Computer Workshop: Ordinal logistic regression

Dr Heini Väisänen

The aims of this workshop are:

- Fit and interpret a ordinal regression model in Stata
- Calculate predicted probabilities and compare these across ordinal and multinomial models

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[1], 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:

[1] 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
caseid	Case identifier (9 digits)
sex	Gender
agegrp7	Age grouped
educat3	Education
wburgl	How worried about having your home broken into?

Descriptive statistics

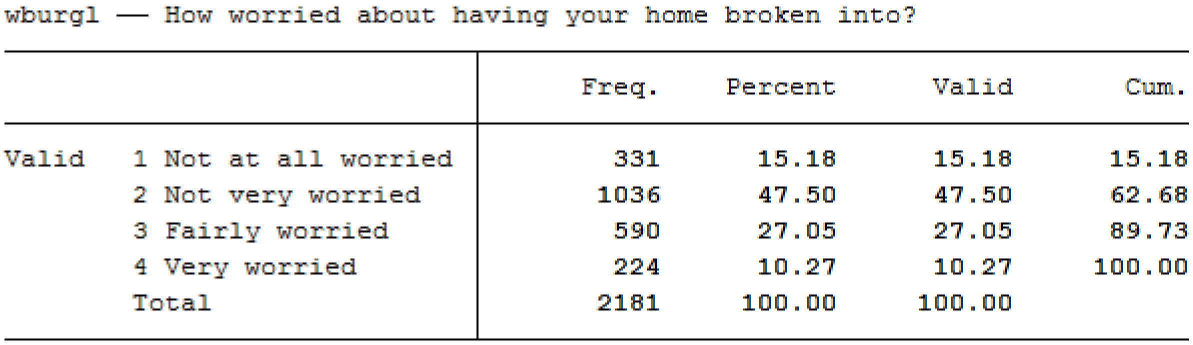
NB! If you have already worked through the multinomial regression computer workshop materials, you can skip this part, as it is the same.

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.



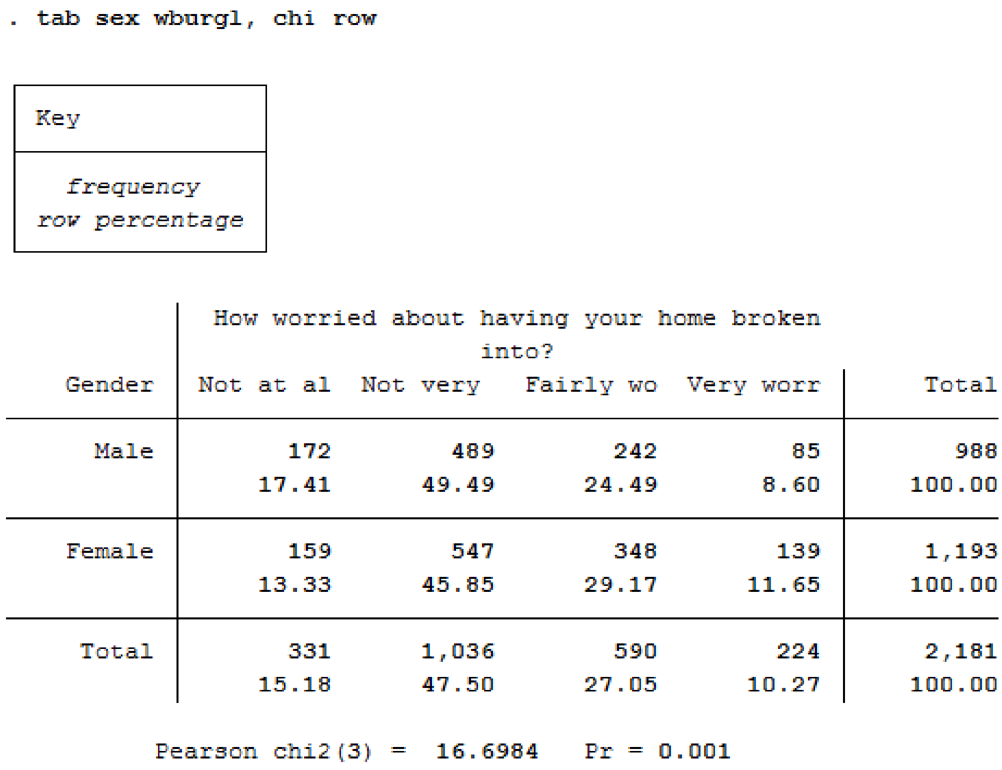
```
wburgl — How worried about having your home broken into?

                                    |  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   |
+-------------------+

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

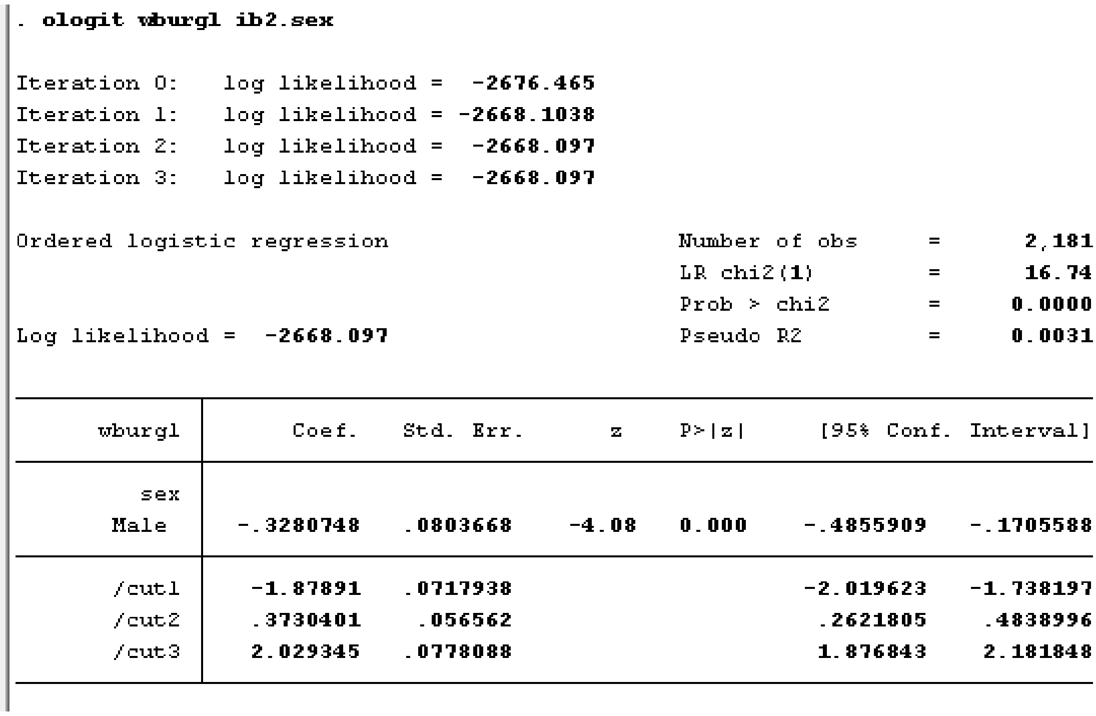
Ordinal Regression with a single predictor

Now we will fit an Ordinal Regression model. The main differences between this model and the multinomial logistic model are:

1. In the multinomial regression, we model the log of the odds between each category and the reference category: $\log\left(\frac{p_{Very}}{p_{Not\ at\ all}}\right)$, and analogously for $\log\left(\frac{p_{Fairly}}{p_{Not\ at\ all}}\right)$, and $\log\left(\frac{p_{Not\ Very}}{p_{Not\ at\ all}}\right)$. In the ordinal regression, on the other hand, we model cumulative logits. In this same example: $\log\left(\frac{p_{(Not\ at\ all)}}{p_{(Not\ very\ or\ Fairly\ or\ Very)}}\right)$, $\log\left(\frac{p_{(Not\ at\ all\ or\ Not\ very)}}{p_{(\ Fairly\ or\ Very)}}\right)$, $\log\left(\frac{p_{(Not\ at\ all\ or\ Not\ very\ or\ Fairly)}}{p_{(Very)}}\right)$.
2. Both models can be written in terms of an equation for each category of the response variable (without including one of the categories). In the multinomial model different intercepts and slopes are allowed for each one of these equations. Meanwhile, in the ordinal model different intercepts are allowed but it is assumed that the slope (the coefficient for each covariate) is the same in all the equations. This assumption is also stated in terms of *proportional odds* or as *parallel lines.*

To fit the model:

- `ologit wburgl ib2.sex`
- `ologit wburgl ib2.sex, or // use or option to obtain odds ratios`
- `estimates store om1`



```
. ologit wburgl ib2.sex

Iteration 0:   log likelihood =  -2676.465
Iteration 1:   log likelihood = -2668.1038
Iteration 2:   log likelihood =  -2668.097
Iteration 3:   log likelihood =  -2668.097

Ordered logistic regression                     Number of obs     =      2,181
                                                LR chi2(1)        =      16.74
                                                Prob > chi2       =     0.0000
Log likelihood =  -2668.097                     Pseudo R2         =     0.0031

------------------------------------------------------------------------------
      wburgl |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |
       Male  |  -.3280748   .0803668    -4.08   0.000    -.4855909   -.1705588
-------------+----------------------------------------------------------------
       /cut1 |   -1.87891   .0717938                     -2.019623   -1.738197
       /cut2 |   .3730401    .056562                      .2621805    .4838996
       /cut3 |   2.029345   .0778088                      1.876843    2.181848
------------------------------------------------------------------------------
```

The coefficient for men is -0.328. As exp(-0.328) is 0.720, it means that *the odds of being in a higher rather than in a lower category of worrying about crime are 28% lower among men than among women*. This is true for all the cumulative odds ratios of this response variable. In practice, this means that men less likely to be in a higher response category (i.e. more worried) than in a lower category (i.e. less worried), when compared to women.

The three equations that characterise this model are:

$$\log\left(\frac{p_{(Not\ at\ all)}}{p_{(Not\ very\ or\ Fairly\ or\ Very)}}\right) = -1.879 + 0.328 \times Man,$$

$$\log\left(\frac{p_{(Not\ at\ all\ or\ Not\ very)}}{p_{(\ Fairly\ or\ Very)}}\right) = 0.373 + 0.328 \times Man,$$

$$\log\left(\frac{p_{(Not\ at\ all\ or\ Not\ very\ or\ Fairly)}}{p_{(Very)}}\right) = 2.029 + 0.328 \times Man$$

Note that the coefficient for gender is positive because the model in Stata is defined with a negative B. These equations can be used to calculate the fitted probabilities of the model, or calculated by Stata as below:

- `margins sex`

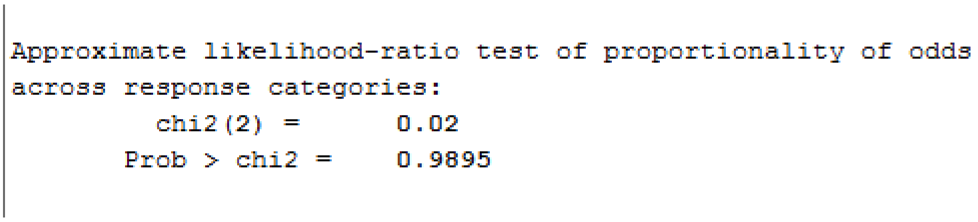
You may wish to calculate some probabilities by hand using the instructions in the first and second videos of this resource and compare whether you get the same results as from Stata.

Test of parallel lines

We need to install the user-written command *ologit* to re-run the model and conduct the test of parallel lines:

- `ssc install omodel`
- `tab sex, gen(gender) // omodel command does not accept the i. notation, so we need to create dummy-variables for gender`
- `omodel logit wburg gender1`

Look at the table containing the test of parallel lines printed below the parameter estimates.



```
Approximate likelihood-ratio test of proportionality of odds
across response categories:
         chi2(2) =      0.02
       Prob > chi2 =    0.9895
```

The null hypothesis is that the coefficient of the slope is the same for all the categories of the response variable. In this sense, the equations determined by the linear predictor are all parallel because they only differ by the intercept. Moreover, note that if there is only one slope for all categories, all the cumulative odds ratios are proportional with proportionality constant e^{B} as shown above. For this reason, this test is also called test of the proportional odds.

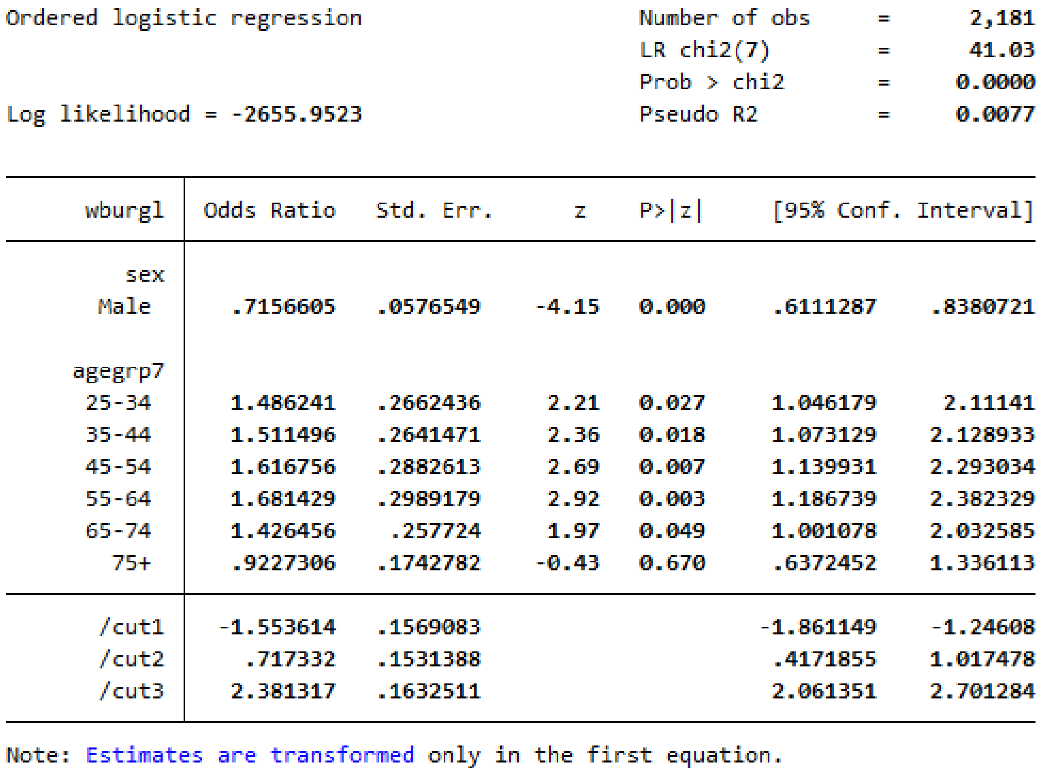
Using this dataset, we are not able to reject the null hypothesis at 5% of significance, i.e., the model with only one slope is reasonable. Had we rejected this hypothesis, we would have to stay with the multinomial model.

Multivariate Ordinal models

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

- `ologit wburgl ib2.sex i.agegrp7`
- `ologit wburgl ib2.sex i.agegrp7, or`
- `estimates store om2`
- `lrtest om1 om2`

The output from the *lrtest* command gives you the likelihood ratio test of nested models (with and without age) and shows you that age should be included in the model. According to the likelihood ratio test, age was statistically significant at the 1% level (LR=24.29, p=0.0005). The results of the model are shown below.



```
Ordered logistic regression                     Number of obs     =      2,181
                                                LR chi2(7)        =      41.03
                                                Prob > chi2       =     0.0000
Log likelihood = -2655.9523                     Pseudo R2         =     0.0077

------------------------------------------------------------------------------
      wburgl | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |
        Male |   .7156605   .0576549    -4.15   0.000     .6111287    .8380721
             |
     agegrp7 |
       25-34 |   1.486241   .2662436     2.21   0.027     1.046179     2.11141
       35-44 |   1.511496   .2641471     2.36   0.018     1.073129    2.128933
       45-54 |   1.616756   .2882613     2.69   0.007     1.139931    2.293034
       55-64 |   1.681429   .2989179     2.92   0.003     1.186739    2.382329
       65-74 |   1.426456    .257724     1.97   0.049     1.001078    2.032585
         75+ |   .9227306   .1742782    -0.43   0.670     .6372452    1.336113
-------------+----------------------------------------------------------------
       /cut1 |  -1.553614   .1569083                     -1.861149    -1.24608
       /cut2 |    .717332   .1531388                      .4171855    1.017478
       /cut3 |   2.381317   .1632511                      2.061351    2.701284
------------------------------------------------------------------------------
Note: Estimates are transformed only in the first equation.
```

Let’s interpret the effect of age in this model. When interpreting the results of a dummy-variable with a large number of categories, it is rarely of interest to report every single odds ratio for that variable. It is preferable to say something about the overall trend, and support that with an example.

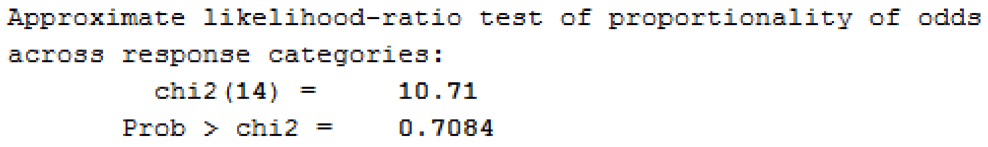
Those in age groups between 25 and 74 were more likely to be more worried about burglary (i.e. in the higher categories of the outcome) than the reference group of 16-24 years. For instance, those aged 25-34 had 1.5 times the odds of being more worried than those age 24 years or less. The oldest age group (75 years or more) did not differ statistically significantly from the youngest group (p=0.670).

To test the parallel lines assumption, we first create the dummy-variables for age and then run the ‘omodel’ command:

➢ `tab agegrp7, gen(agec)`

➢ `omodel logit wburgl gender1 agec1 agec2 agec3 agec4 agec5 agec6`

As shown below, the test of parallel lines is not statistically significant (p=0.708) suggesting that we cannot reject the null hypothesis that the slope coefficients are the same across response categories. Therefore, we can use ordinal regression.



```
Approximate likelihood-ratio test of proportionality of odds
across response categories:
         chi2(14) =     10.71
       Prob > chi2 =     0.7084
```

We can also calculate predicted probabilities:

➢ `ologit ib2.sex i.agegrp7`

- `margins agegrp7, at(sex==2)`
- `marginsplot, legend(order(1 "Not at all worried" 2 "Not very worried" 3 "Fairly worried" 4 "Very worried"))`

If you compare the predicted probabilities from the ordinal model to the multinomial one, you can get more information about how well your model fits.

- `mlogit wburgl ib2.sex i.agegrp7, b(1)`
- `margins agegrp7, at(sex==2)`
- `marginsplot, legend(order(1 "Not at all worried" 2 "Not very worried" 3 "Fairly worried" 4 "Very worried"))`

There are some differences, particularly among the ‘not at all’ and ‘not very’ categories, but none of the probabilities are very far from the more precise multinomial model, which suggests that our ordinal model fits reasonably well.