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.

R Packages for Today

```
library(tidyverse) # for most things
library(stevemisc) # for %nin% and formatting
library(stevedata) # for ?election_turnout and ?gss_abortion
library(modelsummary) # for tables
library(kableExtra) # for prettying up tables
```

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).

OLS

OLS has a ton of nice properties.

- Best linear unbiased estimator (BLUE)
- Simple to execute and interpret.

It'd be a *shame* if something were to happen to one of your assumptions.

The Problem of Binary DVs

The biggest problem you'll encounter will concern your DV.

- OLS assumes the DV is distributed normally.

You'll most often encounter DVs that are binary.

- Candidate won/lost.
- Citizen voted/did not vote.
- Program succeeded/failed.
- War happened/did not happen.

Most social/political phenomena are typically "there"/"not there."

The Implications of OLS When Assumptions are Violated

1. Your errors will be **heteroskedastic.**
2. Your $\hat{y}$s will often make no sense.

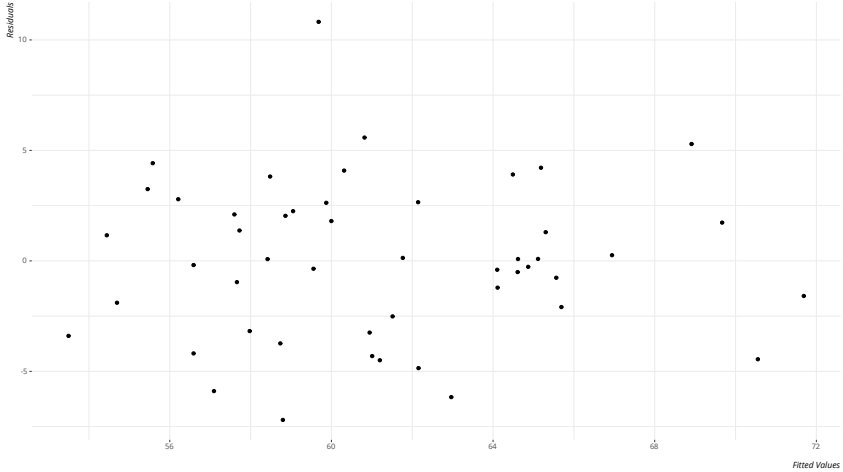
A Tale of Two Regressions

```r
# what was the turnout?
# We'll omit Hawaii and DC for being odd observations here.
M1 <- lm(turnoutho ~ percoled + ss,
         data=subset(election_turnout, state %nin% c("Hawaii", "District of Columbia")))

# did Trump win (1) or not (0)?
M2 <- lm(trumpw ~ percoled + ss,
         data=subset(election_turnout, state %nin% c("Hawaii", "District of Columbia")))
```

Here's What A Fitted-Residual Plot Should Look Like

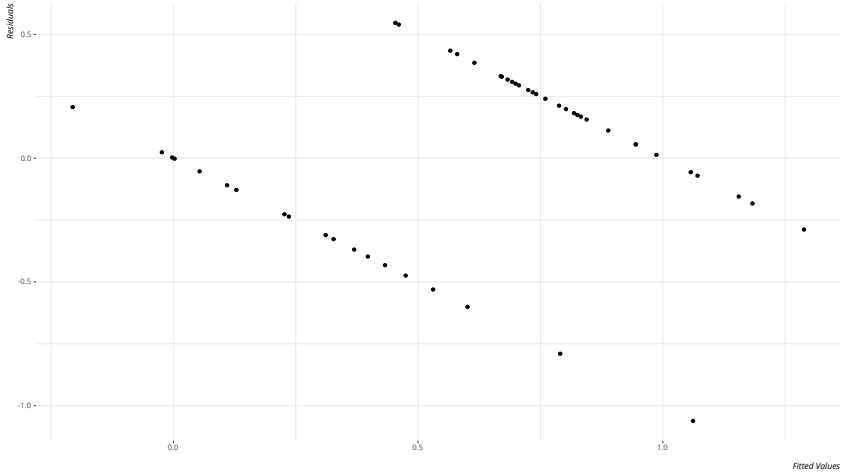
The variation between what we estimate (fit) and the error from it (residuals) is effectively normal.



Data: election_turnout in stevedata. A simple linear model regressing turnout on college education and swing state.

No Fitted Residual Plot Should Look Like This

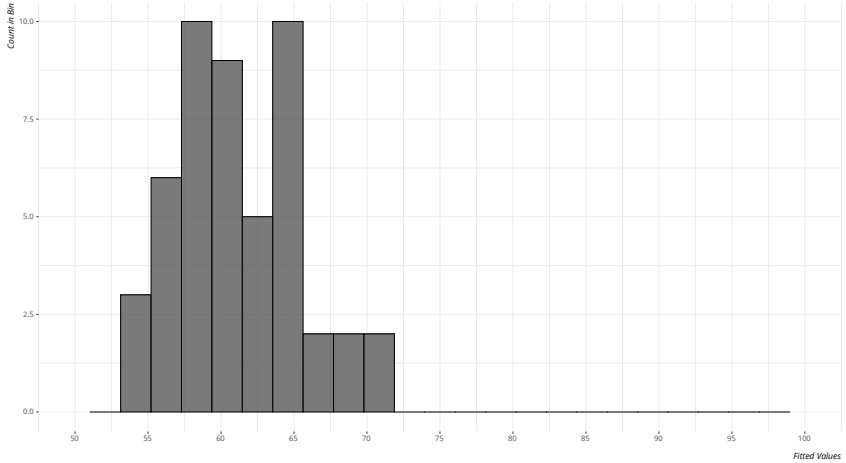
If you see a clear pattern(s) emerge like this, OLS is likely not the model you want.



Data: election_turnout in stevedata. A simple linear model regressing whether Trump won on college education and swing state.

The Estimated Values From Your Linear Model Should Be Plausible

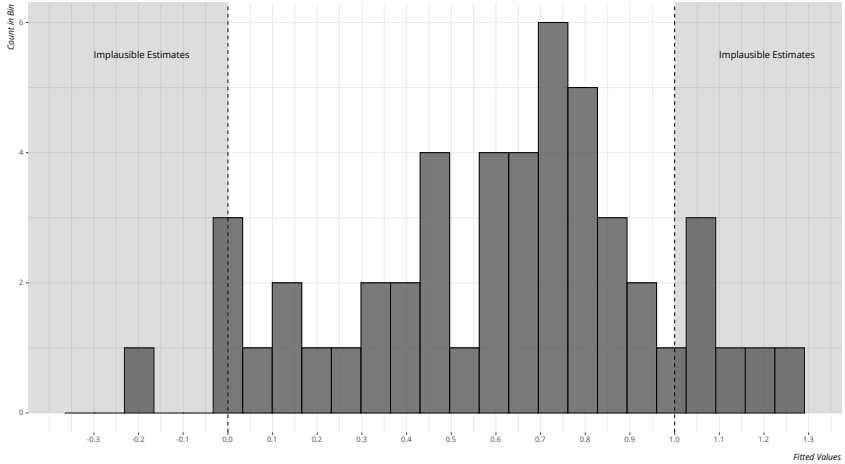
In this example, they are. Turnout everywhere ranged from low 50s to mid 70s, which we are estimating.



Data: election_turnout in stevedata. A simple linear model regressing turnout on college education and swing state. Note that scaling and bin numbers are arbitrary. I know. I also generally don't like histograms for continuous data.

The Estimated Values From Your Linear Model Should Be Plausible

In this example, they are not. Probability is hard-bound between 0 and 1 and our estimates are falling outside the scale.



Data: election_turnout in stevedata. A simple linear model regressing whether Trump won in 2016 on college education and swing state.
Note that scaling and bin numbers are arbitrary. I know. I also generally don't like histograms for continuous data.

What Estimates Fell Out of Bounds?

```
election_turnout %>%
  filter(state %nin% c("Hawaii", "District of Columbia")) %>%
  mutate(fitted = fitted(M2)) %>%
  filter(fitted > 1 | fitted < 0) %>%
  select(state, trumpw, percoled, ss, fitted)
```

```
## # A tibble: 9 x 5
##   state         trumpw percoled    ss   fitted
##   <chr>          <int>    <dbl> <int>    <dbl>
## 1 Arkansas           1     21.1     0  1.16
## 2 Connecticut        0     37.6     0 -0.00264
## 3 Kentucky           1     22.3     0  1.07
## 4 Louisiana          1     22.5     0  1.06
## 5 Maryland           0     37.9     0 -0.0237
## 6 Massachusetts      0     40.5     0 -0.206
## 7 Mississippi        1     20.7     0  1.18
## 8 Nevada             0     23       1  1.06
## 9 West Virginia      1     19.2     0  1.29
```

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.

Table 1: Individual-Level Education and Voting (Hypothetical Data)

Vote	0. Low	1. Mid-low	2. Middle	3. Mid-high	4. High	Total
Yes	6	20	50	80	94	250
No	94	80	50	20	6	250
N	100	100	100	100	100	500
p(vote)	.06	.20	.50	.80	.94	.50

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

Table 2: Probability/Odds of Education and Voting (Hypothetical Data)

Education	p(vote)	Odds of Voting
0. Low	.06	.06/.94 = .06
1. Mid-low	.20	.20/.80 = .25
2. Middle	.50	.50/.50 = 1
3. Mid-high	.80	.80/.20 = 4
4. High	.94	.94/.06 = 16

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 the table.

- 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*. Let's Add to Table 2 now.

Table 3: Probability, Odds, and Logits of Education and Voting (Hypothetical Data)

Education	p(vote)	Odds of Voting	Logged Odds
0. Low	.06	.06/.94 = .06	-2.8
1. Mid-low	.20	.20/.80 = .25	-1.4
2. Middle	.50	.50/.50 = 1	0
3. Mid-high	.80	.80/.20 = 4	1.4
4. High	.94	.94/.06 = 16	2.8

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 the table.

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 $x = 0$).

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

...or in R

```
exp(-2.8)/(1 + exp(-2.8))
```

```
## [1] 0.05732418
```

```
plogis(-2.8) # do this instead
```

```
## [1] 0.05732418
```

Using Actual Data

What would this look like with actual data? Let's consider this example from the 2016 and 2018 waves of GSS data.

- y: should women be able to obtain legal abortion for any reason?

Table 4: Should Woman Be Able to Have Abortion for Any Reason? (GSS, 2016-2018)

Response	Number of Observations	Percentage
No	1733	51.98%
Yes	1601	48.02%

Explanatory Variables

- Female (1 = women)
- Fixed effects for race (whites [omitted], black, other)
- Hispanic ethnicity (1 = yes)
- Education levels (years in school [0:20])
- Party ID (strong D to strong R [0:6])
- Religious activity (never to several times a day [0:10])

```r
gss_abortion %>%
  filter(year >= 2016) %>%
  mutate(race = fct_relevel(race, "White"),
         relactiv = relactiv - 1,
         female = ifelse(sex == "Female", 1, 0)) -> Data

M3 <- glm(abany ~ female + factor(race) + hispanic + educ +
            pid7 + relactiv, data=Data,
          family=binomial(link="logit"))

modelsummary(list("Should Abortion Be Allowed for Any Reason?" = M3), output="latex",
             title = "Attitudes About Abortion in the GSS (2016-2018)",
             stars = TRUE, gof_omit = "IC|F|Log.|R2$",
             coef_map = c("female" = "Female",
                          "factor(race)Black" = "Race = Black",
                          "factor(race)Other" = "Race = Other",
                          "hispanic" = "Hispanic",
                          "educ" = "Years of Education",
                          "pid7" = "Party ID (D to R)",
                          "relactiv" = "Religiosity/Religious Activity",
                          "(Intercept)" = "Intercept"),
             align = "lc")
```

Table 5: Attitudes About Abortion in the GSS (2016-2018)

	Should Abortion Be Allowed for Any Reason?
Female	-0.087
	(0.078)
Race = Black	-0.388***
	(0.111)
Race = Other	-0.181
	(0.143)
Hispanic	-0.354***
	(0.125)
Years of Education	0.153***
	(0.014)
Party ID (D to R)	-0.276***
	(0.022)
Religiosity/Religious Activity	-0.196***
	(0.018)
Intercept	-0.895***
	(0.218)
Num.Obs.	3175

* p < 0.1, ** p < 0.05, *** p < 0.01

Interpreting Table 5

- Women are no different than men.
- Black respondents are less likely to think women should be able to have an abortion for any reason than white respondents.
    - No difference between other races (i.e. effectively: Asians) and white respondents.
- Hispanics are less likely to favor abortion for any reason.
- Education has a discernible positive effect.
- Unsurprisingly: Republican partisanship and religiosity decrease likelihood of answering “yes.”

Extracting Some Quantities of Interest

Let's say we're interested in the effect of Hispanics.

- Exp(-0.354) = $e^{-0.354}$ = .701. This is the odds ratio.
- 100*($e^{-0.354}$ - 1) = -29.81%. This is the percentage change in the odds of voting.

What about the effect of increasing education levels.

- Odds ratio: $e^{.153}$ = 1.165.
- Percentage change in odds: 100*($e^{.153}$ - 1) = 16.53%

Our intercept ($\hat{a}$ = -.895) is meaningful too.

- It tells us logged odds of answering "yes" for a white, non-hispanic dude who has never been educated, is a strong Democrat, and never goes to religious services.
- Predicted probability of a "yes" from that guy: .290

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