Introduction to Machine Learning for Social Science

Class 3: Multivariate Regression

Rochelle Terman

Postdoctoral Fellow
Center for International Security Cooperation
Stanford University

January 16th, 2018

Loose Threads

1) Reducible vs. Irreducible error.
   - Even if you estimate f perfectly, don't expect perfect predictions.
2) Least squared versus absolute errors.
   - Some cases do apply least absolute deviations (L1 regression), which is robust to outliers.
   - Mathematical convenience: Easier to compute the derivative of a polynomial than absolute value.
   - By Gauss Markov, OLS is BLUE

Questions?

Predicting Election Results

Goal: Predict Incumbent Vote Share (create prediction function)

- Use relationship in prior elections to predict future election
- Training data (In sample) $\rightsquigarrow$ Testing data (Out of sample)

Method: Linear Regression (Least squares)
Evaluation (Focus of next lectures):

1) In sample fit (training data)
2) Out of sample fit (test data)

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

Our Estimated Prediction Function

Our Estimated Prediction Function

$$\text{Vote}_i \quad = \quad \widehat{\beta}_0 + \widehat{\beta}_1 \text{Approval}_i + \epsilon_i$$

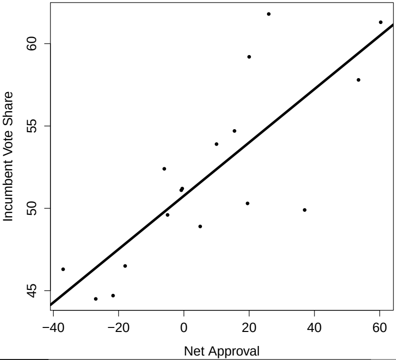
Our Estimated Prediction Function

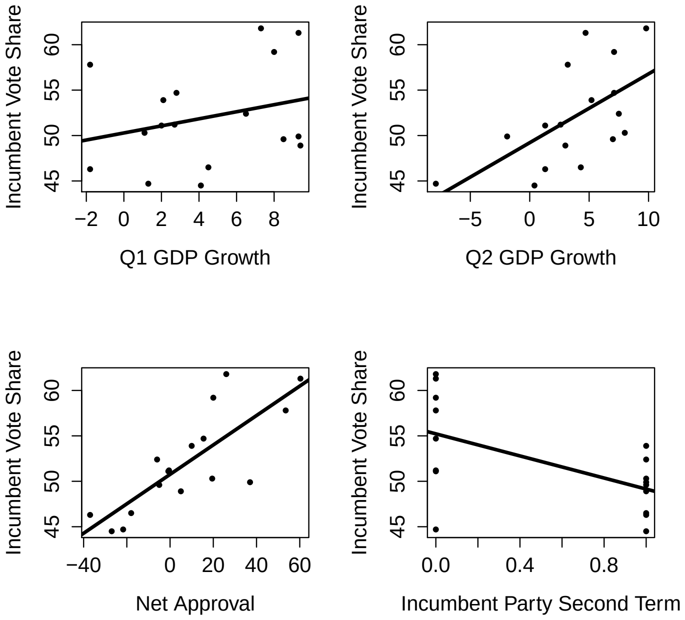
$$
\begin{aligned}
\text{Vote}_i &= \widehat{\beta}_0 + \widehat{\beta}_1 \text{Approval}_i + \epsilon_i \\
\text{Vote}_i &= \underbrace{50.76 + 0.16 \times \text{Approval}_i}_{\widehat{\text{Vote}_i}} + \epsilon_i
\end{aligned}
$$

Our Estimated Prediction Function

$$\begin{aligned}
\text{Vote}_i &= \widehat{\beta}_0 + \widehat{\beta}_1 \text{Approval}_i + \epsilon_i \\
\text{Vote}_i &= \underbrace{50.76 + 0.16 \times \text{Approval}_i}_{\widehat{\text{Vote}_i}} + \epsilon_i
\end{aligned}$$

R Code!




Incumbent Vote Share
45
50
55
60
−2
0
2
4
6
8
Q1 GDP Growth
Incumbent Vote Share
45
50
55
60
−5
0
5
10
Q2 GDP Growth
Incumbent Vote Share
45
50
55
60
−40
0
20
40
60
Net Approval
Incumbent Vote Share
45
50
55
60
0.0
0.4
0.8
Incumbent Party Second Term

What is Linear Regression?

Multiple linear regression:

- Assumes a linear relationship between quantitative response Y and **multiple** variables $X_1, X_2, ...X_p$.
- Also called multivariate regression because there are > 2 variables.

Multivariate Regression: Function Perspective

Function f such that

$$\text{Vote}_i = f(\text{Approval}_i, \text{Q1 GDP}_i, \text{Q2 GDP}_i, \text{Inc 2nd Term}_i) + \epsilon_i$$

Multivariate Regression: Function Perspective

Function f such that

$$\begin{aligned} \text{Vote}_i &= f(\text{Approval}_i, \text{Q1 GDP}_i, \text{Q2 GDP}_i, \text{Inc 2nd Term}_i) + \epsilon_i \\ &= \beta_0 + \beta_1 \text{Approval}_i + \beta_2 \text{Q1 GDP}_i \\ &\quad + \beta_3 \text{Q2 GDP}_i + \beta_4 \text{Inc 2nd Term}_i + \epsilon_i \end{aligned}$$

Multivariate Regression: Function Perspective

Function f such that

$$\begin{aligned}
\text{Vote}_i &= f(\text{Approval}_i, \text{Q1 GDP}_i, \text{Q2 GDP}_i, \text{Inc 2nd Term}_i) + \epsilon_i \\
&= \beta_0 + \beta_1 \text{Approval}_i + \beta_2 \text{Q1 GDP}_i \\
&\quad + \beta_3 \text{Q2 GDP}_i + \beta_4 \text{Inc 2nd Term}_i + \epsilon_i
\end{aligned}$$

$$\begin{aligned}
\widehat{\text{Vote}_i} &= \beta_0 + \beta_1 \text{Approval}_i + \beta_2 \text{Q1 GDP}_i \\
&\quad + \beta_3 \text{Q2 GDP}_i + \beta_4 \text{Inc 2nd Term}_i
\end{aligned}$$

Multivariate Regression: Function Perspective

Function f such that

$$\begin{aligned} \text{Vote}_i &= f(\text{Approval}_i, \text{Q1 GDP}_i, \text{Q2 GDP}_i, \text{Inc 2nd Term}_i) + \epsilon_i \\ &= \beta_0 + \beta_1 \text{Approval}_i + \beta_2 \text{Q1 GDP}_i \\ &\quad + \beta_3 \text{Q2 GDP}_i + \beta_4 \text{Inc 2nd Term}_i + \epsilon_i \end{aligned}$$

$$\begin{aligned} \widehat{\text{Vote}_i} &= \beta_0 + \beta_1 \text{Approval}_i + \beta_2 \text{Q1 GDP}_i \\ &\quad + \beta_3 \text{Q2 GDP}_i + \beta_4 \text{Inc 2nd Term}_i \end{aligned}$$

We interpret β_j as the average effect on Y of a one unit increase in X_j , holding all other predictors fixed.

Fitting a Multivariate Regression

Fitting a Multivariate Regression

Choose β's to minimize:

Fitting a Multivariate Regression

Choose β's to minimize:

$$\sum_{i=1948}^{2012} \epsilon_i^2 \quad = \quad \sum_{i=1948}^{2012} \left(\text{Vote}_i - \widehat{\text{Vote}_i}\right)^2$$

Fitting a Multivariate Regression

Choose β's to minimize:

$$\begin{aligned}\sum_{i=1948}^{2012} \epsilon_i^2 &= \sum_{i=1948}^{2012} \left(\text{Vote}_i - \widehat{\text{Vote}_i}\right)^2 \\ &= \sum_{i=1948}^{2012} (\text{Vote}_i - \\ &\quad \beta_0 - \beta_1 \text{Approval}_i - \beta_2 \text{Q1 GDP}_i - \beta_3 \text{Q2 GDP}_i - \beta_4 \text{Inc 2nd Term}_i)^2\end{aligned}$$

Fitting a Multivariate Regression

Choose β's to minimize:

$$
\begin{aligned}
\sum_{i=1948}^{2012} \epsilon_i^2 &= \sum_{i=1948}^{2012} \left(\text{Vote}_i - \widehat{\text{Vote}_i}\right)^2 \\
&= \sum_{i=1948}^{2012} (\text{Vote}_i - \\
&\quad \beta_0 - \beta_1 \text{Approval}_i - \beta_2 \text{Q1 GDP}_i - \beta_3 \text{Q2 GDP}_i - \beta_4 \text{Inc 2nd Term}_i)^2
\end{aligned}
$$

`to the R code! (Sections 1 and 2)`

Multivariate Regression: Our Estimated Prediction Function

$$\text{Vote}_i = \underbrace{51.01 + 0.10 \times \text{Q1 GDP} + 0.57 \times \text{Q2 GDP} + 0.10 \times \text{Approval} - 4.35 \times \text{Inc 2nd Term}}_{\widehat{\text{Vote}}_i} + \epsilon_i$$

Multivariate Regression: Our Estimated Prediction Function

$$\text{Vote}_i = \underbrace{\begin{aligned}&51.01 + 0.10 \times \text{Q1 GDP} + 0.57 \times \text{Q2 GDP} \\ &+0.10 \times \text{Approval} - 4.35 \times \text{Inc 2nd Term}\end{aligned}}_{\widehat{\text{Vote}}_i} + \epsilon_i$$

Bonus: Why does Approval have a smaller effect in the multivariate regression (.10) compared to the bivariate regression (.16)?
(Hint: Sharks and Ice-cream)

Multivariate Regression: Our Estimated Prediction Function

$$\text{Vote}_i = \underbrace{\underbrace{51.01 + 0.10 \times \text{Q1 GDP} + 0.57 \times \text{Q2 GDP}}_{} \quad + 0.10 \times \text{Approval} - 4.35 \times \text{Inc 2nd Term}}_{\widehat{\text{Vote}}_i} + \epsilon_i$$

Bonus: Why does Approval have a smaller effect in the multivariate regression (.10) compared to the bivariate regression (.16)?
(Hint: Sharks and Ice-cream)

`R Code (Section 3)!`

Linear Algebra

Vectors

Vector: ordered list of n numbers

- 1
- π
- e
- $(1, 2)$
- $(0, 0)$
- (π, e)
- $(3.1, 4.5, 6.11132)$
- $(\beta_0, \beta_1, \beta_2, \beta_3)$
- Incumbent vote share across elections, Net Approval across election

We will write vectors with bold ($\boldsymbol{\beta}$)

Inner Product

Consider two vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ and they are the same length. The define their inner product, $\boldsymbol{u} \cdot \boldsymbol{v}$, as

$$\boldsymbol{u} \cdot \boldsymbol{v} \quad = \quad u_1 v_1 + u_2 v_2 + \ldots + u_n v_n$$

Inner Product

Consider two vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ and they are the same length. The define their inner product, $\boldsymbol{u} \cdot \boldsymbol{v}$, as

$$\begin{aligned}\boldsymbol{u} \cdot \boldsymbol{v} &= u_1 v_1 + u_2 v_2 + \ldots + u_n v_n \\ &= \sum_{i=1}^{N} u_i v_i\end{aligned}$$

Rewriting our Prediction

Define:

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= (\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2, \widehat{\beta}_3)) \\ &= (51.01, 0.10, 0.57, 0.10, -4.35) \end{aligned}$$

Rewriting our Prediction

Define:

$$
\begin{aligned}
\widehat{\boldsymbol{\beta}} &= (\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2, \widehat{\beta}_3)) \\
&= (51.01, 0.10, 0.57, 0.10, -4.35) \\
\boldsymbol{x}_i &= (1, \text{Q1 GDP}_i, \text{Q2 GDP}_i, \text{Approval}_i, \text{Inc 2nd Term}_i)
\end{aligned}
$$

Rewriting our Prediction

Define:

$$\begin{aligned}\widehat{\boldsymbol{\beta}} &= (\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2, \widehat{\beta}_3)) \\ &= (51.01, 0.10, 0.57, 0.10, -4.35) \\ \boldsymbol{x}_i &= (1, \text{Q1 GDP}_i, \text{Q2 GDP}_i, \text{Approval}_i, \text{Inc 2nd Term}_i)\end{aligned}$$

Then, we can write our prediction as:

$$\widehat{\text{Vote}_i} = \widehat{\boldsymbol{\beta}} \cdot \boldsymbol{x}_i$$

`R Code (Sections 4  5)`

Potential Problems

See *Introduction to Statistical Learning* 3.3.3 (p. 92–102)

1. **Non-linearity of response-predictor relationships**: true relationship is not linear.
2. **Correlation of error terms:** ϵ_i provides information about ϵ_{i+1}, as in time-series.
3. **Non-constant variance of error terms (heteroscedasticity)**: variances of error terms change with value of response, making a funnel shape.
4. **Outliers and high-leverage points**: unusual values for Y or X.
5. **Collinearity**: two or more predator variables are closely related to one another.

Key Terms:

- Linear Regression, Simple Regression, Multiple Regression
- Cost function
- Sum of Squared Residuals
- In sample, Out of Sample

Key Techniques and `R` Functions

- Linear algebra operations and terms
  - Inner product
  - Matrix
- `lm`, `plot` , `%*%`

Up next:
Classification and Logistic Regression