# Ordinal regression

Harry Ganzeboom

RESMA course Data Analysis and Report #7

March 19 2009

# Binomial logistic regression

- Three SPSS program will produce the same model:
  - LOGIST Binomial logistic regresssion
  - NOMREG Multinomial logistic regression
  - PLUM Ordered logistic regression
- All these programs do:
  - Automatic dummy variable coding
  - Automatic interaction terms
- However, they vary in the choice of reference categories for both X and Y variables. It is easy to loose your way here.

# LOGISTIC

- X-variable reference: lowest, cannot be changed.
- Y-variable reference: lowest, cannot be changes (except by recoding)

# NOMREG

- X-variables reference: highest, cannot be changed
- Y-variable reference: lowest (first), can be changed.

# PLUM

- X-variables reference: Highest, cannot be changed
- Y-variable reference: lowest.
- “Location”: regression coefficient.
- “Threshhold”: intercept (= log odds for reference categories).

# Multinomial: more than two categories

- When we have more than two categories, an obvious way to go is multinomial regression (NOMREG). Conceptually, this is the same as doing a series of binomial logistics regression on the categories organized as a set of dummy variables with a fixed reference category.
- NOMREG adds to this the constraint that the expected probabilities should sum to 100%. Separate binomials will show slight deviations.

# Multinomial: pitfalls

- Multinomial coefficients are hard to interpret, because there are so many.
- The choice of the reference category can be problem: in particular when it is small and/or not at an ‘extreme’.
- It helps when you (implicitly) order the dependent variable. Note that the model does not use this rank-order information.
- The multinomial model uses many degrees of freedom and tends to have low statistical power (nothing is significant). Use it only with lots of data.

# Cumulative odds / logits

- In the most commonly used ordered regression model, the dependent variable is sliced in binomial parts at each category.
- Conceptually, the model is:
  - Separate binomial regression for each of the slices
  - In which all coefficients are constrained to be equal.
- The model has multiple intercepts (thresholds), but the same sort of regression coefficients as OLS. The can be interpreted as log odd of making a next step in the underlying process.

# Cumulative odds: remarks

- The cumulative odds model is very powerful and easy to use as an alternative to OLS.
- The model is fit for processes in which the choice between categories is not sequentially (but simultaneously made).
- The power of the model derived from the assumption of ‘parallel regressions’. This can be tested, and often fails to pass.

# Sequential logits

- The sequential logit model organizes the categories of the dependent variables in sequential binary transitions (see figure).
- This expresses developmental processes in which decisions are taken sequentially:
  - Referral processes
  - Careers
  - Education: the ‘Mare’ model.
- You then estimate a series of logistic regression models.

# Sequential logits: problems

- Of course, you need to know about rank-order, before you can use sequences. Some processes (such as education) or only partially ordered.
- Sequential logits organizes the data in segments with different amounts of variation in the independent variables.
- There is usually very little association in the last transistion. This is at the same time the true state of the world and a problem. We would want to know causal effects, if selectivity would be controlled.
- Sequential logits tend to concentrate power in the early stages.

# Adjacent odds

- An alternative to cumulative odds is modelling binary choice between adjacent odds.
- Behaviorally, this has the same assumption: the choice are not sequentially made.
- We can constrain al binomial regression to be equal (=uniform association),
- But can also estimate distances between categories that will scale the binamials (=scaled uniform association).
- These models cannot be estimated in SPSS, for continuous data. However, LOGLIN can be employed to estimate this models for categorized data.