Binary and Nominal Dependent Variables

Applying MLE

Jan Rovny

February 5, 2023

Introduction

- Political scientists are often interested in binary or nominal outcomes:
  - Did a respondent vote or not?
  - Is a respondent employed or not?
  - Was there war between country A and B in 1967?
  - Is a respondent below the poverty line or not?
  - Which party did a respondent vote for?
- These outcomes cannot be operationalized as continuous variables and thus cannot be estimated using OLS.
- We must turn to MLE

Binary outcomes and OLS

Did a respondent vote in the last election?

- We could attempt to estimate this using OLS
  - $Pr(y = 1|\mathrm{x}) = \beta_0 + \mathrm{x}\boldsymbol{\beta}$
- But that that would violate OLS assumptions in a number of ways:

Binary outcomes and OLS

Did a respondent vote in the last election?

- We could attempt to estimate this using OLS
  - $Pr(y = 1|\mathrm{x}) = \beta_0 + \mathrm{x}\boldsymbol{\beta}$
- But that that would violate OLS assumptions in a number of ways:
  - It would be heteroscedastic

Binary outcomes and OLS

Did a respondent vote in the last election?

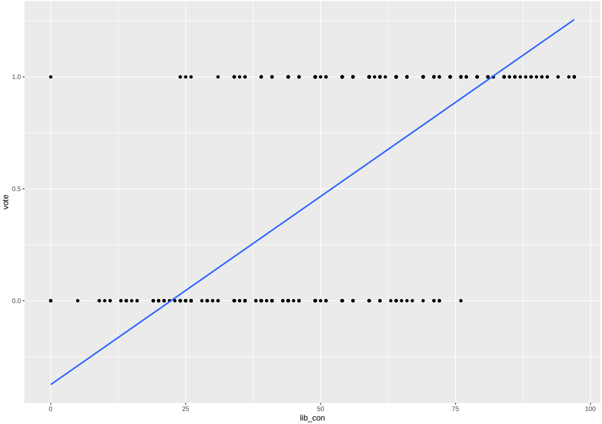
- We could attempt to estimate this using OLS
  - $Pr(y = 1|\mathrm{x}) = \beta_0 + \mathrm{x}\boldsymbol{\beta}$
- But that that would violate OLS assumptions in a number of ways:
  - It would be heteroscedastic
  - The probabilities would not be bounded by 0 and 1

Binary outcomes and OLS

Did a respondent vote in the last election?

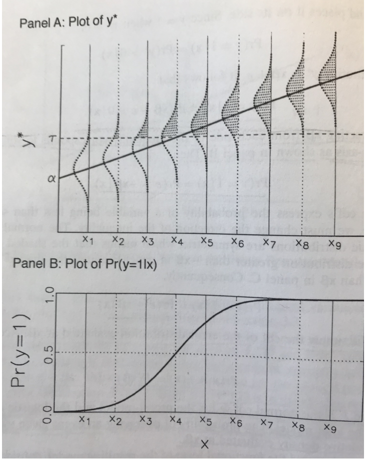
- We could attempt to estimate this using OLS
  - $Pr(y = 1|\mathrm{x}) = \beta_0 + \mathrm{x}\boldsymbol{\beta}$
- But that that would violate OLS assumptions in a number of ways:
  - It would be heteroscedastic
  - The probabilities would not be bounded by 0 and 1
  - It would predict a linear function (no diminishing marginal effects, poor prediction of middle cases)

OLS prediction of binary outcome

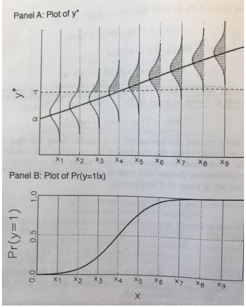


An alternative: Logit/Probit

- Imagine the binary outcome y_i as a manifestation of an unobserved continuous latent variable y_i*
- y_i* can be understood as propensity to choose $y = 1$



The Logic of Logit/Probit mathematically

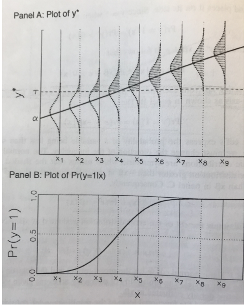


$$y_i = \begin{cases} 1 & \text{if } y_i* > \tau \\ 0 & \text{if } y_i* \leq \tau \end{cases}$$

- y_i* can be understood as a continuous function of $\boldsymbol{x}$ plus ϵ.
- Thus: $y* = \boldsymbol{x\beta} + \epsilon$
  (this is a normal linear function)
- if $\tau = 0$, then $y = 1$ when $y* > 0$.

- We can write that:
  $Pr(y = 1|\boldsymbol{x}) = Pr(\boldsymbol{x\beta} + \epsilon > 0|\boldsymbol{x})$
- If we subtract $\boldsymbol{x\beta}$ from both sides of the inequality, we get:
  $Pr(y = 1|\boldsymbol{x}) = Pr(\epsilon > -\boldsymbol{x\beta}|\boldsymbol{x})$
- Given the symmetry of the distributions ($p > -\boldsymbol{x\beta} = p \leq \boldsymbol{x\beta}$),
  Consequently:
  $Pr(y = 1|\boldsymbol{x}) = Pr(\epsilon \leq \boldsymbol{x\beta}|\boldsymbol{x})$

The Logic of Logit/Probit mathematically



What error distribution should we assume?

(What is the distribution of the error curves in Panel A above?)

- Logit $\epsilon \sim L(0, \pi^2/3)$
- Probit $\epsilon \sim N(0,1)$, then

$$Pr[y_i = 1] = Pr[\epsilon > -\mathbf{x_i}\beta] = Pr[\epsilon < \mathbf{x_i}\beta]^1 = F(\mathbf{x_i}\beta)$$

- where F is either standard logistic CDF (logit) or standard normal CDF (probit)

[1] This is because both logit and normal distributions are symmetrical. See Long p.45

Estimation

- Estimation of logit and probit requires MLE
- Assume that we have a sample of N independent observations
- We have $y = 1$ and $y = 0$, where 1s occur with probability π and 0s with probability $1 - \pi$
- The likelihood function is:

$$\mathcal{L} = \prod_{y_i=1} \pi \prod_{y_i=0} (1 - \pi) \qquad (1)$$

$$\mathcal{L} = \prod_{y_i=1} F(\mathbf{x_i}\beta) \prod_{y_i=0} [1 - F(\mathbf{x_i}\beta)] \qquad (2)$$

$$\mathcal{L} = \prod_{i=1}^{N} [F(\mathbf{x_i}\beta)]^{y_i} [1 - F(\mathbf{x_i}\beta)]^{1-y_i} \qquad (3)$$

- where F is either the standard logistic CDF (logit) or the standard normal CDF (probit)

Logit

- The logit function refers to log odds. That is, the logged odds of an outcome are:

$$ln\left(\frac{Pr(y=1)}{Pr(y=0)}\right) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + \epsilon$$

- This can be written as:

$$Pr(y=1|x_1, x_2, ...x_k) = \frac{e^{(\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_k x_k)}}{1+e^{(\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_k x_k)}} =$$

$$= \frac{1}{1+exp(-\boldsymbol{x_i}\boldsymbol{\beta})}$$

Probit

- An alternative to logit is probit

$$Pr(y = 1|x_1, x_2, ...x_k) = \Phi(x_1, x_2, ...x_k)$$

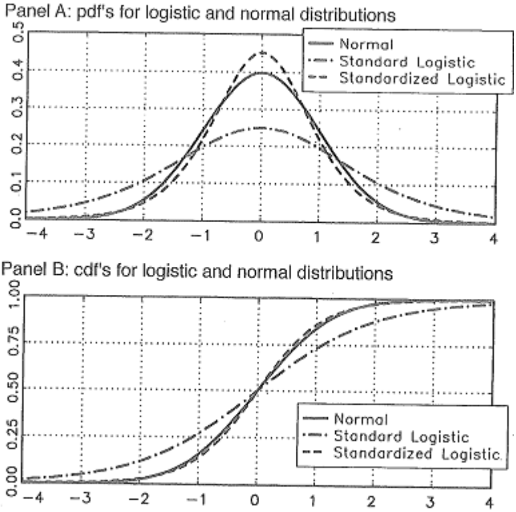
- here $\Phi(\cdot)$ is the cumulative distribution function (cdf) of the normal distribution, so

$$Pr(y = 1|x_1, x_2, ...x_k) = G(x\beta)$$

- where

$$G(x\beta) = \int_{-\infty}^{x\beta} \frac{1}{\sqrt{2\pi}} exp(-\frac{\nu^2}{2}) d\nu$$

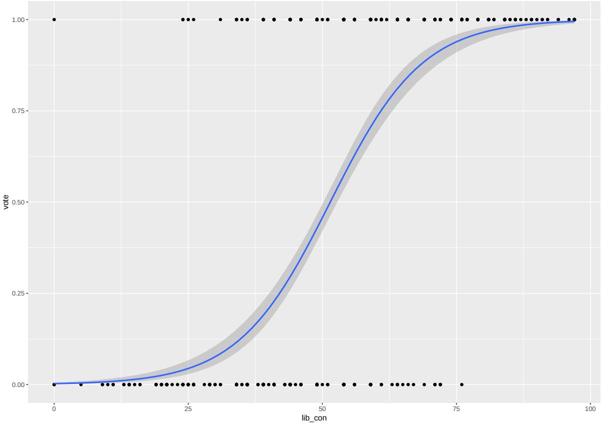
PDFs and CDFs



Logit and Probit

- Logit v. Probit:
  - The differences between logit and probit are minior
  - The main difference is in their computation, where logit is easier.
  - But this has been erased with computer power.
- Logit and Probit have desirable properties:
  - Have constant error variance (by definition) logit $\pi^2/3$, probit 1
  - Their predictions are bounded between 0 and 1
  - Follow an S-shape

Logit and Probit



Example

- The dependent variable inlf is coded 1 or 0 for whether a woman is in the labour force or not. (Data: Mroz.dta)
- The predictors are:

Variable	Description	Mean
nwifeinc	non-wife income	20.13
educ	education	12.29
exper	work experience in years	10.63
expersq	squared work experience	178.04
age	age in years	42.54
kidslt6	number of children $<$ 6 years old	0.24
kidsge6	number of children $\geq$ 6 years old	1.35

In R

```
logit<-glm(inlf~nwifeinc+educ+exper+expersq+age+kidslt6+kidsge6, data=D, family = binomial(link=logit))
summary(logit)

Call:
glm(formula = inlf ~ nwifeinc + educ + exper + expersq + age +
    kidslt6 + kidsge6, family = binomial(link = logit), data = D)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1770  -0.9063   0.4473   0.8561   2.4032
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.425452   0.860365   0.495  0.62095
nwifeinc    -0.021345   0.008421  -2.535  0.01126 *
educ         0.221170   0.043439   5.091 3.55e-07 ***
exper        0.205870   0.032057   6.422 1.34e-10 ***
expersq     -0.003154   0.001016  -3.104  0.00191 **
age         -0.088024   0.014573  -6.040 1.54e-09 ***
kidslt6     -1.443354   0.203583  -7.090 1.34e-12 ***
kidsge6      0.060112   0.074789   0.804  0.42154
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1029.75  on 752  degrees of freedom
Residual deviance:  803.53  on 745  degrees of freedom
AIC: 819.53

Number of Fisher Scoring iterations: 4
```

Interpreting logit coefficients

- Cannot interpret βs in the same way as in OLS! They are not linear!
- However, the logit – the **log** odds – are written linearly:

$$ln\left(\frac{Pr(y=1)}{Pr(y=0)}\right) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + \epsilon$$

- We can thus interpret this as:
  "for a unit change in x_k, the logit changes by β_k, all else constant"
- The problem is that we do not have an intuitive sense of what the logit is...

Odds ratios

- We can, however, exponentiate both sides of the equation:

$$exp(ln\left(\frac{Pr(y=1)}{Pr(y=0)}\right)) = exp(\beta_0 + \beta_1 x_1 + ... + \beta_k x_k + \epsilon)$$

$$\left(\frac{Pr(y=1)}{Pr(y=0)}\right) = exp(\beta_0 + \beta_1 x_1 + ... + \beta_k x_k + \epsilon)$$

- Now we can read this as: "*a unit change in x_k changes the odds by a factor of $exp(\beta_k)$*"
- Odds are centered around 1, values < 1 suggests decreasing effect
- values > 1 suggest increasing effect

Odds ratios

- To assess the effect of a variable in terms of odds ratios as x changes by δ units:

$$exp(\beta_k * \delta) - 1$$

- Multiply by 100 to get percentage change:

$$100(exp(\beta_k * \delta) - 1)$$

- This will tell you the *"percentage change in the likelihood of $y = 1$ as x_k changes by δ units"*

Odds ratios – example

From the coefficients of the model above:

nwifeinc	educ	exper	expersq	age	kidslt6	kidsge6
-0.021	0.221	0.205	-0.003	-0.088	-1.443	0.060

- For each additional small child (< 6 years), the likelihood of a woman working decreases by 76 percent.

$$100(exp(-1.443 * 1) - 1) = -76.378$$

- For additional 5 years of education, the likelihood of a woman working increases by 202 percent – she is twice as likely to work.

$$100(exp(0.221 * 5) - 1) = 201.922$$

Predicted probabilities

- The best way to assess effects in logit/probit models is to calculate predicted probabilities:

$$Pr(y = 1|x) = \frac{exp(x\beta)}{1 + exp(x\beta)} =$$

$$= \frac{exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k)}{1 + exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k)}$$

- We generally want to assess the effects of key variables, such as x_1
- To keep the *ceteris paribus* condition, we maintain other (control) variables at some constant value
  - For continuous variables, most usually, the mean.
  - For categorical variables, usually the mode.

Predicted probabilities example

- From the data above, assess the probability that a woman is in the labour force:
  - as a function of her education,
  - and of the number of small children she has,
  - while other variables are held constant.
- This means:
  - setting all other variables at some constant values,
  - while varying education and number of small children from their min to max,
  - and then calculating the predicted probabilities (using the equation above).

[see demonstration in R]

Goodness of Fit

- Logit and Probit obviously cannot estimate an R^2
- One alternative is any of a number of pseudo R^2 measures, mostly based on the log-likelihood: $1 - \frac{L_{ur}}{L_r}$
  - McFadden's R^2 in R: `pR2()` `{pscl}`
- A better alternative: share of observations correctly predicted
  - Each value has a predicted probability of scoring 1
  - Assign each observation with $pp \geq 0.5$ to 1, otherwise 0
  - Compare the predicted ones and zeros to the actual reported outcomes.
  - Best to report both the percentage of negatives and positives correctly predicted

[see demonstration in R]

For more information on model fit, see here