Cramming Sam's Tips for Chapter 19: Logistic regression

Issues in logistic regression

- In logistic regression, we assume the same things as ordinary regression.
- The linearity assumption is that each predictor has a linear relationship with the log of the outcome variable.
- If we created a table that combined all possible values of all variables then we should ideally have some data in every cell of this table. If we don't then we must watch out for big standard errors.
- If the outcome variable can be predicted perfectly from one predictor variable (or a combination of predictor variables) then we have *complete separation*. This problem creates large standard errors too.
- *Overdispersion* is where the variance is larger than expected from the model. This can be caused by violating the assumption of independence. This problem makes the standard errors too small.

Model fit

- Build your model systematically and choose the most parsimonious model as the final one.
- The overall fit of the model is shown by $-2LL$ and its associated chi-square statistic. If the significance of the chi-square statistic is less than .05, then the model is a significant fit of the data.
- Check the table labelled *Variables in the Equation* to see the regression parameters for any predictors you have in the model.
- For each variable in the model, look at the Wald statistic and its significance (which again should be below .05). More important, though, use the odds ratio, $Exp(B)$, for interpretation. If the value is greater than 1 then as the predictor increases, the odds of the outcome occurring increase. Conversely, a value less than 1 indicates that as the predictor increases, the odds of

the outcome occurring decrease. For the aforementioned interpretation to be reliable the confidence interval of *Exp*(*B*) should not cross 1.

Diagnostic statistics

- Look for cases that might be influencing the logistic 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.
- Look for absolute values of DFBeta greater than 1.