UL SSRM Quant Stream: Unit 2 labs, Categorical Data Analysis

Brendan Halpin

May/June 2022

Contents

Note: available online at http://teaching.sociology.ul.ie/ssrm/unitb2/unitb2lab.html

1 Lab 1

1.1 Measures of association

1.1.1 Tables and Chi-squared

This is data showing the relation between class of origin and highest educational qualification for a UK sample in 2001:

```
Social Class of Origin and Whether has Degree
```

class	univ No	Yes	Total
Prof/Man	2,333	1,025	3,358
Routine non-manual	1,400	124	1,524
Skilled manual	947	31	978
Semi/unskilled	1,077	18	1,095
Total	5,757	1,198	6,955

Social Class of Origin and Qualification

class	qual			Total
	Univ	2nd level	Incomplet	
Prof/Man	1025	1566	767	3358
Routine non-manual	124	687	713	1524
Skilled manual	31	483	464	978
Semi/unskilled	18	361	716	1095
Total	1198	3097	2660	6955

Source: British Household Panel Survey 2001

Use the first panel to calculate the odds ratio (use a calculator or a spreadsheet) comparing prof/man versus semi/unskilled in their chances of having compared with not having a university education. Interpret it.

Do the same for routine-non-manual versus semi/unskilled and skilled versus semi/unskilled. Is there a pattern in the three ORs?

Enter the table in Stata:

```
input class qual count
1 1 1025
1 2 1566
1 3 767
2 1 124
2 2 687
2 3 713
3 1 31
3 2 483
3 3 464
4 1 18
4 2 361
4 3 716
end

label define cl  1 "Prof/Man" ///
 2 "Routine non-manual" ///
 3 "Skilled manual" ///
 4 "Semi/unskilled"
label define ql  1 "Univ"  2 "2nd level" 3 "Incomplete 2nd"
label values class cl
label values qual ql

// We can get the table using weights
tab class qual [freq=count]

// Or we can "expand" the data (watch out for n=0)
expand count

tab class qual
```

- Analyse the pattern of percentages (`tab class qual, row`)
- Compare the observed and expected values – use the Stata add-on, `tabchi class qual`
- Use `tabchi` with the `res` and `adj` options to generate the raw and adjusted residuals; examine these
- Run the χ^2 test and interpret, (`tab class qual [freq=count], chi`)
- Run and interpret the gamma test (`tab class qual [freq=count], gamma`)
- What does it tell you? Compare with the pattern of association shown in the percentages with the gamma,and consider which gives you the better summary.

1.2 Logistic Regression

1.2.1 Binary logistic regression: Shuttle disaster data

The following data represents measurements of o-ring failures in Space Shuttle launches (an o-ring failure caused one shuttle to explode shortly after launch, and the suspicion was that weather conditions were partly responsible for the failure. The data are air temperature (in degrees F) and whether or not there was o-ring damage.

T	y/n	T	y/n	T	y/n
66	0	70	0	5	7 1
72	0	76	0	5	3 1
70	1	69	0	7	6 0
75	0	70	0	6	7 0
75	1	67	0	6	3 1
70	1	81	0	6	7 0
73	0	58	1	7	9 0
78	0	68	0		

Source: Agresti & Finlay 3rd edition, q15.4

```
input temp failure
 66  0
 70  0
 57  1
 72  0
 76  0
 53  1
 70  1
 69  0
 76  0
 75  0
 70  0
 67  0
 75  1
 67  0
 63  1
 70  1
 81  0
 67  0
 73  0
 58  1
 79  0
 78  0
```

```
 68  0
end
```

Fit a binary logistic regression with temperature as the explanatory variable, and answer the following questions:

1. Report the regression equation, specifying what the effect of temperature is
2. Test the hypothesis that temperature has an effect on failure
3. Calculate the predicted probability of failure at 32°F, 60°F and 75°F
4. At what temperature is the predicted probability of failure 50 percent?

1.3 Odds ratios and parameter estimates

Use the command

```
do http://teaching.sociology.ul.ie/ws/dpind.do
```

to load a small data set that looks at race and sentences in capital murder cases, from the US. It breaks down whether the death penalty was handed down, by race of defendent and victim.

Calculate the overall odds ratio for white vs black defendents, and then fit a logistic regression with defendent's race as the explanatory variable. Calculate the odds ratio from the regression coefficient and compare with the one you already calculated.

Repeat the exercise for white victims only:

```
tab def pen if vic==1
logit pen def if vic==1
```

Note what happens when you do the same for black victims.

2 Lab 2: Logistic Regression continued

2.1 Classification table

With the credit-card data used in Agresti & Finlay, fit a logistic regression and use `estat class` to generate the classification table.

```
do http://teaching.sociology.ul.ie/ws/creditcard.do
logit card income
estat class
```

Compare the results with those for the null model (i.e., no explanatory variables).

2.2 Binary logistic continued

2.2.1 Getting to university

Data from the British Household Panel Survey covering educational qualifications and some potential predictor variables, are available with the following command:

```
use http://teaching.sociology.ul.ie/ws/data/bhpsqual, clear
```

Explore the data set and find a good model to predict having a university qualification. Consider using a non-linear formulation of age. Interpret the results.

Generate and interpret the Hosmer-Lemeshow statistic:

```
logit univ i.sex c.age##c.age // One possible model, can you find a better one?
estat gof, group(10)
```

Does this model "fit"?

3 Lab 3 Multinomial and Ordinal Logistic Regression

3.1 Multinomial and ordinal logistic regression

3.1.1 Multinomial logistic regression

Use the following command to load the BHPS excerpt:

```
use http://teaching.sociology.ul.ie/ws/data/bhpsqual
```

`vote` has four categories. Examine bivariate relationships between vote and some of the other variables, and then fit a multinomial logistic regession:

```
mlogit vote income i.sex, baseoutcome(1)
```

I suggest using `baseoutcome(1)` to make "Conservative" the base category, as the default would be the highest category, "Nationalist/Other", which is too mixed. Search for a good model, and interpret it.

Using the variables in your model, select two or three "typical" cases and generate predicted values for them.

Repeat the exercise using `qual` as the dependent variable.

3.1.2 Ordinal logistic regression

`qual` is an ordinal variable. In your previous analysis, did you observe patterns in the parameter estimates? Fit a proportional odds model with syntax such as the following:

```
ologit qual i.sex age
```

Compare the ordinal logistic results with the multinomial results you have already produced. Do they tell the same story?

For one or two "typical" persons, calculate the predicted probabilities of being "higher" rather than "lower" across the three contrasts.

3.1.3 Probit and logistic regression

Search for a reasonable logistic model predicting having a university qualification. Repeat the exercise using probit regression (replace `logit` with `probit` in the estimation command). What differences do you see? Do they matter?

Use Stata to generate predicted values from each model, and compare them.