Ordinal and Multinomial Logistic Regression

Andy Grogan-Kaylor

14 May 2020

Meta-Background

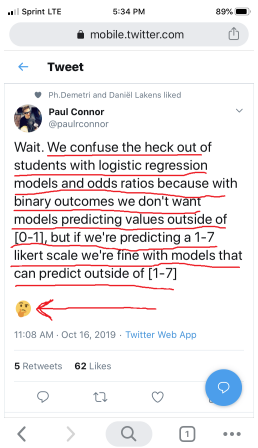

Figure 1: Tweet About Ordinal Models

Key Concepts and Commands

- Implementations differ; formulas are our friends
- Extensions to logistic model: ordinal and multinomial logit

$$F(y) = \beta_0 + \beta x_1 + \beta x_2 + ...$$

- Ordinal model

$y(1,\ 2,\ 3,\ \text{etc.}) = \beta_0 + \beta x_1 + \beta x_2 + ...$

- Multinomial model

$y(2 \text{ vs. } 1) = \beta_0 + \beta x_1 + \beta x_2 + ...$
$y(3 \text{ vs. } 1) = \beta_0 + \beta x_1 + \beta x_2 + ...$

- Think about OR's, predicted probabilities, non-linearity

Get The Data (General Social Survey)

```
. clear all

. set maxvar 10000 // increase number of allowable variables

. use "/Users/agrogan/Box Sync/DATA WAREHOUSE/General Social Survey/GSS7218_R1.DTA", clear

. keep sex maeduc paeduc age degree

. save GSSsmall.dta, replace
file GSSsmall.dta saved

. describe // describe the data
Contains data from GSSsmall.dta
  obs:        64,814
 vars:             5                          14 May 2020 22:41
 size:       324,070
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
age             byte    %8.0g      AGE        age of respondent
paeduc          byte    %8.0g      LABK       highest year school completed, father
maeduc          byte    %8.0g      LABK       highest year school completed, mother
degree          byte    %8.0g      LABL       r´s highest degree
sex             byte    %8.0g      SEX        respondents sex
-------------------------------------------------------------------------------
Sorted by:
```

Thinking About Your Data and Data Wrangling

It is always good to think about your data and what the values of different variables represent. In Stata, however, there is very little additional data wrangling to prepare the data. In R, there is considerable data wrangling since we have to employ special commands just to get *variable* and *value* labels, and to ensure that *numeric dependent* variables are recoded as *factors*. In Stata there are no such issues!!!

Descriptive Statistics

```
. summarize
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         age |     64,586    46.09936     17.5347        18         89
      paeduc |     45,837    10.71026    4.342689         0         20
      maeduc |     53,870    10.85365    3.768792         0         20
      degree |     64,641     1.35858    1.175289         0          4
         sex |     64,814    1.558521    .4965673         1          2
```

```
. tabulate degree
   r´s highest |
        degree |      Freq.     Percent        Cum.
---------------+-----------------------------------
lt high school |     13,587       21.02       21.02
   high school |     33,195       51.35       72.37
junior college |      3,668        5.67       78.05
      bachelor |      9,475       14.66       92.70
      graduate |      4,716        7.30      100.00
---------------+-----------------------------------
         Total |     64,641      100.00
```

The Ordinal Model (*k categories*)

$$ln(\frac{p(y \leq k)}{p(y > k)}) = \beta_0 + \beta_1 x + ...$$

Ordinal Regression

```
. ologit degree sex age paeduc maeduc
Iteration 0:   log likelihood = -56160.846
Iteration 1:   log likelihood = -50678.236
Iteration 2:   log likelihood = -50453.401
Iteration 3:   log likelihood = -50452.782
Iteration 4:   log likelihood = -50452.782
Ordered logistic regression                     Number of obs     =     42,583
                                                LR chi2(4)        =   11416.13
                                                Prob > chi2       =     0.0000
Log likelihood = -50452.782                     Pseudo R2         =     0.1016
```

degree	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
sex	-.0756255	.0188243	-4.02	0.000	-.1125204	-.0387307
age	.0124686	.0006014	20.73	0.000	.0112899	.0136474
paeduc	.151748	.0031156	48.71	0.000	.1456416	.1578545
maeduc	.157931	.0036724	43.00	0.000	.1507332	.1651288
/cut1	1.686014	.0565978			1.575084	1.796944
/cut2	4.710994	.06085			4.59173	4.830258
/cut3	5.061419	.0614286			4.941021	5.181817
/cut4	6.542017	.0645181			6.415564	6.66847

Many commands for regression of categorical dependent variables in R do not provide p values, and an extra step has to be taken to get p values. This is *not* a problem in Stata!

Exponentiating Coefficients: e^{β}

```
. ologit degree sex age paeduc maeduc, or
Iteration 0:   log likelihood = -56160.846
Iteration 1:   log likelihood = -50678.236
Iteration 2:   log likelihood = -50453.401
Iteration 3:   log likelihood = -50452.782
Iteration 4:   log likelihood = -50452.782
Ordered logistic regression                     Number of obs     =     42,583
                                                LR chi2(4)        =   11416.13
                                                Prob > chi2       =     0.0000
Log likelihood = -50452.782                     Pseudo R2         =     0.1016
```

degree	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
sex	.9271633	.0174532	-4.02	0.000	.8935791	.9620098
age	1.012547	.000609	20.73	0.000	1.011354	1.013741
paeduc	1.163867	.0036261	48.71	0.000	1.156782	1.170996
maeduc	1.171085	.0043007	43.00	0.000	1.162686	1.179545
/cut1	1.686014	.0565978			1.575084	1.796944
/cut2	4.710994	.06085			4.59173	4.830258
/cut3	5.061419	.0614286			4.941021	5.181817
/cut4	6.542017	.0645181			6.415564	6.66847

```
Note: Estimates are transformed only in the first equation.
```

The Proportional Odds Assumption