Introduction to Machine Learning for Social Scientists

Class 4: More on OLS

Edgar Franco Vivanco

Stanford University
Department of Political Science

edgarf1@stanford.edu

Summer 2018

Homework 2 Due July 11 at 1:30pm

At which you point, you get another one.

Questions?

New classroom

- Larger classroom!
- **Building 250** (History Corner- Main Quad)
- Room 305
- Starting next class

Final: Friday, August 17, 2018.
3:30-6:30 p.m.

Today's Goals

1. Review general concepts
2. Review Bivariate regression
3. Introduction to multivariate regression
4. Introduction to model testing

Concepts

Prediction vs Inference

Two main reasons that we might wish to create a model:

1. *Prediction*
   - $\hat{Y} = \hat{f}(X)$
   - $\hat{f}$ is treated as a *black box*.
   - Better model = more accurate predictions of $\hat{Y} \approx Y$

Prediction vs Inference

Two main reasons that we might wish to create a model:

1. *Prediction*
   - $\hat{Y} = \hat{f}(X)$
   - $\hat{f}$ is treated as a *black box*.
   - Better model = more accurate predictions of $\hat{Y} \approx Y$
2. *Inference*
   - How is Y affected as $X_1, X_2, \ldots X_p$ change?
   - $\hat{f}$ no longer treated as a *black box*.
   - Better model = more interpretable

Prediction vs Inference

Two main reasons that we might wish to create a model:

1. *Prediction*
    - $\hat{Y} = \hat{f}(X)$
    - $\hat{f}$ is treated as a *black box*.
    - Better model = more accurate predictions of $\hat{Y} \approx Y$
2. *Inference*
    - How is Y affected as $X_1, X_2, \ldots X_p$ change?
    - $\hat{f}$ no longer treated as a *black box*.
    - Better model = more interpretable

ML is mainly about prediction

Supervised Learning vs Unsupervised Learning

1. *Supervised*
   We have access to:
   - p features X_1, X_2,..., X_p measured on n observations.
   - And a response Y (label)

   The goal is to predict Y using X_1, X_2,..., X_p

Supervised Learning vs Unsupervised Learning

1. *Supervised*
   We have access to:
   - p features X_1, X_2,..., X_p measured on n observations.
   - And a response Y (label)

   The goal is to predict Y using X_1, X_2,..., X_p
2. *Unsupervised*
   We only know:
   - p features X_1, X_2,..., X_p measured on n observations.

   The goal is to is to discover interesting things about the measurements on X_1, X_2,..., X_p

Regression vs Classification

1. *Regression*
   - Quantitative responses
   - Example: Age, height, salary, price, vote share, etc.

Regression vs Classification

1. *Regression*
   - Quantitative responses
   - Example: Age, height, salary, price, vote share, etc.
2. *Classification*
   - Qualitative responses
   - Example: Election result (win, lose), fake news (yes, no, maybe).

Regression vs Classification

1. *Regression*
   - Quantitative responses
   - Example: Age, height, salary, price, vote share, etc.
2. *Classification*
   - Qualitative responses
   - Example: Election result (win, lose), fake news (yes, no, maybe).

NOTE: The distinction is not always that clear-cut: Logistic regression (a type of non-linear regression) is often used for classification.

Bivariate vs Multivariate

Regression can be:

1. *Bivariate*
    - A single predictor X_1
    - Advantages:
    - Disadvantages:

Bivariate vs Multivariate

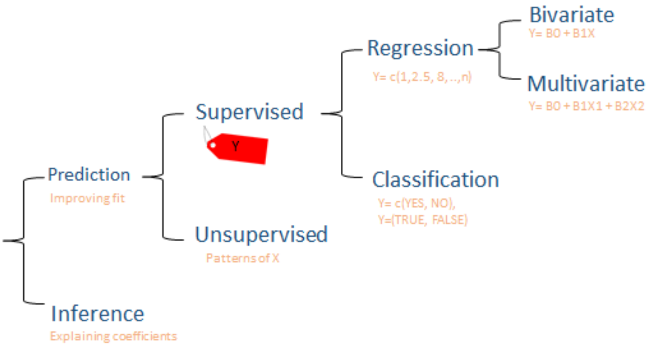
Regression can be:

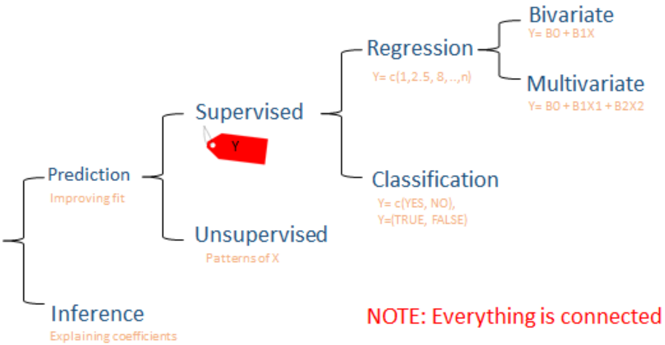
1. *Bivariate*
   - A single predictor X_1
   - Advantages:
   - Disadvantages:
2. *Multivariate*
   - Multiple predictors X_1, X_2,..., X_p
   - Advantages:
   - Disadvantages:

Bivariate vs Multivariate

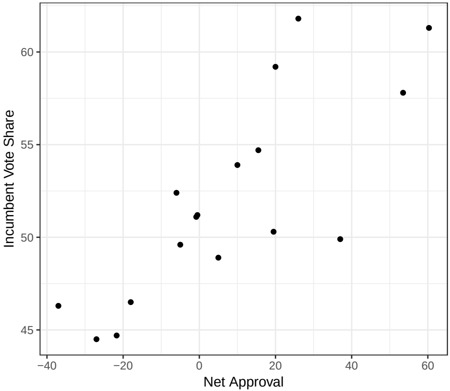
Regression can be:

1. *Bivariate*
   - A single predictor X_1
   - Advantages:
   - Disadvantages:
2. *Multivariate*
   - Multiple predictors X_1, X_2,..., X_p
   - Advantages:
   - Disadvantages:


Bivariate
Y= B0 + B1X
Regression
Y= c(1,2.5, 8, ..,n)
Multivariate
Y= B0 + B1X1 + B2X2
Supervised
Y
Prediction
Improving fit
Classification
Y= c(YES, NO),
Y=(TRUE, FALSE)
Unsupervised
Patterns of X
Inference
Explaining coefficients

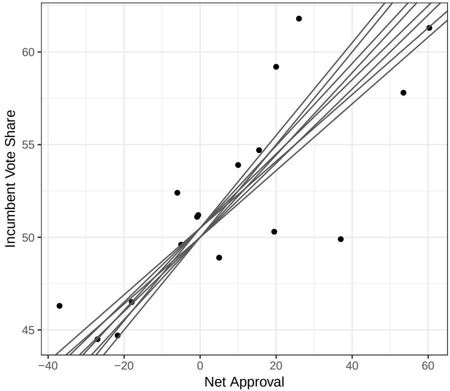

Bivariate
Y= B0 + B1X
Regression
Y= c(1,2.5, 8, .., n)
Multivariate
Y= B0 + B1X1 + B2X2
Supervised
Y
Prediction
Improving fit
Classification
Y= c(YES, NO),
Y=(TRUE, FALSE)
Unsupervised
Patterns of X
Inference
Explaining coefficients
NOTE: Everything is connected

Bi-variate regression



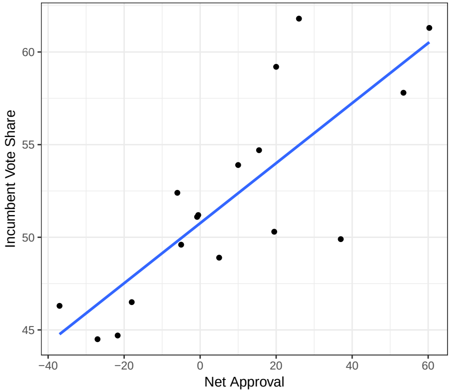
- Relating Y (output) and X (input).

Bi-variate regression



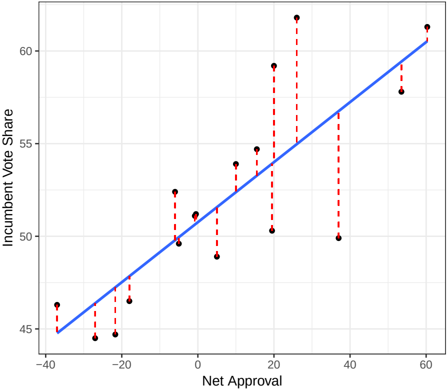
- Relating Y (output) and X (input).
- Many possible mappings $Y = f(X)$

Bi-variate regression



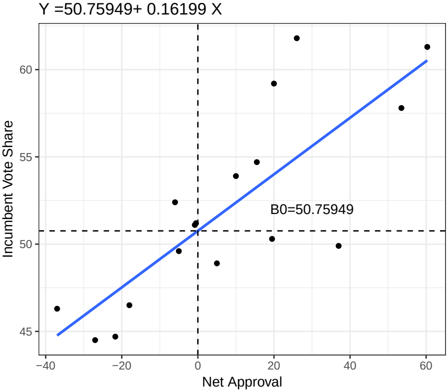
- Relating Y (output) and X (input).
- Many possible mappings $Y = f(X)$
- Least squares finds a relationship $Y = \beta_0 + \beta_1 X + \epsilon$

Bi-variate regression



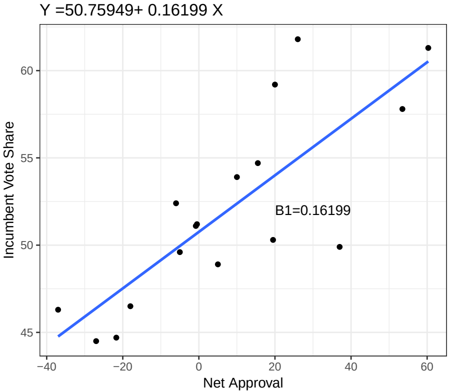
- Relating Y (output) and X (input).
- Many possible mappings $Y = f(X)$
- Least squares finds a relationship $Y = \beta_0 + \beta_1 X + \epsilon$
- $\epsilon_i = y_i - \hat{y}_i$ represents the i residual.
- OLS choses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the cost function: $\sum_{i=1}^{N} \epsilon_i$

Bi-variate regression



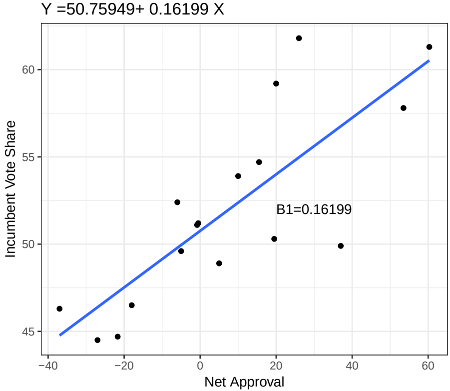
- The Intercept $\hat{\beta}_0$ represents the value of $\hat{y}_i$ when $x_i = 0$

Bi-variate regression



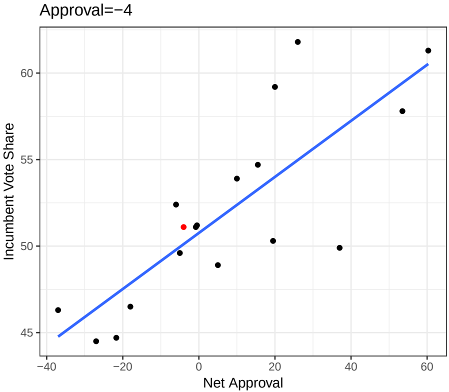
- The Intercept $\hat{\beta}_0$ represents the value of $\hat{y}_i$ when $x_i = 0$
- The slope $\hat{\beta}_1$ represents the direction and intensity of the relationship

Bi-variate regression



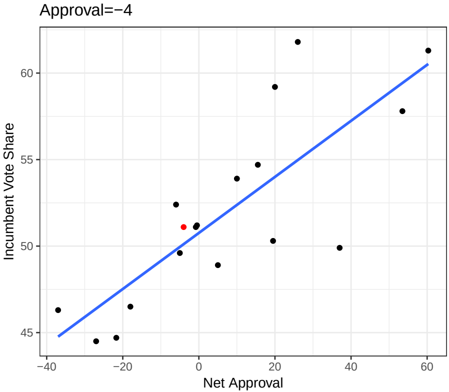
- The Intercept $\hat{\beta}_0$ represents the value of $\hat{y}_i$ when $x_i = 0$
- The slope $\hat{\beta}_1$ represents the direction and intensity of the relationship
- Since $\hat{\beta}_0$ and $\hat{\beta}_1$ are estimates there is a degree of uncertainty, represented by the standard errors. Those are used to create confidence intervals.

Bi-variate regression



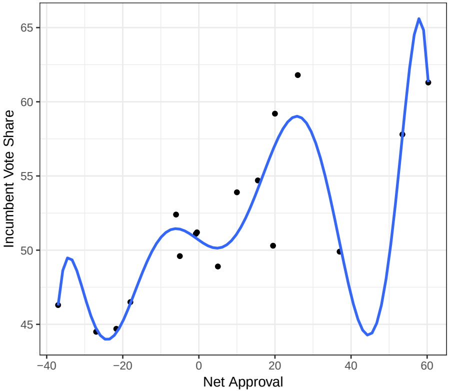
- We can obtain the predictions $\hat{Y}$ with the same data we used to create the model (training test). We call these: in-sample predictions.

Bi-variate regression



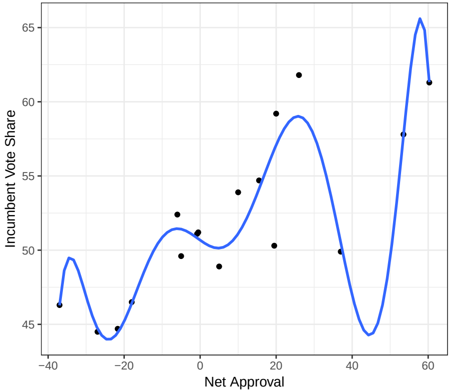
- We can obtain the predictions $\hat{Y}$ with the same data we used to create the model (training test). We call these: in-sample predictions.
- Alternatively we could use new data with the same parameters. We call these: out-of-sample predictions.

Bi-variate regression



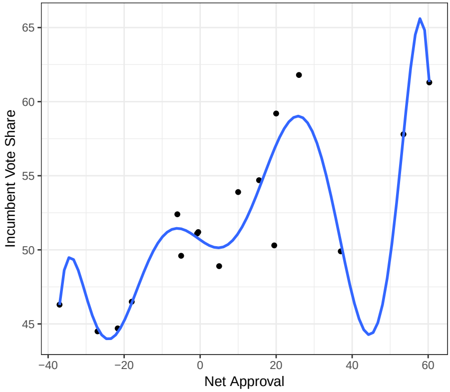
- Why don't create a more flexible model with RSS equal or close to zero?

Bi-variate regression



- Why don't create a more flexible model with RSS equal or close to zero?
- OVERFITTING!!!

Bi-variate regression



- Why don't create a more flexible model with RSS equal or close to zero?
- OVERFITTING!!!
- We could end up estimating noise.
- The model will be perfect for in-sample predictions but it will perform poorly for out-of-sample predictions.

Calculating parameters

Slope:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Intercept:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

Potential issues

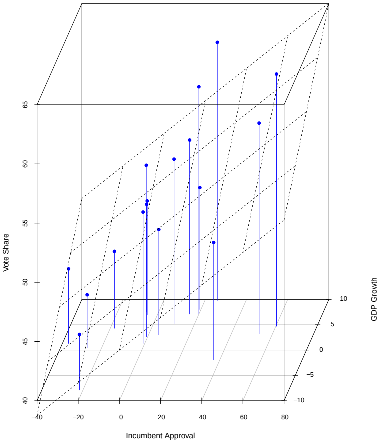
- Non-linearity of the response-predictor relationship
- Outliers (unusual value of Y)
- High leverage points (unusual value of X)
- Collinearity" Refers to the situation in which two or more predictor variables are closely related to one another.

Concepts

- Y: Output, dependent variable, response; X: Input, independent, predictor.
- $f()$: Mapping from X to Y
- Ordinary Least Squares (OLS): Find a linear relationship between X and Y that minimizes RSS (SSE): $Y = \beta_0 + \beta_1 X + \epsilon$
- Parameters: Intercept $\hat{\beta}_0$ and slope β_1
- We can use those parameters to create predictions.
- Overfitting.

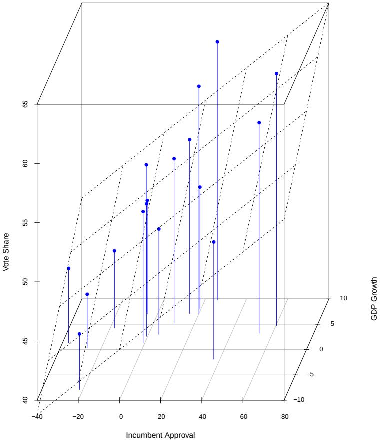
R!

Multivariate regression



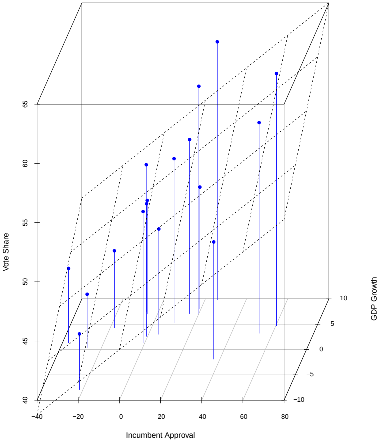
- Same logic but with additional dimensions

Multivariate regression



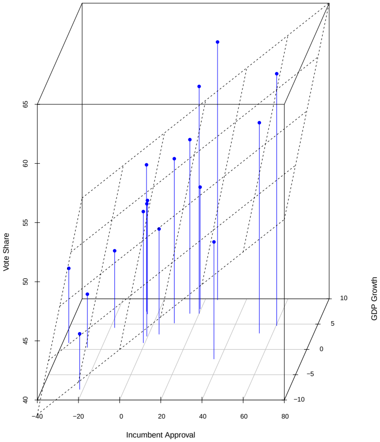
- Same logic but with additional dimensions
- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$

Multivariate regression



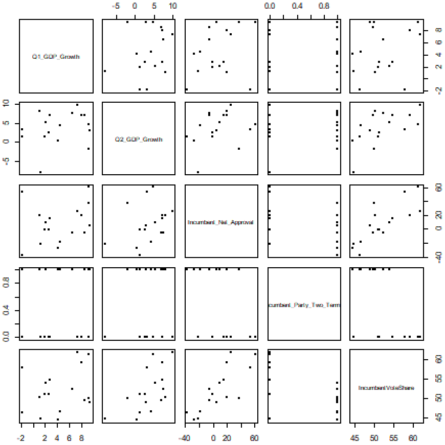
- Same logic but with additional dimensions
- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$
- Again, we use $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$ to predict $\hat{Y}$

Multivariate regression



- Same logic but with additional dimensions
- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$
- Again, we use $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$ to predict $\hat{Y}$
- In this case: Vote Share $= \hat{\beta}_0 + \hat{beta}_1$ Approval + $\hat{beta}_2$ GDP

Multivariate



Multivariate

$$\begin{aligned} IncumbentVoteShare = {} & Incumbent\_Net\_Approval + \\ & Incumbent\_Party\_Two\_Terms + \\ & Q1\_GDP\_Growth + Q2\_GDP\_Growth \end{aligned}$$

Some important questions

- Is at least one of the predictors X_1, X_2,..., X_p useful in predicting the response?
- Do all the predictors help to explain Y, or is only a subset of the predictors useful?
- How well does the model fit the data?
- Given a set of predictor values, what response value should we predict, and how accurate is our prediction?

Some important questions

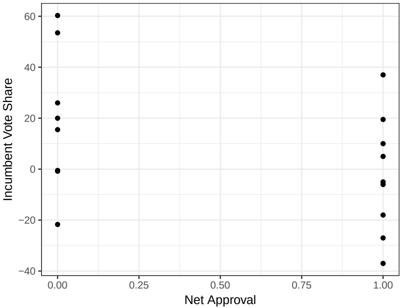
- Is at least one of the predictors X_1, X_2,..., X_p useful in predicting the response?
- Do all the predictors help to explain Y, or is only a subset of the predictors useful?
- How well does the model fit the data?
- Given a set of predictor values, what response value should we predict, and how accurate is our prediction?

Some important questions

- Is at least one of the predictors X_1, X_2,..., X_p useful in predicting the response?
- Do all the predictors help to explain Y, or is only a subset of the predictors useful?
- How well does the model fit the data?
- Given a set of predictor values, what response value should we predict, and how accurate is our prediction?

Extensions and Other Considerations

- Qualitative predictors
- Interaction Terms:
  $Y = X1 + X2 + X1 * X2$
- Non-linear relationships
  $Y = X1 + X1^2$



Comparison of RSS

- RSS Bi-variate Model: 170.0881
- RSS Multivariate Model: 45.72899

Comparison of RSS

- RSS Bi-variate Model: 170.0881
- RSS Multivariate Model: 45.72899

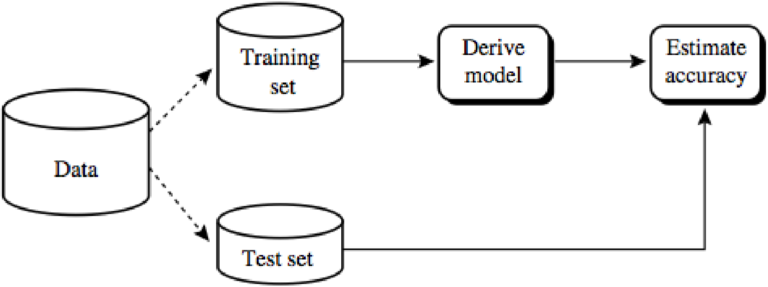
What does this mean?

R!

Divide your data

- **Training set:** A set of examples used to fit the parameters and learn. These are already classified by human coders, or produced in a semi-automated way, to create a 'gold standard'.
  Conventionally 80% of available data.
- **Test set:** A set that follows the same probability distribution and is used to test de model.
  Conventionally 20% of available data.

Divide your data



NEXT

- More on model testing
- Classification
- Remember the new room!