Practical For Session 8: Categorical Outcomes

29/11/2022

1 Practical For Session 8: Categorical Outcomes

Datasets

The datasets that you will use in this practical can be accessed via http from within stata. However, the directory in which they are residing has a very long name, so you can save yourself some typing if you create a global macro for this directory. You can do this by entering

```
global basedir http://personalpages.manchester.ac.uk/staff/mark.lunt
global datadir $basedir/stats/8_categorical/data
```

(In theory, the global variable `datadir` could have been set with a single command, but fitting the necessary command on the page would have been tricky. Far easier to use two separate commands as shown above). If you wish to run the practical on a computer without internet access, you would need to:

1. Obtain copies of the necessary datasets

2. Place them in a directory on your computer

3. Define the global macro `$datadir` to point to this directory.

1.1 Binomial & Multinomial Logistic Regression

The data used for this section was collected as part of a survey of alligator food choices in 4 lakes in Florida. The largest contributor to the volume of the stomach contents was used as the outcome variable `food`, and the charactertics of the alligators are their length (dichotomised as ≤ 2.3m and > 2.3m), their gender and which of the four lakes they were caught in.

1.1   Load the alligators data into stata with the command `use $datadir/alligators`, and familiarise yourself with the values used for each of the variables and their meanings with the command `label list`

1.2   Create a new variable `invertebrate` which takes the value 0 if the main food was fish, 1 if the main food was invertebrates and missing if the main food was anything else. This can be done with the command `gen invertebrate = food - 1 if food < 3`

1.3   Produce a cross-tabulation of food against length, with the command

        `tabulate invertebrate size, co`

        You should see that whilst fish and invertebrates are equally common in the smaller alligators, the larger ones are more likely to eat fish than invertebrates.

1 Practical For Session 8: Categorical Outcomes

1.4   Obtain an odds ratio for the effect of size on the probability that the main food is either fish or invertebrates with

```
logistic invertebrate size
```

  Is size a significant predictor of food choice ?

1.5   Now create another outcome variable which compares the probability that the main food is reptiles to the probability that the main food is fish with

```
gen reptile = (food == 3) if (food == 1) | (food == 3)
```

1.6   Obtain an odds ratio for the effect of size on the probability that the main food is either fish or reptiles with

```
logistic reptile size
```

  Is size a significant predictor of this food choice ?

1.7   Now use `mlogit food size, rrr` to get the odds ratios for the effect of size on all food choices. Which food category is the comparison group ?

1.8   Check that the odds ratios for the invertebrate vs. fish and reptile vs. fish comparisons are the same as before.

1.9   Are larger alligators more likely to choose reptiles rather than invertebrates ? You can test this with

```
lincom [Reptile]size - [Invertebrate]size, eform
```

     What is the odds ratio for size in this food choice ?

1.10   Generate a new variable to enable you to check this result using a single logistic regression model (`gen rep_inv = food == 3 if food == 3 | food == 2`). Perform the logistic regression with

```
logistic rep_inv size
```

  Are the results the same as you got with `lincom` ?

1.11   Now we are going to look at the influence of the lakes on the food choices. Produce a table of main food choice against lake with

```
tabulate food lake, co chi2
```

     Does the primary food differ between the 4 lakes ?

1.12   What proportion of alligators from Lake Hancock had invertebrates as their main food choice ?

1.13   How does this proportion compare to the other three lakes ?

4

1.14   Now fit a multinomial logistic regression model with

```
mlogit food i.lake, rrr
```

Look at the LR χ^2 statistic at the top: does this suggest that the primary food differs between the lakes ?

1.15   What is the odds ratio for preferring invertebrates to fish in lake Oklawaha compared to Lake Hancock ? Does this agree with what you saw in the table ?

1.16   Confirm your answer to the previous question by using the command `logistic invertebrate i.lake`

1.2   Using `mlogit`

This section uses the dataset `$datadir/politics`, which contains information on the effect of gender and race on political party identification.

2.1   Use `label list` to find out the meanings of the variables

2.2   Use `mlogit party race, rrr` to determine the effect of race on party affiliation. How does being black affect the odds of being a republican rather than a democrat ?

2.3   How does being black affect the odds of being an independent rather than a democrat ?

2.4   Use `tabulate party race, co` to confirm that your answers to the previous questions are sensible.

2.5   What is the odds ratio for being a republican rather than a democrat for women compared to men (use `mlogit party gender, rrr` to find out).

2.6   Fit a multinomial model in which party identification is predicted from both race and gender (`mlogit party race gender, rrr`).

2.7   Add the interaction between race and gender, to see if the race influence differs between men and women. Is this difference statistically significant ?

1.3   Ordinal Models

This section uses the data in `$datadir/housing`. This data concerns levels of satisfaction among tenants of different types of housing, according how much contact they have with other residents and how much influence they feel they have over the management of their housing.

3.1   Use `label list` to find out the meanings of the variables.

1 Practical For Session 8: Categorical Outcomes

3.2   Does the degree of satisfaction depend on which type of housing the tenant lives in ? (Use `ologit satisfaction i.housing` to find out).

3.3   Of which type of housing are the tenants most satisfied ?

3.4   Test whether `influence` and `contact` are significant predictors of satisfaction

3.5   Create a multivariate model for predicting satisfaction from all of the variables that were significant univariately. Are these predictors all independently significant ? (You may need to use `testparm` for categorical predictors).

3.6   Does the effect of influence depend on which type of housing a subject lives in ? (Fit an interaction term and use `testparm` to test its significance).