Nominal Outcomes

Ordinal Variables

Statistical Modelling in Stata: Categorical Outcomes

Mark Lunt
Arthritis Research UK Epidemiology Unit
University of Manchester

20/11/2018

R by C Table: Example

<table>
<thead>
<tr>
<th></th>
<th>Females</th>
<th>Males</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indemnity</td>
<td>234 (51%)</td>
<td>60 (40%)</td>
<td>294 (48%)</td>
</tr>
<tr>
<td>Prepaid</td>
<td>196 (42%)</td>
<td>81 (53%)</td>
<td>277 (45%)</td>
</tr>
<tr>
<td>No Insurance</td>
<td>32 (7%)</td>
<td>13 (8%)</td>
<td>45 (7%)</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>462 (100%)</strong></td>
<td><strong>154 (100%)</strong></td>
<td><strong>616 (100%)</strong></td>
</tr>
</tbody>
</table>

\[ \chi^2 = 6.32, \ p = 0.04 \]

tab insure male, co chi2
Analysing an R by C Table

- $\chi^2$-test: says if there is an association
- Need to assess what that association is
- Can calculate odds ratios for each row compared to a baseline row

Odds Ratios from Tables

- Prepaid vs Indemnity
  - OR for males = $\frac{81 \times 234}{60 \times 196} = 1.61$
- No Insurance vs Indemnity
  - OR for males = $\frac{13 \times 234}{60 \times 32} = 1.58$

Multiple Logistic Regression Models

- Previous results can be duplicated with 2 logistic regression models
  - Prepaid vs Indemnity
  - No Insurance vs Indemnity
- Logistic regression model can be extended to more predictors
- Logistic regression model can include continuous variables

Multiple Logistic Regression Models: Example

```
.logistic insure1 male
           insure1 | Odds Ratio Std. Err.    z  P>|z|     [95% Conf. Interval]
-------------+--------------------------------------------------
    male | 1.611735   .3157844  2.44  0.015     1.09779   2.36629
-------------
.logistic insure2 male
           insure2 | Odds Ratio Std. Err.   z  P>|z|     [95% Conf. Interval]
-------------+--------------------------------------------------
    male | 1.584375   .5633029  1.28  0.200     .7934322   3.204163
```
It would be convenient to have a single analysis give all the information.

Can be done with multinomial logistic regression.

Also provides more efficient estimates (narrower confidence intervals) in most cases.

Multinomial Regression Example

```
. mlogit insure male, rrr

Multinomial logistic regression
Number of obs = 616
LR chi2(2) = 6.38
Prob > chi2 = 0.0413
Log likelihood = -553.40712 Pseudo R2 = 0.0057

------------------------------------------------------------------------------
insure | RRR Std. Err. z P>|z| [95% Conf. Interval]
-------------+----------------------------------------------------------------
Prepaid |        
male | 1.611735  .3157844  2.44 0.015 1.09779  2.36629
-------------+----------------------------------------------------------------
Uninsured |        
male | 1.584375  .5693021  1.28 0.200 .7834329  3.20416
------------------------------------------------------------------------------
(Outcome insure==Indemnity is the comparison group)
```

Multinomial Regression in Stata

Command `mlogit`

Option `rrr` (Relative risk ratio) gives odds ratios, rather than coefficients.

Option `basecategory` sets the baseline or reference category.

Using `predict` after `mlogit`

Can predict probability of each outcome

- Need to give `k` variables
  - `predict p1-p3, p`

Can predict probability of one particular outcome

- Need to specify which with `outcome` option
  - `predict p2, p outcome(2)`
Using `predict` after `mlogit`: Example

```
. by male: summ p1-p3

-> male = 0

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>p1</td>
<td>477</td>
<td>.50649</td>
<td>0</td>
<td>.50649</td>
<td>.50649</td>
</tr>
<tr>
<td>p2</td>
<td>477</td>
<td>.42424</td>
<td>0</td>
<td>.42424</td>
<td>.42424</td>
</tr>
<tr>
<td>p3</td>
<td>477</td>
<td>.06926</td>
<td>0</td>
<td>.06926</td>
<td>.06926</td>
</tr>
</tbody>
</table>

-> male = 1

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>p1</td>
<td>167</td>
<td>.38961</td>
<td>0</td>
<td>.38961</td>
<td>.38961</td>
</tr>
<tr>
<td>p2</td>
<td>167</td>
<td>.52597</td>
<td>0</td>
<td>.52597</td>
<td>.52597</td>
</tr>
<tr>
<td>p3</td>
<td>167</td>
<td>.08442</td>
<td>0</td>
<td>.08442</td>
<td>.08442</td>
</tr>
</tbody>
</table>
```

Using `lincom` after `mlogit`

- Can use `lincom` to
  - test if coefficients are different
  - calculate odds of being in a given outcome category
- Need to specify which outcome category we are interested in
- Normally, use the option `eform` to get odds ratios, rather than coefficients

```
. lincom [Prepaid]male - [Uninsure]male
( 1) [Prepaid]male - [Uninsure]male = 0

------------------------------------------------------------------------------
insure | Coef. Std. Err. z P>|z| [95% Conf. Interval]
------------------------------------------------------------------------------
(1) | .017121 .3544299 0.05 0.961 -.6775487 .7117908
------------------------------------------------------------------------------
```

Using `lincom` after `mlogit`

- Can ignore ordering, use multinomial model
- Can use a test for trend
- Can use an ordered logistic regression model

```
. lincom [Prepaid]male - [Uninsure]male
( 1) [Prepaid]male - [Uninsure]male = 0

------------------------------------------------------------------------------
insure | Coef. Std. Err. z P>|z| [95% Conf. Interval]
------------------------------------------------------------------------------
(1) | .017121 .3544299 0.05 0.961 -.6775487 .7117908
------------------------------------------------------------------------------
```
- \( \chi^2 \)-test tests for any differences between columns (or rows)
- Not very powerful against a linear change in proportions
- Can divide the \( \chi^2 \)-statistic into two parts: linear trend and variations around the linear trend.
- Test for trend more powerful against a trend
- Has no power to detect other differences
- Often used for ordinal predictors

### Test for Trend: Example

<table>
<thead>
<tr>
<th></th>
<th>Treatment A</th>
<th>Treatment B</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Healed</td>
<td>12 (38%)</td>
<td>5 (16%)</td>
<td>17 (27%)</td>
</tr>
<tr>
<td>Improved</td>
<td>10 (31%)</td>
<td>8 (25%)</td>
<td>18 (28%)</td>
</tr>
<tr>
<td>No Change</td>
<td>4 (13%)</td>
<td>8 (25%)</td>
<td>12 (19%)</td>
</tr>
<tr>
<td>Worse</td>
<td>6 (19%)</td>
<td>11 (34%)</td>
<td>17 (27%)</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>32 (100%)</strong></td>
<td><strong>32 (100%)</strong></td>
<td><strong>34 (100%)</strong></td>
</tr>
</tbody>
</table>

### Test for Trend: Results

```
. ptrendi 12 5 1 \ 10 8 2 \ 4 8 3 \ 6 11 4

+------------------------+
| r nr _prop x |      |
+------------------------+
1. | 12 5 0.706 1.00 |
2. | 10 8 0.556 2.00 |
3. | 4 8 0.333 3.00 |
4. | 6 11 0.353 4.00 |
+------------------------+

Trend analysis for proportions

Regression of p = r/(r+nr) on x:
Slope = -.12521, std. error = .0546, Z = 2.293
```

### Test for Trend: Caveat

- Test for trend only tests for a linear association between predictors and outcome.
- U-shaped or inverted U-shaped associations will not be detected.
Test for Trend in Stata

- Test for trend often used, should know about it
- Not implemented in base Stata:
  - see http://www.stata.com/support/faqs/stat/trend.html
- Very rarely the best thing to do:
  - If trend variable is the outcome, use ordinal logistic regression
  - If trend variable is a predictor:
    - fit both categorical & continuous, testparm categorical variables
    - If non-significant, use continuous variable
    - If significant, use categorical variables

Fitting an ordinal predictor

```
. regress write oread i.oread
  note: 6.oread omitted because of collinearity

Source | SS df MS Number of obs = 200
-------------+------------------------------ F( 5, 194) = 22.77
Model | 6612.82672 5 1322.56534 Prob > F = 0.0000
Residual | 11266.0483 194 58.0724138 R-squared = 0.3699
-------------+------------------------------ Adj R-squared = 0.3536
Total | 17878.875 199 89.843593 Root MSE = 7.6205
-------------+------------------------------
write | Coef. Std. Err. t P>|t| [95% Conf. Interval]
-------------+----------------------------------------------------------------
oread | 3.288889 1.606548 2.05 0.042 .1203466 6.457431
| 2 | -6.669841 6.339542 -1.05 0.294 -19.17311 5.833432
| 3 | -3.666385 4.761676 -0.77 0.442 -13.05768 5.724914
| 4 | .3641026 3.568089 0.10 0.919 -6.673124 7.401329
| 5 | .4233918 2.825015 0.15 0.881 -5.148294 5.995078
| 6 | 0 (omitted)
| _cons | 42.71111 9.158732 4.66 0.000 24.64764 60.77458
-------------+----------------------------------------------------------------
. testparm i.oread
( 1) 2.oread = 0
( 2) 3.oread = 0
( 3) 4.oread = 0
( 4) 5.oread = 0
F( 4, 194) = 1.36
Prob > F = 0.2497
```

Dose Response

- Don’t confuse trend with dose response
- All three models may have significant trend test
- Only first model has a dose-response effect
- Other models better fitted using categorical variables

```
<table>
<thead>
<tr>
<th>Genetic Model</th>
<th>Genotype</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>aa</td>
</tr>
<tr>
<td>Additive (dose-response)</td>
<td>0</td>
</tr>
<tr>
<td>Dominant</td>
<td>0</td>
</tr>
<tr>
<td>Recessive</td>
<td>0</td>
</tr>
</tbody>
</table>
```
Ordinal Regression: Example

<table>
<thead>
<tr>
<th></th>
<th>Treatment A</th>
<th>Treatment B</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Healed</td>
<td>12 (38%)</td>
<td>5 (16%)</td>
<td>17 (27%)</td>
</tr>
<tr>
<td>Improved</td>
<td>10 (31%)</td>
<td>8 (25%)</td>
<td>18 (28%)</td>
</tr>
<tr>
<td>No Change</td>
<td>4 (13%)</td>
<td>8 (25%)</td>
<td>12 (19%)</td>
</tr>
<tr>
<td>Worse</td>
<td>6 (19%)</td>
<td>11 (34%)</td>
<td>17 (27%)</td>
</tr>
<tr>
<td>Total</td>
<td>32 (100%)</td>
<td>32 (100%)</td>
<td>34 (100%)</td>
</tr>
</tbody>
</table>

Ordinal Regression: Using Tables

- Dichotomise outcome to “Better” or “Worse”
- Can split the table in three places
- This produces 3 odds ratios
- Suppose these three odds ratios are estimates of the same quantity
- Odds of being in a worse group rather than a better one

Ordered Polytomous Logistic Regression

\[
\log\left( \frac{p_i}{1 - p_i} \right) = \alpha_i + \beta x
\]

Where
- \( p_i \) = probability of being in a category up to and including the \( i^{th} \)
- \( \alpha_i \) = Log-odds of being in a category up to and including the \( i^{th} \) if \( x = 0 \)
- \( \beta \) = Log of the odds ratio for being in a category up to and including the \( i^{th} \) if \( x = 1 \), relative to \( x = 0 \)
- ologit fits ordinal regression models
- Option or gives odds ratios rather than coefficients
- Can compare likelihood to mlogit model to see if common odds ratio assumption is valid
- predict works as after mlogit

### Ordinal Regression in Stata: Example

```
. ologit outcome treat, or
Iteration 3: log likelihood = -85.2492
Ordered logit estimates
Number of obs = 64
LR chi2(1) = 5.49
Prob > chi2 = 0.0191
Log likelihood = -85.2492
Pseudo R2 = 0.0312
------------------------------------------------------------------------------
outcome | Odds Ratio Std. Err. z P>|z| [95% Conf. Interval]
-------------+--------------------------------------------------
treat | 2.932028 1.367427 2.31 0.021 1.175407 7.31388
------------------------------------------------------------------------------
```

### Ordinal Regression Caveats

- Assumption that same $\beta$ fits all outcome categories should be tested
  - AIC, BIC or LR test compared to mlogit model
- User-written gologit2 can also be used
  - Allows for some variables to satisfy proportional odds, others not
  - Option autofit() selects variables that violate proportional odds
- There are a variety of other, less widely used, ordinal regression models: see Sander Greenland: Alternative Models for Ordinal Logistic Regression, Statistics in Medicine, 1994, pp1665-1677.