Chapter 9

9.1

A binary (having only two categories) dependent variable means that most regression methods are worthless. If there is a binary independent variable, dummy coding can be used, but this is not an option when it is the dependent variable that is binary. Using OLS regression in such a situation is also not possible, as it is based on the assumption of a linear relationship.

Luckily the use of logistic regression still allows prediction to be made about a binary dependent variable. This only counts if the dependent variable includes 2 categories, however. For three or more categories multinomial logistic regression may be used, but that will not be touched on here.

9.2

As we know, in regression information about the independent variable(s) is used to predict values of the dependent variable. This is always recapped in the regression equation. However, the regression equation of a logistic regression is a bit more complicated. In order to understand how this regression equation works it is first important to understand the following three concepts: probability, odds, and logit.

Let's start with probability. These are generated values between 0 and 1, that, as the name implies, predict the probability of that data point falling into either of the categories of the dependent variable. This allows prediction through regression to be possible, but it does constrict all meaningful values to the range of 0 to 1. So we move on to odds.

Odds can be defined as the ratio of the probability of the two categories of the dependent variable. It is calculated as follows:

$$Odds(Y=1) = \frac{P(Y=1)}{1-P(Y=1)}$$

When using odds the range of meaningful values is increased to contain all positive numbers. This means, however, that values under 0 are still meaningless. This is when we turn to logit.

Logit is obtained by taking the natural logarithm of odds, and creates a value of the dependent variable that below and above zero (to infinity if needed). The logit is then calculated as follows:

$$Logit(Y) = ln\frac{P(Y=1)}{1-P(Y=1)}$$

Interpretation of logit equations is actually very similar to the interpretation of OLS: For each one-unit change in the independent variable, the logistic regression coefficients show the related change in the dependent variable.

Here that change in the dependent variable takes the form of the predicted log odds of being in either category of the dependent variable. Interestingly, this measure change is not constant in a logistic regression model. Namely because the natural log linearizes the s-shaped curve of the model, a one-unit change has a larger effect on the change of probabilities when it is localized in the centre.

In contrast to OLS regression, standardizing coefficients is not usual practice in logistic regression models. This is the case because an interpretation of a standard deviation change may make sense when the dependent variable is continuous, but not when it is dichotomous. Besides this, such practice is not very compatible with the use of log odds.

As explained above, interpretation of the logistic equation is not too difficult, as it is quite similar to OLS, but log odds themselves are a more complicated metric to work with. For this reason, when trying to understand the specific relation between a predictor and the dependent variable, we convert the log odds back to odds. This is done by exponentiation, as can be seen in the following formula:

$$Odds(Y=1)=e^{logit(Y)}=e^{ln[Odds(Y=1)]}=e^{\alpha+\beta 1X1+\beta 2X2+\ldots+\beta mXm}=\left(e^{\alpha}\right)\left(e^{\beta 1X1}\right)\left(e^{\beta 2X2}\right)\ldots.\left(e^{\beta mXm}\right)$$

This exponentiation makes the equation multiplicative (rather than additive), which changes the interpretation of the coefficients. The value of the odds (and thus the outcome) will not change if the coefficient is 1. A coefficient greater than 1 will cause an increase in the odds, and accordingly a coefficient less than 1 decreases the odds. The greater the distance the value is from one, the larger the change in the odds will be.

Odds can also be changed back into probability, with the following formula:

$$P(Y=1)=\frac{Odds(Y=1)}{1+Odds(Y=1)}=\frac{e^{\alpha+\beta 1X1+\beta 2X2+\ldots+\beta mXm}}{1+e^{\alpha+\beta 1X1+\beta 2X2+\ldots+\beta mXm}}$$

When it comes to probability, the closer a value is to 1, the higher the probability is.

9.3

The next step is estimating the equation and determining whether the model fits. When it comes to logistic regression we use the Maximum likelihood (ML) estimation in order to do this. To estimate the values of the parameters (or: logistic coefficients) Ml is applied to the model and then estimates the odds of occurrence after these are transformed into the logit. Or more simply put: ML estimates the most likely parameters given the patterns of the sample data.

This ML estimation results in the log of the likelihood (LL) function. The LL reflects the likelihood of observing these sample statistics in light of the population parameters. In this way the LL shows how much the model doesn't yet explain with the current estimation of the parameters. Due to this the LL is used to indicate how well the model fits. Values of the LL are on a range of 0 and downwards (to infinity if necessary), where values closer to 0 indicate better model fit.

ML estimation usually starts with conservative estimates, and then continuously creates better parameters in order to create larger likelihoods. This process stops when the increases in likelihood are too small to hold any real value.

9.4

In logistic regression there are two tests of significance to be done. The first is testing the significance of the overall regression model, which evaluates how well the model fits and to which extent the predicted values represent the observed values accurately. This can be done in several ways:

The likelihood ratio test: A test based on the change in the LL function that occurs when changing from a small model to a larger model (including more predictors). This test can thus be used to analyse changes in model fit between different fitted models. This test is similar to the overall F test in OLS regression, testing the null hypothesis that all the regression coefficients are equal to 0. Calculations are done as follows:

$$\chi^2 = -2(LL_{model} - LL_{baseline})$$

The smaller model is always used as the baseline model. The larger the difference between the LL values of the baseline and the model, the better the model fits. This test does make the assumption of nested models, however, so all values of the baseline model must also be included in the fitted model.

The Hosmer-Lemeshow Goodness-of-fit test: For this test a Hosmer-Lemeshow (HL) statistic needs to be computed, which is done by dividing cases into 10 groups (deciles) based on predicted probabilities. A chi-square value is then computed based on the observed and expected frequencies. A non-statistically significant result from this test indicates an acceptable fit (because this means that the model is not statistically significant from the observed values).

This model is easy to use but has received critique for being conservative, likely to indicate fit when 5 groups or less are used, and offering little diagnostics in the case of poor model fit.

The pseudo-variance explained index: This is an index very similar to multiple R2 in OLS regression. These values are considered as pseudo-variance explained because the variance in a dichotomous outcome is of course different that those in a continuous outcome (such as in OLS). There are many ways to compute this, several of these ways are:

Cox & Snell (automatically computed by SPSS): Computed as the ratio of the likelihood values, to the power of 2/n.

Nagelkerke (automatically computed by SPSS): Same as Cox & Snell, but adjusted so that the maximum value is 1.

Hosmer & Lemeshow: Computed by the ratio of the model to the baseline of -2LL. the value ranges from 0 to 1, and provides an indication of how the fit of the model is improved by including more predictors in the fitted model.

$$R_L^2 = \frac{-2LL_{model}}{-2LL_{baseline}}$$

Harell: Same as Homser & Lemeshow, but adjusted for the number of parameters in the model.

$$R_{LA}^2 = \frac{(-2LL_{model}) - 2m}{-2LL_{baseline}}$$

Aldrich & Nelson: Gives a value that is equivalent to the squared contingency coefficient.

Traditional R2: Computed by correlating the observed values of the dichotomous dependent variable with the predicted values of the logistic regression model. This correlated value is then squared.

Predicting group membership: Here testing is done by evaluating the difference between predicted and observed group membership. With a cut value of 0.5, a predicted probability of 0.5 or more means the predicted value is given as 1, and a predicted probability under 0.5 means the predicted value is given as 0 (creating a dichotomy again). Then the predicted group membership and observed group membership are compared to see whether the predicted values are correctly classified. This comparison leads to a frequency and percentage of correctly classified cases. A perfect model gets a 100% score, while a model with a 50% score is worthless, as its predictions are no better than chance. As a formal test of classification accuracy Press's Q (a chi-square statistic) can be used:

$$Q = \frac{[N - (n\,K)]^2}{N(K-1)}$$

Here N is the total sample size; n is the number of correctly classified cases; and K is the number of groups.

Weaknesses of this test are that it is (1) sensitive to sample size, and (2) might overlook unacceptable classifications of one or more groups. It is therefore important to also evaluate the classification of each individual group. Matters to keep an eye out for:

- Sensitivity: The probability that a case is coded as 1 with regards to the dependent variable is coded correctly (aka the percentage of correct predictions of 1).
- Specificity: The probability that a case coded as 0 with regards to the dependent variable is coded correctly (aka the percentage of correct predictions of 0).
- False positive rate: The probability that a case coded as 0 with regards to the dependent variable is coded incorrectly (aka the percentage of cases with incorrect predictions of 1, while it was actually 0).
- False negative rate: The probability that a case coded as 1 with regards to the dependent variable is coded incorrectly (aka the percentage of cases with incorrect predictions of 0, while it was actually 1).
- Cross validation: A recommended practice in logistic regression, though it can only be used if the sample size is sufficient. Cross validation involves testing the model on two samples, a primary sample (about 75-80% of your original sample) and a holdout sample (about 20-25% of the original sample). If the differences in classification accuracy are 10% or less, then the utility of the logistic regression model has been proven.

The second significance test involves testing the significance of the logistic regression coefficients. SPSS here uses the Wald statistic (which follows chi-square distribution) as the test statistic for regression coefficients. With continuous independent variables/predictors this is calculated as follows (squaring the ratio of the regression coefficient and dividing it by its standard error):

$$W = \frac{\beta_k^2}{SE_{\hat{\imath}}} 2^k$$

A downside to this test are that due to rounding errors, large regression coefficients can create imprecision in the estimation of the standard errors. This leads in inaccuracies when testing the null hypothesis, and increased Type 2 errors (failing to reject the null hypothesis while it is false).

An alternative to the Wald test is the log likelihood (LL) test described earlier.

Another alternative is the Bayesian information criterion (BIC), proposed by Raferty, which represents the difference between the chi-square value and the natural log of the sample size, but which can also be used to test logistic regression coefficients. The formula is as follows:

$$BIC = \chi^2 - \ln n$$

The BIC needs to be positive in order to reject the null hypothesis.

After determining the statistical significance of individual predictors, it might also be worth it to asses which predictors are adding most to the model. Sadly SPSS has no standardized regression coefficients for logistic regression, but luckily they are easy to calculate. You just need to standardise the predictors before you generate the logistic regression model, and then run the model. The logistic regression coefficients can then be interpreted as standardized regression coefficients (just like in OLS).

Another option is to form a confidence interval (CI) around the logistic regression coefficient (bk). This CI formula is the same as in OLS regression:

$$CI\left(b_k\right) = b \pm t_{(n-m-1)} s_b$$

9.5

The assumptions of logistic regression are a bit more relaxed than those of OLS. There are, however, still four primary assumptions to be considered:

- Non-collinearity. This assumption is applicable to any regression model with multiple predictors. Multicollinearity can be detected by creating an OLS regression model in SPSS with the same variables as the logistic regression model, and requesting collinearity statistics. Tolerance statistics of less than 0.2 suggests that there is multicollinearity, and values less than 0.1 suggest serious multicollinearity. Any value above 10 indicates that there is a violation of the noncollinearity assumption. (See chapter 8 on more details about this assumption).
- Linearity. In logistic regression this assumption refers to linearity between the logit of the dependent variable and the continuous independent variables. This assumption can be tested in multiple ways, but the easiest is the Box-Tidwell transformation. This is done by creating a logistic regression model including all independent variables of interest, each coupled with an interaction term. This interaction term is created by multiplying the continuous independent variable and its natural log. Nonlinearity is suggested by statistically significant interaction terms. Violation of this assumption can lead to biased parameter estimates, and the expected changes of Y not being constant over X. Important: This assumption is only applicable for continuous predictors.
- Independence of errors. This assumption is applicable to logistic regression model in the same manner that it is applicable to OLS. Violation of this assumption can lead to underestimated standard errors, with various consequences. (See chapter 7 and 8 for more details).
- Values of X are fixed. This assumption is no different in logistic regression than in OLS. (See chapter 7 and 8 for more details).

Logistic regression also needs to adhere to the following conditions:

- Nonzero cell counts (only in the case of nominal independent variables). A zero cell count occurs when the outcome is constant for one or more categories of the nominal variable. As it means that entire groups of individuals have odds of 0 or 1, it leads to high standard errors. Several ways of removing zero cell counts are: (1) recoding the categories or (2) adding a constant. Zero cell counts may be retained if it does not generally impact the relationship between the predictors and the dependent variable, but its presence should remain recognized in this situation.
- Non-separation of data. When the dependent variable is predicted perfectly, complete separation may occur, which results in an inability to estimate the models. If the separation is less than complete this is called quasi-complete separation, which results in very large coefficients and standard errors. These conditions may arise when the number of variables is equal (or nearly equal) to the number of cases in the dataset.
- Lack of influential points. Just as in OLS outliers and influential cases are problematic in logistic regression. The same means, like residual analysis and other diagnostic tests, as are used in OLS can be applied here. (See chapter 7 an 8 for more details).
- Sufficient sample size. Logistic regression is best used with large samples. In order to accurately conduct tests of significance samples of size 100 or larger are needed here.

9.6

It is important to consider the statistic odds ratio (OR), which can be used as an effect size index (similar to R2). The OR is computed by calculating by exponentiation the logistic regression coefficient, ebk. An OR of 1 would indicate that there is no relationship between the dependent and the independent variable (the predictor). Thus, when testing the effect size we want to know whether OR is significantly different from 1. When the OR is larger than 1, the independent variable increases the odds in the occurrence of the dependent variable, when the OR is smaller than 1, the independent variable decreases the odds of occurrence.

SPSS output labels OR as “Exp(B)” under “Variables in the Equation”.

OR values can also be converted to Cohen’s d in the following manner:

$$d = \frac{ln\left(OR\right)}{1.81}$$

9.7

With logistic regression three types of model building can be used.

The first is simultaneous logistic regression, where all independent variables of interest are simultaneously included in the model. This method is usually used when it is not thought that some predictors would be more important than others. In this manner all independent variables can be evaluated as if they were the most recent addition to the model. This may, however, lead to the correlations between an independent and a dependent variable being seen as stronger than they actually are (because the independent variables strengthen each other). In SPSS this method is referred to as “Enter”.

The second is stepwise logistic regression, where the computer adds and removes variables in a stepwise manner, as discussed earlier.

The third is hierarchical regression, where the predictor variables are added/removed according to a sequence that is specified by the researcher beforehand. This sequence can be forward, backward, or stepwise. (See chapter 8 for more details).

9.8

View pages 437-456 for an extensive overview of how to handle logistic regression models in SPSS.

9.9

View pages 456-458 for a brief overview of using G*Power in SPSS to determine a priori and post hoc power in logistic regression models.

9.10

View pages 459-461 for a guide on how to talk about logistic regression models in reports, including explanations of the APA-style write-up.