

Practical 9: Modelling Counts

Mark Lunt

22/11/2011

1 Practical For Session 9: Counts

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/9_Counts/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 Poisson Regression

In this section you will be analysing the dataset `$datadir/ships`. This is data from Lloyds of London concerning the rate at which damage occurred at different times to different types of ship. There are 5 types of ship (labelled “A” to “E”), which could have been built in any one of 4 time periods, and sailed during one of two time periods. The aggregate duration of operation of each type of ship is given by `months`, and the number of incidents of damage is given by `damage`.

- 1.1 Familiarise yourself with the the meanings of each of the variables with the command

```
label list
```

Set the reference categories for `type` and `time built` to E and 1975-1979 respectively with the commands

```
char type[omit] 5
char built[omit] 4
```

- 1.2 Are there any differences in the rates at which damage occurs according to the type of ship ? The command to test this is

```
xi: poisson damage i.type, exposure(months) irr
```

1 Practical For Session 9: Counts

- 1.3 Are there any differences in the rates at which damage occurs according to the time at which the ship was built ? The command to test this is

```
xi: poisson damage i.built, exposure(months) irr
```

- 1.4 Are there any differences in the rates at which damage occurs according to the time in which the ship was operated ? (You can work out this command for yourself).

- 1.5 Now add all three variables into a multivariate poisson model. Use

```
testparm _Itype*
```

to test if type is still significant after adjusting for the other predictors.

- 1.6 Use

```
predict pred_n
```

to obtain predicted numbers of damage incidents. Compare the observed and predicted numbers of incidents with

```
list type built sailed damage pred_n
```

For which type of ship and which time periods are predicted values furthest from the observed values ?

- 1.7 Use `estat gof` to test whether the model is adequate.

- 1.8 Add a term for the interaction between ship type and year of operation (`i.type*i.built`). Use `testparm` to determine whether this term is statistically significant.

- 1.9 Does this term affect the adequacy of the model as determined by `estat gof` ?

1.2 Negative Binomial Regression

This section used data concerning childhood mortality in three cohorts, from the dataset `$datadir/nbreg`. The children were divided into 7 age-bands, and the number of deaths, and the persons-months of exposure are recorded in `deaths` and `exposure` respectively.

- 2.1 Fit a poisson regression model using only cohort as a predictor:

```
xi: poisson deaths i.cohort, exposure(exposure) irr
```

Are there differences in mortality rate between the cohorts ?

- 2.2 Use `estat gof` to test whether the poisson model was appropriate

- 2.3 Fit a negative binomial regression model to test the same hypothesis:
`xi: nbreg deaths i.cohort, exposure(exposure) irr`
 Do you reach the same conclusion about the role of cohort ?
- 2.4 What is the value of the parameter α , and its 95% confidence interval ?
- 2.5 Fit a constant dispersion negative binomial regression model with
`xi: nbreg deaths i.cohort, exposure(exposure) dispersion(constant) irr`
 Is δ significantly greater than 0 in this model ?
- 2.6 Does this model suggest any different conclusions as to whether the mortality rate differs between cohorts ?
- 2.7 On possible source of the extra variation is a change in mortality with age. Fit a model to test whether mortality varies with age with
`xi: nbreg deaths i.age_mos, exposure(exposure) irr`
 Is age a significant predictor of mortality ?
- 2.8 Would it be appropriate to use Poisson regression to fit this model ?
- 2.9 Now fit a negative binomial regression model with both age and cohort as predictors. Use `testparm` to determine whether both age and cohort are independently significant predictors of mortality.
- 2.10 Is α significantly greater than 0 in this model ?
- 2.11 Fit the same model using `poisson`. Does this model agree with the negative binomial model ?
- 2.12 Use `estat gof` to test the adequacy of this model. Is using a Poisson regression model appropriate in this case ?

1.3 Using constraints

This section uses the data on damage to ships from the dataset `$datadir/ships` again.

- 3.1 Refit the final Poisson regression model we considered with
`xi: poisson damage i.type i.built i.sailed, irr exposure(months)`
 Which of the incidence rate ratios are not significantly different from 1 ?
- 3.2 Create predicted numbers of damage incidents with the command
`predict pred_n`

1 Practical For Session 9: Counts

- 3.3 Define a constraint to force the incidence rate ratio for ships of type D to be equal to 1 with

```
constraint define 1 _Itype_4 = 0
```

(Note that the constraints are defined on the *coefficients* of the model, rather than the incidence rate ratios. If the coefficient is 0, the incidence rate ratio is 1.)

- 3.4 Fit this model with the command

```
xi: poisson damage i.type i.built i.sailed, irr exposure(months)
constr(1)
```

How does the output of this command differ from that of the previous Poisson regression command ?

- 3.5 Use `estat gof` to test the adequacy of this model. How does the constrained model compare to the unconstrained model ?

- 3.6 Define a second constraint to force the incidence rate ratio for ships of type E to be equal to 1 with

```
constraint define 2 _Itype_5 = 0
```

- 3.7 Fit a Poisson regression model with both of these constraints using the command

```
xi: poisson damage i.type i.built i.sailed, irr exposure(months)
constr(1 2)
```

(The above command should be entered on one line.)

- 3.8 How does the adequacy of this model compare to that of the previous one ?

- 3.9 It appears that the incidence rate ratio for being built in 1965-1969 is very similar to the incidence rate ratio for being built in 1970-1974. Define a new constraint to force these parameters to be equal with

```
constraint define 3 _Ibuilt_2 = _Ibuilt_3
```

Fit a Poisson regression model with all three constraints using the command

```
xi: poisson damage i.type i.built i.sailed, irr exposure(months)
constr(1 2 3)
```

(The above command should be entered on one line.) Notice that the lines for `_Ibuilt_2` and `_Ibuilt_3` are now identical. In what way do these two lines differ from the lines for the other constrained values ?

- 3.10 What do you think is the reason for the difference you have just observed ?

- 3.11 Use `estat gof` to test the adequacy of this constrained model. Have the constraints that you have applied to the model had a serious detrimental effect on the fit of the model.
- 3.12 Obtain predicted counts from this constrained model with the command
- ```
predict pred_cn
```
- 3.13 Compare the predictions from the constrained model and the unconstrained model to each other and to the observed values with
- ```
corr damage pred_n pred_cn
```
- How has the fit of the model been affected by the constraints ?
- 3.14 If you wish, you can examine the observed and predicted values directly with
- ```
list type built sailed damage pred_n pred_cn
```
- Does this list confirm your answer to the previous question ?

## 1.4 Constraints in Multinomial Logistic Regression

Constraints can be applied to many different types of regression model. However, applying constraints when using `mlogit` can be tricky because there are several equations. The syntax is then similar to the syntax we saw last week for `lincom`. For this part of the practical, we are using the same `$datadir/alligators` dataset that we saw last week.

- 4.1 Use
- ```
label list
```
- to remind yourself of what the variables mean.
- 4.2 Fit a multinomial logistic regression model to predict food choice from lake with the command
- ```
xi: mlogit food i.lake, rrr
```
- Are there significant differences between lakes in the primary food choice ?
- 4.3 What are the odds ratios for preferring invertebrates to fish in Lakes Oklawaha, Trafford and George ?
- 4.4 It appears that for the choice of invertebrates rather than fish, there is no significant difference between Lake Oklawaha and Lake Trafford. Define the constraint that corresponds to this with
- ```
constraint define 1 [Invertebrate]_l1lake_2 = [Invertebrate]_l1lake_3
```
- Fit the model again with this constraint using
- ```
xi: mlogit food i.lake, rrr const(1)
```

1 *Practical For Session 9: Counts*

- 4.5 Even Lake George does not appear to be significantly different from Lake Oklawaha and Lake Trafford. Define a new constraint with

```
constraint define 2 [Invertebrate]_llake_4 = [Invertebrate]_llake_3
```

Fit a multinomial logistic regression model with both of these constraints with

```
xi: mlogit food i.lake, rrr const(1 2)
```

How does the common odds ratio for all three lakes compare to the 3 separate odds ratios you calculated previously ?