[a]

Case Number	Std. Residual	Social Anxiety (SPAI)	Predicted Value	Residual
8	2.645	134	60.39	73.612
16	-2.205	16	77.37	-61.370
45	2.708	137	61.65	75.354
49	-2.368	-5	60.91	-65.914
120	2.244	134	71.55	62.452
121	-2.002	12	67.71	-55.707
127	-2.074	-3	54.72	-57.717

a. Dependent Variable: Social Anxiety (SPAI)

Output 5

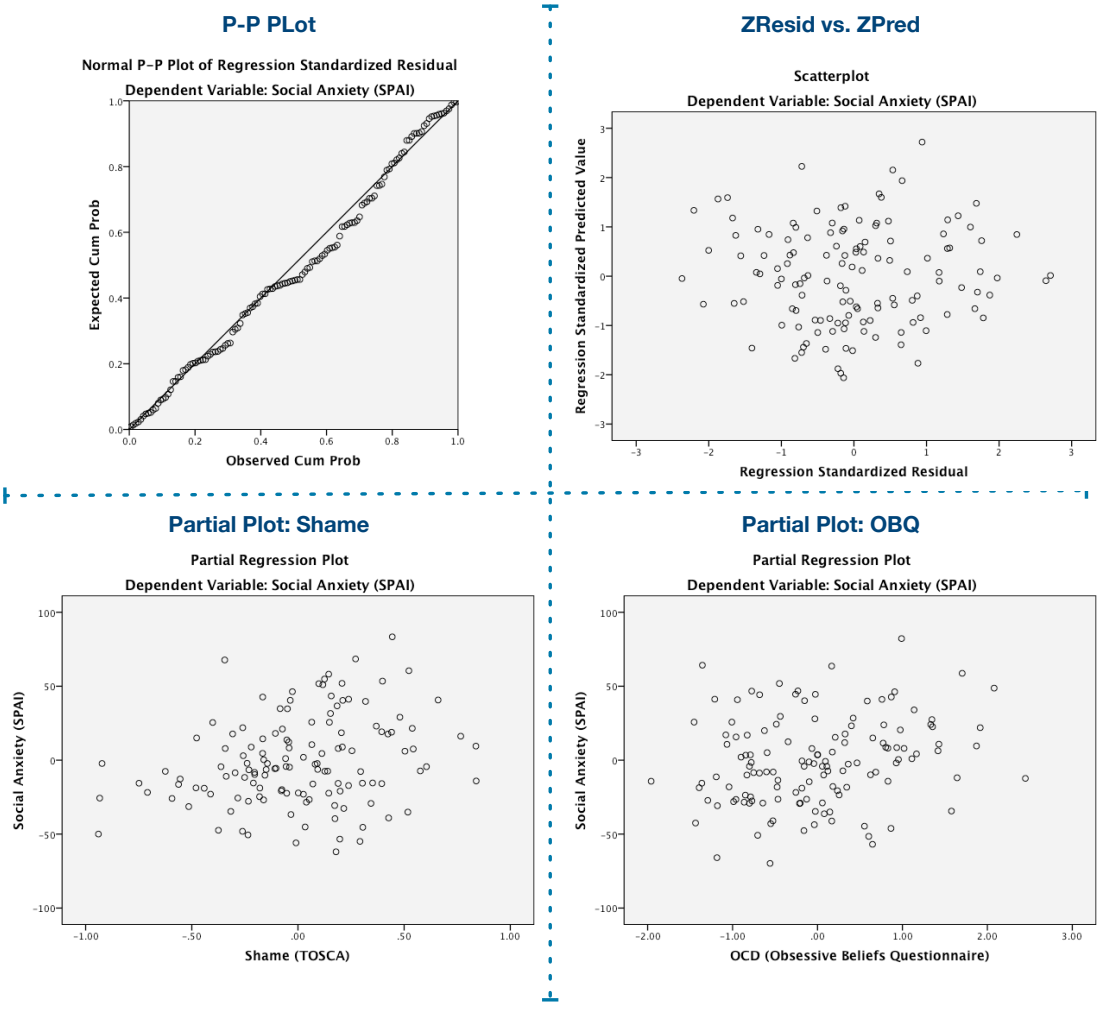


Figure 8: P-P plot (top left), a plot of standardized residuals vs. standardized predicted values (top right), and partial plots of social anxiety against shame (bottom left) and OBQ (bottom right)

We can use histograms and P-P plots to look for normality of the residuals. Figure 8 (top left) shows the P-P plot for our model. The dots hover fairly close to the diagonal line indicating normality in the residuals. We can look for heteroscedasticity and non-linearity using a plot of standardized residuals against standardized predicted values. If everything is OK then this graph should look like a random array of dots, if the graph funnels out then that is a sign of

heteroscedasticity and any curve suggest nonlinearity (see earlier). Figure 8 (top right) shows the plot for our model. Note how the points are randomly and evenly dispersed throughout the plot. This pattern is indicative of a situation in which the assumptions of linearity and homoscedasticity have been met. Compare this with the examples in Figure 1.

Figure 8 also shows the partial plots, which are scatterplots of the residuals of the outcome variable and each of the predictors when both variables are regressed separately on the remaining predictors. Obvious outliers on a partial plot represent cases that might have undue influence on a predictor's regression coefficient and that non-linear relationships and heteroscedasticity can be detected using these plots as well. For shame (Figure 8 bottom left) the partial plot shows the positive relationship to social anxiety. There are no obvious outliers on this plot, but the cloud of dots is a bit funnel-shaped, possibly indicating some heteroscedasticity. For OBQ (Figure 8, bottom right) the plot again shows a positive relationship to social anxiety. There are no obvious outliers on this plot.

Finally, the VIF values are well below 10 which reassures us that multicollinearity is not a problem.

Writing Up Multiple Regression Analysis

If your model has several predictors than you can't really beat a summary table as a concise way to report your model. As a bare minimum report the betas, their confidence interval, significance value and some general statistics about the model (such as the R^2). The standardized beta values and the standard errors are also very useful. So, basically, you want to reproduce the table labelled *Coefficients* from the SPSS output and omit some of the non-essential information. For the example in this chapter we might produce a table like that in.

See if you can look back through the SPSS output in this chapter and work out from where the values came. Things to note are: (1) I've rounded off to 2 decimal places throughout because this is a reasonable level of precision given the variables measured; (2) for the standardized betas there is no zero before the decimal point (because these values shouldn't exceed 1) but for all other values less than 1 the zero is present; (3) often you'll see the significance of the variable is denoted by an asterisk with a footnote to indicate the significance level being used but it's better practice to report exact *p*-values; (4) the R^2 for the initial model and the change in R^2 (denoted as ΔR^2) for each subsequent step of the model are reported below the table; and (5) in the title I have mentioned that confidence intervals and standard errors in the table are based on bootstrapping: this information is important for readers to know

Table 1: Linear model of predictors of social anxiety (SPAI). 95% confidence intervals reported in parentheses.

	b	SE *B*	β	*p*
Step 1				
Constant	−54.37 (−111.00, 2.27)	28.62		*p* = .06
Shame (TOSCA)	27.45 (14.08, 40.81)	6.75	.34	*p* < .001
Step 2				
Constant	−51.49 (−107.08, 4.09)	28.09		*p* = .069
Shame (TOSCA)	22.05 (8.24, 35.86)	6.98	.27	*p* = .002
OCD (OBQ)	6.92 (1.35, 12.49)	2.82	.21	*p* = .015

Note. R^2 = .12 for Step 1: ΔR^2 = .04 for Step 2 (*p*s < .05).

Tasks

Task 1

A fashion student was interested in factors that predicted the salaries of catwalk models. She collected data from 231 models. For each model she asked them their salary per day on days when they were working (**salary**), their age (**age**), how many years they had worked as a model (**years**), and then got a panel of experts from modelling agencies to rate the attractiveness of each model as a percentage with 100% being perfectly attractive (**beauty**). The data are in the file **Supermodel.sav** on the course website. Conduct a multiple regression to see which factors predict a model's salary? (Answers to this task can be found at www.uk.sagepub.com/field4e/study/smartalex/chapter8.pdf).

How much variance does the final model explain?

Your Answers:

Which variables significantly predict salary?

Your Answers:

Fill in the values for the following APA format table of the results:

	b	*SE b*	β	*p*
Constant				
Age				
Years as a Model				
Attractiveness				

Note. R^2 =

Write out the regression equation for the final model.

Your Answers:

Are the residuals as you would expect for a good model?

Your Answers:

Is there evidence of normality of errors, homoscedasticity and no multicollinearity?

Your Answers:

Task 2

Coldwell, Pike and Dunn (2006) investigated whether household chaos predicted children's problem behaviour over and above parenting. They collected data from 118 two-parent families. For each family they recorded the age and gender of both the older and younger sibling; **age_child1**, **gender_child1, age_child1** and **gender_child2** respectively. They then interviewed each child about their relationship with their parent's using the Berkeley Puppet Interview (BPI). The interview measured each child's relationship with each parent along two dimensions: (1) warmth/enjoyment, and (2) anger/hostility. Higher scores indicate more anger/hostility and warmth/enjoyment respectively. Each parent was then interviewed about their relationship with each of their children using The Parent-child Relationship Scale. This resulted in scores for parent-child relationship positivity and parent-child relationship negativity. Overall, these measures result in a lot of variables:

	Mum		*Dad*	
Measures	Child 1	Child 2	Child 1	Child 2
Warmth/Enjoyment	**mum_warmth_child1**	**mum_warmth_child2**	**dad_warmth_child1**	**dad_warmth_child2**
Anger/Hostility	**mum_anger_child1**	**mum_anger_child2**	**dad_anger_child1**	**dad_anger_child2**
Positive Relationship	**mum_pos_child1**	**mum_pos_child2**	**dad_pos_child1**	**dad_pos_child2**
Negative Relationship	**mum_neg_child1**	**mum_neg_child2**	**dad_neg_child1**	**dad_neg_child2**

Household chaos (**chaos**) was assessed using the Confusion, Hubbub, And Order Scale (CHAOS). There were two outcome variables (one for each child) that measured children's adjustment (**sdq_child1** and **sdq_child2**) using the Strengths and Difficulties Questionnaire: the higher the score, the more problem behaviour the child is reported to be displaying.

The data are in the file **CHAOS.sav** on the course website. To test whether household chaos was predictive of children's problem behaviour over and above parenting, conduct four hierarchical regressions:

(1) Maternal relationship with child 1
(2) Maternal relationship with child 2
(3) Paternal relationship with child 1
(4) Paternal relationship with child 2

Each hierarchical regression consists of three steps. First, enter child age and child gender as control variables. In the second step add the variables measuring parent-child positivity, parent-child negativity, parent-child warmth, parent-child anger. Finally, in the third step, chaos should be added. The crucial test of the hypothesis lies in the final step. To confirm that household chaos is predictive of children's problem behaviour over and above parenting, this third step must result in a significant R^2 change.

What conclusions can you draw from these analyses?

Your Answers:

Look at Coldwell, J., Pike, A. & Dunn, J. (2006). Household chaos - links with parenting and child behaviour. *Journal of Child Psychology and* Psychiatry, *47*, 1116-1122. (On the course website). How do your results and interpretation compare to those reported? Reflect upon how you have used regression as a tool to answer an important psychological question.

Your Answers:

Fill in the values for the following APA format table of the results:

	Mother-child relationship				Father-child relationship			
	Older sibling SDQ		Younger sibling SDQ		Older sibling SDQ		Younger sibling SDQ	
	Total R^2 =		Total R^2 =		Total R^2 =		Total R^2 =	
	β	ΔR^2	β	ΔR^2	β	ΔR^2	β	ΔR^2
Step 1								
Child age								
Child gender								

Step 2

Child age

Child gender

Child rpt parent-child positivity

Child rpt parent-child negativity

Parent rpt parent-child positivity

Parent rpt parent-child negativity

Step 3

Child age

Child gender

Child rpt parent-child positivity

Child rpt parent-child negativity

Parent rpt parent-child positivity

Parent rpt parent-child negativity

CHAOS

---

* $p < .05$, ** $p < .01$, *** $p < .001$

Task 3

Complete the multiple choice questions for **Chapter 8** on the companion website to Field (2013): https://studysites.uk.sagepub.com/field4e/study/mcqs.htm. If you get any wrong, re-read this handout (or Field, 2013, Chapter 8) and do them again until you get them all correct.

Task 4

Go back to the output for last week's task (does listening to heavy metal predict suicide risk). Is the model valid (i.e. are all of the assumptions met?)?

References

Berry, W. D. (1993). Understanding regression assumptions. Sage university paper series on quantitative applications in the social sciences, 07–092. Newbury Park, CA: Sage.

Cook, R. D., & Weisberg, S. (1982). Residuals and influence in regression. New York: Chapman & Hall.

Durbin, J., & Watson, G. S. (1951). Testing for serial correlation in least squares regression, II. Biometrika, 30, 159-178.

Field, A. P. (2013). Discovering statistics using IBM SPSS Statistics: And sex and drugs and rock 'n' roll (4th ed.). London: Sage.

Terms of Use