Statistical Modelling with Stata: Binary Outcomes

Mark Lunt

Arthritis Research UK Epidemiology Unit
University of Manchester

13/11/2018
Cross-tabulation

<table>
<thead>
<tr>
<th></th>
<th>Exposed</th>
<th>Unexposed</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cases</td>
<td>a</td>
<td>b</td>
<td>a + b</td>
</tr>
<tr>
<td>Controls</td>
<td>c</td>
<td>d</td>
<td>c + d</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>a + c</strong></td>
<td><strong>b + d</strong></td>
<td><strong>a + b + c + d</strong></td>
</tr>
</tbody>
</table>

- Simple random sample: fix $a + b + c + d$
- Exposure-based sampling: fix $a + c$ and $b + d$
- Outcome-based sampling: fix $a + b$ and $c + d$
The $\chi^2$ Test

- Compares observed to expected numbers in each cell
- Expected under null hypothesis: no association
- Works for any of the sampling schemes
Measures of Association

Relative Risk = \( \frac{a}{a+c} \) \( \frac{b}{b+d} \) \( \frac{a(b + d)}{b(a + c)} \)

Risk Difference = \( \frac{a}{a + c} - \frac{b}{b + d} \)

Odds Ratio = \( \frac{a}{c} \) \( \frac{b}{d} \) \( \frac{ad}{cb} \)

- All obtained with cs disease exposure[, or]
- Only Odds ratio valid with outcome based sampling
### Crosstabulation in Stata

In Stata, you can use the command `.cs` to perform a crosstabulation. Here is an example:

```
.cs back_p sex, or
```

<table>
<thead>
<tr>
<th>sex</th>
<th>Exposed</th>
<th>Unexposed</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>637</td>
<td>445</td>
<td>1082</td>
</tr>
<tr>
<td>Noncases</td>
<td>1694</td>
<td>1739</td>
<td>3433</td>
</tr>
<tr>
<td></td>
<td>2331</td>
<td>2184</td>
<td>4515</td>
</tr>
</tbody>
</table>

### Risk and Point Estimate
- **Risk**: 
  - Exposed: 0.2732733
  - Unexposed: 0.2037546
  - Total: 0.2396456

### Confidence Intervals
- **Risk difference**: 0.0695187
  - [95% Conf. Interval]: 0.044767 to 0.0942704
- **Risk ratio**: 1.341188
  - [95% Conf. Interval]: 1.206183 to 1.491304
- **Attr. frac. ex.**: 0.2543926
  - [95% Conf. Interval]: 0.1709386 to 0.329446
- **Attr. frac. pop**: 0.1497672
- **Odds ratio**: 1.469486
  - [95% Conf. Interval]: 1.27969 to 1.68743 (Cornfield)

### Chi-Square Test
- `chi2(1) = 29.91`  
- `Pr>chi2 = 0.0000`
Limitations of Tabulation

- No continuous predictors
- Limited numbers of categorical predictors
Linear Regression and Binary Outcomes

- Can’t use linear regression with binary outcomes
  - Distribution is not normal
  - Limited range of sensible predicted values
- Changing parameter estimation to allow for non-normal distribution is straightforward
- Need to limit range of predicted values
Example: CHD and Age

![Graph showing CHD and Age relationship]
Example: CHD by Age group

![Graph showing the proportion of subjects with CHD by age group. The x-axis represents mean age, ranging from 20 to 60, and the y-axis represents the proportion of subjects with CHD, ranging from 0 to 0.8. There is a positive correlation between age and the proportion of subjects with CHD.](image-url)
Example: CHD by Age - Linear Fit

Proportion of subjects with CHD vs Age

- Fitted values
- Proportion of subjects with CHD vs Age

20 30 40 50 60 70
0 0.5 1
Generalized Linear Models

- Linear Model

\[ Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon \]

\( \varepsilon \) is normally distributed

- Generalized Linear Model

\[ g(Y) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon \]

\( \varepsilon \) has a known distribution
## Probabilities and Odds

<table>
<thead>
<tr>
<th>Probability</th>
<th>Odds</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p$</td>
<td>$\Omega = p/(1 - p)$</td>
</tr>
<tr>
<td>0.1 = 1/10</td>
<td>0.1/0.9 = 1:9 = 0.111</td>
</tr>
<tr>
<td>0.5 = 1/2</td>
<td>0.5/0.5 = 1:1 = 1</td>
</tr>
<tr>
<td>0.9 = 9/10</td>
<td>0.9/0.1 = 9:1 = 9</td>
</tr>
</tbody>
</table>
Probabilities and Odds

- **Proportion**
  - 0.0
  - 0.2
  - 0.4
  - 0.6
  - 0.8
  - 1.0

- **Log odds**
  - -5
  - 0
  - 5

The graph illustrates the relationship between proportions and log odds.
Advantage of the Odds Scale

- Just a different scale for measuring probabilities
- Any odds from 0 to $\infty$ corresponds to a probability
- Any log odds from $-\infty$ to $\infty$ corresponds to a probability
- Shape of curve commonly fits data
The binomial distribution

- Outcome can be either 0 or 1
- Has one parameter: the probability that the outcome is 1
- Assumes observations are independent
The Logistic Regression Equation

\[ \log \left( \frac{\hat{\pi}}{1 - \hat{\pi}} \right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p \]

\[ Y \sim \text{Binomial}(\hat{\pi}) \]

- \( Y \) has a binomial distribution with parameter \( \pi \)
- \( \hat{\pi} \) is the predicted probability that \( Y = 1 \)
When $x_i$ increases by 1, $\log(\hat{\pi}/(1 - \hat{\pi}))$ increases by $\beta_i$.

Therefore $\hat{\pi}/(1 - \hat{\pi})$ increases by a factor $e^{\beta_i}$.

For a dichotomous predictor, this is exactly the odds ratio we met earlier.

For a continuous predictor, the odds increase by a factor of $e^{\beta_i}$ for each unit increase in the predictor.
Logistic Regression in Stata

```
. logistic chd age

Logistic regression
Number of obs = 100
LR chi2(1) = 29.31
Prob > chi2 = 0.0000
Log likelihood = -53.676546 Pseudo R2 = 0.2145

------------------------------------------------------------------------------
   chd | Odds Ratio Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+---------------------------------------------------------------
    age | 1.117307   .026882  4.61 0.000     1.065842    1.171257
------------------------------------------------------------------------------
```
Predict

- Lots of options for the `predict` command
- `p` gives the predicted probability for each subject
- `xb` gives the linear predictor (i.e. the log of the odds) for each subject
Plot of probability against age

- **Pr(chd)**
- Proportion of subject in each ageband with CHD
Plot of log-odds against age
Other Models for Binary Outcomes

- Can use any function that maps \((−\infty, \infty)\) to \((0, 1)\)
  - Probit Model
  - Complementary log-log
- Parameters lack interpretation
The Log-Binomial Model

- Models $\log(\pi)$ rather than $\log(\pi/(1 - \pi))$
- Gives relative risk rather than odds ratio
- Can produce predicted values greater than 1
- May not fit the data as well
- **Stata command:** `glm varlist, family(binomial) link(log)`
- If association between $\log(\pi)$ and predictor non-linear, lose simple interpretation.
Log-binomial model example

- Logistic predictions
- Log–binomial predictions
- Proportion of subjects with CHD
Logistic Regression Diagnostics

- Goodness of Fit
- Influential Observations
- Poorly fitted Observations
Problems with $R^2$

- Multiple definitions
- Lack of interpretability
- Low values
  - Can predict $P(Y = 1)$ perfectly, not predict $Y$ well at all if $P(Y = 1) \approx 0.5$. 
Hosmer-Lemeshow test

- Very like $\chi^2$ test
- Divide subjects into groups
- Compare observed and expected numbers in each group
- Want to see a *non*-significant result
- Command used is `estat gof`
. estat gof, group(5) table

Logistic model for chd, goodness-of-fit test

(Table collapsed on quantiles of estimated probabilities)

<table>
<thead>
<tr>
<th>Group</th>
<th>Prob</th>
<th>Obs_1</th>
<th>Exp_1</th>
<th>Obs_0</th>
<th>Exp_0</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.1690</td>
<td>2</td>
<td>2.1</td>
<td>18</td>
<td>17.9</td>
<td>20</td>
</tr>
<tr>
<td>2</td>
<td>0.3183</td>
<td>5</td>
<td>4.9</td>
<td>16</td>
<td>16.1</td>
<td>21</td>
</tr>
<tr>
<td>3</td>
<td>0.5037</td>
<td>9</td>
<td>8.7</td>
<td>12</td>
<td>12.3</td>
<td>21</td>
</tr>
<tr>
<td>4</td>
<td>0.7336</td>
<td>15</td>
<td>15.1</td>
<td>8</td>
<td>7.9</td>
<td>23</td>
</tr>
<tr>
<td>5</td>
<td>0.9125</td>
<td>12</td>
<td>12.2</td>
<td>3</td>
<td>2.8</td>
<td>15</td>
</tr>
</tbody>
</table>

number of observations = 100
number of groups = 5
Hosmer-Lemeshow ch2(3) = 0.05
Prob > ch2 = 0.9973
Sensitivity and Specificity

<table>
<thead>
<tr>
<th></th>
<th>Test +ve</th>
<th>Test -ve</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cases</td>
<td>a</td>
<td>b</td>
<td>a + b</td>
</tr>
<tr>
<td>Controls</td>
<td>c</td>
<td>d</td>
<td>c + d</td>
</tr>
<tr>
<td>Total</td>
<td>a + c</td>
<td>b + d</td>
<td>a + b + c + d</td>
</tr>
</tbody>
</table>

**Sensitivity:**
- Probability that a case classified as positive
- \( \frac{a}{a + b} \)

**Specificity:**
- Probability that a non-case classified as negative
- \( \frac{d}{c + d} \)
Sensitivity and Specificity in Logistic Regression

- Sensitivity and specificity can only be used with a single dichotomous classification.
- Logistic regression gives a probability, not a classification.
- Can define your own threshold for use with logistic regression.
- Commonly choose 50% probability of being a case.
- Can choose any probability: sensitivity and specificity will vary.
- Why not try every possible threshold and compare results: ROC curve.
**ROC Curves**

- Shows how sensitivity varies with changing specificity
- Larger area under the curve = better
- Maximum = 1
- Tossing a coin would give 0.5
- Command used is `lroc`
Area under ROC curve = 0.7999
Influential Observations

- Residuals less useful in logistic regression than linear
- Can only take the values $1 - \hat{\pi}$ or $-\hat{\pi}$.
- Leverage does not translate to logistic regression model
- $\Delta \hat{\beta}_i$ measures effect of $i^{th}$ observation on parameters
- Obtained from `dbeta` option to `predict` command
- Plot against $\hat{\pi}$ to reveal influential observations
Plot of $\Delta \hat{\beta}_i$ against $\hat{\pi}$
. logistic chd age if dbeta < 0.2

Logistic regression

|        | Odds Ratio | Std. Err. | z    | P>|z| | [95% Conf. Interval] |
|--------|------------|-----------|------|-----|-----------------------|
| age    | 1.130329   | 0.0293066 | 4.73 | 0.000 | 1.074324 - 1.189254   |
Poorly fitted observations

- Can be identified by residuals
  - Deviance residuals: `predict varname, ddeviance`
  - $\chi^2$ residuals: `predict varname, dx2`
- Not influential: omitting them will not change conclusions
- May need to explain fit is poor in particular area
- Plot residuals against predicted probability, look for outliers
Separation

- Need at least one case and one control in each subgroup
- If you have lots of subgroups, this may not be true
- In which case, log(OR) for that group is $-\infty$ or $\infty$
- Stata will drop all subjects from that group (unless you use the option `asis`)
- Not a problem with continuous predictors