APPLIED STATISTICS INSTRUCTION SHEET

BINARY REGRESSION IN SPSS AND PSPP

Instructions

Binary logistic regression uses a binary dependent variable with any combination of independent variables. The path in both SPSS and PSPP is Analyze> Regression>Binary Logistic.

Instructions for both SPSS and PSPP

- Toggle the dependent variable into the Dependent box on the upper right-hand side of the dialog box.
- Toggle the independent variables into the Independent(s) box on the right-hand side. One can have more than one independent variable in the model, making the model a multiple regression. The independent variables can be either numeric or categorical.
- Categorical variables: Categorical variables must be broken into a block of dummy variables for this analysis. Options available vary by software.
    - Both SPSS and PSPP will accept a block of dummy variables created outside the procedure. Each dummy variable is entered individually into the Independent(s) box with one dummy variable withheld as the reference category.
    - In SPSS, the Categorical button opens a dialog box that allows the specification of a variable as Categorical. While this option saves the work of creating a block of dummy variables, only the first or last category can be specified as the reference category. If the researcher wants a different reference category, it is necessary to use the first option.
- Under the Options button, request a classification table (SPSS only) and designated the cutoff point (both SPSS and PSPP). Unless a specific use of the classification requires otherwise, the cutoff point should be the probability of the occurrence of the event into order to classify respondents/observations by whether the event is high or low probability.

Key Statistics

The Coefficients tables contains parameters of the regression line and is where the results related to most hypothesis testing are found. The key statistics are as follows:

- The column B contains the parameter estimates for the y-intercept (Constant) and individual variables that are used to build the regression equation. The sign of B determines whether the direction is positive or negative.
- The effect is determined by the exponentiated B, also known as the odds ratio. The null hypothesis is the odds ratio is one (not zero). The ratio is determined by its distance from 1.00. An odds ratio of 1.78 is 1.78X higher than the reference category or 78% higher than the occurrence in reference category. An odds ratio of .78 is 22% less than the reference category.
- There is no consensus on the threshold for effect size for an odds ratio, though a common designation is as follows:
    - For odds ratios greater than one: 1.50=small, 3.50=medium, 9.00=strong.

Author: Richard Lee Rogers. Last Update: 30 March 2020. Recommended Citation: Rogers, R. L. (2020). Binary logistic regression in SPSS and PSPP. [Applied statistics instruction sheet.] *Statistics Penguin*, http://statisticpenguin.com.

    - For odds ratios less than one: .65=small, .30=medium, .10=strong.
- The Sig. column gives the p-value used in determining statistical significance. The null hypothesis is that B=0.

The goodness of fit:

- The closest approximation to the R^2 is the Nagelkerke R^2, which is one of the pseudo-R^2 numbers. Pseudo R^2 numbers measure improvements in the model over the model with only the constant, but they are not percentages of the variance explained. There is no effect size for this statistic. The statistic is regarded as very useful when comparing results across models, but as an absolute measure of fit it is more dubious.
- For the classification table, the sensitivity and specificity are commonly reported numbers. There is no set rule for interpreting these numbers, but in general the higher the better. In behavioral research, the classification table often does not make sense if the cut point is set to zero when the probability of the occurrence of the dependent variable is much different.

Written Interpretation

Written comments should highlight the direction, effect size, and statistical significance of the individual coefficients, which are usually discussed in terms of the odds ratio and not the B. The goodness of fit numbers should be provided on the table for the benefit of the reader, but do not necessarily needed discussion because of their vagueness. Avoid interpreting them like OLS Regression R^2 numbers or attributing an effect size.