

# Computer Workshop: Multinomial and ordinal logistic regression

The aims of this workshop are:

- Fit and interpret a multinomial regression model in Stata
- Calculate predicted probabilities

## Download data, open Stata, and set up the do file

- Download the Stata dataset **crime2013-14\_multicat.dta** to a suitable destination. Remember where you saved these files, as we will use this as our “Working Directory” for the rest of the workshop.
- Open Stata and a new do-file (we always recommend using a do-file so that you have a record of your code and can easily re-run the model).
- Set up the do-file by typing the following in the first few rows:

```
capture log close // closes any log files you may have open
```


- type the path to your working directory between the quotation marks, e.g.  

```
cd "C:\statistics\binarylogit"
```

```
log using "NCRM_multinomial_logit.log", text replace
```

```
use "crime2013-14_multicat.dta", clear
```



Finally, click on the  icon in the toolbar (or press CTRL+D) to execute all of the commands that you have typed into the do-file so far. Some output should then appear in the results window.

- Use *describe* to get a feel for the dataset.

In this workshop, we will study the association between a multi-category response variable and a set of predictors using multinomial regression. For doing so, we will continue to use the dataset extracted from the Crime Survey for England and Wales, 2013-2014<sup>1</sup>, but this time we will only use a subset of respondents (N=2181), who answered questions about how much they worry about crime. Our aim is to determine whether there is an association between worrying about having one’s home being broken into (*wburgl*, 1 "Not at all worried" 2 "Not very worried" 3 "Fairly worried" 4 "Very worried") and some socio-demographic characteristics of the respondent. The dataset

includes the following variables:

<sup>1</sup> Office for National Statistics, University of Manchester. Cathie Marsh Institute for Social Research (CMIST). UK Data Service. (2016). *Crime Survey for England and Wales, 2013-2014: Unrestricted Access Teaching Dataset*. [data collection]. UK Data Service. SN: 8011, <http://doi.org/10.5255/UKDA-SN-8011-1>

VARIABLE	DESCRIPTION
<b>caseid</b>	Case identifier (9 digits)
<b>sex</b>	Gender
<b>agegrp7</b>	Age grouped
<b>educat3</b>	Education
<b>wburgl</b>	How worried about having your home broken into?

## Descriptive statistics

First, we will start by displaying the frequencies of the variables of interest.

➤ `fre sex-wburgl`

NB! If Stata does not run the 'fre' command, try typing '`ssc install fre`' first.

Check the **results** window. Scroll down through this output carefully and note what Stata has produced. You will get a first insight of the distribution of each variable and the presence/absence of missing values by taking a look at the tables. One example is shown below. You can see, for instance, that 10.3% of the respondents report being very worried about burglary, whereas 15.2% are not at all worried about it and that there are no missing values for this variable.

`wburgl` — How worried about having your home broken into?

		Freq.	Percent	Valid	Cum.
Valid	1 Not at all worried	331	15.18	15.18	15.18
	2 Not very worried	1036	47.50	47.50	62.68
	3 Fairly worried	590	27.05	27.05	89.73
	4 Very worried	224	10.27	10.27	100.00
	Total	2181	100.00	100.00	

Now, let's study the relationship between the response variable and each one of the potential predictors (age, gender and education) by producing some cross tabulations and chi-square tests of independence for each of the three explanatory variables and the outcome separately. You can use the command below by replacing the text <variable> with the relevant variable name.

➤ `tab <variable> wburgl, chi row`

An example of the output is shown below. It tells you that women are more often than men worried or very worried about their houses being broken into. For instance, 11.7% of women are very worried compared to 8.6% of men. The association is statistically significant at 1% level ( $p=0.001$ ) according to the Chi-squared test.

```
. tab sex wburgl, chi row
```

Key
frequency row percentage

Gender	How worried about having your home broken into?				Total
	Not at all	Not very	Fairly wo	Very worr	
Male	172 17.41	489 49.49	242 24.49	85 8.60	988 100.00
Female	159 13.33	547 45.85	348 29.17	139 11.65	1,193 100.00
Total	331 15.18	1,036 47.50	590 27.05	224 10.27	2,181 100.00

Pearson chi2(3) = 16.6984 Pr = 0.001

Take a look at all the other tables you have produced as well to get familiar with the data and the associations between each explanatory variable and the outcome.

## Fitting a Multinomial Logistic Regression with a Single predictor variable

We are interested in modelling the probability of being worried about burglary. To start, we will consider as predictor the gender of the respondent. Use *ib2.sex* to set the reference level of the sex variable to the second level (Female), and the option *b(1)* to set the base level for the regression model as the first level (Not at all worried). The *rrr* option now gives us the relative risk ratios (i.e. odds ratios).

- `mlogit wburgl ib2.sex, b(1)`
- `mlogit wburgl ib2.sex, b(1) rrr`

```
. mlogit wburgl ib2.sex, b(1) rrr
```

```
Iteration 0:  log likelihood = -2676.465
Iteration 1:  log likelihood = -2668.1022
Iteration 2:  log likelihood = -2668.0864
Iteration 3:  log likelihood = -2668.0864
```

```
Multinomial logistic regression      Number of obs   =    2,181
                                      LR chi2(3)      =    16.76
                                      Prob > chi2     =    0.0008
Log likelihood = -2668.0864          Pseudo R2      =    0.0031
```

wburgl	RRR	Std. Err.	z	P> z	[95% Conf. Interval]	
<b>Not_at_all_worried</b> (base outcome)						
<b>Not_very_worried</b>						
sex						
Male	.8263998	.1044551	-1.51	0.131	.6450601	1.058718
_cons	3.440252	.3099562	13.71	0.000	2.883364	4.104695
<b>Fairly_worried</b>						
sex						
Male	.6428428	.0888637	-3.20	0.001	.4902736	.8428903
_cons	2.188679	.2095067	8.18	0.000	1.814273	2.640351
<b>Very_worried</b>						
sex						
Male	.565292	.0996297	-3.24	0.001	.4001778	.7985325
_cons	.8742138	.1015126	-1.16	0.247	.6962693	1.097635

The output from the *mlogit* command above shows you the number of observations, the likelihood ratio statistics, and the McFadden pseudo R-squared value, along with the relative risk ratios, confidence intervals, and p-values for the variables in the model. Remember that in the case of multinomial logistic regression, a different model is being formulated for each one of the J-1 categories of the response variable.

The following commands help with model selection:

- `qui mlogit wburgl ib2.sex, b(1)`
- `estimates store m1`
- `testparm ib2.sex // wald-test`
- `qui mlogit wburgl, b(1) // Run model without gender`
- `lrtest m1 // Test if gender is significant using Likelihood ratio test`

The output from the *testparm* command shows you the joint Wald-test results for the sex variable and *lrtest* the Likelihood-ratio test results.

For our response variable with four categories, three equations have been fitted:

$$\log\left(\frac{p_{not\ very}}{p_{Not\ at\ all}}\right) = 1.236 - 0.191 \times Man$$

$$\log\left(\frac{p_{fairly}}{p_{Not\ at\ all}}\right) = 0.783 - 0.442 \times Man$$

$$\log\left(\frac{p_{\text{very}}}{p_{\text{Not at all}}}\right) = -0.134 - 0.570 \times \text{Man}$$

Where  $p_{\text{Not at all}}$  is the probability of not being at all worried about burglary,  $p_{\text{not very}}$  is the probability of not being very worried, and so on. In analogous way,  $\log\left(\frac{p_{\text{not very}}}{p_{\text{Not at all}}}\right)$  is the log-odds of not being very worried the reference category, not at all worried.

In analogy with what happens with the binary logistic regression model, the  $\exp(B)$  corresponds to the estimated odds ratio of having a particular level of worry instead of not at all worried (reference category for the response), given that the respondent is a man rather than a woman (reference category for the covariate).

If the covariate were continuous instead of categorical,  $\exp(B)$  would represent how much the corresponding odds increase or reduce by unit of change in the covariate.

However, the interpretation of the model is not particularly intuitive in terms of the odds or odds ratio. From the equation, it is possible to estimate the predicted probabilities of observing each value of the response variable given the gender of the respondent. Using the three equations defined for this model we have, for men:

$$p_{\text{very}|\text{man}} = \frac{e^{B_{0|\text{very}}+B_{\text{man}|\text{very}}}}{1 + e^{B_{0|\text{very}}+B_{\text{man}|\text{very}}} + e^{B_{0|\text{fairly}}+B_{\text{man}|\text{fairly}}} + e^{B_{0|\text{not very}}+B_{\text{man}|\text{not very}}}}$$

The fitted probability of being very worried and a man is calculated as follows:

$$p_{\text{very}|\text{man}} = \frac{e^{-0.134-0.57}}{1 + e^{-0.134-0.57} + e^{0.783-0.442} + e^{1.236-0.191}} = 0.0860$$

Analogously,

$$p_{\text{fairly}|\text{man}} = 0.245, \quad p_{\text{not very}|\text{man}} = 0.495 \quad \text{and} \quad p_{\text{not at all}|\text{man}} = 0.174.$$

The estimated probabilities for women are calculated by excluding the term corresponding to gender in each equation (this because woman is the reference category). We have,

$$p_{\text{very}|\text{woman}} = \frac{e^{-0.134}}{1 + e^{-0.134} + e^{0.783} + e^{1.236}} = 0.117$$

Analogously,

$$p_{\text{fairly}|\text{woman}} = 0.292, \quad p_{\text{not very}|\text{woman}} = 0.459 \quad \text{and} \quad p_{\text{not at all}|\text{woman}} = 0.133.$$

You can ask Stata to calculate these probabilities for you by using the commands below. Convince yourself that these probabilities are the same as calculated above.

- `estimates restore m1 // to activate the model results`
- `margins sex // calculates probabilities by gender and outcome category`

## Multivariate Multinomial regression

We will add another predictor in addition to gender. We are interested in whether respondent's age is associated with the outcome variable:

- `mlogit wburg1 ib2.sex i.agegrp7, b(1)`
- `mlogit wburg1 ib2.sex i.agegrp7, b(1) rrr`
- `estimates store m2`
- `lrtest m1 m2`
- `testparm ib1.sex`
- `testparm i.agegrp7`

The output from the `lrtest` command gives you the likelihood ratio test of nested models, `testparm` the joint Wald-test results. The likelihood ratio test shows a small p-value of 0.01, suggesting that age should be included in the model.

We can use the `margins` command again to calculate predicted probabilities. If we wanted to investigate the probabilities of being worried by age group among women, we would type:

- `margins agegrp7, at(sex==2)`

The `marginsplot` command below gives you a graphical presentation of the results, which can make interpretation easier.

- `marginsplot, legend(order(1 "Not at all worried" 2 "Not very worried" 3 "Fairly worried" 4 "Very worried"))`

The table below the results from the above Stata output organised in a more reader friendly manner.

➤

Table: Predicted probabilities of how worried about burglary women are by age, multinomial logistic model.

	<b>How worried about burglary</b>			
<b>Age</b>	<i>Not at all</i>	<i>Not very</i>	<i>Fairly</i>	<i>Very</i>
<i>16-24</i>	0.174	0.494	0.260	0.072
<i>25-34</i>	0.116	0.471	0.292	0.121
<i>35-44</i>	0.106	0.486	0.283	0.125
<i>45-54</i>	0.106	0.454	0.327	0.113
<i>55-64</i>	0.114	0.436	0.315	0.136
<i>65-74</i>	0.146	0.436	0.281	0.137
<i>75+</i>	0.213	0.447	0.261	0.079

Those in the youngest and the oldest age groups are least worried about crime: their probability of being in the 'not at all worried' group is higher (17% for those aged 16-24 and 21% for those aged 75 or more) than the others' (most around 10-11%), and their probability of being in the 'very worried' group is lower (7% for those aged 16-24 and 8% for those aged 75 or more) than for others (between 11 and 14%). The differences between the other age groups were quite small.