Selected Quantitative Methods

Lecture 4

Multinomial Logistic Regression

Harry B.G. Ganzeboom

VU-FSW Master Social Research

September 14-16 2011

MLR

- Multinomial Logistic Regression (MLR) is an extension of Binomial Logistic Regression (BLR) and shares most of its principles.
- Als BLR as a special case of MLR, any MLR program can estimate also BLR.
- In SPSS it is quite useful to use NOMREG or PLUM to do BLR.

NOMREG

- NOMREG is the SPSS program for MLR.
- By restricting the dependent variable to two categories, you can also use it for BLR.
- Like in UNIANOVA, NOMREG distinguished nominal independent variables (BY factors) and scaled independent variables (WITH covariates); it will generate dummy variable coding for factors, and interaction terms do not have to be constructed.
- However, you will often become happier by using your self-defined dummy variables and interaction terms. Use these as covariates.

Reference category in Y

- MLR have a reference category in Y. Every effect in the model is compared to this baseline.
- NOMREG allows you to choose either the first or the last category.
- As always, it is wise to choose a large category that “is easy to talk about”.
- Choice of the reference category does not change the model intrinsically (same expected frequencies), but it changes all the parameter estimates. This can create much confusion.

Reference categories in X

- I have not found a way in NOMREG to designate the reference category in factors.
- Of course you can make your own dummy variables and choose your own reference category by omitting the dummy variable of your preference.

Mixing covariates and factors

- One of the very convenient features of NOMREG (and UNIANOVA and GENLIN) is the ease of changing from categorical to scaled, vice versa.
- You can even specify the same variable as a factor and as a covariate, using copies.
- My favorite modeling strategy is to start at a saturated model, with all categorical variables, and then simplify the model by restricting factors to follow scaled models (make is a covariate).

Expected probabilities

- MLR can produce a table of expected (and observed) counts / probabilities for the cross-tabulation implied by the factors.
- Comparison of observed and expected counts is the basis ingredient of the “Goodness of Fit” test. If it is not significant, the expected counts resemble the observed counts with sampling variation.

Fit statistics

- Goodness-of-fit: this looks at the difference between observed and expected probabilities. This should be not significant (P>.05).
- Pseudo R2: there are three varieties, of which Cox-Snell and Nagelkerke are often reported (Stata reports McFadden).

Interpretation of effects

- To interpret effects it is most useful to think about MLR as a series of BLR, each with the same reference categories.
- If you used NOMREG to produce BLR, you get the same results as in LOGISTIC (provided you have the same data and model specifications).
- There and usually slight differences between coefficients / expected probabilities in the two methods. Notice that the SE’s in the simultaneous model are systematically smaller than in the seperate equations model.

Testing of effects

- NOMREG produces a Likelihood Ratio Test that will give an over-all test of the contribution of factors and covariates to the variations in counts / probabilities.
- These should be significant (P<=.05).

MLR in an expanded file

- Once you realize that MLR is nothing else but a bunch of BLR put together, it becomes interesting to estimate the model in an expanded file.
- In an expanded file we stack the separate BLR on top of one another. We need to add a variable that distinguishes the separate equations.
- Using BLR, we can now estimate the MLR model.
- The virtue of this procedure is that it gives us more flexibility in dropping parts of the predictors from the model.
- Moreover, it also gets us used to the ideas how ordered logistics regression and conditional logistic regression work.

Organizing the data

- O’Connell’s Table 4.1 is rather verbose, but worth studying to find out how different models are specified.
- You can do all of this by stacking equations and estimate the appropriate model with logistic.
- In many models with stacked equations you need to correct for duplication of data. In SPSS this can be done by specifying the person ID in GENLIN (Generalized Estimating Equations) or by specifying a clustered sampling plan in Complex Samples. (Stata: cluster correction).

Ordinal logistic regression

- OLR is a simple, but powerful idea. In ordinal data, you can define K-1 splits.
- Expand the file to K-1 parts, each covering one of the splits.
- The dependent variable is binomial (0,1) and denotes whether subject has passed certain split.
- Run a single logistic regression on the 0,1 dependent variable, but splits (nominal) as a control.

Parallel lines assumption

- In standard OLR we assume that the effect of all the X-variables is the same on all splits.
- The submodels are different by intercept (‘Threshholds’) but not by slopes (‘Location’).
- This assumption is referred as ‘parallel lines’ or ‘proportional odds’.
- In a expanded data-file it is easy to think about a test: specify and test a split*X interaction.
- Interesting models can sometimes have partly (multiple) non-proportional odds, and partly a single proportional odds effect.