Statistical Modelling with Stata: Binary Outcomes

Mark Lunt

Centre for Epidemiology Versus Arthritis
University of Manchester

12/12/2023

Cross-tabulation

	Exposed	Unexposed	Total
Cases	a	b	a + b
Controls	c	d	c + d
Total	a + c	b + d	a + b + c + d

- 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 χ^2 Test

- Compares observed to expected numbers in each cell
- Expected under null hypothesis: no association
- Works for any of the sampling schemes
- Says that there is a difference, not what the difference is

Measures of Association

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

$$\text{Risk Difference} = \frac{a}{a+c} - \frac{b}{b+d}$$

$$\text{Odds Ratio} = \frac{\frac{a}{c}}{\frac{b}{d}} == \frac{ad}{cb}$$

- All obtained with `cs disease exposure[, or]`
- Only Odds ratio valid with outcome based sampling

CENTRE FOR EPIDEMIOLOGY VERSUS ARTHRITIS

Crosstabulation in stata

```
. cs back_p sex, or

                 | sex                    |
                 |   Exposed   Unexposed  |      Total
-----------------+------------------------+------------
           Cases |       637         445  |       1082
        Noncases |      1694        1739  |       3433
-----------------+------------------------+------------
           Total |      2331        2184  |       4515
                 |                        |
            Risk |  .2732733    .2037546  |   .2396456
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
 Risk difference |         .0695187       |     .044767    .0942704
      Risk ratio |         1.341188       |    1.206183    1.491304
 Attr. frac. ex. |         .2543926       |    .1709386     .329446
 Attr. frac. pop |         .1497672       |
      Odds ratio |         1.469486       |     1.27969     1.68743 (Cornfield)
                 +-------------------------------------------------
                               chi2(1) =    29.91  Pr>chi2 = 0.0000
```

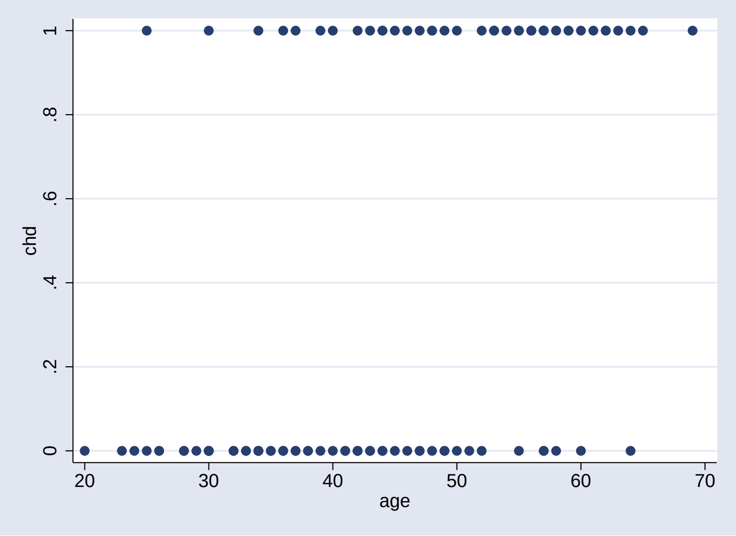
Limitations of Tabulation

- No continuous predictors
- Limited numbers of categorical predictors

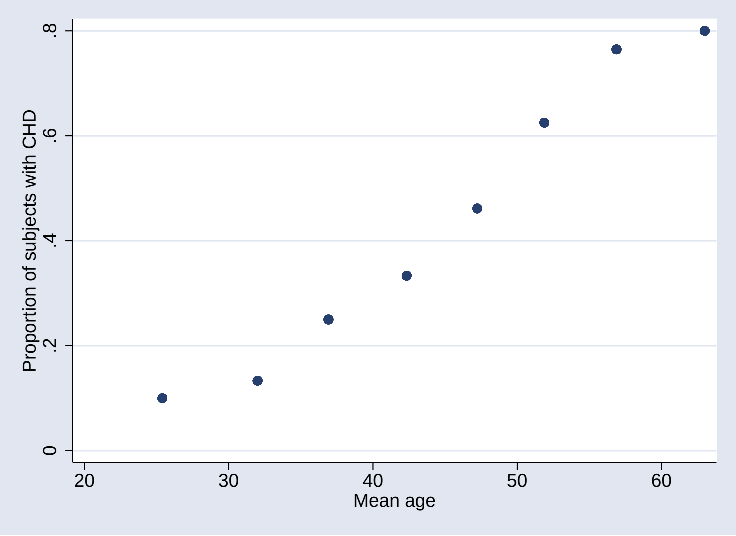
Linear Regression and Binary Outcomes

- 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

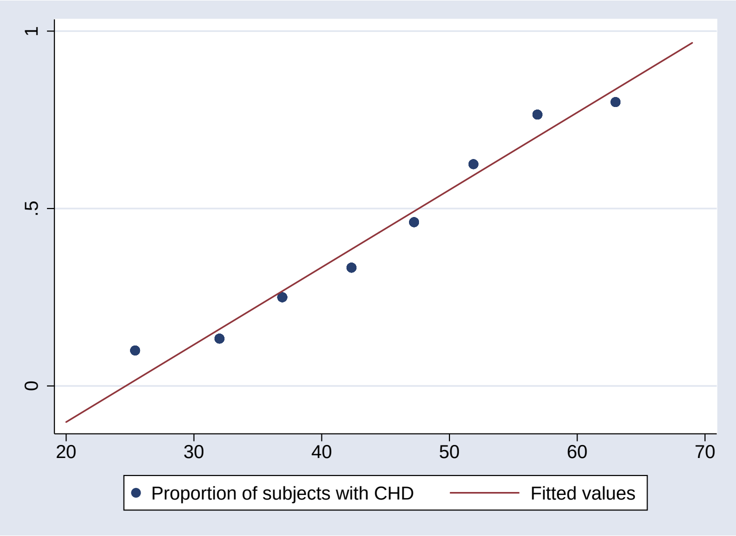
Example: CHD and Age



Example: CHD by Age group



Example: CHD by Age - Linear Fit



Generalized Linear Models

- Linear Model

$$
\begin{aligned}
Y &= \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon \\
\varepsilon &\text{ is } \text{ normally distributed}
\end{aligned}
$$

Generalized Linear Models

- Linear Model

$$\begin{aligned} Y &= \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon \\ \varepsilon &\text{ is } \text{ normally distributed} \end{aligned}$$

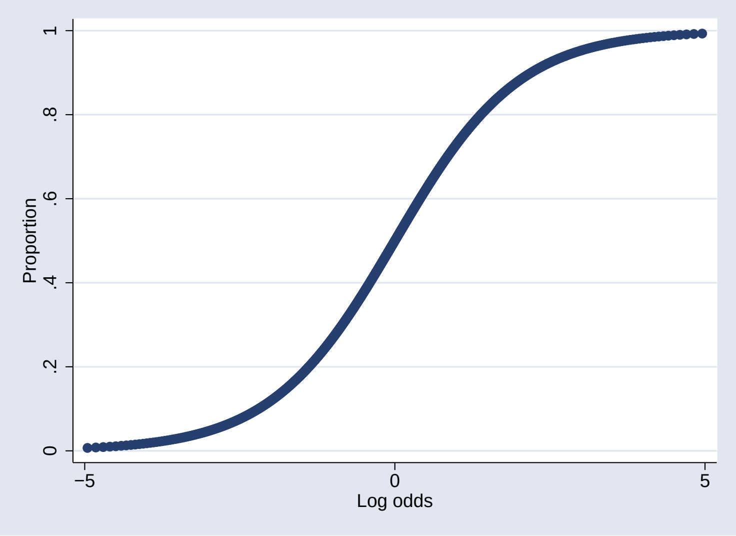
- Generalized Linear Model

$$\begin{aligned} g(Y) &= \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon \\ \varepsilon &\text{ has } \text{ a known distribution} \end{aligned}$$

Probabilities and Odds

Probability	Odds	
p	$\Omega = p/(1-p)$	
0.1 = 1/10	0.1/0.9 =	1:9 = 0.111
0.5 = 1/2	0.5/0.5 =	1:1 = 1
0.9 = 9/10	0.9/0.1 =	9:1 = 9

Probabilities and Odds



Advantage of the Odds Scale

- Just a different scale for measuring probabilities
- Any odds from 0 to ∞ corresponds to a probability
- Any log odds from $-\infty$ to ∞ 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

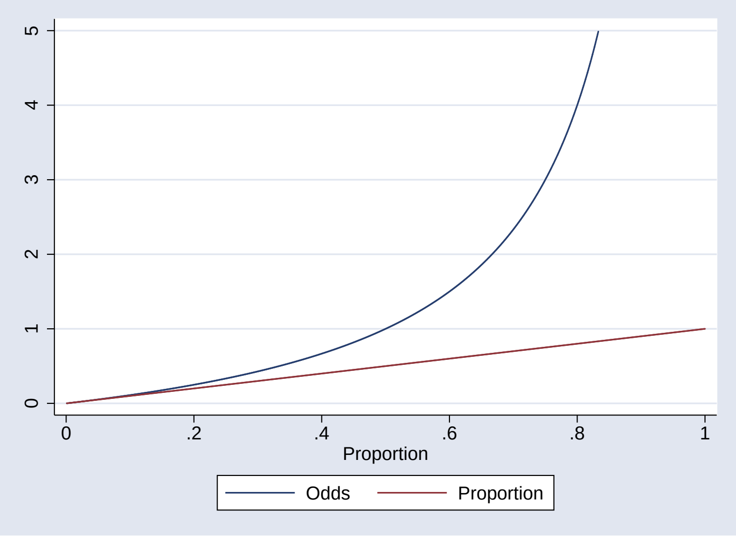
$$\begin{aligned} \log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) &= \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p \\ Y &\sim \text{Binomial}(\hat{\pi}) \end{aligned}$$

- Y has a binomial distribution with parameter π
- $\hat{\pi}$ is the predicted probability that $Y = 1$

Parameter Interpretation

- When x_i increases by 1, $\log\left(\hat{\pi}/(1-\hat{\pi})\right)$ increases by β_i
- Therefore $\hat{\pi}/(1-\hat{\pi})$ increases by a factor e^{β_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^{β_i} for each unit increase in the predictor

Odds Ratios and Relative Risks



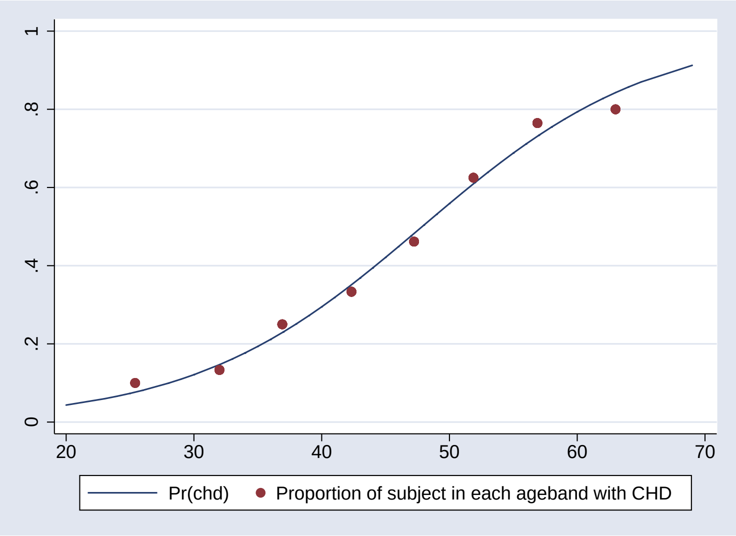
Logistic Regression in Stata

```
. 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   .0268822     4.61   0.000     1.065842    1.171257
------------------------------------------------------------------------------
```

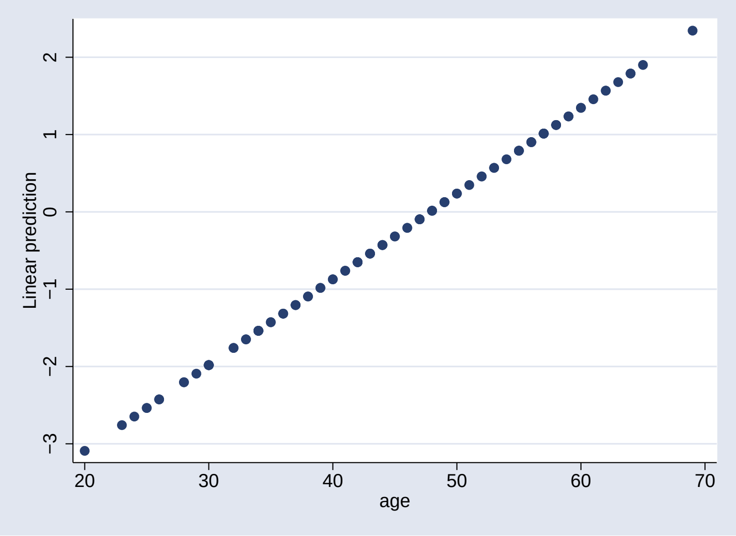
Predict

- 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



Plot of log-odds against age



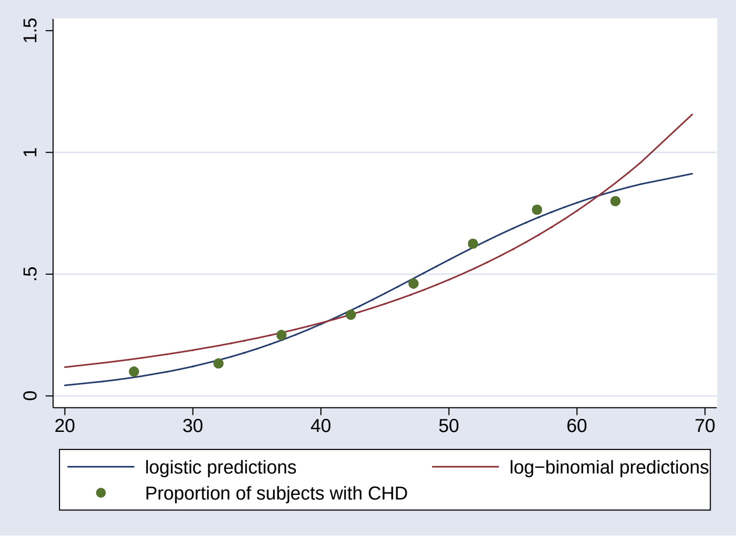
Other Models for Binary Outcomes

- 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 if outcome is not rare
- Stata command: `glm` *`varlist`*`, family(binomial) link(log)`
- If association between $\log(\pi)$ and predictor non-linear, lose simple interpretation.

Log-binomial model example



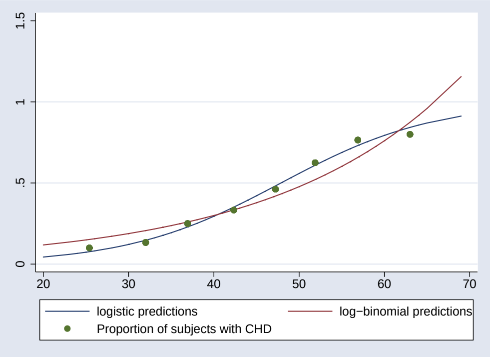
Logistic Regression Diagnostics

- Discrimination and Calibration
- Goodness of Fit
- Influential Observations
- Poorly fitted Observations

Discrimination and Calibration

Discrimination Subjects with higher predicted probabilities more likely to have the event

Calibration Predicted probability is a good measure of probability of the event.



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

- Detects lack of calibration
- Very like χ^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`
- Can always improve model by adding non-linear or interaction terms

Hosmer-Lemeshow test example

```
. estat gof, group(5) table

Logistic model for chd, goodness-of-fit test

  (Table collapsed on quantiles of estimated probabilities)
  +--------------------------------------------------------+
  | Group |   Prob | Obs_1 | Exp_1 | Obs_0 | Exp_0 | Total |
  |-------+--------+-------+-------+-------+-------+-------|
  |     1 | 0.1690 |     2 |   2.1 |    18 |  17.9 |    20 |
  |     2 | 0.3183 |     5 |   4.9 |    16 |  16.1 |    21 |
  |     3 | 0.5037 |     9 |   8.7 |    12 |  12.3 |    21 |
  |     4 | 0.7336 |    15 |  15.1 |     8 |   7.9 |    23 |
  |     5 | 0.9125 |    12 |  12.2 |     3 |   2.8 |    15 |
  +--------------------------------------------------------+

       number of observations =       100
             number of groups =         5
      Hosmer-Lemeshow chi2(3) =         0.05
                  Prob > chi2 =         0.9973
```

Sensitivity and Specificity

	Test +ve	Test -ve	Total
Cases	a	b	a + b
Controls	c	d	c + d
Total	a + c	b + d	a + b + c + d

- Sensitivity:
  - Probability that a case classified as positive
  - $a/(a+b)$
- Specificity:
  - Probability that a non-case classified as negative
  - $d/(c+d)$

Sensitivity and Specificity in Logistic Regression

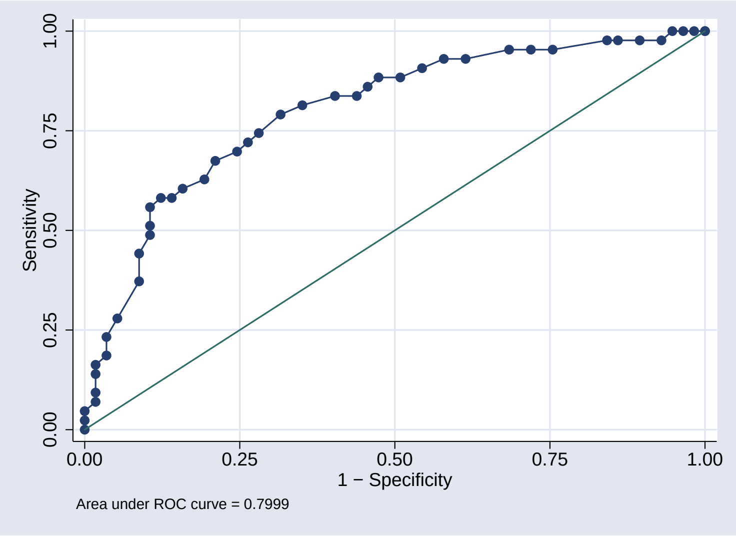
- Sensitivity and specificity can only be used with a single dichotomous classification.
- Logistic regression gives a probability, not a classification
- Can define your own threshold for use with logistic regression
- Commonly choose 50% probability of being a case
- Can choose any probability: sensitivity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

CENTRE FOR EPIDEMIOLOGY VERSUS ARTHRITIS

ROC Curves

- Shows how sensitivity varies with changing specificity
- Gives a measure of discrimination
- Larger area under the curve = better
- Maximum = 1
- Tossing a coin would give 0.5
- Command used is `lroc`

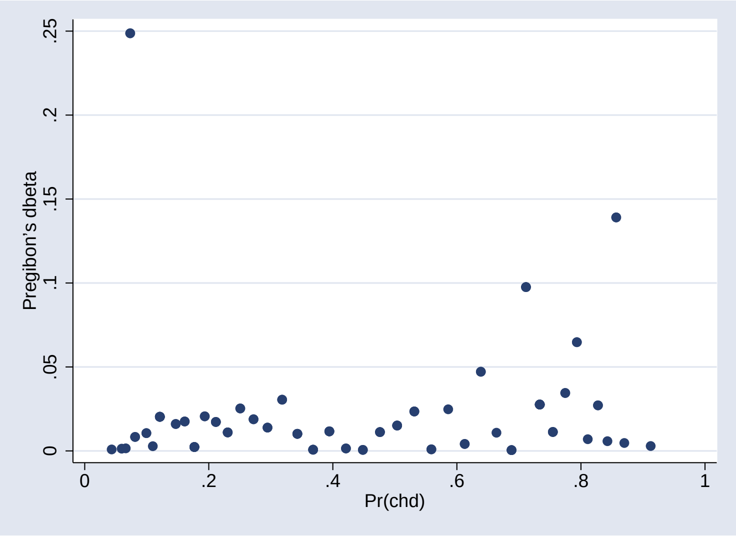
ROC Example


Sensitivity
0.00
0.25
0.50
0.75
1.00
0.00
0.25
0.50
0.75
1.00
1 − Specificity
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 $-\hat{\pi}$.
- Grouping by covariate pattern may help: observed outcome can now lie between 0 and 1 if multiple observations have same pattern
- 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 $\hat{\pi}$



Effect of removing influential observation

```
. logistic chd age if dbeta < 0.2

Logistic regression                               Number of obs   =         98
                                                  LR chi2(1)      =      32.12
                                                  Prob > chi2     =     0.0000
Log likelihood = -50.863658                       Pseudo R2       =     0.2400

------------------------------------------------------------------------------
         chd | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.130329   .0293066     4.73   0.000     1.074324    1.189254
------------------------------------------------------------------------------
```

Poorly fitted observations

- Can be identified by residuals
  - Deviance residuals: `predict` *`varname`*`, ddeviance`
  - χ^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 to calculate odds for that subgroup
- If you have lots of subgroups, this may not be true
- In which case, log(OR) for that group is $-\infty$ or ∞
- Stata will drop all subjects from that group (unless you use the option `asis`)
- Not a problem with continuous predictors