Introduction to Machine Learning for Social Scientists

Class 3: Relating Variables and Creating Models for Prediction

Edgar Franco Vivanco

Stanford University
Department of Political Science

edgarf1@stanford.edu

Summer 2018

Homework 1 Due July 4 at 1:30pm

At which you point, you get another one.

Questions?

So far...

- R basics
- R Markdown
- Principles of Supervised Learning

Today's goals:

- Basic relationship between variables
- Models
- Bi-variate regression
- Prediction

Today's goals:

- Basic relationship between variables
- Models
- Bi-variate regression
- Prediction
- R:
  - Create random variables
  - Run a linear model with the 'lm()' function
  - Create your own functions

Today's goals:

- Basic relationship between variables
- Models
- Bi-variate regression
- Prediction
- R:
  - Create random variables
  - Run a linear model with the 'lm()' function
  - Create your own functions

TA: Haemin Jee

Overview

Logistics

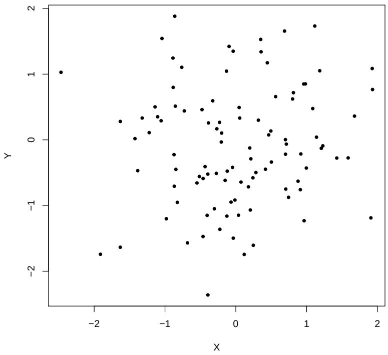
Relationship between variables

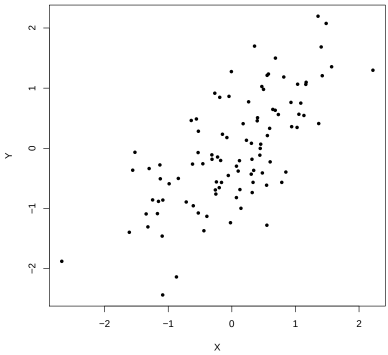
Models

Bi-variate regression and prediction

Relating variables

- How can we say that two variables are related?
- How can we differentiate between a strong or a weak relationship?
- Is the relationship linear?
- How can we use existing information to predict future data?


Y
X
−2
−1
0
1
2


Y
X
−2
−1
0
1
2

Correlation

$$\rho_{X,Y} = Corr(X, Y) = \frac{Cov(X, Y)}{\sigma_X, \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X, \sigma_Y} \quad (1)$$

Where E is the expected value operator, *cov* means covariance, σ are the standard deviations, μ the mean expected values for each variable, and *corr* is a widely used alternative notation for the correlation coefficient.

This value goes from -1 to 1. The Pearson correlation is 1 in the case of a perfect direct (increasing) linear relationship (correlation), -1 in the case of a perfect decreasing (inverse) linear relationship. If the correlation coefficient is 0 then we say that the two variables are independent.

R!

Predicting Election Results

A ML problem: If we know the relationship between variables, can we use this information to forecast future outcomes?

Goal: forecast election winner

Potential predictors

1) GDP Growth
2) Polling data (Incumbent Presidential Popularity)

Conjecture: use relationship in prior elections to predict future election

GDP Growth, popularity $\rightsquigarrow$ input variables.

- *predictors, independent variables, features, attributes, variables*
- X, X_1 (GDP growth), X_2 (popularity)

Election winner $\rightsquigarrow$ output variables.

- *response, dependent variable, outcome, target, labels*
- Y
    - Quantitative (e.g, 15, 3.14, -82000) $\rightsquigarrow$ Regression
    - Categorical (e.g, Republican/Democrat, 0/1, High/Medium/Low) $\rightsquigarrow$ Classification

We assume some relationship between Y and $X = (X_1, X_2, ..., X_p)$, such that:

$$Y = f(X) + \epsilon$$

where f is fixed but unknown function of $X_1, ..., X_p$, and ϵ is a random error term that is is independent of X and has mean zero. .

We assume some relationship between Y and $X = (X_1, X_2, ..., X_p)$, such that:

$$Y = f(X) + \epsilon$$

where f is fixed but unknown function of $X_1, ..., X_p$, and ϵ is a random error term that is is independent of X and has mean zero. .

Machine learning: estimating f with $\hat{f}$.

Why Estimate f?

Two main reasons that we may wish to estimate f:

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

Why Estimate f?

Two main reasons that we may wish to estimate f:

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 $X_1, X_2, ...X_p$ change?
   - $\hat{f}$ no longer treated as a *black box*.
   - Better model = more interpretable

Why Estimate f?

Two main reasons that we may wish to estimate f:

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

We'll focus mostly on prediction in this class.

How Do We Estimate f?

1. Collect a set of n data points with p predictors, called training data.

Year	Incumbent net approval	Incumbent vote share
2012	-.08	51.1
2008	-37	46.3

How Do We Estimate f?

1. Collect a set of n data points with p predictors, called training data.

Year	Incumbent net approval	Incumbent vote share
2012	-.08	51.1
2008	-37	46.3

2. Select a model or method of estimating f.

How Do We Estimate f?

1. Collect a set of n data points with p predictors, called training data.

Year	Incumbent net approval	Incumbent vote share
2012	-.08	51.1
2008	-37	46.3

2. Select a model or method of estimating f.
3. Use the training data to train or fit $\hat{f}$ (our prediction function).

How Do We Estimate f?

1. Collect a set of n data points with p predictors, called training data.

Year	Incumbent net approval	Incumbent vote share
2012	-.08	51.1
2008	-37	46.3

2. Select a model or method of estimating f.
3. Use the training data to train or fit $\hat{f}$ (our prediction function).
4. Use $\hat{f}$ to predict values for Y on previous previously unseen observations.

Year	Incumbent net approval	Incumbent vote share
2016	3	???

How Do We Estimate f?

1. Collect a set of n data points with p predictors, called training data.

Year	Incumbent net approval	Incumbent vote share
2012	-.08	51.1
2008	-37	46.3

2. Select a model or method of estimating f.
3. Use the training data to train or fit $\hat{f}$ (our prediction function).
4. Use $\hat{f}$ to predict values for Y on previous previously unseen observations.

Year	Incumbent net approval	Incumbent vote share
2016	3	???

BUT!! There are many models and methods for estimating f!!!

How Do We Estimate f?

1. Collect a set of n data points with p predictors, called training data.

Year	Incumbent net approval	Incumbent vote share
2012	-.08	51.1
2008	-37	46.3

2. Select a model or method of estimating f.
3. Use the training data to train or fit $\hat{f}$ (our prediction function).
4. Use $\hat{f}$ to predict values for Y on previous previously unseen observations.

Year	Incumbent net approval	Incumbent vote share
2016	3	???

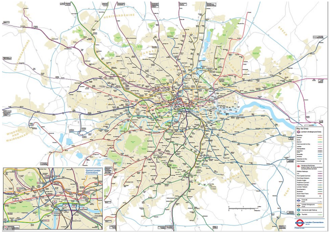
BUT!! There are many models and methods for estimating f!!!

5. Compare predicted response value ($\hat{Y}$) with true response value (Y) for observations in test / validation data to

There is no free lunch in statistics: no one method dominates all others over all possible data sets. Selecting the best method is hard.

All models are wrong... but some are useful.

Models



Test your knowledge

Explain whether each scenario is a classification or regression problem, and indicate whether we are most interested in inference or prediction. Finally, provide n and p.

We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.

Test your knowledge

Explain whether each scenario is a classification or regression problem, and indicate whether we are most interested in inference or prediction. Finally, provide n and p.

We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.

Answer: Regression, inference, $n = 500, p = 3$.

Test your knowledge

Explain whether each scenario is a classification or regression problem, and indicate whether we are most interested in inference or prediction. Finally, provide n and p.

We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.

Test your knowledge

Explain whether each scenario is a classification or regression problem, and indicate whether we are most interested in inference or prediction. Finally, provide n and p.

We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.

Answer: Classification, prediction, $n = 20, p = 13$.

Test your knowledge

What is the difference between $f(X)$ and $\hat{f}(X)$?

Test your knowledge

What is the difference between $f(X)$ and $\hat{f}(X)$?

Answer: f is the true function that maps X onto Y. $\hat{f}(X)$ is the estimated / prediction function trained on sample data, mapping observed X onto observed Y.

Overview

Logistics

Relationship between variables

Models

Bi-variate regression and prediction

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: Simple (today) and Multiple (next week)

Evaluation (Focus of next lectures):

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

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` , `%*%`

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

R!

What is Linear Regression?

Linear regression is a simple approach for supervised learning.

- Around since 1800s.
- Still a widely used tool for predicting quantitative response.
- Good jumping-off point for newer approaches.
- We need to understand it before moving on!

What is Linear Regression?

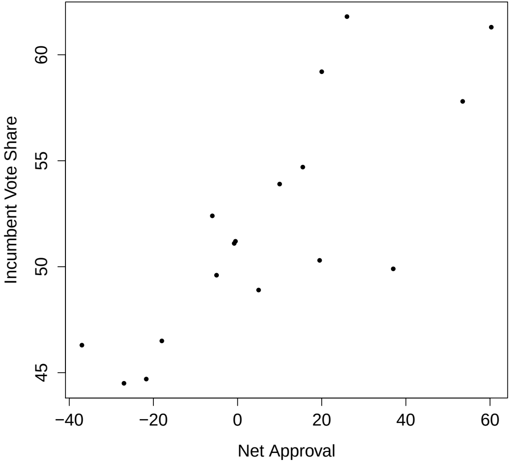
Linear regression is a simple approach for supervised learning.

- Around since 1800s.
- Still a widely used tool for predicting quantitative response.
- Good jumping-off point for newer approaches.
- We need to understand it before moving on!

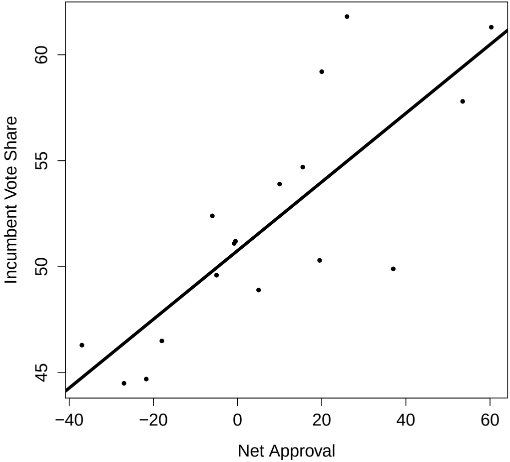
Simple linear regression:

- Assumes a linear relationship between quantitative response Y and a **single** variable X.
- Also called bivariate regression because there are two variables (X and Y)

Bivariate Regression: Geometric Perspective



Bivariate Regression: Geometric Perspective



Bivariate Regression: Function Perspective

For each election i, $(i = 1948, 1952, \ldots, 2012)$,

Bivariate Regression: Function Perspective

For each election i, ($i = 1948, 1952, \ldots, 2012$),
Let:

Approval_i = Incumbent Net Approval in election i

Vote_i = Incumbent Vote Share in election i.

Bivariate Regression: Function Perspective

For each election i, $(i = 1948, 1952, \ldots, 2012)$,
Let:

Approval_i = Incumbent Net Approval in election i
Vote_i = Incumbent Vote Share in election i.

Find a function f such that relates Approval_i to Vote_i vote share

Bivariate Regression: Function Perspective

For each election i, ($i = 1948, 1952, \ldots, 2012$),
Let:

Approval_i = Incumbent Net Approval in election i
Vote_i = Incumbent Vote Share in election i.

Find a function f such that relates Approval_i to Vote_i vote share

$$\text{Vote}_i = f(\text{Approval}_i) + \epsilon_i$$

Bivariate Regression: Function Perspective

For each election i, ($i = 1948, 1952, \ldots, 2012$),
Let:

Approval_i = Incumbent Net Approval in election i
Vote_i = Incumbent Vote Share in election i.

Find a function f such that relates Approval_i to Vote_i vote share

$$\begin{aligned} \text{Vote}_i &= f(\text{Approval}_i) + \epsilon_i \\ \text{Vote}_i &= \beta_0 + \beta_1 \text{Approval}_i + \epsilon_i \end{aligned}$$

Bivariate Regression: Function Perspective

For each election i, $(i = 1948, 1952, \ldots, 2012)$,
Let:

$$\text{Approval}_i = \text{Incumbent Net Approval in election } i$$
$$\text{Vote}_i = \text{Incumbent Vote Share in election } i.$$

Find a function f such that relates Approval_i to Vote_i vote share

$$\begin{aligned} \text{Vote}_i &= f(\text{Approval}_i) + \epsilon_i \\ \text{Vote}_i &= \beta_0 + \beta_1 \text{Approval}_i + \epsilon_i \end{aligned}$$

β_0 and β_1 are two unknown constraints known as the model coefficients or parameters.

Bivariate Regression: Function Perspective

For each election i, $(i = 1948, 1952, \ldots, 2012)$,
Let:

$$\text{Approval}_i = \text{Incumbent Net Approval in election } i$$
$$\text{Vote}_i = \text{Incumbent Vote Share in election } i.$$

Find a function f such that relates Approval_i to Vote_i vote share

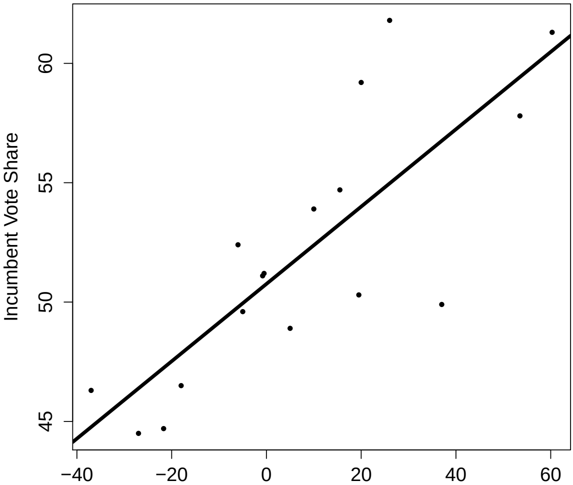
$$\begin{aligned}\text{Vote}_i &= f(\text{Approval}_i) + \epsilon_i \\ \text{Vote}_i &= \beta_0 + \beta_1 \text{Approval}_i + \epsilon_i\end{aligned}$$

β_0 and β_1 are two unknown constraints known as the model coefficients or parameters.

We use our training data to produce estimates:

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

Geometric and Function Perspective Combined



1) What corresponds to β_0? (intercept)
2) What corresponds to β_1? (slope)
3) What corresponds to ϵ_i ? (residual)
4) What is an in-sample estimate and what is an out-of-sample estimates?

Fitting a Bivariate Regression

Goal: Obtain coefficient estimates β_0 and β_1 such that the linear model fits the available data **well**, i.e. **close**.

Fitting a Bivariate Regression

Goal: Obtain coefficient estimates β_0 and β_1 such that the linear model fits the available data **well**, i.e. **close**.

$$\text{Vote}_i = \beta_0 + \beta_1 \text{Approval}_i + \epsilon_i$$

Fitting a Bivariate Regression

Goal: Obtain coefficient estimates β_0 and β_1 such that the linear model fits the available data **well**, i.e. **close**.

$$
\begin{aligned}
\text{Vote}_i &= \beta_0 + \beta_1 \text{Approval}_i + \epsilon_i \\
\text{Vote}_i &= \underbrace{\beta_0 + \beta_1 \text{Approval}_i}_{\widehat{\text{Vote}_i}} + \epsilon_i
\end{aligned}
$$

Fitting a Bivariate Regression

Goal: Obtain coefficient estimates β_0 and β_1 such that the linear model fits the available data **well**, i.e. **close**.

$$\begin{aligned}
\text{Vote}_i &= \beta_0 + \beta_1 \text{Approval}_i + \epsilon_i \\
\text{Vote}_i &= \underbrace{\beta_0 + \beta_1 \text{Approval}_i}_{\widehat{\text{Vote}_i}} + \epsilon_i \\
\epsilon_i &= \text{Vote}_i - \widehat{\text{Vote}_i}
\end{aligned}$$

Fitting a Bivariate Regression

Goal: Obtain coefficient estimates β_0 and β_1 such that the linear model fits the available data **well**, i.e. **close**.

$$\begin{aligned}
\text{Vote}_i &= \beta_0 + \beta_1 \text{Approval}_i + \epsilon_i \\
\text{Vote}_i &= \underbrace{\beta_0 + \beta_1 \text{Approval}_i}_{\widehat{\text{Vote}_i}} + \epsilon_i \\
\epsilon_i &= \text{Vote}_i - \widehat{\text{Vote}_i} \\
\sum_{i=1948}^{2012} \epsilon_i &= \sum_{i=1948}^{2012} \left(\text{Vote}_i - \widehat{\text{Vote}_i} \right)
\end{aligned}$$

Fitting a Bivariate Regression

Goal: Obtain coefficient estimates β_0 and β_1 such that the linear model fits the available data **well**, i.e. **close**.

$$\begin{aligned}
\text{Vote}_i &= \beta_0 + \beta_1 \text{Approval}_i + \epsilon_i \\
\text{Vote}_i &= \underbrace{\beta_0 + \beta_1 \text{Approval}_i}_{\widehat{\text{Vote}_i}} + \epsilon_i \\
\epsilon_i &= \text{Vote}_i - \widehat{\text{Vote}_i} \\
\sum_{i=1948}^{2012} \epsilon_i &= \sum_{i=1948}^{2012} \left(\text{Vote}_i - \widehat{\text{Vote}_i}\right)
\end{aligned}$$

Goal: choose β_0 and β_1 to minimize $\sum_{i=1948}^{2012} \epsilon_i$?

Fitting a Bivariate Regression

Fitting a Bivariate Regression

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

Fitting a Bivariate Regression

$$\epsilon_i^2 = \left(\text{Vote}_i - \widehat{\text{Vote}_i}\right)^2$$

$$\sum_{i=1}^{N} \epsilon_i^2 = \sum_{i=1}^{N} \left(\text{Vote}_i - \widehat{\text{Vote}_i}\right)^2$$

Fitting a Bivariate Regression

$$
\begin{aligned}
\epsilon_i^2 &= \left(\text{Vote}_i - \widehat{\text{Vote}_i}\right)^2 \\
\sum_{i=1}^{N} \epsilon_i^2 &= \sum_{i=1}^{N} \left(\text{Vote}_i - \widehat{\text{Vote}_i}\right)^2 \\
\sum_{i=1}^{N} \epsilon_i^2 &= \sum_{i=1}^{N} (\text{Vote}_i - \beta_0 - \beta_1 \text{Approval}_i)^2
\end{aligned}
$$

Fitting a Bivariate Regression

$$\epsilon_i^2 = \left(\text{Vote}_i - \widehat{\text{Vote}_i}\right)^2$$

$$\sum_{i=1}^{N} \epsilon_i^2 = \sum_{i=1}^{N} \left(\text{Vote}_i - \widehat{\text{Vote}_i}\right)^2$$

$$\sum_{i=1}^{N} \epsilon_i^2 = \sum_{i=1}^{N} (\text{Vote}_i - \beta_0 - \beta_1 \text{Approval}_i)^2$$

Goal: choose β_0 and β_1 to minimize $\sum_{i=1948}^{2012} \epsilon_i^2$?

Fitting a Bivariate Regression

Fitting a Bivariate Regression

- $\sum_{i=1948}^{2012} \epsilon_i^2 =$ Sum of Squared Residuals or Residual Sum of Squares or Sum of Squared Error

Fitting a Bivariate Regression

- $\sum_{i=1948}^{2012} \epsilon_i^2 =$ Sum of Squared Residuals or Residual Sum of Squares or Sum of Squared Error

- Choose β_0 and β_1 to minimize $\sum_{i=1948}^{2012} \epsilon_i^2 \rightsquigarrow$ cost function

Fitting a Bivariate Regression

- $\sum_{i=1948}^{2012} \epsilon_i^2 =$ Sum of Squared Residuals or Residual Sum of Squares or Sum of Squared Error
- Choose β_0 and β_1 to minimize $\sum_{i=1948}^{2012} \epsilon_i^2 \rightsquigarrow$ cost function

Two ways to minimize cost function:

- Calculus!
- Gradient Descent

Fitting a Bivariate Regression

- $\sum_{i=1948}^{2012} \epsilon_i^2$ = Sum of Squared Residuals or Residual Sum of Squares or Sum of Squared Error
- Choose β_0 and β_1 to minimize $\sum_{i=1948}^{2012} \epsilon_i^2 \rightsquigarrow$ cost function

Two ways to minimize cost function:

- Calculus!
- Gradient Descent

R!

Our Estimated Prediction Function

Our Estimated Prediction Function

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

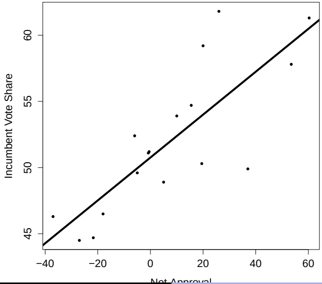
Our Estimated Prediction Function

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

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!`



Predicting 2016

Gallup (1/3/2016-1/5/2016):

Predicting 2016

Gallup (1/3/2016-1/5/2016):

- 46% Approve
- 50% Disapprove
- -4% Net Approval

Predicting 2016

Gallup (1/3/2016-1/5/2016):

- 46% Approve
- 50% Disapprove
- -4% Net Approval

$$\widehat{\text{Vote}_{2016}} = 50.76 + 0.16 \times -4$$

Predicting 2016

Gallup (1/3/2016-1/5/2016):

- 46% Approve
- 50% Disapprove
- -4% Net Approval

$$\begin{aligned} \widehat{\text{Vote}_{2016}} &= 50.76 + 0.16 \times -4 \\ &= 50.76 - 0.64 = 50.12 \end{aligned}$$

Predicting 2016

Gallup (1/3/2016-1/5/2016):

- 46% Approve
- 50% Disapprove
- -4% Net Approval

$$\begin{aligned}\widehat{\text{Vote}_{2016}} &= 50.76 + 0.16 \times -4 \\ &= 50.76 - 0.64 = 50.12\end{aligned}$$

Actual share: 51.1

Predicting 2016

Gallup (1/3/2016-1/5/2016):

- 46% Approve
- 50% Disapprove
- -4% Net Approval

$$\begin{aligned} \widehat{\text{Vote}_{2016}} &= 50.76 + 0.16 \times -4 \\ &= 50.76 - 0.64 = 50.12 \end{aligned}$$

Actual share: 51.1
Residual / Error: 51.1 - 50.12 = 0.98

Predicting 2016

Gallup (1/3/2016-1/5/2016):

- 46% Approve
- 50% Disapprove
- -4% Net Approval

$$\begin{aligned}\widehat{\text{Vote}_{2016}} &= 50.76 + 0.16 \times -4 \\ &= 50.76 - 0.64 = 50.12\end{aligned}$$

Actual share: 51.1
Residual / Error: 51.1 - 50.12 = 0.98

What went wrong?

Why was our prediction wrong?

What went wrong?

Why was our prediction wrong?

- reducible error difference between $\hat{f}$ (observed) and f (unobserved)

What went wrong?

Why was our prediction wrong?

- reducible error difference between $\hat{f}$ (observed) and f (unobserved)
- $\widehat{\beta}_0, \widehat{\beta}_1 \approx \beta_0, \beta_1$

What went wrong?

Why was our prediction wrong?

- reducible error difference between $\hat{f}$ (observed) and f (unobserved)
- $\widehat{\beta}_0, \widehat{\beta}_1 \approx \beta_0, \beta_1$
- Tools for assessing accuracy of coefficient estimates: confidence intervals, t statistics, p values, etc

What went wrong?

Why was our prediction wrong?

- reducible error difference between $\hat{f}$ (observed) and f (unobserved)
- $\widehat{\beta}_0, \widehat{\beta}_1 \approx \beta_0, \beta_1$
- Tools for assessing accuracy of coefficient estimates: confidence intervals, t statistics, p values, etc
- irreducible error ϵ (catch-all for what we miss with this simple model)

Caution: function is defined for all values!

Caution: function is defined for all values!

- Net approval: 100%

Caution: function is defined for all values!

- Net approval: 100%
- Net approval: -100%

NEXT

- NO CLASS ON WEDNESDAY (July 4th)
- Multivariate Regression
- Dividing sets for prediction
- **Homework 2:** Released on Wed July 4th and due on Wed July 11h

  In section:

  - More on 'predict'
  - Your own functions
  - For loops