Multinomial/Polytomous Logistic Regression

Todd BenDor

You will be using the STATA data file ("travel2.dta") available in the Lab #3 (http://todd.bendor.org/wuhan).

Here, Greene & Hensher (1997) collected data on choice of different travel mode (n=456; Variable: **Mode**). We wish to predict which mode of transportation is chosen based on a number of independent variables.

1. Total time
2. In-vehicle time
3. Generalized cost
4. In-vehicle cost
5. Household income
6. Size of traveling party

Which modes are available? What are the independent variables? Hint: use the **`describe`** command to look further into the dataset. How many observations are there? How many variables are there?

Here, you will perform several separate regression analyses while attempting to form a well-fitted, statistically significant model relating a number of determinants of travel modes to the independent variables.

1) Plot the independent variables vs. the dependent variable. You can do this by going to "Graphics -> Two Way Graphs" (you can also run a "Scatterplot Matrix") or by typing "twoway (scatter yvar xvar)" in the STATA command window. What do you see? Could you draw a straight line through this relationship as we did in linear regression?

2) Try running a **bivariate linear regression** model with each independent variable.
   Theoretical question: What assumptions does this violate regarding linear regression?

3) Now, try running a binary logistic regression model for each of the possible outcome sets. Create new variables by recoding two chosen travel modes into 0 and 1 and the third mode into a missing data (denoted by a period [**.**]) There will be 3 outcome sets for this model.

4) What variables are significant in bivariate regression? What do we lose by running multiple binary regressions?

5) Determine the regression equations used to estimate your data. What are the y-hats? How many intercepts are produced by each bivariate model? How many coefficients? How many equations (models) are generated by doing this (in each binary regression and in all of them)?

   Theoretical question: What are you estimating in these binary regressions? Why are you estimating this instead of the probability or odds directly? How does the fact that you are artificially creating binary data affect your findings in each of these models?

Now, try running a series of **multinomial logistic regressions** on your data, playing with adding and removing independent variables (an informal forward and backward stepwise regression process; can you figure out a way to automate this? See the handouts/examples on Sakai):

1) Compute and interpret variable coefficients in terms of their type of relationship to the relative odds/risk of someone using different modes of transportation vs. other modes.
2) Do these relationships make real-world sense? What variables are significant?

3) Why can't you talk about this in terms of odds ratios?

Theoretical question: What is the probabilistic relationship between more than two outcomes? Is this an odds ratio, or are you estimating something else?

4) Interpret the pseudo-R^2 value. Are there other goodness-of-fit measures we could use? Try looking these up in the STATA help menu and try installing plugins to help you do this (i.e. see the STATA overview examples and homework).

5) Graph the data and the predicted probabilities of the modes of transportation based on each independent variable (use the regression equation and given values of the independent variable to predict Y values). **Hint**: Type "help predict" to get more info on how to do this! How many new variables do you need to create?

- **Try running the regression with all of the independent variables, except for the ID variable. What happens? Why?**

  !SPOILER ALERT!: The regression goes bonkers because we have independent variables that perfectly predict the dependent variables (dummies for car, train, and bus). There *may* also be multicollinearirty in the model. **TO GET OUT OF THIS SITUATION, JUST PRESS THE BIG RED X BUTTON "BREAK"**

- Does there appear to be multicollinearity in the independent variables above? What variables are collinear with other variables (multicollinearity)?

- Try running the model with only the following independent variables: **invc choice hinc psize time.** How would you interpret this model, including all of the elements listed above (elements 1-5 on the last page).

- Use the likelihood ratio test to test the significance of your final model versus the model without the least significant variable (i.e with the highest p-value - ex: p=0.04 is higher than p=0.02). To explore this, first save the your base model by typing **`estimates store A`** (this saves the model as 'A' for future reference). Add other variables and re-run your model. Save again using **`estimates store B`**.

- Try testing these additions using the likelihood ratio test (**Statistics > Post-estimation > Tests > Likelihood ratio test**). Here, you can compare models that you've saved (or recently run) using the likelihood ratio chi-squared test. This test is a better test than either the Wald's z or χ^2 for determining significance of individual variables.

- You can also do this by hand. How would you do this? How many degrees of freedom would you use for this test?

- How many different relative risk ratios are contained in this model? Hint: use **`listcoef`** command - you can find this by searching for the test13_ado package (**`findit spost9_ado`**). This command displays all of the comparisons between all model outcomes, where coefficients determine the effect on the relative odds of one outcome vs. another outcome.

  Why does the listcoef function give you a different number of 'models' than the the mlogit output? Try looking through the STATA **`listcoef`** help function to find out more options to understand your models.