150B/355B
Introduction to Machine Learning for Social Science
TA Section 2

Haemin Jee and Tongtong Zhang

January 19, 2018

Road Map

1. Multivariate Regression

Road Map

1. Multivariate Regression
2. Linear Algebra: Inner Product

Road Map

1. Multivariate Regression
2. Linear Algebra: Inner Product
3. Classification
   - Linear Probability Model (LPM)

Multivariate Regression

Time for Change Model (Abramowitz, Linzer)

Multivariate Regression

Time for Change Model (Abramowitz, Linzer)

Predict Incumbent Vote Share with political and economic fundamentals

Multivariate Regression

Time for Change Model (Abramowitz, Linzer)

Predict Incumbent Vote Share with political and economic fundamentals

1. GDP Growth
2. Incumbent Presidential Popularity
3. Incumbent Party

Time for Change Model

Bivariate model: $Vote_i = \hat{\beta}_0 + \hat{\beta}_1 Approval_i + \epsilon_i$

Time for Change Model

Bivariate model: $Vote_i = \hat{\beta}_0 + \hat{\beta}_1 Approval_i + \epsilon_i$

```
rm(list=ls())
setwd("~/Dropbox (IPL)/150B Machine Learning/Lecture/Lecture 3")
d<-read.csv("TimeChange.csv")

#Bivariate model of VoteShare on Incumbency Approval
bivariate <- lm(IncumbentVoteShare~Incumbent_Net_Approval, data = d)
bivariate$coefficients

##             (Intercept) Incumbent_Net_Approval
##              50.7594886              0.1619931
```

Time for Change Model

Multivariate model:

Time for Change Model

Multivariate model:

$$Vote_i = f(Approval_i, Q1GDP_i, Q2GDP_i, Inc2ndTerm_i) + \epsilon_i$$
$$Vote_i = \hat{\beta}_0 + \hat{\beta}_1 Approval_i + \hat{\beta}_2 Q1GDP_i + \hat{\beta}_3 Q2GDP_i + \hat{\beta}_4 Inc2ndTerm_i + \epsilon_i$$

Time for Change Model

Multivariate model:

$$Vote_i = f(Approval_i, Q1GDP_i, Q2GDP_i, Inc2ndTerm_i) + \epsilon_i$$
$$Vote_i = \hat{\beta}_0 + \hat{\beta}_1 Approval_i + \hat{\beta}_2 Q1GDP_i + \hat{\beta}_3 Q2GDP_i + \hat{\beta}_4 Inc2ndTerm_i + \epsilon_i$$

How do we interpret $\hat{\beta}_1$ in this regression?

Time for Change Model

Multivariate model:

$$Vote_i = f(Approval_i, Q1GDP_i, Q2GDP_i, Inc2ndTerm_i) + \epsilon_i$$

$$Vote_i = \hat{\beta}_0 + \hat{\beta}_1 Approval_i + \hat{\beta}_2 Q1GDP_i + \hat{\beta}_3 Q2GDP_i + \hat{\beta}_4 Inc2ndTerm_i + \epsilon_i$$

How do we interpret $\hat{\beta}_1$ in this regression?

One unit increase in Approval is associated with $\hat{\beta}_1$ units increase in Vote share, **holding all other predictors constant**.

Time for Change Model

Multivariate model:

```r
#Multivariate model
multivariate <- lm(IncumbentVoteShare~Incumbent_Net_Approval + Q1_GDP_Growth +
                     Q2_GDP_Growth + Incumbent_Party_Two_Terms, data = d)
multivariate$coefficients

##               (Intercept)    Incumbent_Net_Approval
##               51.01161540                0.09511158
##             Q1_GDP_Growth             Q2_GDP_Growth
##                0.10335203                0.57188740
## Incumbent_Party_Two_Terms
##               -4.35308941
```

Time for Change Model

Bivariate model: $Vote_i = 50.76 + 0.16 * Approval_i + \epsilon_i$

Multivariate model: $Vote_i = 51.01 + 0.10 * Approval_i + 0.10 * Q1GDP_i + 0.57 * Q2GDP_i - 4.35 * Inc2ndTerm_i + \epsilon_i$

Time for Change Model

Bivariate model: $Vote_i = 50.76 + 0.16 * Approval_i + \epsilon_i$

Multivariate model: $Vote_i = 51.01 + 0.10 * Approval_i + 0.10 * Q1GDP_i + 0.57 * Q2GDP_i - 4.35 * Inc2ndTerm_i + \epsilon_i$

Q: Why does the coefficient on Approval change between the two models?

Time for Change Model

Bivariate model: $Vote_i = 50.76 + 0.16 * Approval_i + \epsilon_i$

Multivariate model: $Vote_i = 51.01 + 0.10 * Approval_i + 0.10 * Q1GDP_i + 0.57 * Q2GDP_i - 4.35 * Inc2ndTerm_i + \epsilon_i$

Q: Why does the coefficient on Approval change between the two models?

A: Because omitting other explanatory variables in the bivariate model leads to a positive bias on the coefficient on Approval.

Time for Change Model

Bivariate model: $Vote_i = 50.76 + 0.16 * Approval_i + \epsilon_i$

Multivariate model: $Vote_i = 51.01 + 0.10 * Approval_i + 0.10 * Q1GDP_i + 0.57 * Q2GDP_i - 4.35 * Inc2ndTerm_i + \epsilon_i$

Q: Why does the coefficient on Approval change between the two models?

A: Because omitting other explanatory variables in the bivariate model leads to a positive bias on the coefficient on Approval. BUT, bias can also be negative! Example: effect of years of education on income controlling for PhD degree.

Time for Change Model

Bivariate model: $Vote_i = 50.76 + 0.16 * Approval_i + \epsilon_i$

Multivariate model: $Vote_i = 51.01 + 0.10 * Approval_i + 0.10 * Q1GDP_i + 0.57 * Q2GDP_i - 4.35 * Inc2ndTerm_i + \epsilon_i$

Q: Why does the coefficient on Approval change between the two models?

A: Because omitting other explanatory variables in the bivariate model leads to a positive bias on the coefficient on Approval. BUT, bias can also be negative! Example: effect of years of education on income controlling for PhD degree.

R Code (Question 1, Section 3)!

Inner Product

$$\mu = (\mu_1, \mu_2, ..., \mu_n)$$
$$\mathbf{v} = (v_1, v_2, ..., v_n)$$

$$\mu \cdot \mathbf{v} = \mu_1 v_1 + \mu_2 v_2 + ... + \mu_n v_n$$

Inner Product

$$\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)$$
$$= (51.01, 0.1, 0.1, 0.57, -4.35)$$
$$\mathbf{x_i} = (1, Approval_i, Q1GDP_i, Q2GDP_i, Inc2ndTerm_i)$$

Inner Product

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3) \\ &= (51.01, 0.1, 0.1, 0.57, -4.35) \\ \mathbf{x_i} &= (1, Approval_i, Q1GDP_i, Q2GDP_i, Inc2ndTerm_i)\end{aligned}$$

Then, we can write our prediction as:

$$\begin{aligned}\hat{Vote_i} &= \hat{\boldsymbol{\beta}} \cdot \mathbf{x_i} \\ \hat{Vote_i} &= 51.01 + 0.1 * Approval_i + 0.1 * Q1GDP_i \\ &= +0.57 * Q2GDP_i - 4.35 * Inc2ndTerm_i\end{aligned}$$

Inner Product

$$\begin{aligned}\hat{\beta} &= (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3) \\ &= (51.01, 0.1, 0.1, 0.57, -4.35) \\ \mathbf{x_i} &= (1, Approval_i, Q1GDP_i, Q2GDP_i, Inc2ndTerm_i)\end{aligned}$$

Then, we can write our prediction as:

$$\begin{aligned}\hat{Vote_i} &= \hat{\beta} \cdot \mathbf{x_i} \\ \hat{Vote_i} &= 51.01 + 0.1 * Approval_i + 0.1 * Q1GDP_i \\ &= +0.57 * Q2GDP_i - 4.35 * Inc2ndTerm_i\end{aligned}$$

R code (Question 1, Sections 4 and 5)

Classification

Goal: predict Iraq vote (probability of yes, classify senators as for and against).

Classification

Goal: predict Iraq vote (probability of yes, classify senators as for and against).
Method: Linear Probability Model

Classification

Goal: predict Iraq vote (probability of yes, classify senators as for and against).

Method: Linear Probability Model

- Dependent variable: $Vote_i$ (1 or 0)
- Independent variable: Senator characteristics (party, vote for Gore, etc.)

Linear Probability Model

Two estimation goals:

Linear Probability Model

Two estimation goals:

Estimate:

- Probability of voting yes: $Pr(\widehat{Vote_i = 1}|\mathbf{x_i})$

Linear Probability Model

Two estimation goals:

Estimate:

- Probability of voting yes: $\widehat{Pr(Vote_i = 1|\mathbf{x_i})}$

- Classification of vote: $\widehat{Vote_i} = \mathbf{I}(\widehat{Pr(Vote_i = 1|\mathbf{x_i})} > t)$, where t is a threshold and $\mathbf{I}$ is an indicator function

Linear Probability Model

$$Vote_i = \beta \cdot \mathbf{x_i} + \epsilon_i$$

$$\widehat{Pr(Vote_i = 1|\mathbf{x_i})} = \hat{\beta} \cdot \mathbf{x_i}$$

$$V\hat{o}te_i = 1 \text{ if } \hat{\beta} \cdot \mathbf{x_i} > t$$
$$V\hat{o}te_i = 0 \text{ if } \hat{\beta} \cdot \mathbf{x_i} \leq t$$

Linear Probability Model

$$Vote_i = \beta \cdot \mathbf{x_i} + \epsilon_i$$

$$\widehat{Pr(Vote_i = 1|\mathbf{x_i})} = \hat{\beta} \cdot \mathbf{x_i}$$

$$V\hat{o}te_i = 1 \text{ if } \hat{\beta} \cdot \mathbf{x_i} > t$$
$$V\hat{o}te_i = 0 \text{ if } \hat{\beta} \cdot \mathbf{x_i} \leq t$$

$\mathbf{x_i}$: Party, vote share for Gore in 2000

Linear Probability Model

$$Vote_i = \beta \cdot \mathbf{x_i} + \epsilon_i$$

$$\widehat{Pr(Vote_i = 1|\mathbf{x_i})} = \hat{\beta} \cdot \mathbf{x_i}$$

$$\hat{Vote_i} = 1 \text{ if } \hat{\beta} \cdot \mathbf{x_i} > t$$
$$\hat{Vote_i} = 0 \text{ if } \hat{\beta} \cdot \mathbf{x_i} \leq t$$

$\mathbf{x_i}$: Party, vote share for Gore in 2000

R code (Question 2)