Introduction to Linear Regression

Samuel C. Gutekunst

Cornell University
scg94@cornell.edu

June 5, 2018

Example: Gingles Criteria

(Thornburg v Gingles, 1986)

To demonstrate racial gerrymandering, show that all three of the following hold:

1. A minority group is “sufficiently numerous and compact to form a majority in a single-member district”; and
2. **The minority group is "politically cohesive"; and**
3. **The "majority votes sufficiently as a bloc to enable it ... usually to defeat the minority’s preferred candidate."’**

Example: Gingles Criteria

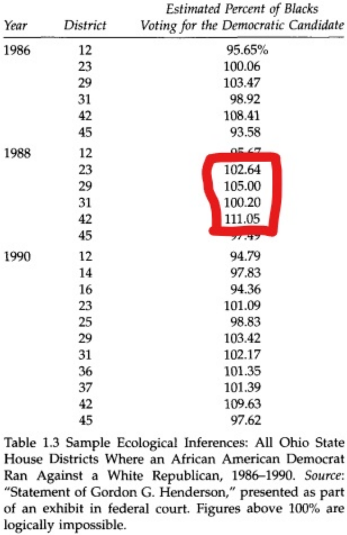
(Thornburg v Gingles, 1986)

To demonstrate racial gerrymandering, show that all three of the following hold:

1. A minority group is “sufficiently numerous and compact to form a majority in a single-member district”; and
2. **The minority group is "politically cohesive"; and**
3. **The "majority votes sufficiently as a bloc to enable it ... usually to defeat the minority’s preferred candidate."’**

$$\begin{aligned}\#\text{ Dem Votes in Precinct i} \approx\ & (\text{Minority Pop})\,(\text{Minority Voting Rate})\\ & + (\text{Majority Pop})\,(\text{Majority Voting Rate})\end{aligned}$$

Example: Gingles Criteria



Year	*District*	*Estimated Percent of Blacks Voting for the Democratic Candidate*
1986	12	95.65%
	23	100.06
	29	103.47
	31	98.92
	42	108.41
	45	93.58
1988	12	95.67
	23	102.64
	29	105.00
	31	100.20
	42	111.05
	45	97.49
1990	12	94.79
	14	97.83
	16	94.36
	23	101.09
	25	98.83
	29	103.42
	31	102.17
	36	101.35
	37	101.39
	42	109.63
	45	97.62

Table 1.3 Sample Ecological Inferences: All Ohio State House Districts Where an African American Democrat Ran Against a White Republican, 1986–1990. *Source*: "Statement of Gordon G. Henderson," presented as part of an exhibit in federal court. Figures above 100% are logically impossible.

Figure from King, 97.

Example: Gingles Criteria

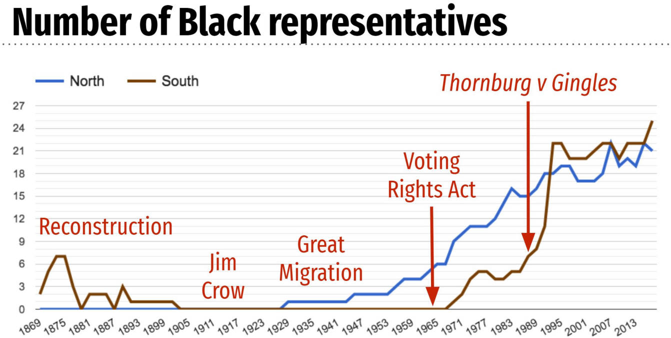
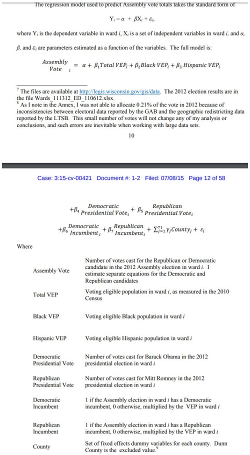


Figure from Mira Bernstein.

Example: Predicting Total Votes



The regression model used to predict Assembly vote totals takes the standard form of

$$Y_i = \alpha + \beta X_i + \varepsilon_i,$$

where Y_i is the dependent variable in ward *i*, X_i is a set of independent variables in ward *i*, and α, β, and ε_i are parameters estimated as a function of the variables. The full model is:

$$\underset{i}{\begin{matrix}\textit{Assembly}\\ \textit{Vote}\end{matrix}} = \alpha + \beta_1 \textit{Total VEP}_i + \beta_2 \textit{Black VEP}_i + \beta_3\ \textit{Hispanic VEP}_i$$

[7] The files are available at http://legis.wisconsin.gov/gis/data. The 2012 election results are in the file Wards_111312_ED_110612.xlsx.

[8] As I note in the Annex, I was not able to allocate 0.21% of the vote in 2012 because of inconsistencies between electoral data reported by the GAB and the geographic redistricting data reported by the LTSB. This small number of votes will not change any of my analysis or conclusions, and such errors are inevitable when working with large data sets.

10

Case: 3:15-cv-00421 Document #: 1-2 Filed: 07/08/15 Page 12 of 58

$$+\beta_4 \begin{matrix}\textit{Democratic}\\ \textit{Presidential Vote}_i\end{matrix} + \beta_5 \begin{matrix}\textit{Republican}\\ \textit{Presidential Vote}_i\end{matrix}$$

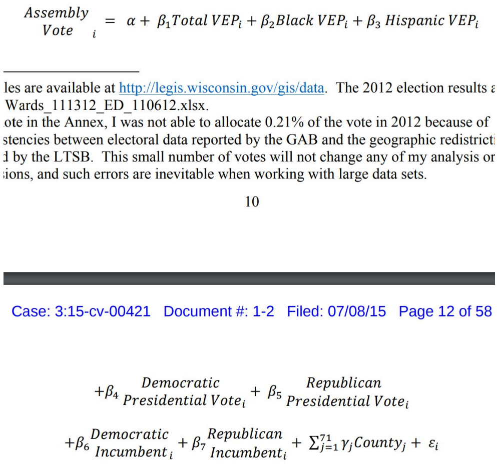
$$+\beta_6 \begin{matrix}\textit{Democratic}\\ \textit{Incumbent}_i\end{matrix} + \beta_7 \begin{matrix}\textit{Republican}\\ \textit{Incumbent}_i\end{matrix} + \sum_{j=1}^{71} \gamma_j \textit{County}_j + \varepsilon_i$$

Where

Assembly Vote	Number of votes cast for the Republican or Democratic candidate in the 2012 Assembly election in ward *i*. I estimate separate equations for the Democratic and Republican candidates
Total VEP	Voting eligible population in ward *i*, as measured in the 2010 Census
Black VEP	Voting eligible Black population in ward *i*
Hispanic VEP	Voting eligible Hispanic population in ward *i*
Democratic Presidential Vote	Number of votes cast for Barack Obama in the 2012 presidential election in ward *i*
Republican Presidential Vote	Number of votes cast for Mitt Romney in the 2012 presidential election in ward *i*
Democratic Incumbent	1 if the Assembly election in ward *i* has a Democratic incumbent, 0 otherwise, multiplied by the VEP in ward *i*
Republican Incumbent	1 if the Assembly election in ward *i* has a Republican incumbent, 0 otherwise, multiplied by the VEP in ward *i*
County	Set of fixed effects dummy variables for each county. Dunn County is the excluded value.[9]

From Kenneth Mayer, *Analysis of the Efficiency Gaps of Wisconsin's Current Legislative District Plans and Plaintiffs' Demonstration Plan*

Example: Predicting Total Votes



$$\underset{i}{\begin{matrix}Assembly\\ Vote\end{matrix}} = \alpha + \beta_1 Total\ VEP_i + \beta_2 Black\ VEP_i + \beta_3\ Hispanic\ VEP_i$$

les are available at http://legis.wisconsin.gov/gis/data. The 2012 election results a
Wards_111312_ED_110612.xlsx.
ote in the Annex, I was not able to allocate 0.21% of the vote in 2012 because of
stencies between electoral data reported by the GAB and the geographic redistricti
d by the LTSB. This small number of votes will not change any of my analysis or
ions, and such errors are inevitable when working with large data sets.

10

Case: 3:15-cv-00421 Document #: 1-2 Filed: 07/08/15 Page 12 of 58

$$+\beta_4 \begin{matrix}Democratic\\ Presidential\ Vote_i\end{matrix} + \beta_5 \begin{matrix}Republican\\ Presidential\ Vote_i\end{matrix}$$

$$+\beta_6 \begin{matrix}Democratic\\ Incumbent_i\end{matrix} + \beta_7 \begin{matrix}Republican\\ Incumbent_i\end{matrix} + \sum_{j=1}^{71} \gamma_j County_j + \varepsilon_i$$

From Kenneth Mayer, *Analysis of the Efficiency Gaps of Wisconsin's Current Legislative District Plans and Plaintiffs' Demonstration Plan*

This Talk

Sole Goal

Become comfortable using and understanding R for multilinear regression.

- Understand political science papers
- Replicate analysis
- Do your own analysis
- Convince a lay person that you're doing reasonable things

- Modeling framework
- Terminology
- Statistical reasoning

This Talk

Sole Goal

Become comfortable using and understanding R for multilinear regression.

- Understand political science papers
- Replicate analysis
- Do your own analysis
- Convince a lay person that you're doing reasonable things

1. If you're good with R, linear regression, and statistical modeling, you should probably leave.
2. If you know linear regression and statistical modeling but not R, do the computational lab.
3. Else stay here, then do the lab!

Outline

1. Introduction: Lot's About the Simple Linear Regression

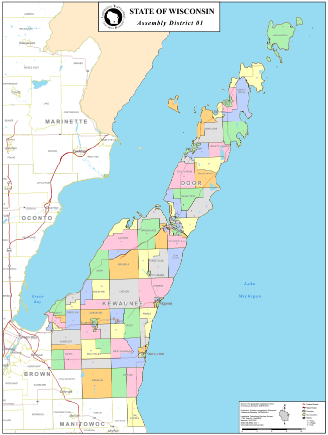
2. Checking Assumptions of the Linear Model

3. Extending the Simple Linear Model: More Predictors

4. Inference Using the Linear Model

5. Generalizations of Linear Regression

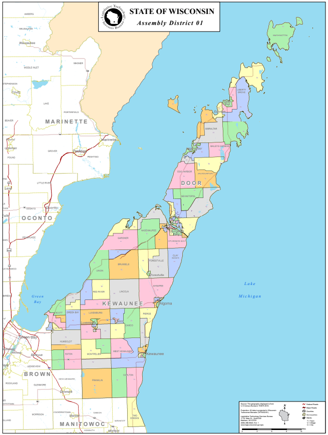
Motivating Example: WI District 1



Wards in District 1. From `legis.wisconsin.gov`

- Say we want to estimate the total democratic votes in a future election

Motivating Example: WI District 1



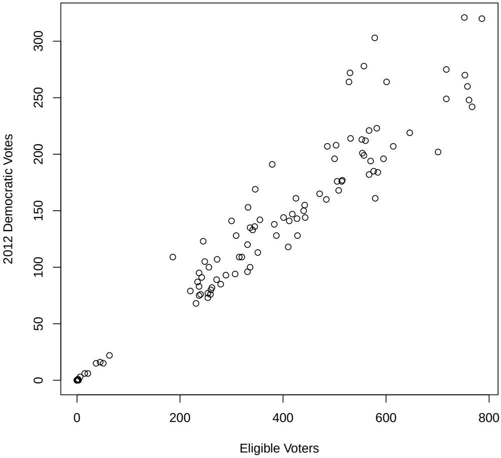
Wards in District 1. From `legis.wisconsin.gov`

- Say we want to estimate the total number of democratic votes in a future election
- **Question:** Within this district, does the total democratic vote grow linearly with the voting eligible population?

Motivating Example: WI District 1

Sanity Check

Is this reasonable?? Zoom in on the ∼103 wards in district 1!



Notational Setup

- We have $n = 103$ **paired data points**

  $$(y_1, x_1), (y_2, x_2), ..., (y_{103}, x_{103})$$

  where $y_1, ..., y_n$ are measurements of a **response variable** and $x_1, ..., x_n$ are corresponding measurements of an **explanatory variable**.
- Our goal is to understand how the response variable (democratic vote) depends on the explanatory variable (voting eligible population).

The Statistical Modeling Process

Given data and a research question, the modeling process typically involves:

1. Assume a model that specifies the "type" of relationship between the variables you're interested in.
2. Find the version of the model that "fits best."
3. Check to see if the model "is good."
4. If the model "is good," draw inferences from the model.

The Statistical Modeling Process

Given data and a research question, the modeling process typically involves:

1. Assume a model that specifies the "type" of relationship between the variables you're interested in.
2. Find the version of the model that "fits best."
3. Check to see if the model "is good."
4. If the model "is good," draw inferences from the model.
   - 4.1 Think if the model and inferences make sense!
   - 4.2 Communicate conclusions to people who may have a fear and hatred of mathematics!

Think!

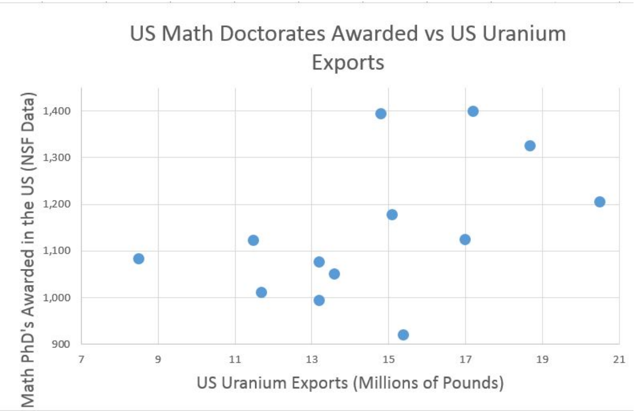


Figure: Data Source: `SpuriousCorellations`

Setting up a Statistical Model: Informally

- In the simple linear model, we assume a "mostly linear" relationship between the response and explanatory variable. Here:

$$\#\ \text{Dem. Votes} = b_0 + b_1(\#\ \text{Eligible Voters in Ward}) + \text{Random Error.}$$

  - b_0 and b_1 are **unknown** parameters we would like to estimate
  - We assume the random error follows a "bell curve" and is, on average, zero.

Setting up a Statistical Model: Formally

- In the simple linear model, we assume a “mostly linear” relationship between the response and explanatory variable.
- Mathematically,

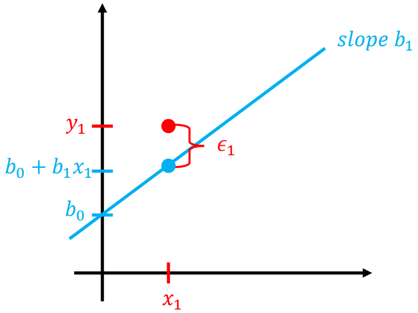
$$Y_i = b_0 + b_1 X_i + \epsilon_i, \qquad i = 1, ..., 103.$$

  - Y_i is the number of democratic votes in the ith ward of Assembly District 1 (for $i = 1, ..., 103$)
  - X_i is the voting eligible population of the ith ward of Assembly District 1
  - b_0 and b_1 are **unknown** parameters we would like to estimate
  - ϵ_i is random noise (formally, $\epsilon_i \sim N(0, \sigma^2)$ are independent and identically distributed normal random variables with mean zero)

Pictures

- In the simple linear model, we assume a "mostly linear" relationship between the response and explanatory variable.
- Mathematically,

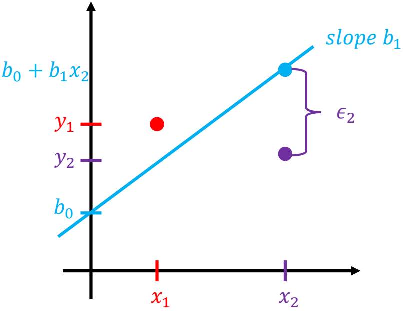
$$Y_i = b_0 + b_1 X_i + \epsilon_i, \qquad i = 1, \ldots, 103.$$



Pictures

- In the simple linear model, we assume a “mostly linear” relationship between the response and explanatory variable.
- Mathematically,

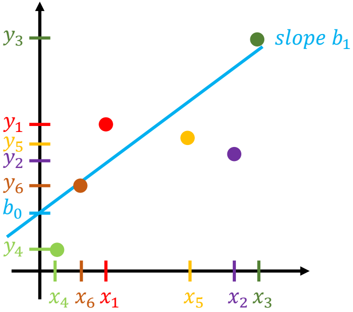
$$Y_i = b_0 + b_1 X_i + \epsilon_i, \qquad i = 1, ..., 103.$$



Pictures

- In the simple linear model, we assume a "mostly linear" relationship between the response and explanatory variable.
- Mathematically,

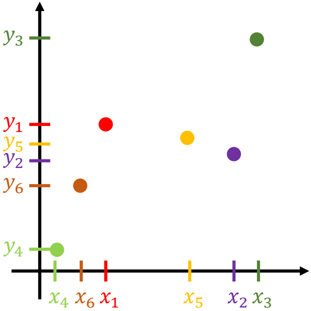
$$Y_i = b_0 + b_1 X_i + \epsilon_i, \quad i = 1, ..., 103.$$



Pictures

- In the simple linear model, we assume a “mostly linear” relationship between the response and explanatory variable.
- Mathematically,

$$Y_i = b_0 + b_1 X_i + \epsilon_i, \qquad i = 1, ..., 103.$$



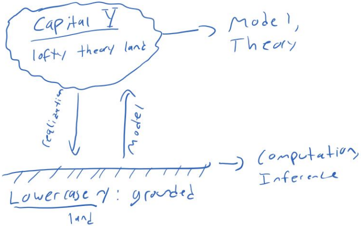
Pictures

- In the simple linear model, we assume a “mostly linear” relationship between the response and explanatory variable.
- Mathematically,

$$Y_i = b_0 + b_1 X_i + \epsilon_i, \qquad i = 1, ..., 103.$$

Capitalization Matters

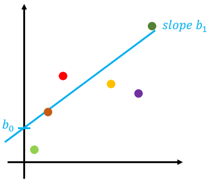
Capital letters refer to an underlying model/random variables. Lowercase letters refer to realizations of that model/real data.

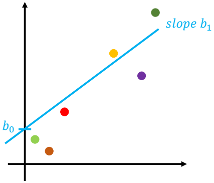


Pictures

- In the simple linear model, we assume a “mostly linear” relationship between the response and explanatory variable.
- Mathematically,

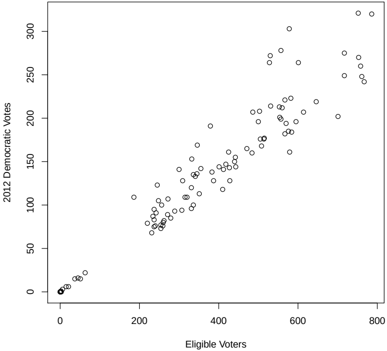
$$Y_i = b_0 + b_1 X_i + \epsilon_i, \qquad i = 1, ..., 103.$$





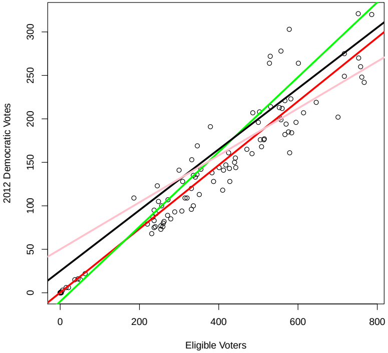
Finding b_0 and b_1

- Our mathematical model says $Y_i = b_0 + b_1 X_i + \epsilon_i, \quad i = 1, ..., 103.$
- We have one realization $(y_1, x_1), ..., (y_{103}, x_{103})$ and want to **reverse engineer** b_0 and b_1.



Finding b_0 and b_1

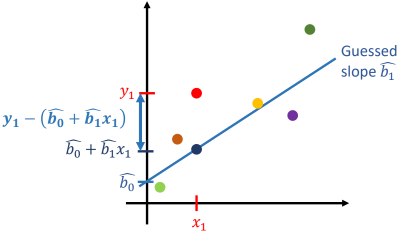
- Our mathematical model says
  $Y_i = b_0 + b_1 X_i + \epsilon_i, \quad i = 1, ..., 103.$
- We have one realization $(y_1, x_1), ..., (y_{103}, x_{103})$ and want to **reverse engineer** b_0 and b_1.



Finding b_0 and b_1

- Our mathematical model says $Y_i = b_0 + b_1 X_i + \epsilon_i, \quad i = 1, ..., 103.$
- We have one realization $(y_1, x_1), ..., (y_{103}, x_{103})$ and want to **reverse engineer** b_0 and b_1.

Idea: Minimize "Squared Error Loss"



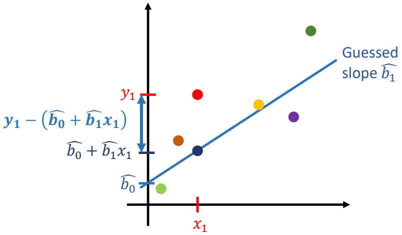
Guessed slope $\widehat{b_1}$
y_1
$\boldsymbol{y_1 - (\widehat{b_0} + \widehat{b_1} x_1)}$
$\widehat{b_0} + \widehat{b_1} x_1$
$\widehat{b_0}$
x_1

Finding b_0 and b_1

- Our mathematical model says $Y_i = b_0 + b_1 X_i + \epsilon_i, \quad i = 1, ..., 103.$
- We have one realization $(y_1, x_1), ..., (y_{103}, x_{103})$ and want to **reverse engineer** b_0 and b_1.

Idea: Minimize "Squared Error Loss." Guess $\hat{b_0}$ and $\hat{b_1}$ that minimize

$$\sum_{i=1}^{103} \left(y_i - \left(\hat{b_0} + \hat{b_1} x_i\right)\right)^2.$$


Guessed slope $\widehat{b_1}$
y_1
$y_1 - (\widehat{b_0} + \widehat{b_1} x_1)$
$\widehat{b_0} + \widehat{b_1} x_1$
$\widehat{b_0}$
x_1

Finding b_0 and b_1

- Our mathematical model says $Y_i = b_0 + b_1 X_i + \epsilon_i, \quad i = 1, ..., 103.$
- We have one realization $(y_1, x_1), ..., (y_{103}, x_{103})$ and want to **reverse engineer** b_0 and b_1.

Idea: Minimize "Squared Error Loss." Guess $\hat{b_0}$ and $\hat{b_1}$ that minimize

$$\sum_{i=1}^{103} \left(y_i - \left(\hat{b_0} + \hat{b_1} x_i \right) \right)^2 .$$

- This handles signs and penalizes "extreme" differences
- Gauss and Legendre said so
- This has nice statistical properties (see: MLE, Gauss-Markov Theorem, linear, unbiased, etc.)
- It's "easy" to do computationally
- It has a nice geometric interpretation

Ordinary Least Squares Regression

Choosing $\hat{b_0}$ and $\hat{b_1}$ to minimize the squared error loss, our model is

$$Y_i = 0.26 + 0.37X_i + \epsilon_i.$$

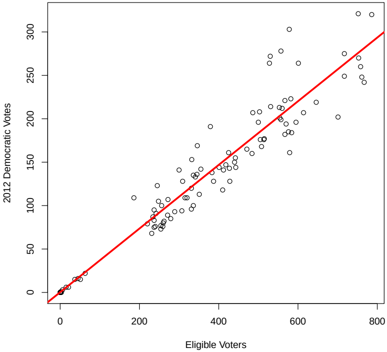
In the Style of Mayer

$$\# \text{ Democratic Votes}_i = 0.26 + 0.37 \text{ Ward Voting Pop}_i + \epsilon_i.$$

Ordinary Least Squares Regression

Choosing $\hat{b_0}$ and $\hat{b_1}$ to minimize the squared error loss, our model is

$$Y_i = 0.26 + 0.37X_i + \epsilon_i.$$



Ordinary Least Squares Regression

Choosing $\hat{b_0}$ and $\hat{b_1}$ to minimize the squared error loss, our model is

$$Y_i = 0.26 + 0.37X_i + \epsilon_i.$$

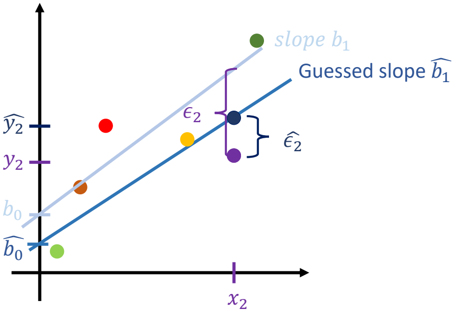
Prediction

Because we assume that the ϵ_i are "on average zero," we can guess the number of votes if a wards population changes using

$$\#\text{ Democratic Votes} = 0.26 + 0.37 \text{ Ward Voting Pop}$$

Some More Words

- The **fitted values** are $\hat{y_i} = \hat{b_0} + \hat{b_1} x_i$.
- The **residuals** are $\hat{\epsilon_i} = y_i - \hat{y_i}$.



Summary so far

- We assumed a convenient (and at least somewhat plausible) model: guessing that the number of democratic votes in a ward grows linearly with the population.
- We used ordinary least squares regression (OLS) to “fit the model” (guess b_0, b_1).
- The fitted model lets us interpret the data and we can use it to predict the number of democratic votes in a future election (with different ward populations)

What's Missing?

- How do we check that this model is reasonable?
- How do we extend to multiple explanatory variables?
- What if we wanted to have a general model for all assembly districts?
- How do we measure how reliable insights from the model are?

Assumptions in our Simple Model

$$\# \text{ Dem. Votes} = b_0 + b_1(\# \text{ Eligible Voters in Ward}) + \text{Random Error}.$$

We are assuming:

1. The relationship is "almost linear": on average

$$\mathbb{E}[Y_i] = b_0 + b_1 X_i.$$

2. The random errors are independently drawn from a bell curve
3. Our data is accurate

R Lab

The R lab shows you some ways to do these things!

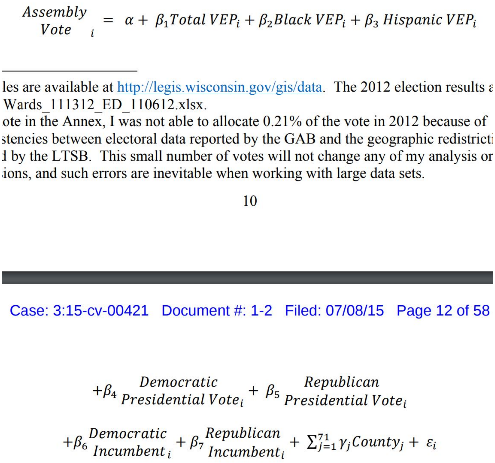
Linear Models: Examples

$$\#\text{ Dem Votes}_i = b_0 + b_1 \#\text{ Voters}_i + \epsilon_i$$

$$\#\text{ Dem Votes}_i = b_0 + b_1 \#\text{ Voters}_i + b_2 \#\text{ Minority Voters}_i + \epsilon_i$$

$$\#\text{ Dem Votes}_i = b_0 + b_1 (\#\text{ Voters}_i) \times (\#\text{ Minority Voters}_i) + \epsilon_i$$

Linear Models: Examples



$$\begin{matrix}\mathit{Assembly}\\ \mathit{Vote}\end{matrix}_{\,i} = \alpha + \beta_1 \mathit{Total\ VEP}_i + \beta_2 \mathit{Black\ VEP}_i + \beta_3\ \mathit{Hispanic\ VEP}_i$$

les are available at http://legis.wisconsin.gov/gis/data. The 2012 election results a
Wards_111312_ED_110612.xlsx.
ote in the Annex, I was not able to allocate 0.21% of the vote in 2012 because of
stencies between electoral data reported by the GAB and the geographic redistricti
d by the LTSB. This small number of votes will not change any of my analysis or
ions, and such errors are inevitable when working with large data sets.

10

Case: 3:15-cv-00421 Document #: 1-2 Filed: 07/08/15 Page 12 of 58

$$+\beta_4 \begin{matrix}\mathit{Democratic}\\ \mathit{Presidential\ Vote}_i\end{matrix} + \ \beta_5 \begin{matrix}\mathit{Republican}\\ \mathit{Presidential\ Vote}_i\end{matrix}$$

$$+\beta_6 \begin{matrix}\mathit{Democratic}\\ \mathit{Incumbent}_{\,i}\end{matrix} + \beta_7 \begin{matrix}\mathit{Republican}\\ \mathit{Incumbent}_i\end{matrix} + \ \sum_{j=1}^{71} \gamma_j \mathit{County}_j + \ \varepsilon_i$$

Linear Models: Generality

A statistical model is a **linear model** if it imposes a linear relationship on *some* set of explanatory variables, up to random error.

More Examples

Let Y_i be the number of democratic votes in the ith ward, X_i be the voting eligible population in the ith ward, and Z_i be the number of hot air balloon companies in the ith ward. Linear models include

$$Y_i = b_0 + b_1 X_i + b_2 Z_i + \epsilon_i$$

$$Y_i = b_0 + b_1 X_i + b_2 X_i^2 + b_3 X_i Z_i + \epsilon_i$$

$$Y_i = b_0 + b_1 e^{X_i Z_i} + b_2 \left[\int_0^2 \sin(\tau^{X_i - Z_i}) d\tau \right] + \epsilon_i.$$

Linear Models: Generality

A statistical model is a **linear model** if it imposes a linear relationship on *some* set of explanatory variables, up to random error.

More Examples

Let Y_i be the number of democratic votes in the ith ward, X_i be the voting eligible population in the ith ward, and Z_i be the number of hot air balloon companies in the ith ward. Linear models include

$$Y_i = b_0 + b_1 X_i + b_2 Z_i + \epsilon_i$$

$$Y_i = b_0 + b_1 X_i + b_2 X_i^2 + b_3 X_i Z_i + \epsilon_i$$

$$Y_i = b_0 + b_1 e^{X_i Z_i} + b_2 \left[\int_0^2 \sin(\tau^{X_i - Z_i}) d\tau \right] + \epsilon_i.$$

In the third example, $e^{X_i Z_i}$ and $\left[\int_0^2 \sin(\tau^{X_i - Z_i}) d\tau \right]$ are input data: they're just numbers we can compute.

Linear Models: Generality

A statistical model is a **linear model** if it imposes a linear relationship on *some* set of explanatory variables, up to random error.

Messy Example, continued

Let Y_i be the number of democratic votes in the ith ward, X_i be the voting eligible population in the ith ward, and Z_i be the number of hot air balloon companies in the ith ward. Let $W_i = e^{X_i Z_i}$ and $Q_i = \left[\int_0^2 \sin(\tau^{X_i - Z_i}) d\tau\right]$. Then

$$\begin{aligned} Y_i &= b_0 + b_1 e^{X_i Z_i} + b_2 \left[\int_0^2 \sin(\tau^{X_i - Z_i}) d\tau\right] + \epsilon_i \\ &= b_0 + b_1 W_i + b_2 Q_i + \epsilon_i \end{aligned}$$

Linear Models: Formally

Let $\vec{x_i} = (x_{i1}, x_{i2}, \cdots, x_{ip})$ be a vector of measurements of explanatory variables corresponding to response variable y_i. Our **paired data** is $(\vec{x}_1, y_1)$, ..., $(\vec{x}_n, y_n)$. The linear model is

$$Y_i = b_1 X_{i_1} + b_2 X_{i_2} + ... + b_p X_{i_p} + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2) \text{ independent.}$$

Example

If $\vec{x}_i = (x_{i1}, x_{i2}, x_{i3})$ where $x_{i1} = 1$, x_{i2} is the total voting population, and x_{i3} is the average income in the ith ward, we have

$$\#\text{Dem Votes} = b_1 + b_2\#\text{Voters} + b_3\text{Average Income} + \text{Random Error}$$

Linear Models: Formally

Let $\vec{x_i} = (x_{i1}, x_{i2}, \cdots, x_{ip})$ be a vector of measurements of explanatory variables corresponding to response variable y_i. Our **paired data** is $(\vec{x}_1, y_1)$, ..., $(\vec{x}_n, y_n)$. The linear model is

$$Y_i = b_1 X_{i_1} + b_2 X_{i_2} + ... + b_p X_{i_p} + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2) \text{ independent.}$$

Matrix-Vector Notation

Setting $\vec{X}_i = (X_{i1}, ..., X_{ip})$ and $\vec{\beta} = \begin{pmatrix} b_1 \\ \vdots \\ b_p \end{pmatrix}$, the model is

$$Y_i = \vec{X}_i \vec{\beta} + \epsilon_i.$$

Linear Models: Formally

Let $\vec{x_i} = (x_{i1}, x_{i2}, \cdots, x_{ip})$ be a vector of measurements of explanatory variables corresponding to response variable y_i. Our **paired data** is $(\vec{x}_1, y_1), ..., (\vec{x}_n, y_n)$. The linear model is

$$Y_i = b_1 X_{i_1} + b_2 X_{i_2} + ... + b_p X_{i_p} + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2) \text{ independent.}$$

Matrix-Vector Notation

Setting $\vec{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_p \end{pmatrix}$, $X = \begin{pmatrix} \vec{X}_1 \\ \vdots \\ \vec{X}_n \end{pmatrix}$ $\vec{\beta} = \begin{pmatrix} b_1 \\ \vdots \\ b_p \end{pmatrix}$, and $\vec{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$, the model is

$$\vec{Y} = X\vec{\beta} + \vec{\epsilon}$$

Linear Models: Deja-Vu

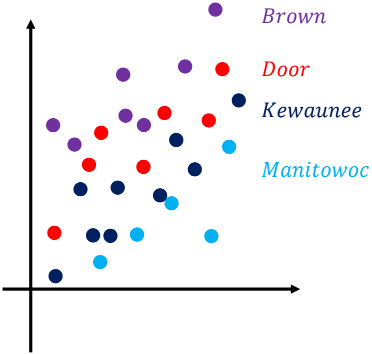
When we take a model like,

$$\#\text{Dem Votes} = b_0 + b_1\#\text{Voters} + b_2\text{Average Income} + \text{Random Error}.$$

1. We assume an underlying linear relationship
2. We guess “the best” values of $b_0, b_1, b_2, \ldots$ using OLS
3. Doing so gives lots of nice properties

Categorical Variables

Assembly District 1 has wards from four counties: Door, Kewaunee, Brown, and Manitowoc. Maybe we want to assume people act differently in each county



Categorical Variables

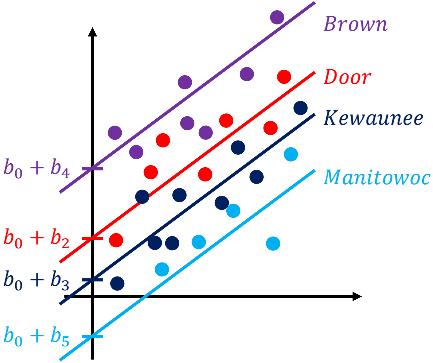
Assembly District 1 has wards from four counties: Door, Kewaunee, Brown, and Manitowoc. Maybe we want to assume people act differently in each county

$$\begin{aligned}\#\text{Dem Votes}_i = b_0 &+ b_1\#\text{ Voters}\\ &+ b_2\text{Is ward i in Door?} + b_3\text{Is ward i in Kewaunee?}\\ &+ b_4\text{Is ward i in Brown?} + b_5\text{Is ward i in Manitowoc?}\\ &+ \text{Random Error}\end{aligned}$$

This is still linear, just in a larger set of explanatory variables. (See R practical for a slight correction to this slide)

Categorical Variables

Assembly District 1 has wards from four counties: Door, Kewaunee, Brown, and Manitowoc. Maybe we want to assume people act differently in each county



Summary so far

- We use the linear model for any situation where we expect a (mostly) linear relationship between a response variable and explanatory variables....
- But those explanatory variables are quite general so that the linear model captures a whole bunch of interesting situations.
- And everything is basically the same as in the simple linear model!

Inference in the Linear Model

Suppose we fit

$$\#\text{ Democratic Votes}_i = 0.26 + 0.37\text{ Ward Voting Pop}_i + \epsilon_i.$$

- We know how to predict democratic votes if the population of a ward changes: plug into the formula and assume zero error.

Inference in the Linear Model

Suppose we fit

$$\#\ \text{Democratic Votes}_i = 0.26 + 0.37\ \text{Ward Voting Pop}_i + \epsilon_i.$$

- We know how to predict democratic votes if the population of a ward changes: plug into the formula and assume zero error.
- How do we know that 0.37 “is meaningful?”

p-Values by Contrived Example

Suppose Mathcamp has 100 students and you suspect that it's been invaded by spies from Historycamp. If Historycamp spies reaches critical mass (more than 10% of the students), you will shut down Mathcamp.

p-Values by Contrived Example

Suppose Mathcamp has 100 students and you suspect that it's been invaded by spies from Historycamp. If Historycamp spies reaches critical mass (more than 10% of the students), you will shut down Mathcamp. Two scenarios:

- H_0 : 10% are spies (the conservative option, or **null hypothesis**)
- H_a :> 10% are spies (the **alternative hypothesis**).

p-Values by Contrived Example

Suppose Mathcamp has 100 students and you suspect that it's been invaded by spies from Historycamp. If Historycamp spies reaches critical mass (more than 10% of the students), you will shut down Mathcamp. Two scenarios:

- H_0 : 10% are spies (the conservative option, or **null hypothesis**)
- H_a :> 10% are spies (the **alternative hypothesis**).

To test, survey 12 campers. Find 3 are spies.

Question

Is this sufficient evidence to shut down Mathcamp?

p-Values by Contrived Example

Question

You survey 12 campers and find 3 are spies from history camp. Is this sufficient evidence that strictly more than 10% of Mathcamp students are spies?

Answer: Look at a Probability

Doing some boring computations you can compute:

$$\mathbb{P}[\text{at least 3 of 12 surveyd are spies} \mid \text{10\% of campers are spies}] = 0.11.$$

Here, 0.11 is a **p-value**: it is the probability of seeing data at least as extreme as what we observed assuming H_0 is true.

p-Values by Contrived Example

Question

You survey 12 campers and find 3 are spies from history camp. Is this sufficient evidence that strictly more than 10% of Mathcamp students are spies?

Answer: Look at a Probability

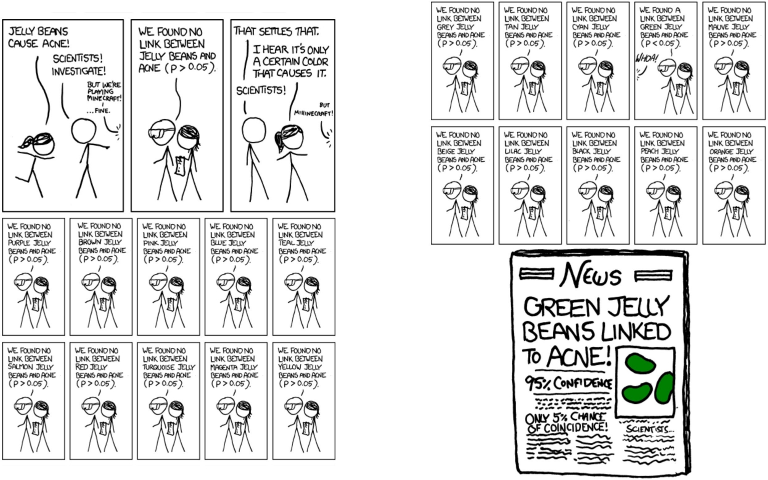
Doing some boring computations you can compute:

$$\mathbb{P}[\text{at least 3 of 12 surveyd are spies} \mid \text{10\% of campers are spies}] = 0.11.$$

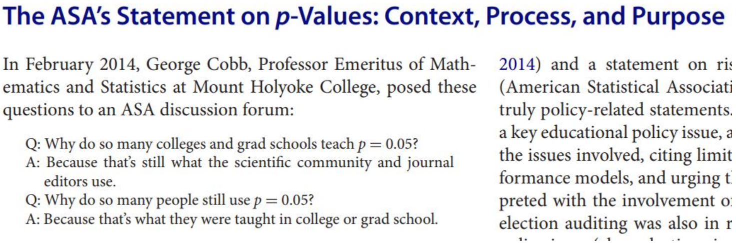
Here, 0.11 is a **p-value**: it is the probability of seeing data at least as extreme as what we observed assuming H_0 is true.

- Generally, a p-value less than 0.05 is thought to indicate that our data is sufficiently unlikely to contradict the null hypothesis.
- **A *p*-value is NOT the probability that H_0 is true**.
- Always report *p*-values in full

XKCD Interlude


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
...FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIINECRAFT!
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).
NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

ASA Interlude



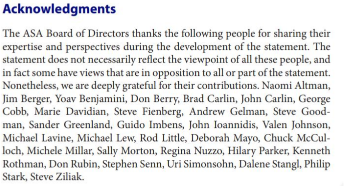
The ASA's Statement on *p*-Values: Context, Process, and Purpose

In February 2014, George Cobb, Professor Emeritus of Mathematics and Statistics at Mount Holyoke College, posed these questions to an ASA discussion forum:

> Q: Why do so many colleges and grad schools teach $p = 0.05$?
> A: Because that's still what the scientific community and journal editors use.
> Q: Why do so many people still use $p = 0.05$?
> A: Because that's what they were taught in college or grad school.

2014) and a statement on ris
(American Statistical Associati
truly policy-related statements.
a key educational policy issue, a
the issues involved, citing limit
formance models, and urging t
preted with the involvement o
election auditing was also in r

ASA Interlude



Acknowledgments

The ASA Board of Directors thanks the following people for sharing their expertise and perspectives during the development of the statement. The statement does not necessarily reflect the viewpoint of all these people, and in fact some have views that are in opposition to all or part of the statement. Nonetheless, we are deeply grateful for their contributions. Naomi Altman, Jim Berger, Yoav Benjamini, Don Berry, Brad Carlin, John Carlin, George Cobb, Marie Davidian, Steve Fienberg, Andrew Gelman, Steve Goodman, Sander Greenland, Guido Imbens, John Ioannidis, Valen Johnson, Michael Lavine, Michael Lew, Rod Little, Deborah Mayo, Chuck McCulloch, Michele Millar, Sally Morton, Regina Nuzzo, Hilary Parker, Kenneth Rothman, Don Rubin, Stephen Senn, Uri Simonsohn, Dalene Stangl, Philip Stark, Steve Ziliak.

- *Benjamin, Daniel J, and Berger, James O:* A simple alternative to p-values
- *Benjamini, Yoav:* It's not the p-values' fault
- *Berry, Donald A:* P-values are not what they're cracked up to be
- *Carlin, John B:* Comment: Is reform possible without a paradigm shift?
- *Cobb, George:* ASA statement on p-values: Two consequences we can hope for
- *Gelman, Andrew:* The problems with p-values are not just with p-values
- *Goodman, Steven N:* The next questions: Who, what, when, where, and why?
- *Greenland, Sander:* The ASA guidelines and null bias in current teaching and practice
- *Ioannidis, John P.A.:* Fit-for-purpose inferential methods: abandoning/changing P-values versus abandoning/changing research
- *Johnson, Valen E.:* Comments on the "ASA Statement on Statistical Significance and P-values" and marginally significant p-values
- *Lavine, Michael, and Horowitz, Joseph:* Comment
- *Lew, Michael J:* Three inferential questions, two types of P-value
- *Little, Roderick J:* Discussion
- *Mayo, Deborah G:* Don't throw out the error control baby with the bad statistics bathwater
- *Millar, Michele:* ASA statement on p-values: some implications for education
- *Rothman, Kenneth J:* Disengaging from statistical significance
- *Senn, Stephen:* Are P-Values the Problem?
- *Stangl, Dalene:* Comment
- *Stark, P.B.:* The value of p-values
- *Ziliak, Stephen T:* The significance of the ASA statement on statistical significance and p-values

ASA Interlude: Quotes

*I*nformally, a p-value is the probability under a specified statistical model that a statistical summary of the data (e.g., the sample mean difference between two compared groups) would be equal to or more extreme than its observed value. – ASA

1. P-values can indicate how incompatible the data are with a specified statistical model
2. P-values do not measure the probability that the studied hypothesis is true...
3. Proper inference requires full reporting and transparency

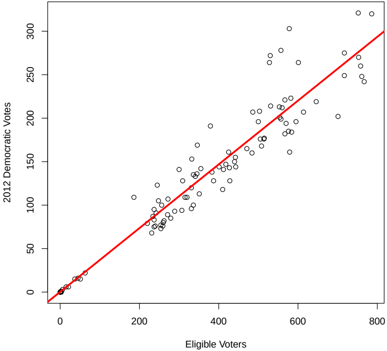
p-Values in Linear Regression

- In linear regression, we test
  - $H_0 : b_i = 0$ (the conservative option, or **null hypothesis**)
  - $H_a : b_i \neq 0$ (the **alternative hypothesis**).
- R computes a p-value that measures "how likely we'd see data at least as linearly related to X_i if $b_i = 0$."

In Our Example

In the Style of Mayer

$$\# \text{ Democratic Votes}_i = 0.26 + 0.37 \text{ Ward Voting Pop}_i + \epsilon_i.$$



The p-value for $b_1 \neq 0$ is 2×10^{-16}. The p-value for $b_0 \neq 0$ is 0.96.

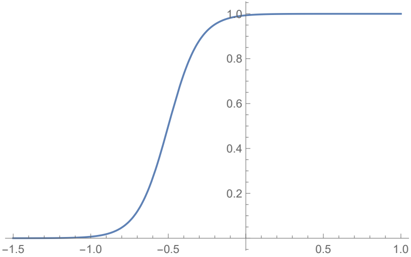
Logistic Regression

- In the simple linear model, $Y_i \sim N(b_0 + b_1 x_i, \sigma^2)$.
- Suppose instead we want $Y_i \in \{0, 1\}$ (e.g., "do the democrats win ward i?"). **Logistic Regression** says

$$Y_i \sim Ber(p_i), \qquad \log \frac{p_i}{1 - p_i} \sim b_0 + b_1 x_i.$$

- This says

$$p_i = \frac{1}{1 + e^{-(b_0 + b_1 x_i)}}$$



Penalized Regression

- Simple linear regression minimizes

$$\sum_{i=1}^{103}\left(y_i - \left(\hat{b_0} + \hat{b_1}x_{1i} + \hat{b_2}x_{2i} + \cdots + \hat{b_p}x_{pi}\right)\right)^2.$$

- We might include a whole bunch of possible predictors (make p large) and then search for sparse solutions. E.g. minimize

$$\sum_{i=1}^{103}\left(y_i - \left(\hat{b_0} + \hat{b_1}x_{1i} + \hat{b_2}x_{2i} + \cdots + \hat{b_p}x_{pi}\right)\right)^2 + \lambda\left(\sum_{j=1}^{p} bj^2\right).$$

Final Summary

- Simple linear regression makes sense when we have good reason to suspect an underlying linear relationship between a response and explanatory variable.
- The simple linear model generalizes to the linear model, a powerful, ubiquitous tool in political science.
- Basically everything works the same as in simple linear regression.
- p-values can be used for inference, and are often reported, and R has nice tools to sanity check models.

Thanks!

Questions?