Logistic Regression

POSC 3410 – Quantitative Methods in Political Science

Steven V. Miller

Department of Political Science

Goal for Today

Discuss logistic regression for binary variables.

Going Further with Applied Statistics

We are now answering our own questions in political science.

- We already know about concepts, measures, and variables.
- We believe variation in y can be attributed to variation in x.
- After controlling for rival explanations (z), our linear regression produces a partial effect of x on y.

Linear regression (OLS) draws lines of "best fit" through the data.

- "Best fit": minimizes the sum of squared differences (hence: OLS).

Limitations with OLS

OLS regression has a lot of nice properties. Use it if you can.

- However, *any* model is useful only when the assumptions of the model are ment.

When y is not interval, OLS will not suffice.

- Prediction errors are not constant (i.e. **heteroscedasticity**).
- Further, they follow a binomial (and not normal) distribution.

Limitations with OLS

Substantively, regression coefficients become misleading.

- Recall: OLS coefficients assume constant linear effects of x on y.
- When we have only 0s and 1s, linear effects are not immediately intuitive.

Logistic Regression

We will deal with the problem of binary DVs with **logistic regression**.

- This tells us the effect of a unit change in x on the *natural logged odds of y.*

We'll start with an understanding of what "natural logged odds of *y*" mean.

Odds

You typically hear of **odds** in the world of sports betting.

- It's closely linked with probability

Given some probability p of an event occurring, the odds of the event equal:

$$\text{Odds} = \frac{p}{1-p}$$

Ever hear of something like "the odds are 4 to 1 against" an event occurring?

- Translation: for every five trials, we expect 1 occurrence to 4 non-occurrences, on average.

Education and Turnout

Let's look at hypothetical education-turnout data.

Table 9-1 Education and the Probability of Voting

	Education					
Did respondent vote?	*0. Low*	*1. Middle-low*	*2. Middle*	*3. Middle-high*	*4. High*	Total
1. Yes, voted	6	20	50	80	94	250
0. No, did not vote	94	80	50	20	6	250
Total (n)	100	100	100	100	100	500
Probability of voting	.06	.20	.50	.80	.94	.50

Note: Hypothetical data.

Education and Turnout

You're obviously seeing a positive relationship.

- i.e. more educated people are more likely to vote.

You're also seeing the probability of non-linearity in discrete DVs.

- The effect of 0 to 1 in x is a change of .14 in the probability of voting.
- From 1 to 2 in x: change of .30.
- From 2 to 3 in x: change of .30 again.
- From 3 to 4 in x: change of .14.

Think of the issue as analogous to a "tipping point".

Visualizing Odds

Now, let's look at the odds of voting.

Table 9-2 Probability of Voting, Odds of Voting, and Logged Odds of Voting at Five Levels of Education

Education (x)	Probability of voting (y)	Odds of voting (y)	Logged odds of voting (y)
0. Low	.06	.06/.94 = .06	–2.8
1. Middle-low	.20	.20/.80 = .25	–1.4
2. Middle	.50	.50/.50 = 1	0
3. Middle-high	.80	.80/.20 = 4	+1.4
4. High	.94	.94/.06 = 16	+2.8

Note: Hypothetical data.

Visualizing Odds

The middle column, odds of voting, translates probabilities to odds.

- e.g. $\frac{p}{1-p}$ when $x = 0 = \frac{.06}{.94} = .06382979$.
- Once we get to the middle education category, the odds become integers.
    - When the odds are 1, we expect one voter for every non-voter.

Odds Ratio

How can we use just odds to answer the question we have of how x affects y?

- One preliminary answer is the **odds ratio**.

Odds and Odds Ratios

Take a look at Table 9-2.

- Odds of voting in low education category: .06.
- Odds of voting in middle-low education category: .25.

The odds of voting for the middle-low category is more than four times the odds of voting for the low category.

- $\frac{.25}{.06} = 4.1\overline{6}$
- Do this for all other values and the odds ratio is four each time.

$$\text{Odds ratio} = \frac{1}{.25} = \frac{4}{1} = \frac{16}{4} = 4$$

Percentage Change in Odds

We can also calculate the **percentage change in odds**.

Percentage Change in Odds

Consider, again, the odds of voting in the bottom two categories.

- Calculate the unit increase (here: .25 - .06 = .19).
- Divide that over the odds of the lower value (here: .06).
- This gets you a value of $3.1\overline{6}$.
- Multiply that by 100 to get a percentage change.

If we did that for all other values, we'd get values of 3 (or 300%).

- Translation: the odds of voting increase 300% for each unit increase in education.

Logits (Natural Logged Odds of y)

We have seen that each unit change in x does not solicit a consistent change in y.

- However, the effect of change in the odds ratio and percentage change in odds is consistent.
- The next step is to take the natural logarithmic transformation of the odds, or **logit**.

Natural Logarithmic Transformation

The key term here is "*natural* logarithmic transformation".

- Contrast this with a base-10 algorithm engineers commonly use.
- In calculus, the natural logarithm with base e is much more common.

Natural Logarithmic Transformation

e is an irrational number with an interesting history.

- The Pythagoreans put one of their own (Hippasus) to death for postulating the existence of an irrational number like e.

Jacob Bernoulli touched on it in his discovery of the limit of the now famous compound interest formula.

A Question

$$f(x) = (1 + \frac{1}{x})^x$$

What happens to this formula when x goes to infinity?

- *Note*: this was when compound interest was calculated continuously rather than at set intervals.

Natural Logarithmic Transformation

When x goes to infinity, the exponent goes to infinity.

- However, the denominator does as well.

Meaning: you'd be taking an exponential of infinity for a value close to 1, which would result in (basically) 1.

- Bernoulli discovered the limit must be between 2 and 3.

Leonhard Euler proposed the answer is e (an irrational number) and can be denoted as $e = 2.7182818284$, approximately.

Natural Logarithmic Transformation

Take the natural log for our odds of y. Revisit Table 9-2.

Table 9-2 Probability of Voting, Odds of Voting, and Logged Odds of Voting at Five Levels of Education

Education (x)	Probability of voting (y)	Odds of voting (y)	Logged odds of voting (y)
0. Low	.06	.06/.94 = .06	–2.8
1. Middle-low	.20	.20/.80 = .25	–1.4
2. Middle	.50	.50/.50 = 1	0
3. Middle-high	.80	.80/.20 = 4	+1.4
4. High	.94	.94/.06 = 16	+2.8

Note: Hypothetical data.

Logistic Regression

Our y is not simply 0s and 1s now, but logit functions applied to the odds of 0s and 1s for all values of x. Formally:

$$\text{Logged odds of } y = \hat{a} + \hat{b}(x)$$

What would this look like in our simple case?

Logistic Regression

$$\text{Logged odds of voting} = -2.8 + 1.4(x)$$

Recall:

- $\hat{a}$ is our estimate of the logged odds of y when $x = 0$ (thus: -2.8).
- 1.4 is our $\hat{b}$ we observe from the right column from Table 9-2.

Interpreting a Logistic Regression

Saying "each unit increase in x leads to a 1.4 increase in the logged odds of y" is the correct interpretation.

- It's also not that intuitive.

How do we get more digestible, substantive results?

- Simple: start reversing your tracks.

Interpreting a Logistic Regression.

"Un-log" (i.e. exponentiate) your regression coefficient.

$$\text{Exp}(\hat{b}) = \text{Exp}(1.4) = e^{1.4} = 4$$

Does this look familiar?

Interpreting a Logistic Regression

It's the odds ratio.

- Recall: your regression coefficient is the estimate of effect size from one unit to the next highest, across the range of x.

Interpreting a Logistic Regression

We can also get the percentage change in odds.

$$\text{Percentage change in odds of } y = 100 * (\text{Exp}(\hat{b}) - 1)$$

With this data, this is unsurprisingly 300. Each unit increase in x (here: education) increases the odds of voting by 300 percent.

Interpreting a Logistic Regression

We can also get probabilities too (say: when $x = 0$).

$$\text{Probability} = \frac{\text{Odds}}{1 + \text{Odds}} = \frac{e^{-2.8}}{1 + e^{-2.8}} = .06$$

Using Actual Data

What would this look like with actual data? See Table 9-3:

- y: did respondent vote in 1996 Presidential election.
- x: education in years (0:20).

Education and Voting

Table 9-3 Education and Voting: Logistic Regression Coefficients and Related Statistics

Logged odds (voting)	=	Intercept $\hat{a}$	+	Education $\hat{b}$
Coefficient estimate		–1.581		.180
Standard error				.016
Wald				128.85
Significance				.000
Exp(b)				1.197

Source: James A. Davis, Tom W. Smith, and Peter V. Marsden, General Social Surveys, 1972–2002 (Chicago: National Opinion Research Center [producer], 2003; Storrs, Conn: Roper Center for Public Opinion Research, University of Connecticut/Ann Arbor: Inter-university Consortium for Political and Social Research [distributors], 2003).

Note: Displayed data are from the 1998 General Social Survey. $N = 2{,}605$. The dependent variable is reported turnout in the 1996 presidential election. The independent variable is number of years of formal schooling.

Interpreting Table 9-3

The coefficient is statistically significant. How else do we interpret it?

- Exp(.180) = $e^{.180}$ = 1.197. This is the odds ratio.
- 100*($e^{.180}$ - 1) = 19.7%. This is the percentage change in the odds of voting.

Our intercept ($\hat{a}$ = -1.581) is meaningful too.

- Predicted probability of a person voting who was never educated is .170.

Conclusion

Binary DVs violate the assumptions of OLS and produce misleading estimates.

- This leads us to logistic regression.
- The process of inference is the same, but the coefficients communicate something a bit different.
- It's the same old regression, just on a transformed DV.

Computers do heavy lifting for us, but it's important to understand what the computer is doing here.

Table of Contents