SO5032 Lecture 11

Brendan Halpin

April 17, 2023

Outline

SO5032 Lecture 11: Multinomial and ordinal regression

SO5032 Lecture 11: Multinomial and ordinal regression

Outline

Summary

- Binary logistic regression is for 2 outcomes (yes/no)
- With more than two outcomes:
    - Multinomial logistic regression (nominal outcomes)
    - Ordinal logistic regression (ordinal outcomes)

SO5032 Lecture 11: Multinomial and ordinal regression

Many categories

What if we have multiple possible outcomes, not just two?

- Logistic regression is binary: yes/no
- Many interesting dependent variables have multiple categories
  - voting intention by party
  - first destination after second-level education
  - housing tenure type
- We can use binary logistic by
  - recoding into two categories
  - dropping all but two categories
- But that would lose information

Multinomial logistic regression

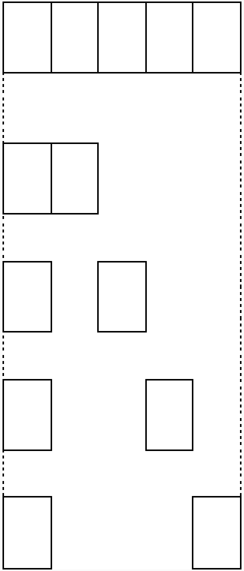
- Another idea:
- Pick one of the J categories as baseline
- For each of $J-1$ other categories, fit binary models contrasting that category with baseline
- Multinomial logistic effectively does that, fitting $J-1$ models simultaneously

$$\log \frac{P(Y=j)}{P(Y=J)} = \alpha_j + \beta_j X, \ \ j = 1, \ldots, c-1$$

- Which category is baseline is not critically important, but better for interpretation if it is reasonably large and coherent (i.e. "Other" is a poor choice)

Multinomial logit: $J - 1$ contrasts

Each category except one is compared against a baseline, and a single model is fitted in one go



Example

- Let’s attempt to predict housing tenure
  - Owner occupier
  - Local authority renter
  - Private renter
- using age and employment status
  - Employed
  - Unemployed
  - Not in labour force
- `mlogit ten3 age i.eun`

Stata output

```
. mlogit ten3 age i.eun

Iteration 0:   log likelihood =  -7222.352
Iteration 1:   log likelihood = -6837.8941
Iteration 2:   log likelihood = -6795.5044
Iteration 3:   log likelihood = -6795.3972
Iteration 4:   log likelihood = -6795.3972

Multinomial logistic regression                 Number of obs     =     11,770
                                                LR chi2(8)        =     853.91
                                                Prob > chi2       =     0.0000
Log likelihood = -6795.3972                     Pseudo R2         =     0.0591

------------------------------------------------------------------------------
           ten3 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
----------------+-------------------------------------------------------------
Owner_occupier  |  (base outcome)
----------------+-------------------------------------------------------------
Social_renter   |
            age |  -.0008792   .0027744    -0.32   0.751     -.006317    .0045587
                |
            eun |
     Unemployed |   2.197923   .1401941    15.68   0.000     1.923148    2.472698
      Not in LM |   1.818469   .0736188    24.70   0.000     1.674179    1.962759
        Retired |   1.068702   .0975851    10.95   0.000     .8774384    1.259965
                |
          _cons |  -2.425975    .135135   -17.95   0.000    -2.690835   -2.161115
----------------+-------------------------------------------------------------
Private_renter  |
            age |    -.02291   .0043864    -5.22   0.000    -.0315072   -.0143128
                |
            eun |
     Unemployed |   1.209508   .2153007     5.62   0.000     .7875264     1.63149
      Not in LM |   .8079265    .111692     7.23   0.000     .5890142    1.026839
        Retired |   .3597836    .158331     2.27   0.023     .0494605    .6701067
                |
          _cons |  -1.747756   .1999509    -8.74   0.000    -2.139653   -1.355859
------------------------------------------------------------------------------
```

sociology
UNIVERSITY OF LIMERICK

Interpretation

- Stata chooses category 1 (owner) as baseline
- Each panel is similar in interpretation to a binary regression on that category versus baseline
- Effects are on the log of the odds of being in category j versus the baseline

Inference

- At one level inference is the same:
    - Wald test for $H_o : \beta_k = 0$
    - LR test between nested models
- However, each variable has $J - 1$ parameters
- Better to consider the LR test for dropping the variable across all contrasts: $H_0 : \beta_1 k = \beta_2 k = ... = \beta_j k = 0$
- Thus retain a variable even for contrasts where it is insignificant as long as it has an effect overall
- Which category is baseline affects the parameter estimates but not the fit (log-likelihood, predicted values, LR test on variables)

SO5032 Lecture 11: Multinomial and ordinal regression

Ordinal logit

Predicting ordinal outcomes

- While `mlogit` is attractive for multi-category outcomes, it is imparsimonious
- For nominal variables this is necessary, but for ordinal variables there should be a better way
- We consider one useful model (others exist)
    - Proportional odds logit

sociology
UNIVERSITY OF LIMERICK

SO5032 Lecture 11: Multinomial and ordinal regression

Proportional odds

The proportional odds model

- The most commonly used ordinal logistic model has another logic
- It assumes the ordinal variable is based on an unobserved latent variable
- Unobserved cutpoints divide the latent variable into the groups indexed by the observed ordinal variable
- The model estimates the effects on the log of the odds of being higher rather than lower across the cutpoints

The model

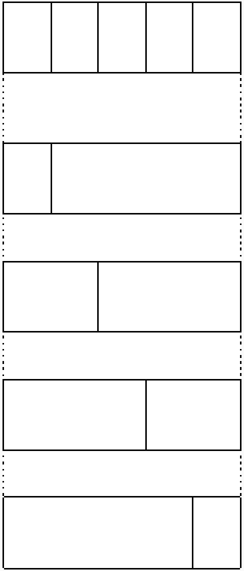
- For $j = 1$ to $J - 1$,

$$\log \frac{P(Y > j)}{P(Y <= j)} = \alpha_j + \beta x$$

- Only one β per variable, whose interpretation is the effect on the odds of being higher rather than lower
- One α per contrast, taking account of the fact that there are different proportions in each one

$J - 1$ contrasts again, but different

But rather than compare categories against a baseline it splits into high and low, with all the data involved each time



An example

- Using data from the BHPS, we predict the probability of each of 5 ordered responses to the assertion "homosexual relationships are wrong"
- Answers from 1: strongly agree, to 5: strongly disagree
- Sex and age as predictors – descriptively women and younger people are more likely to disagree (i.e., have high values)

First approach: just use `mlogit`

```
. mlogit ropfamr i.rsex rage, baseoutcome(1)

Iteration 0:   log likelihood = -18924.158
Iteration 1:   log likelihood = -17839.541
Iteration 2:   log likelihood = -17781.073
Iteration 3:   log likelihood = -17780.905
Iteration 4:   log likelihood = -17780.905

Multinomial logistic regression                         Number of obs =  12,725
                                                        LR chi2(8)    = 2286.51
                                                        Prob > chi2   =  0.0000
Log likelihood = -17780.905                             Pseudo R2     =  0.0604

------------------------------------------------------------------------------------------
                 ropfamr | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------------------+----------------------------------------------------------------
strongly_agree           |  (base outcome)
-------------------------+----------------------------------------------------------------
agree                    |
                    rsex |
                 female  |   .3920172    .084704     4.63   0.000     .2260005     .558034
                    rage |  -.0019587   .0022428    -0.87   0.382    -.0063546    .0024371
                   _cons |    .050326   .1303924     0.39   0.700    -.2052385    .3058905
-------------------------+----------------------------------------------------------------
neither_agree_nor_dis~e  |
                    rsex |
                 female  |   .8480555   .0699274    12.13   0.000     .7110004    .9851106
                    rage |   -.016104   .0018436    -8.74   0.000    -.0197173   -.0124906
                   _cons |   1.808773   .1055106    17.14   0.000     1.601976     2.01557
-------------------------+----------------------------------------------------------------
disagree                 |
                    rsex |
                 female  |   1.228169   .0728716    16.85   0.000     1.085343    1.370995
                    rage |  -.0370249   .0019475   -19.01   0.000    -.0408418   -.0332079
                   _cons |   2.354661   .1077832    21.85   0.000      2.14341    2.565912
-------------------------+----------------------------------------------------------------
strongly_disagree        |
                    rsex |
                 female  |   1.697925   .0796096    21.33   0.000     1.541894    1.853957
                    rage |  -.0671478   .0022283   -30.13   0.000    -.0715151   -.0627804
                   _cons |   2.884952   .1143069    25.24   0.000     2.660915     3.10899
------------------------------------------------------------------------------------------
```

sociology
UNIVERSITY OF LIMERICK

Ordered logistic: Stata output

```
Ordered logistic regression                     Number of obs   =      12725
                                                LR chi2(2)      =    2244.14
                                                Prob > chi2     =     0.0000
Log likelihood = -17802.088                     Pseudo R2       =     0.0593

------------------------------------------------------------------------------
     ropfamr |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      2.rsex |   .8339045    .033062    25.22   0.000     .7691041    .8987048
        rage |  -.0371618   .0009172   -40.51   0.000    -.0389595   -.035364
-------------+----------------------------------------------------------------
       /cut1 |  -3.833869   .0597563                     -3.950989   -3.716749
       /cut2 |  -2.913506   .0547271                      -3.02077   -2.806243
       /cut3 |  -1.132863   .0488522                     -1.228612   -1.037115
       /cut4 |   .3371151   .0482232                      .2425994    .4316307
------------------------------------------------------------------------------
```

Interpretation

- The betas are straightforward:
  - The effect for women is .8339. The OR is $e^{.8339}$ or 2.302
  - Women's odds of being on the "disagree" rather than the "agree" (high values of the variable) side of each contrast are 2.302 times as big as men's
  - Each year of age reduced the log-odds by .03716 (OR 0.964).
- The intercepts are odd: Stata sets up the model in terms of cutpoints in the latent variable, so they are actually $-\alpha_j$

Linear predictor

- Thus the $\alpha + \beta X$ or linear predictor for the contrast between strongly agree (1) and the rest is (2-5 versus 1)

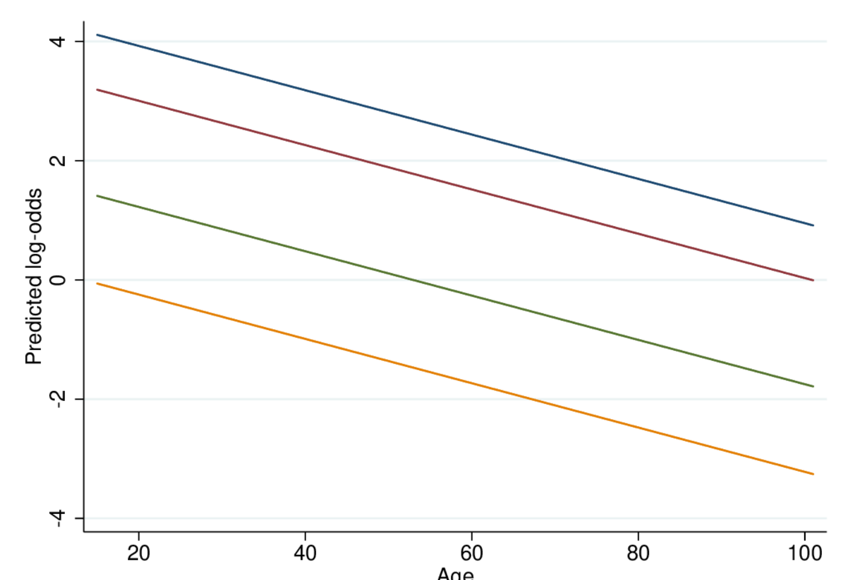
$$3.834 + 0.8339 \times \text{female} - 0.03716 \times \text{age}$$

- Between strongly disagree (5) and the rest (1-4 versus 5)

$$-0.3371 + 0.8339 \times \text{female} - 0.03716 \times \text{age}$$

and so on.

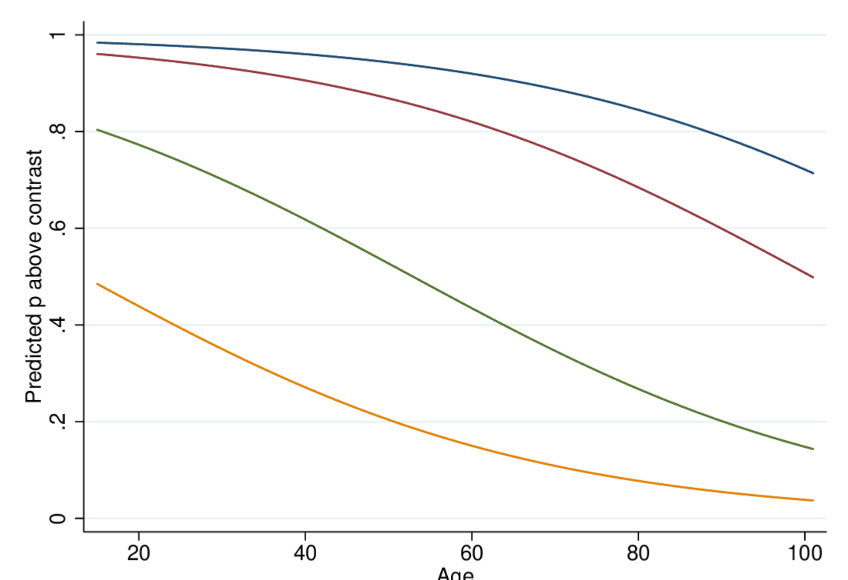
Predicted log odds



Predicted log odds per contrast

- The predicted log-odds lines are straight and parallel
- The highest relates to the 1-4 vs 5 contrast
- Parallel lines means the effect of a variable is the same across all contrasts
- Exponentiating, this means that the multiplicative effect of a variable is the same on all contrasts: hence "proportional odds"
- This is a key assumption

Predicted probabilities relative to contrasts



sociology
UNIVERSITY OF LIMERICK

Predicted probabilities relative to contrasts

- We predict the probabilities of being above a particular contrast in the standard way
- Since age has a negative effect, downward sloping sigmoid curves
- Sigmoid curves are also parallel (same shape, shifted left-right)
- We get probabilities for each of the five states by subtraction

Inference

- The key elements of inference are standard: Wald tests and LR tests
- Since there is only one parameter per variable it is more straightforward than MNL
- However, the key assumption of proportional odds (that there *is* only one parameter per variable) is often wrong.
- The effect of a variable on one contrast may differ from another
- Long and Freese's `SPost` Stata add-on contains a test for this

Compare with linear regression: ologit

```
. ologit ropfamr i.rsex rage

Iteration 0:   log likelihood = -18924.158
Iteration 1:   log likelihood = -17818.231
Iteration 2:   log likelihood = -17802.121
Iteration 3:   log likelihood = -17802.088
Iteration 4:   log likelihood = -17802.088

Ordered logistic regression                         Number of obs =  12,725
                                                    LR chi2(2)    = 2244.14
                                                    Prob > chi2   =  0.0000
Log likelihood = -17802.088                         Pseudo R2     =  0.0593

------------------------------------------------------------------------------
     ropfamr | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
        rsex |
     female  |   .8339045    .033062    25.22   0.000     .7691041    .8987048
        rage |  -.0371618   .0009172   -40.51   0.000    -.0389595   -.035364
-------------+----------------------------------------------------------------
       /cut1 |  -3.833869   .0597563                     -3.950989   -3.716749
       /cut2 |  -2.913506   .0547271                      -3.02077   -2.806243
       /cut3 |  -1.132863   .0488522                     -1.228612   -1.037115
       /cut4 |   .3371151   .0482232                      .2425994    .4316307
------------------------------------------------------------------------------
```

Compare with linear regression: regression

```
. reg ropfamr i.rsex rage

      Source |       SS           df       MS      Number of obs   =    12,725
-------------+----------------------------------   F(2, 12722)     =   1157.61
       Model |  2675.45318         2  1337.72659   Prob > F        =    0.0000
    Residual |  14701.4919    12,722  1.15559597   R-squared       =    0.1540
-------------+----------------------------------   Adj R-squared   =    0.1538
       Total |  17376.9451    12,724  1.36568257   Root MSE        =     1.075

------------------------------------------------------------------------------
     ropfamr | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
        rsex |
     female  |   .4938903   .0191483    25.79   0.000     .4563568    .5314238
        rage |  -.0208292   .0005083   -40.98   0.000    -.0218255   -.0198329
       _cons |   4.073714   .0274276   148.53   0.000     4.019952    4.127476
------------------------------------------------------------------------------
```