Research Methods for Political Science PO3110 (TCD)

HT: Tutorial 7 - Week 8

Letcia Meniconi Barbabela

University College Dublin,
https://github.com/letmeni/research-methods

18 March 2020

Today's topics

- Finish last week's topics:
  - Bootstraping (Stats HT05);
  - Non-parametric tests (Stats HT06).
- This week's topic: Logistic regression (Stats HT08).

Basic concepts

When we run a logistic regression, we estimate the natural logarithm of the odds. But first, let's review some basic concepts:

- **Probability:** the likelihood of something happening:

Basic concepts

When we run a logistic regression, we estimate the natural logarithm of the odds. But first, let's review some basic concepts:

- **Probability:** the likelihood of something happening:
  - Probability of rain: 0.2 (20% chance);

Basic concepts

When we run a logistic regression, we estimate the natural logarithm of the odds. But first, let's review some basic concepts:

- **Probability:** the likelihood of something happening:
  - Probability of rain: 0.2 (20% chance);
  - Probability of not rain: 1-0.2 $\rightarrow$ 0.8 (80% chance);

Basic concepts

When we run a logistic regression, we estimate the natural logarithm of the odds. But first, let's review some basic concepts:

- **Probability:** the likelihood of something happening:
  - Probability of rain: 0.2 (20% chance);
  - Probability of not rain: 1-0.2 $\rightarrow$ 0.8 (80% chance);
  - Always range between 0 and 1.

Basic concepts

When we run a logistic regression, we estimate the natural logarithm of the odds. But first, let's review some basic concepts:

- **Probability:** the likelihood of something happening:
  - Probability of rain: 0.2 (20% chance);
  - Probability of not rain: 1-0.2 $\rightarrow$ 0.8 (80% chance);
  - Always range between 0 and 1.
- **Odds:** Ratio of the probability that an event will occur divided by the probability that an even will not occur:

Basic concepts

When we run a logistic regression, we estimate the natural logarithm of the odds. But first, let's review some basic concepts:

- **Probability:** the likelihood of something happening:
  - Probability of rain: 0.2 (20% chance);
  - Probability of not rain: 1-0.2 $\rightarrow$ 0.8 (80% chance);
  - Always range between 0 and 1.
- **Odds:** Ratio of the probability that an event will occur divided by the probability that an even will not occur:
  - $\frac{P(rain)}{P(\text{not rain})} = \frac{0.2}{0.8} = \frac{1}{4} = 0.25$

Basic concepts

When we run a logistic regression, we estimate the natural logarithm of the odds. But first, let's review some basic concepts:

- **Probability:** the likelihood of something happening:
  - Probability of rain: 0.2 (20% chance);
  - Probability of not rain: 1-0.2 $\rightarrow$ 0.8 (80% chance);
  - Always range between 0 and 1.
- **Odds:** Ratio of the probability that an event will occur divided by the probability that an even will not occur:
  - $\frac{P(rain)}{P(\text{not rain})} = \frac{0.2}{0.8} = \frac{1}{4} = 0.25$
  - Odds may be ≥ 1.

Basic concepts

When we run a logistic regression, we estimate the natural logarithm of the odds. But first, let's review some basic concepts:

- **Probability:** the likelihood of something happening:
  - Probability of rain: 0.2 (20% chance);
  - Probability of not rain: 1-0.2 $\rightarrow$ 0.8 (80% chance);
  - Always range between 0 and 1.
- **Odds:** Ratio of the probability that an event will occur divided by the probability that an even will not occur:
  - $\frac{P(rain)}{P(\text{not rain})} = \frac{0.2}{0.8} = \frac{1}{4} = 0.25$
  - Odds may be ≥ 1.
  - An 80% probability of rain has odds of $\frac{0.8}{0.2} = 4$.

Basic concepts

When we run a logistic regression, we estimate the natural logarithm of the odds. But first, let's review some basic concepts:

- **Probability:** the likelihood of something happening:
  - Probability of rain: 0.2 (20% chance);
  - Probability of not rain: 1-0.2 $\rightarrow$ 0.8 (80% chance);
  - Always range between 0 and 1.
- **Odds:** Ratio of the probability that an event will occur divided by the probability that an even will not occur:
  - $\frac{P(rain)}{P(\text{not rain})} = \frac{0.2}{0.8} = \frac{1}{4} = 0.25$
  - Odds may be ≥ 1.
  - An 80% probability of rain has odds of $\frac{0.8}{0.2} = 4$.
  - A 50% chance of rain has odds of 1.

Basic concepts

When we run a logistic regression, we estimate the natural logarithm of the odds. But first, let's review some basic concepts:

- **Probability:** the likelihood of something happening:
  - Probability of rain: 0.2 (20% chance);
  - Probability of not rain: 1-0.2 $\rightarrow$ 0.8 (80% chance);
  - Always range between 0 and 1.
- **Odds:** Ratio of the probability that an event will occur divided by the probability that an even will not occur:
  - $\frac{P(rain)}{P(\text{not rain})} = \frac{0.2}{0.8} = \frac{1}{4} = 0.25$
  - Odds may be ≥ 1.
  - An 80% probability of rain has odds of $\frac{0.8}{0.2} = 4$.
  - A 50% chance of rain has odds of 1.
- **Logit:** the natural logarithm (ln) of the odds:

Basic concepts

When we run a logistic regression, we estimate the natural logarithm of the odds. But first, let's review some basic concepts:

- **Probability:** the likelihood of something happening:
  - Probability of rain: 0.2 (20% chance);
  - Probability of not rain: 1-0.2 $\rightarrow$ 0.8 (80% chance);
  - Always range between 0 and 1.
- **Odds:** Ratio of the probability that an event will occur divided by the probability that an even will not occur:
  - $\frac{P(rain)}{P(\text{not rain})} = \frac{0.2}{0.8} = \frac{1}{4} = 0.25$
  - Odds may be ≥ 1.
  - An 80% probability of rain has odds of $\frac{0.8}{0.2} = 4$.
  - A 50% chance of rain has odds of 1.
- **Logit:** the natural logarithm (ln) of the odds:
  - If there is a 20% chance of rain, then $ln(0.20) = -1.386$

Some basic differences between logit and linear regression

- Instead of predicting the value of Y, predict the probability of Y ocurring;

Some basic differences between logit and linear regression

- Instead of predicting the value of Y, predict the probability of Y ocurring;
- Instead of being continuous the dependent variable is dichotomous;

Some basic differences between logit and linear regression

- Instead of predicting the value of Y, predict the probability of Y ocurring;
- Instead of being continuous the dependent variable is dichotomous;
- Instead of assuming a linear relationship between the IV and DV, we assume that the logit transformation of the DV has a linear relationship with the IV's;

Some basic differences between logit and linear regression

- Instead of predicting the value of Y, predict the probability of Y ocurring;
- Instead of being continuous the dependent variable is dichotomous;
- Instead of assuming a linear relationship between the IV and DV, we assume that the logit transformation of the DV has a linear relationship with the IV's;
- Estimation: instead of using OLS, Maximum likelihood estimation.

Example: Voting for Trump

$$ln\left[\frac{P(\text{Vote Trump})}{P(\text{Not Vote Trump})}\right] = \beta_0 + \beta_1 \times partyid + \beta_2 \times education \tag{1}$$

Example: Voting for Trump

$$ln\left[\frac{P(\text{Vote Trump})}{P(\text{Not Vote Trump})}\right] = \beta_0 + \beta_1 \times partyid + \beta_2 \times education \qquad (1)$$

To change this from the ln to the odds, we need to take the exponential function of the natural logarithm function:

Example: Voting for Trump

$$ln\left[\frac{P(\text{Vote Trump})}{P(\text{Not Vote Trump})}\right] = \beta_0 + \beta_1 \times partyid + \beta_2 \times education \qquad (1)$$

To change this from the ln to the odds, we need to take the exponential function of the natural logarithm function:

$$\frac{P(\text{Vote Trump})}{P(\text{Not Vote Trump})} = e^{\beta_0} + e^{\beta_1 \times partyid} + e^{\beta_2 \times education} \qquad (2)$$

Example: Voting for Trump

$$ln\left[\frac{P(\text{Vote Trump})}{P(\text{Not Vote Trump})}\right] = \beta_0 + \beta_1 \times partyid + \beta_2 \times education \qquad (1)$$

To change this from the ln to the odds, we need to take the exponential function of the natural logarithm function:

$$\frac{P(\text{Vote Trump})}{P(\text{Not Vote Trump})} = e^{\beta_0} + e^{\beta_1 \times partyid} + e^{\beta_2 \times education} \qquad (2)$$

If we want to get the (predicted) probability of voting for trump, we need to rewrite the formula again.

$$\frac{P(\text{Vote Trump})}{P(\text{Not Vote Trump})} = \frac{1}{1 + e^{-\beta_0} + e^{-\beta_1 \times partyid} + e^{-\beta_2 \times education}} \qquad (3)$$

Logistic Regression in SPSS

- Data: `https://tinyurl.com/anes16sav`
- Codebook: `https://www.electionstudies.org/wp-content/uploads/2016/02/anes_pilot_2016_CodebookUserGuide.pdf`
- **Dependent variable:** turnout12
- **Independent variables:** birthyr, gender, newsint
- Clean data:
  - Recode missing values into system missing;
  - Reverse coding when necessary.
- Run the logit;
- Interpret

Interpreting results

- Divide by four (Stats HT08);
- Odds ratio: `https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-how-do-i-interpret-odds-ratios-in-logistic-regression/`
  - OR > 1: positive relationship
  - OR $= 1$: no relationship
  - OR < 1: negative relationship
  - **Remember:** Constant and different than probabilities!