Cramming Sam's Tips for Chapter 8: Regression

Simple regression

- Simple regression is a way of predicting values of one variable from another.
- We do this by fitting a statistical model to the data in the form of a straight line.
- This line is the line that best summarizes the pattern of the data.
- We have to assess how well the line fits the data using:
    - R^2, which tells us how much variance is explained by the model compared to how much variance there is to explain in the first place. It is the proportion of variance in the outcome variable that is shared by the predictor variable.
    - *F*, which tells us how much variability the model can explain relative to how much it can't explain (i.e., it's the ratio of how good the model is compared to how bad it is).
    - the *b*-value, which tells us the gradient of the regression line and the strength of the relationship between a predictor and the outcome variable. If it is significant (*Sig.* < .05 in the SPSS table) then the predictor variable significantly predicts the outcome variable.

Descriptive statistics

- Use the descriptive statistics to check the correlation matrix for multicollinearity – that is, predictors that correlate too highly with each other, r > .9.

The model summary

- The fit of the regression model can be assessed using the *Model Summary* and *ANOVA* tables from SPSS.
- Look for the R^2 to tell you the proportion of variance explained by the model.
- If you have done a hierarchical regression then assess the improvement of the model at each stage of the analysis by looking at the change in R^2 and whether this change is significant (look for values less than .05 in the column labelled *Sig F Change*).
- The ANOVA also tells us whether the model is a significant fit of the data overall (look for values less than .05 in the column labelled *Sig.)*.
- The assumption that errors are independent is likely to be met if the Durbin–Watson statistic is close to 2 (and between 1 and 3).

Model parameters

- The individual contribution of variables to the regression model can be found in the *Coefficients* table from SPSS. If you have done a hierarchical regression then look at the values for the final model.
- For each predictor variable, you can see if it has made a significant contribution to predicting the outcome by looking at the column labelled *Sig.* (values less than .05 are significant).
- The standardized beta values tell you the importance of each predictor (bigger absolute value = more important).
- The tolerance and VIF values will also come in handy later on, so make a note of them.

Multicollinearity

- To check for multicollinearity, use the VIF values from the table labelled *Coefficients* in the SPSS output.
- If these values are less than 10, then there probably isn't cause for concern.
- If you take the average of VIF values, and it is not substantially greater than 1, then there's also no cause for concern.

Residuals

You need to look for cases that might be influencing the regression model:

- Look at standardized residuals and check that no more than 5% of cases have absolute values above 2, and that no more than about 1% have absolute values above 2.5. Any case with a value above about 3 could be an outlier.
- Look in the data editor for the values of Cook's distance: any value above 1 indicates a case that might be influencing the model.
- Calculate the average leverage (the number of predictors plus 1, divided by the sample size) and then look for values greater than twice or three times this average value.
- For Mahalanobis distance, a crude check is to look for values above 25 in large samples (500) and values above 15 in smaller samples (100). However, Barnett and Lewis (1978) should be consulted for more detailed analysis.
- Look for absolute values of DFBeta greater than 1.
- Calculate the upper and lower limit of acceptable values for the covariance ratio, CVR. The upper limit is 1 plus three times the average leverage, while the lower limit is 1 minus three times the average leverage. Cases that have a CVR that falls outside these limits may be problematic.

Model assumptions

- Look at the graph of ZRESID* plotted against ZPRED*. If it looks like a random array of dots then this is good. If the dots seem to get more or less spread out over the graph (look like a funnel) then this is probably a violation of the assumption of homogeneity of variance. If the dots have a pattern to them (i.e., a curved shape) then this is probably a violation of the assumption of linearity. If the dots seem to have a pattern and are more spread out at some points on the plot than others then this probably reflects violations of both homogeneity of variance *and* linearity. Any of these scenarios puts the validity of your model into question. Repeat the above for all partial plots too.

- Look at histograms and P-P plots. If the histograms look like normal distributions (and the P-P plot looks like a diagonal line), then all is well. If the histogram looks non-normal and the P-P plot looks like a wiggly snake curving around a diagonal line then things are less good. Be warned, though: distributions can look very non-normal in small samples even when they are normal.