Ordinal and Multinomial Logistic Regression

Andy Grogan-Kaylor

17 May 2020

Meta-Background

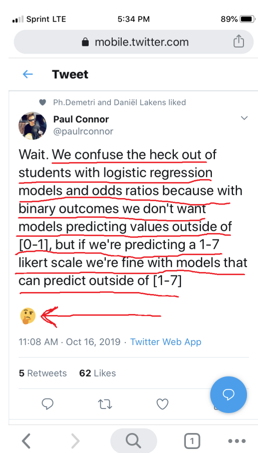

Figure 1: Tweet About Ordinal Models

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                          17 May 2020 20:45
 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

The Brant Test

The Multinomial Model

$$ln(\frac{P(y = y_2)}{P(y = y_1)}) = ln(\frac{P(y = \text{something else})}{P(y = \text{something})})$$

$$= \beta_0 + \beta_1 + ...$$

$$ln(\frac{P(y = y_3)}{P(y = y_1)}) = ln(\frac{P(y = \text{something else altogether})}{P(y = \text{something})})$$

$$= \beta_0 + \beta_1 + ...$$

Estimation

```
. mlogit degree sex age paeduc maeduc
Iteration 0:   log likelihood = -56160.846
Iteration 1:   log likelihood = -50661.077
Iteration 2:   log likelihood = -49974.278
Iteration 3:   log likelihood = -49965.555
Iteration 4:   log likelihood = -49965.546
Iteration 5:   log likelihood = -49965.546
Multinomial logistic regression                 Number of obs     =     42,583
                                                LR chi2(16)       =   12390.60
                                                Prob > chi2       =     0.0000
Log likelihood = -49965.546                     Pseudo R2         =     0.1103
```

degree	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
lt_high_school						
sex	-.1565067	.0315659	-4.96	0.000	-.2183747	-.0946388
age	.0086206	.000955	9.03	0.000	.0067488	.0104925
paeduc	-.1118775	.0050541	-22.14	0.000	-.1217833	-.1019718
maeduc	-.1581699	.005415	-29.21	0.000	-.168783	-.1475568
_cons	1.013337	.085875	11.80	0.000	.8450251	1.181649
high_school	(base outcome)					
junior_college						
sex	.10857	.0419854	2.59	0.010	.0262802	.1908597
age	.0027976	.0013664	2.05	0.041	.0001195	.0054756
paeduc	.0671707	.0069222	9.70	0.000	.0536034	.0807381
maeduc	.0537209	.0084844	6.33	0.000	.0370918	.0703501
_cons	-3.78768	.1379641	-27.45	0.000	-4.058084	-3.517275
bachelor						
sex	-.1383151	.0276789	-5.00	0.000	-.1925648	-.0840654
age	.0159393	.0008977	17.76	0.000	.0141798	.0176989
paeduc	.1430438	.0046993	30.44	0.000	.1338333	.1522543
maeduc	.1164455	.0058259	19.99	0.000	.105027	.127864
_cons	-4.618421	.0963738	-47.92	0.000	-4.807311	-4.429532
graduate						
sex	-.3641641	.0363253	-10.03	0.000	-.4353605	-.2929677
age	.0362201	.0011387	31.81	0.000	.0339882	.038452

paeduc	.1749678	.0061332	28.53	0.000	.1629469	.1869887
maeduc	.1348214	.0076177	17.70	0.000	.1198909	.1497519
_cons	-6.541676	.128908	-50.75	0.000	-6.794331	-6.289021

Exponentiating Coefficients

```
. mlogit, rr

Multinomial logistic regression                 Number of obs     =     42,583
                                                LR chi2(16)       =   12390.60
                                                Prob > chi2       =     0.0000
Log likelihood = -49965.546                     Pseudo R2         =     0.1103
```

degree	RRR	Std. Err.	z	P>|z|	[95% Conf.	Interval]
lt_high_school						
sex	.8551258	.0269928	-4.96	0.000	.8038242	.9097015
age	1.008658	.0009633	9.03	0.000	1.006772	1.010548
paeduc	.8941538	.0045191	-22.14	0.000	.8853402	.903055
maeduc	.8537047	.0046228	-29.21	0.000	.8446922	.8628135
_cons	2.754778	.2365665	11.80	0.000	2.328036	3.259744
high_school	(base outcome)					
junior_college						
sex	1.114683	.0468004	2.59	0.010	1.026629	1.21029
age	1.002801	.0013702	2.05	0.041	1.00012	1.005491
paeduc	1.069478	.0074032	9.70	0.000	1.055066	1.084087
maeduc	1.05519	.0089527	6.33	0.000	1.037788	1.072884
_cons	.0226481	.0031246	-27.45	0.000	.0172821	.0296802
bachelor						
sex	.8708243	.0241035	-5.00	0.000	.8248409	.9193711
age	1.016067	.0009122	17.76	0.000	1.014281	1.017856
paeduc	1.15378	.005422	30.44	0.000	1.143202	1.164456
maeduc	1.123496	.0065453	19.99	0.000	1.110741	1.136398
_cons	.0098684	.0009511	-47.92	0.000	.0081698	.0119201
graduate						
sex	.6947772	.025238	-10.03	0.000	.6470314	.7460462
age	1.036884	.0011807	31.81	0.000	1.034572	1.039201
paeduc	1.191208	.0073059	28.53	0.000	1.176974	1.205614
maeduc	1.144332	.0087172	17.70	0.000	1.127374	1.161546
_cons	.0014421	.0001859	-50.75	0.000	.0011201	.0018566

Note: _cons estimates baseline relative risk for each outcome.

Predicted Probabilities