150B/355B
Introduction to Machine Learning for Social Science
TA Section 3

Haemin Jee and Tongtong Zhang

January 26, 2018

Outline

1. Set up Logistic Regression

Outline

1. Set up Logistic Regression
2. Logistic Regression in `R`

Outline

1. Set up Logistic Regression
2. Logistic Regression in R
3. Model Evaluation

Outline

1. Set up Logistic Regression
2. Logistic Regression in R
3. Model Evaluation
4. Exit Quiz (if time)

Linear Probability Model Set Up

Goal: Predict a probability of an event or outcome happening, given some predictor variables.

Linear Probability Model Set Up

Goal: Predict a probability of an event or outcome happening, given some predictor variables.

Question: In this set up, what are possible values of the dependent variable?

Linear Probability Model Set Up

Goal: Predict a probability of an event or outcome happening, given some predictor variables.

Question: In this set up, what are possible values of the dependent variable?

Answer: 0 or 1!

Linear Probability Model Set Up

Goal: Predict a probability of an event or outcome happening, given some predictor variables.

Question: In this set up, what are possible values of the dependent variable?

Answer: 0 or 1!

$Y_i = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2$
where Y_i is the predicted probability of an event / outcome happening.

Linear Probability Model Set Up

Goal: Predict a probability of an event or outcome happening, given some predictor variables.

Question: In this set up, what are possible values of the dependent variable?

Answer: 0 or 1!

$Y_i = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2$
where Y_i is the predicted probability of an event / outcome happening.

```
fit <- lm(y ~ rep + gorevote, data = iraqVote)
head(fit$fitted.values)
        1         2         3         4         5         6
0.9766924 0.9766924 1.1489597 1.1489597 0.9385758 0.9385758
```

Linear Probability Model Set Up

Goal: Predict a probability of an event or outcome happening, given some predictor variables.

Question: In this set up, what are possible values of the dependent variable?

Answer: 0 or 1!

$Y_i = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2$
where Y_i is the predicted probability of an event / outcome happening.

```
fit <- lm(y ~ rep + gorevote, data = iraqVote)
head(fit$fitted.values)
        1         2         3         4         5         6
0.9766924 0.9766924 1.1489597 1.1489597 0.9385758 0.9385758
```

When we run a LPM in `R`, the fitted values (the $\hat{Y}$) are *predicted probabilities*.

Logistic Regression Set Up

Functions to know:

- $\text{odds}(x) = \frac{x}{1-x}$

Logistic Regression Set Up

Functions to know:

- odds(x) $= \frac{x}{1-x}$
- log odds or logit(x) $= \log \frac{x}{1-x}$

Logistic Regression Set Up

Functions to know:

- $\text{odds}(x) = \frac{x}{1-x}$
- log odds or $\text{logit}(x) = \log\frac{x}{1-x}$
- logistic function or inverse $\text{logit}(x) = \frac{1}{1+\exp(-x)}$

Logistic Regression Set Up

Functions to know:

- $\text{odds}(x) = \frac{x}{1-x}$
- log odds or $\text{logit}(x) = \log \frac{x}{1-x}$
- logistic function or inverse $\text{logit}(x) = \frac{1}{1+\exp(-x)}$
- logistic function or inverse $\text{logit}(x) = \frac{\exp(x)}{1+\exp(x)}$

Logistic Regression Set Up

Goal: Predict a probability of an event or outcome happening, given some predictor variables.

Logistic Regression Set Up

Goal: Predict a probability of an event or outcome happening, given some predictor variables.

Let's call this p_i where i denotes an individual observation.

Logistic Regression Set Up

Goal: Predict a probability of an event or outcome happening, given some predictor variables.

Let's call this p_i where i denotes an individual observation.

$$\log(\frac{\hat{p}_i}{1 - \hat{p}_i}) = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2$$

Logistic Regression Set Up

Goal: Predict a probability of an event or outcome happening, given some predictor variables.

Let's call this p_i where i denotes an individual observation.

$$\log(\frac{\hat{p}_i}{1-\hat{p}_i}) = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2$$

Instead of modeling the *probability* of an event as a linear function of predictor variables (as we did in the LPM), we are modeling the logit of p as a linear function of predictor variables.

Logistic Regression Set Up

Goal: Predict a probability of an event or outcome happening, given some predictor variables.

Let's call this p_i where i denotes an individual observation.

$$\log(\frac{\hat{p}_i}{1-\hat{p}_i}) = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2$$

Instead of modeling the *probability* of an event as a linear function of predictor variables (as we did in the LPM), we are modeling the logit of p as a linear function of predictor variables.

But what we are really interested in is the p - how do we get there?

Interpreting Logistic Regression

Question: How do we interpret β_1 in a logistic regression?

Interpreting Logistic Regression

Question: How do we interpret β_1 in a logistic regression?

Answer: A one unit increase in the value of X_1 results in a β_1 increase / decrease in the *log odds* of an event happening.

Interpreting Logistic Regression

Question: How do we interpret β_1 in a logistic regression?

Answer: A one unit increase in the value of X_1 results in a β_1 increase / decrease in the *log odds* of an event happening.

But we're usually interested in how a one unit change in X_1 will change the *probability* of something happening! How do we calculate that?

LPM vs. Logistic Regression

Why use LPM?

LPM vs. Logistic Regression

Why use LPM?

- We are modeling probability directly - no need for transformation!

LPM vs. Logistic Regression

Why use LPM?

- We are modeling probability directly - no need for transformation!
- That means easier interpretability

LPM vs. Logistic Regression

Why use LPM?

- We are modeling probability directly - no need for transformation!
- That means easier interpretability
- No real difference in substantive results

LPM vs. Logistic Regression

Why use LPM?

- We are modeling probability directly - no need for transformation!
- That means easier interpretability
- No real difference in substantive results

Why use Logistic Regression?

- You will not get unreasonable predicted probabilities

LPM vs. Logistic Regression

Why use LPM?

- We are modeling probability directly - no need for transformation!
- That means easier interpretability
- No real difference in substantive results

Why use Logistic Regression?

- You will not get unreasonable predicted probabilities
- We are making better model assumptions

Logistic regression

`R` Code!

Back to Classification

Remember the big-picture goal of using LPM or Logistic Regression is to *classify* outcomes.

Back to Classification

Remember the big-picture goal of using LPM or Logistic Regression is to *classify* outcomes.

Classification Process:

1. Run a model (LPM, Logistic Regression) of your choice

Back to Classification

Remember the big-picture goal of using LPM or Logistic Regression is to *classify* outcomes.

Classification Process:

1. Run a model (LPM, Logistic Regression) of your choice
2. Find predicted probabilities

Back to Classification

Remember the big-picture goal of using LPM or Logistic Regression is to *classify* outcomes.

Classification Process:

1. Run a model (LPM, Logistic Regression) of your choice
2. Find predicted probabilities
3. Choose a threshold

Back to Classification

Remember the big-picture goal of using LPM or Logistic Regression is to *classify* outcomes.

Classification Process:

1. Run a model (LPM, Logistic Regression) of your choice
2. Find predicted probabilities
3. Choose a threshold
4. Create binary classifications based on the probabilities and thresholds

Back to Classification

Remember the big-picture goal of using LPM or Logistic Regression is to *classify* outcomes.

Classification Process:

1. Run a model (LPM, Logistic Regression) of your choice
2. Find predicted probabilities
3. Choose a threshold
4. Create binary classifications based on the probabilities and thresholds
5. Evalute your model(s)!

Model Evaluation

Accuracy: All of our correct guesses divided by all guesses

Model Evaluation

Accuracy: All of our correct guesses divided by all guesses

$$\text{Accuracy} = \frac{\text{True Yes} + \text{True No}}{\text{True Yes} + \text{True No} + \text{False Yes} + \text{False No}}$$

Model Evaluation

Accuracy: All of our correct guesses divided by all guesses

$$\text{Accuracy} = \frac{\text{True Yes} + \text{True No}}{\text{True Yes} + \text{True No} + \text{False Yes} + \text{False No}}$$

Precision: Among all of our “Yes” guesses, how many were actually true “Yes”?

Model Evaluation

Accuracy: All of our correct guesses divided by all guesses

$$\text{Accuracy} = \frac{\text{True Yes} + \text{True No}}{\text{True Yes} + \text{True No} + \text{False Yes} + \text{False No}}$$

Precision: Among all of our "Yes" guesses, how many were actually true "Yes"?

$$\text{Precision} = \frac{\text{True Yes}}{\text{True Yes} + \text{False Yes}}$$

Model Evaluation

Accuracy: All of our correct guesses divided by all guesses

$$\text{Accuracy} = \frac{\text{True Yes} + \text{True No}}{\text{True Yes} + \text{True No} + \text{False Yes} + \text{False No}}$$

Precision: Among all of our "Yes" guesses, how many were actually true "Yes"?

$$\text{Precision} = \frac{\text{True Yes}}{\text{True Yes} + \text{False Yes}}$$

Recall: Among all the true Yes's, how many were we able to guess correctly?

Model Evaluation

Accuracy: All of our correct guesses divided by all guesses

$$\text{Accuracy} = \frac{\text{True Yes} + \text{True No}}{\text{True Yes} + \text{True No} + \text{False Yes} + \text{False No}}$$

Precision: Among all of our "Yes" guesses, how many were actually true "Yes"?

$$\text{Precision} = \frac{\text{True Yes}}{\text{True Yes} + \text{False Yes}}$$

Recall: Among all the true Yes's, how many were we able to guess correctly?

$$\text{Recall} = \frac{\text{True Yes}}{\text{True Yes} + \text{False No}}$$

Model Evalution

F Score

Model Evalution

F Score

$$\text{F Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Model Evaluation

`R` Code!