POL 213 – Spring 2024

Lecture 6

Binary Choice Models I - Logistic Regression

Lauren Peritz

U.C. Davis

lperitz@ucdavis.edu

May 15, 2024

Table of Contents

Categorical Data in Political Science

Problems with OLS for Categorical DV

Deeper Dive on Motivations

Logistic Regression Model

Categorical Variables

- Categorical variables have a measurement scale that consists of a set of categories
  - Ordinal: Partisanship (strong Democrat, not very strong Democrat, lean Democrat, Independent, lean Republican, not very strong Republican, strong Republican); Presidential approval (strongly approve, somewhat approve, somewhat disapprove, strongly disapprove)
  - Nominal: Democratic regimes (president-parliamentary, premier-presidential, parliamentary, assembly-independent, presidential), Militarized interstate disputes (yes, no), Vote choice (Biden, Trump, Other)
- Main subject of this course is analysis of categorical response or outcome variables
- More than the hard sciences and economics, categorical response variables are ubiquitous in political science

Categorical Variables

- In experiments, researchers manipulate explanatory variables (i.e., features of a social system) and produce responses
- Useful to think of a response as a random variable, even where we cannot manipulate a social system to produce it
- A random variable, Y_i is the assignment of numbers to events in a sample space
    - Y_i is random in that its value will vary across a large number of hypothetical realizations even where explanatory variables do not
- The data, y, are the n observed values of Y_i with each y_i a draw from the random variable Y_i

Table of Contents

OLS for Categorical Dependent Variables

- The output of the classical linear model is a conditional average, $\mathbb{E}(Y|x_i)$. Does this make sense for a dichotomous response? The average for a dichotomous response is a score between 0 and 1 realized by no individual in the population
- For a dichotomous variable, $\mathbb{E}(Y|x_i)$ is simply the proportion of 1s at various values of x_i, or π_i

$$\pi_i \equiv \Pr(Y_i) \equiv \Pr(Y = 1|X = x_i)$$

- We could use nonparametric regression to estimate the conditional proportion for Y at each value of X, but let's try using OLS for a Chilean voters (John Fox's supplementary notes).

Linear Probability Model

Using R, we plot voting intention for Chilean voters against support for the status quo in 1988 election to oust General Pinochet and return country to civilian governance.

```
plot(jitter(vote2, .25) ~ statusquo, mydata,
xlab="Support for Status Quo",
ylab="Voting Intention")
abline(h = 1, lty = 2)
abline(h = 0, lty = 2)
```

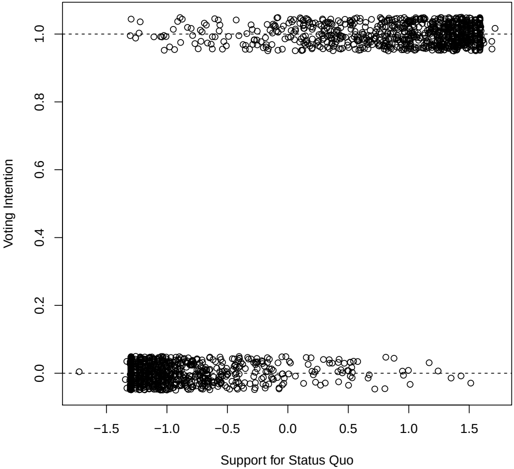
Under the usual assumptions, the linear regression model is:

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

where $\varepsilon_i \sim N(0, \sigma^2)$, ε_i and ε_j are independent for $i \neq j$, and X is independent of ε. Then, we have:

$$E(Y_i) = \pi_i = \alpha + \beta X_i$$

Problems with OLS for Categorical DV



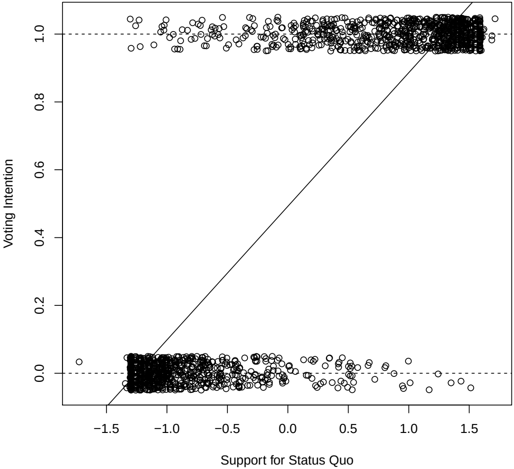
Linear Probability Model

Using the `lm()` command, we estimate a linear model and then use `fitted.values()` to generate fitted values

```
> ols.chile <- lm(vote2 ~ statusquo, mydata)
> summary(ols.chile)
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.492167   0.006203   79.35   <2e-16 ***
statusquo   0.394079   0.005721   68.89   <2e-16 ***

Residual standard error: 0.2598 on 1752 degrees of freedom
Multiple R-squared: 0.7303,     Adjusted R-squared: 0.7302

> yhat.ols <- fitted.values(ols.chile)
> summary(yhat.ols)

Min.   1st Qu.   Median     Mean   3rd Qu.     Max.
-0.18800  0.06449  0.42560  0.49370  0.95420  1.16700
```

The fitted values reveal something is wrong: predictions lie outside the $(0, 1)$ interval

Problems with OLS for Categorical DV



Linear Probability Model

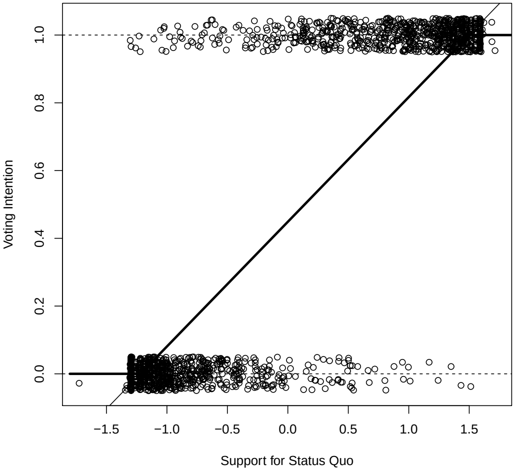
We could constrain π_i to the unit interval, allowing the linear model to characterize π_i within this interval:

$$f(x) = \begin{cases} 0 & \mathit{for} \quad 0 > \alpha + \beta X \\ \alpha + \beta X & \mathit{for} \quad 0 \leq \alpha + \beta X \leq 1 \\ 1 & \mathit{for} \quad \alpha + \beta X > 1 \end{cases}$$

```
segments(min(x), 0, -1.2190848, 0, lty = 1, lwd = 3)
segments(-1.2190848, 0, 1.4973026, 1, lty = 1, lwd = 3)
segments(1.4973026, 1, max(x), 1, lty = 1, lwd = 3)
```

The logic of this model makes some sense, although the slope of the regression line is now discontinuous.

Problems with OLS for Categorical DV



Linear Probability Model

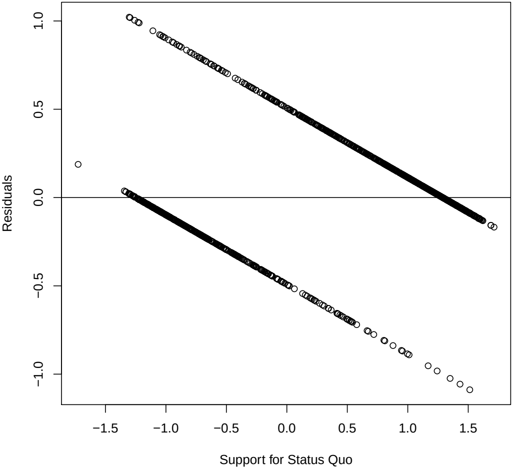
But the linear probability model has other problems that violate OLS assumptions The errors are heteroskedastic, which a plot of residuals shows.

```
resid.ols <- resid(ols.chile)
plot(mydata$statusquo, resid.ols, ylab = "Residuals",
xlab = "Support for Status Quo")
abline(0, 0)
```

Specifically, because *Yi* is either 0 or 1, the conditional distribution of ε_i is dichotomous:

- $Y_i = 1: \qquad \varepsilon_i = 1 - \mathbb{E}(Y_i) = 1 - (\alpha + \beta X)$
- $Y_i = 0: \qquad \varepsilon_i = 1 - \mathbb{E}(Y_i) = 0 - (\alpha + \beta X)$

Problems with OLS for Categorical DV



Linear Probability Model

Plotting the squared residuals suggests another problem

```
sqresid.ols <- resid.ols * resid.ols
plot(mydata$statusquo, sqresid.ols, ylab = "Squared Residuals",
  xlab = "Support for Status Quo")
```

The variance of ε cannot be constant.
$\mathbb{V}(\varepsilon_i) = \pi_i(1-\pi_i)^2 + (1-\pi_i)(-\pi_i)^2 = \pi_i(1-\pi_i)$

The variance depends on π_i itself a function of the unknown parameters α and β. This makes estimation complex, with no guarantees of constraining the fitted values to the unit interval.

Problems with OLS for Categorical DV

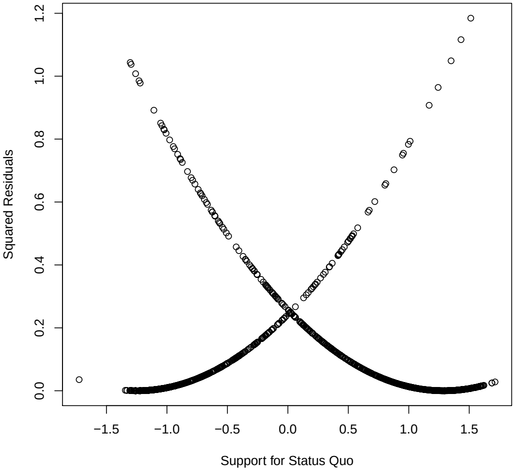


Table of Contents

Categorical Data in Political Science

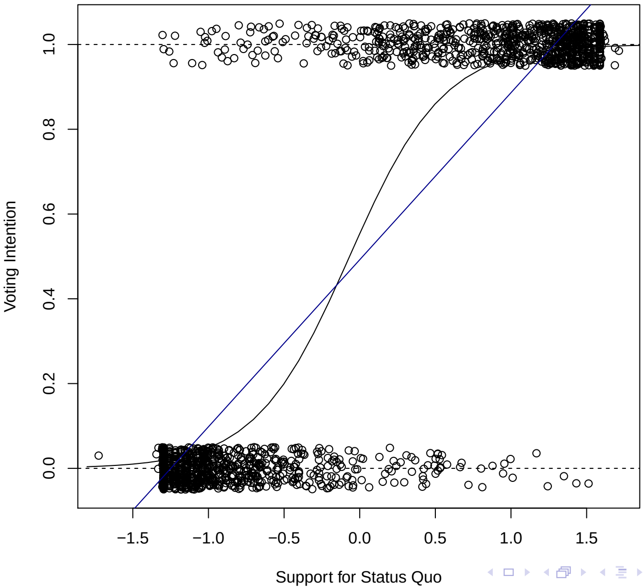
Problems with OLS for Categorical DV

Deeper Dive on Motivations

Logistic Regression Model

Motivating binary choice model

- Pathologies of the linear probability model justify our use of the logit model:
  - OLS failed to constrain π to the unit interval
  - Constant variance, other OLS assumptions untenable
- We used the cdf of the logistic distribution to transform and map the linear predictor into the $[0, 1]$ interval.


Voting Intention
0.0
0.2
0.4
0.6
0.8
1.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
Support for Status Quo

Motivating binary choice model

- The logit fit to the Chilean data tends asymptotically toward 0 at low values and 1 at high values of support for the status quo
    - The fitted values never reach 0 or 1; the model never predicts with certainty
- As the S-shaped curve implies, the logit model is nonlinear in probabilities
    - The rate of change in π depends on the value of x

Motivating binary choice model

- An alternative derivation of the logit (or probit) model posits an underlying regression for a continuous, unobservable response variable ξ, such that:

$$Y_i = \begin{cases} 0 & \text{when} \quad \xi_i \leq 0 \\ 1 & \text{when} \quad \xi_i > 0 \end{cases}$$

- When ξ_i crosses 0, the observed response Y_i changes from 0 to 1. In the Chilean example, ξ_i is the propensity to vote for the plebiscite and Y_i changes from “no” to “yes”.
- We assume this latent variable is a linear function of the explanatory variable and an unobservable error term: $\xi_i = \alpha + \beta X_i - \varepsilon_i$
- We want to estimate α and β but cannot via least square regression because we do not observe the latent response.

Motivating binary choice model

- For a dichotomous variable, recall that $\mathbb{E}(Y|x_i)$ is simply the proportion of 1's at various values of x_i (denoted π_i). Using equations above:

$$\pi_i \equiv \Pr(Y = 1) = \Pr(\xi_i > 0) = \Pr(\alpha+\beta X_i-\varepsilon_i > 0) = \Pr(\varepsilon_i < \alpha+\beta X_i)$$

- Can we determine this probability? If ε_i follows some distribution, we can integrate over that distribution and get the probability that ε_i falls below some value $\alpha + \beta X_i$.
- If the errors ε_i follow the logistic distribution, we get the logit model:

$$\pi_i = \Pr(\varepsilon_i < \alpha + \beta X_i) = \Lambda(\alpha + \beta X_i)$$

Motivating binary choice model

If the errors follow the standard normal distribution ε_i $N(0,1)$, we get the probit model:

$$\pi_i = \Pr(\varepsilon_i < \alpha + \beta X_i) = \Phi(\alpha + \beta X_i)$$

Do this assumption about the error term distribution make sense?

- Since we cannot observe ε_i, we need to make assumptions.
- The normal has appeal due to the CLT; we still fix the variance at 1.
- If we assume the error term follows a **normal distribution**, we get the Probit model.
- The logistic is bell-shaped, symmetric, and similar to the normal.
- The logistic has nice qualities: we can usually ensure the error distribution has a form we want by transforming ξ to make the assumptions are true.

We could make other assumptions... maximum likelihood estimation (MLE) provides a unifying framework for this.

Motivating binary choice model

- Having assumed that ε_i follows a distribution, say the logistic, we can work out $\pi_i \equiv \Pr(Y = 1)$

$$\pi_i = \Lambda(\alpha + \beta X_i) = \frac{1}{1 + exp(-(\alpha + \beta X_i))} = \frac{exp(\alpha + \beta X_i)}{1 + exp(\alpha + \beta X_i)}$$

- Equivalently:

$$\log \frac{\pi_i}{1 - \pi_i} = \alpha + \beta X_i$$

- For a response variable Y_i that takes on two values $\{0, \ 1\}$ with probability π_i and $(1 - \pi_i)$, we can summarize the probability distribution for Y_i simply as:

$$p(y_i) \equiv \Pr(Y_i = y_i) = \pi_i^{y_i}(1 - \pi_i)^{1 - y_i}$$

Mathematics of Logistic Regression Model

- For a sample of N independent observations, the joint probability for the data can be written as:

$$p(y_1, y_2, ..., y_N) = \Pr(Y|\pi) = \prod_{i=1}^{N} \pi_i^{y_i}(1 - \pi_i)^{1-y_i}$$

- substituting from the logit model:

$$\Pr(Y|\alpha, \beta) = \prod_{i=1}^{N}(\frac{1}{1 + exp(-(\alpha + \beta X_i))})^{y_i}(1 - \frac{1}{1 + exp(-(\alpha + \beta X_i))})^{1-y_i}$$

$$= \prod_{i=1}^{N}(1 + exp(-(\alpha + \beta X_i)))^{-y_i}(1 + exp(\alpha + \beta X_i))^{y_i - 1}$$

Mathematics of Logistic Regression Model

- Thinking of this equation as a function of the parameters while treating the data $(y_1, y_2, ..., y_N)$ as fixed gives us the likelihood function for the logit model $L(\alpha, \beta|Y)$ which is proportional to $\Pr(Y|\alpha, \beta)$
- The log-likelihood can be written as:

$$\ln L(\alpha, \beta|y) = \sum_{i=1}^{N} [-y_i \ln[1 + exp(-\alpha - \beta X_i)] - (1 - y_i) \ln[1 + exp(\alpha + \beta X_i)]$$

- Generalizing the logit model to several explanatory variables is straightforward.
  - The Xs can be quantitative variables, transformations of these, dummies for qualitative variables, and/or interactions.
- This log-likelihood function is differentiable, but solving for α and β is not possible through analytical methods.

Table of Contents

Categorical Data in Political Science

Problems with OLS for Categorical DV

Deeper Dive on Motivations

Logistic Regression Model

Transforming the Linear Model

- To constrain π_i to the $(0, 1)$ interval and correct the problems with the linear probability model, we need a positive, monotone function that maps the linear predictor $\eta = \alpha + \beta X$ into the unit interval
- Any cumulative probability density function (cdf) will satisfy this requirement
- Recall that the cdf of a random variable X or its distribution (pdf) evaluated at *x* returns the probability that *X* has a value less than or equal to *x*

$$F(x) = \Pr(X \leq x) = \int_{x_{min}}^{x} f(x)dx$$

- Recall, all pdfs must integrate to 1 over the range of their support $(-\infty, \infty)$

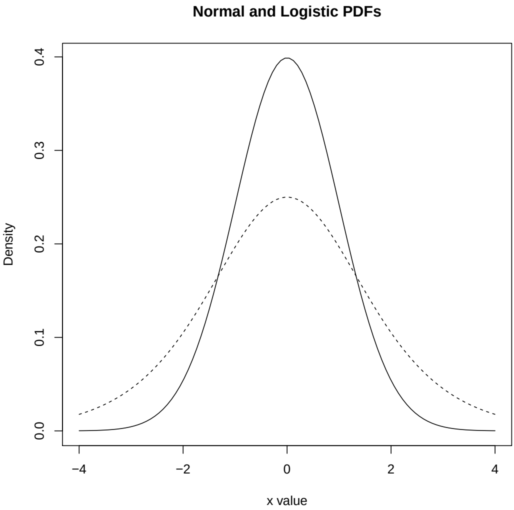
Transforming the Linear Model

The most common function for transforming the linear predictor $\eta = \alpha + \beta X$ is the cdf of the logistic distribution

```
# Plot normal, logistic pdf, cdf
x <- seq(-4,4, length=100)
nx <- dnorm(x)
lx <- dlogis(x)

plot(x, nx, type="l", lty=1, xlab="x value",
ylab="Density", main="Normal and Logistic PDFs"); lines(x,lx, lty=2)
```

The logistic resembles the normal (although the standard normal has smaller variance) and its cdf has a simple functional form:

$$\Lambda(z) = \frac{1}{1 + e^{-z}}$$


Normal and Logistic PDFs
Density
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4
x value

Transforming the Linear Model

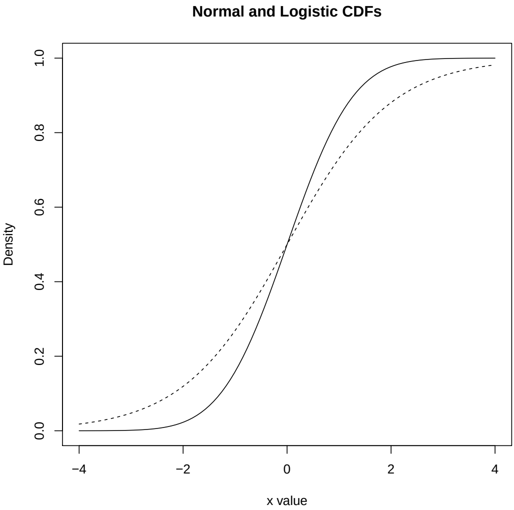
```
ncx <- pnorm(x); lcx <- plogis(x)
plot(x, ncx, type="l", lty=1, xlab="x value",
ylab="Density", main="Normal and Logistic CDFs");
lines(x,lcx, lty=2)
```

Using a logistic transformation yields the familiar logistic model for π_i

$$\pi_i = \Lambda(\alpha + \beta X_i) = \frac{1}{1 + \exp\left(-\alpha - \beta X_i\right)}$$

The logistic is symmetric around 0 and unbounded above and below. Large values of $(\alpha + \beta X_i)$ still map onto $(0, 1)$

Logistic Regression Model



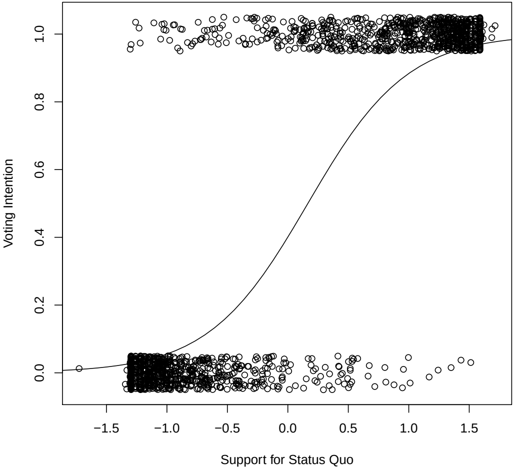
Transforming the Linear Model

We can use the `lm()` command in R to run the logistic regression model by specifying the binomial distribution.

```
> logit.chile <- glm(vote2 ~ statusquo, mydata, family = binomial)
> summary(logit.chile)
Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.21531    0.09964   2.161   0.0307 *
statusquo    3.20554    0.14310  22.401   <2e-16 ***
Null deviance: 2431.28  on 1753  degrees of freedom
Residual deviance:  752.59  on 1752  degrees of freedom
AIC: 756.59

> yhat.logit <- fitted.values(logit.chile)
> summary(yhat.logit)
Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
0.004882 0.036840 0.419100 0.493700 0.981500 0.996700
```

The fitted values are probabilities that are non-linear and range between $(0, 1)$ without reaching the endpoints.

Logistic Regression Model



Transforming the Linear Model

- The logistic regression model does not resemble the linear OLS model, but it is linear in its predictors: $\eta = \alpha + \beta X$
- Transforming the model using the logistic distribution solved the main issues with the linear probability model, particularly by constraining π_i to the unit interval
- The logistic regression model has other nice features. For one, the inverse transformation of the model, $\Lambda'(\pi_i)$, is directly interpretable as a log-odds

$$\frac{\pi_i}{1-\pi_i} = \exp\left(\alpha + \beta X_i\right)$$

  with ${}^{\pi_i}/{}_{1-\pi_i}$ the odds that $Y_i = 1$

- Using the coefficients of the logit model above, we can calculate the odds a voter at the median of support for the status quo will vote yes

$$\frac{\pi_i}{1-\pi_i} = \exp\left(0.21531 + 3.2055 * 0.00396\right) = 1.2560$$

- The odds that the median voter votes yes vs. no on the plebiscite, i.e., to sustain the Pinochet military government, are 1.26 to 1

Next...

We will discuss the interpretation of logistic regression and work on some examples in R.