Descriptive Statistics for Bivariate Regression (Part 1)

August 30, 2022

Outline

Univariate Descriptive Statistics

Descriptive Regression with a Coarse Covariate

Descriptive Linear Regression

What are descriptive statistics?

- Descriptive statistics are summaries of the data we have.
- The data we have might be a sample (e.g., an exit poll) or it might be the whole population (e.g., administrative vote counts), sometimes known as a census.
- Having the whole population is becoming more common in the era of Big Data.
- Note: if you are interested in the causal effects of a treatment (a drug, a policy, etc.), you can never really have the whole population (we will discuss this in the causal inference portion of the course).

What are descriptive statistics?

- Descriptive statistics are summaries of the data we have.
- The data we have might be a sample (e.g., an exit poll) or it might be the whole population (e.g., administrative vote counts), sometimes known as a census.
- Having the whole population is becoming more common in the era of Big Data.
- Note: if you are interested in the causal effects of a treatment (a drug, a policy, etc.), you can never really have the whole population (we will discuss this in the causal inference portion of the course).

What are descriptive statistics?

- Descriptive statistics are summaries of the data we have.
- The data we have might be a sample (e.g., an exit poll) or it might be the whole population (e.g., administrative vote counts), sometimes known as a census.
- Having the whole population is becoming more common in the era of Big Data.
- Note: if you are interested in the causal effects of a treatment (a drug, a policy, etc.), you can never really have the whole population (we will discuss this in the causal inference portion of the course).

What are descriptive statistics?

- Descriptive statistics are summaries of the data we have.
- The data we have might be a sample (e.g., an exit poll) or it might be the whole population (e.g., administrative vote counts), sometimes known as a census.
- Having the whole population is becoming more common in the era of Big Data.
- Note: if you are interested in the causal effects of a treatment (a drug, a policy, etc.), you can never really have the whole population (we will discuss this in the causal inference portion of the course).

What makes a good summary?

What are the two things that make a good summary?

What makes a good summary?

What are the two things that make a good summary?

From the google definition:

1. brief
2. includes main points

What makes a good descriptive statistic?

1. brief: at most a few numbers
2. includes main points:
   - sufficient to describe main characteristics of the data
     - there is a technical definition of sufficiency that I can discuss in office hours for those interested
   - minimizes the information lost by reducing the data
     - I'm using "information" in a loose sense here, will formalize a bit later
     - reducing the data means taking the data, which are often represented by $n * v$ numbers (with n the number of observations and v the number of variables) and reducing to a descriptive statistic that is represented by a few numbers.

Don't worry. Enough forest, trees (examples) are coming!

What makes a good descriptive statistic?

1. brief: at most a few numbers
2. includes main points:
   - sufficient to describe main characteristics of the data
     - there is a technical definition of sufficiency that I can discuss in office hours for those interested
   - minimizes the information lost by reducing the data
     - I'm using "information" in a loose sense here, will formalize a bit later
     - reducing the data means taking the data, which are often represented by $n * v$ numbers (with n the number of observations and v the number of variables) and reducing to a descriptive statistic that is represented by a few numbers.

Don't worry. Enough forest, trees (examples) are coming!

What makes a good descriptive statistic?

1. brief: at most a few numbers
2. includes main points:
   - sufficient to describe main characteristics of the data
     - there is a technical definition of sufficiency that I can discuss in office hours for those interested
   - minimizes the information lost by reducing the data
     - I'm using "information" in a loose sense here, will formalize a bit later
     - reducing the data means taking the data, which are often represented by $n * v$ numbers (with n the number of observations and v the number of variables) and reducing to a descriptive statistic that is represented by a few numbers.

Don't worry. Enough forest, trees (examples) are coming!

What makes a good descriptive statistic?

1. brief: at most a few numbers
2. includes main points:
   - sufficient to describe main characteristics of the data
     - there is a technical definition of sufficiency that I can discuss in office hours for those interested
   - minimizes the information lost by reducing the data
     - I'm using "information" in a loose sense here, will formalize a bit later
     - reducing the data means taking the data, which are often represented by $n * v$ numbers (with n the number of observations and v the number of variables) and reducing to a descriptive statistic that is represented by a few numbers.

Don't worry. Enough forest, trees (examples) are coming!

What makes a good descriptive statistic?

1. brief: at most a few numbers
2. includes main points:
   - sufficient to describe main characteristics of the data
     - there is a technical definition of sufficiency that I can discuss in office hours for those interested
   - minimizes the information lost by reducing the data
     - I'm using "information" in a loose sense here, will formalize a bit later
     - reducing the data means taking the data, which are often represented by $n * v$ numbers (with n the number of observations and v the number of variables) and reducing to a descriptive statistic that is represented by a few numbers.

Don't worry. Enough forest, trees (examples) are coming!

"Hold on, where's the forest again?"

It is easy in a statistics class to focus on the specifics. I want you to remember why we are doing what we are doing.

Outline

Univariate Descriptive Statistics

Descriptive Regression with a Coarse Covariate

Descriptive Linear Regression

Binary Data

Table: Population

i	1	2	3	4	...	625	$\bar{y}_{625}$
y_i	1	1	0	1	...	0	.68

Without covariates (and if the indexes aren't otherwise informative), n and p contain all the information.

What about data with more values?

Example: Epstein and Mershon SCOTUS data

- Data on 27 justices from the Warren ('53 - '69), Burger ('69 - '86), and Rehnquist ('86 - '05) courts
- Data can be interpreted as a census for the 1953 - 2005 era.
- Percentage of votes in liberal direction for each justice in civil liberties cases

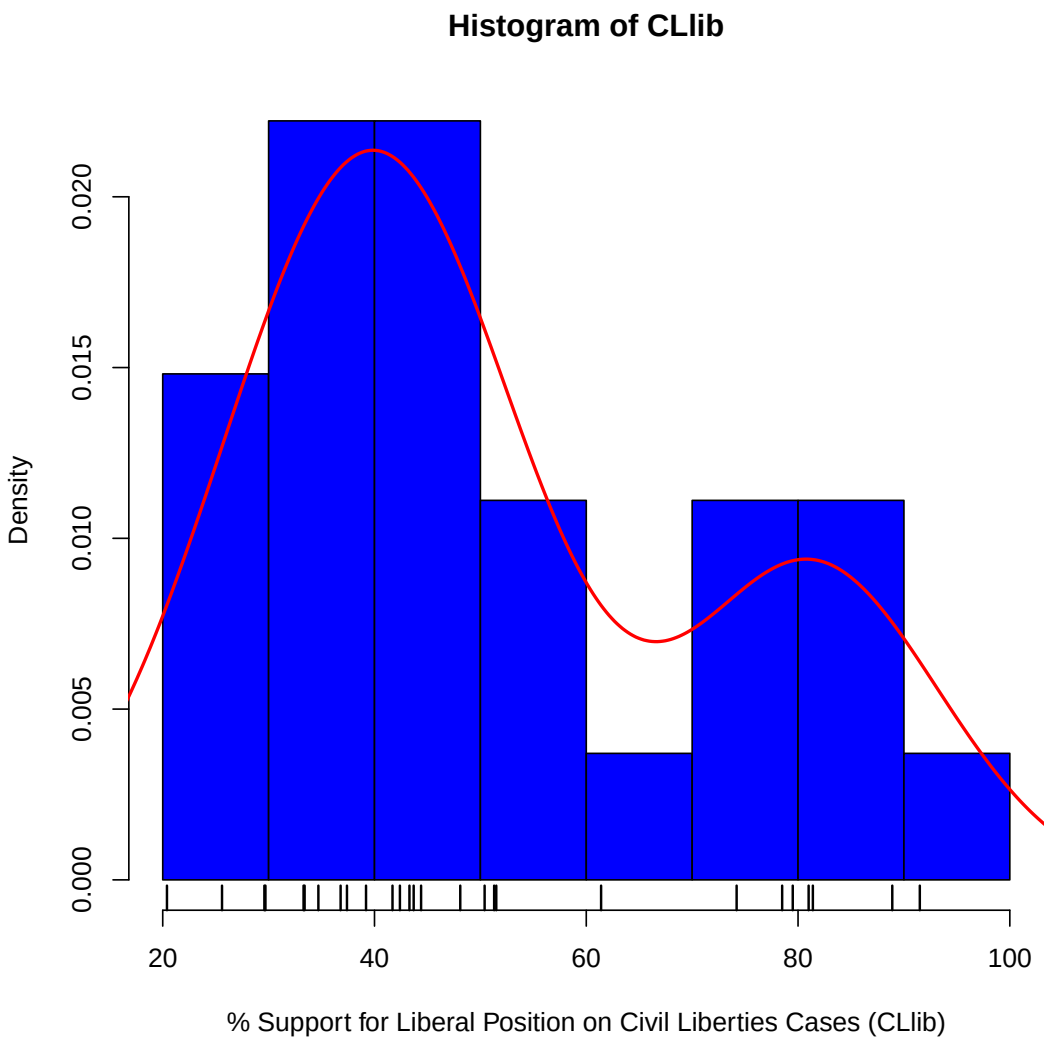
Example: Epstein and Mershon SCOTUS data

- Data on 27 justices from the Warren ('53 - '69), Burger ('69 - '86), and Rehnquist ('86 - '05) courts
- Data can be interpreted as a census for the 1953 - 2005 era.
- Percentage of votes in liberal direction for each justice in civil liberties cases

Example: Epstein and Mershon SCOTUS data

- Data on 27 justices from the Warren ('53 - '69), Burger ('69 - '86), and Rehnquist ('86 - '05) courts
- Data can be interpreted as a census for the 1953 - 2005 era.
- Percentage of votes in liberal direction for each justice in civil liberties cases

Describing Univariate Data



Summarization

- Even for univariate data, we may want to focus on one or two characteristics of the population instead of describing it in its entirety.
- We typically focus on measures of center and spread.
- We will discuss the rational for this over the coming weeks.

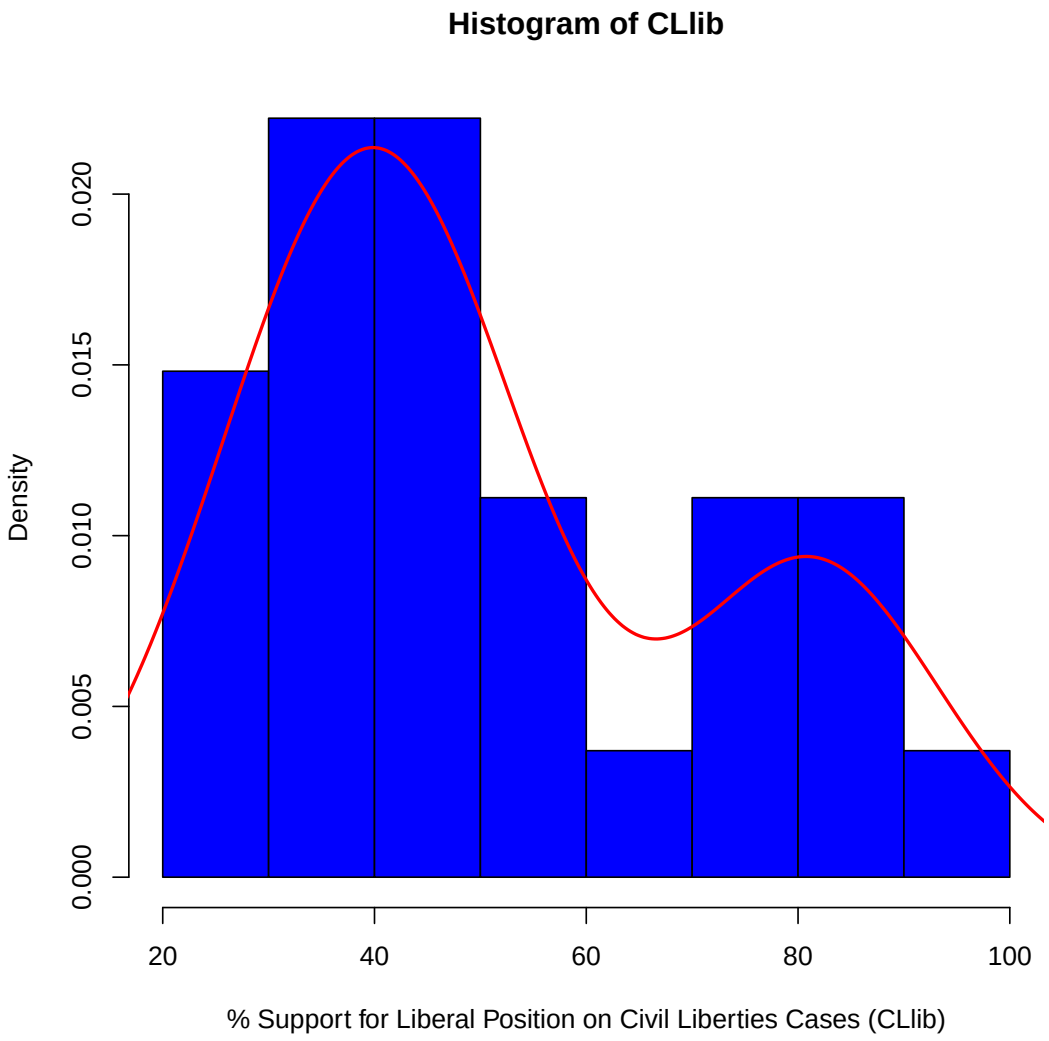
Summarization

- Even for univariate data, we may want to focus on one or two characteristics of the population instead of describing it in its entirety.
- We typically focus on measures of center and spread.
- We will discuss the rational for this over the coming weeks.

Summarization

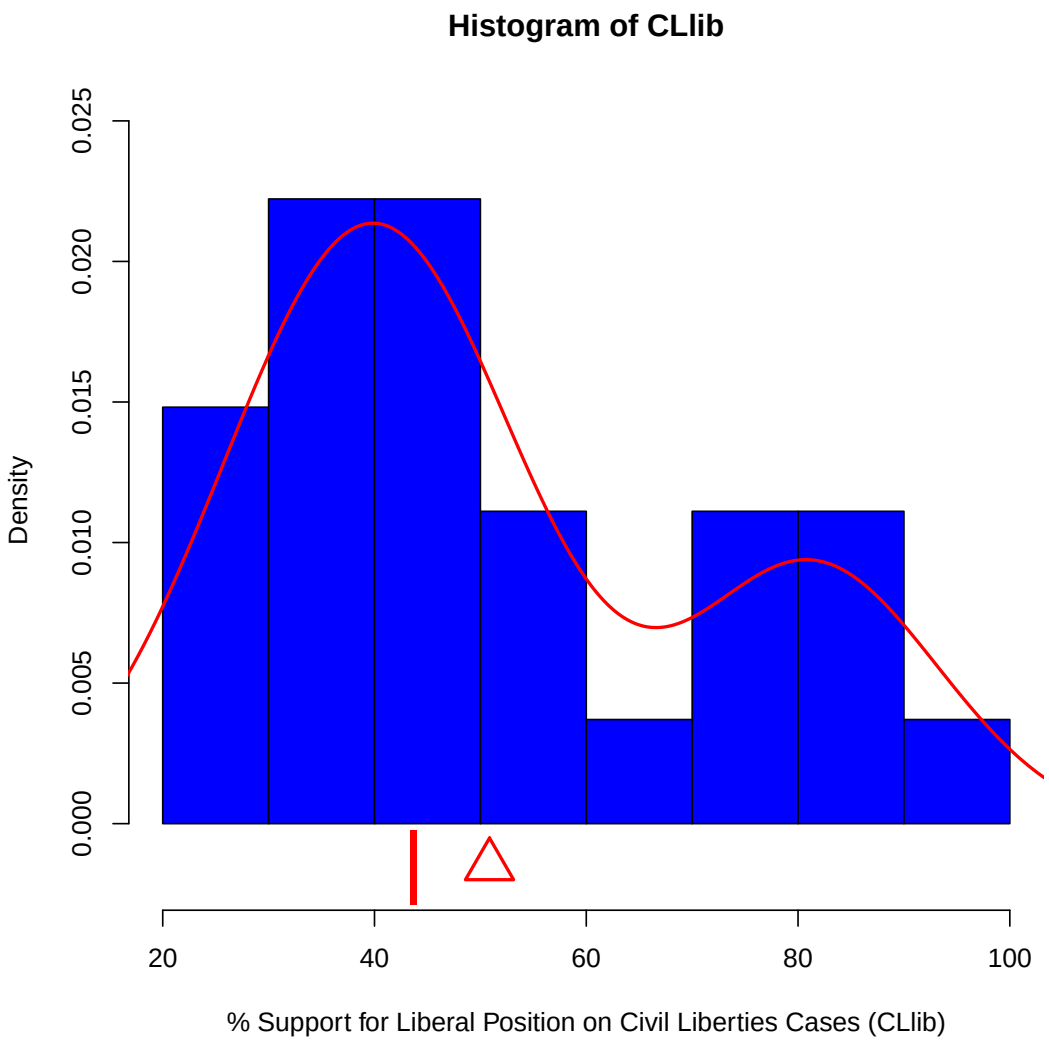
- Even for univariate data, we may want to focus on one or two characteristics of the population instead of describing it in its entirety.
- We typically focus on measures of center and spread.
- We will discuss the rational for this over the coming weeks.

Univariate Activity

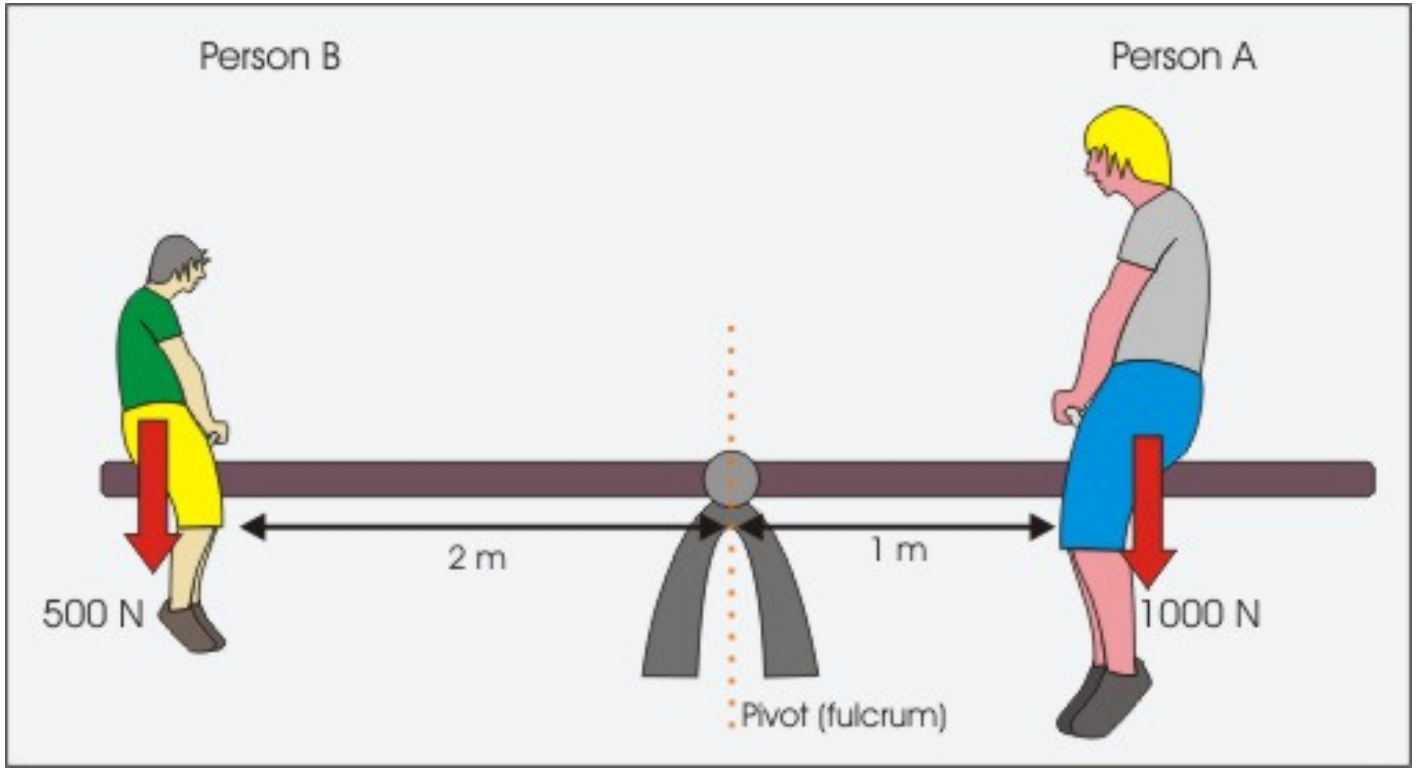


Find the rough location of the mean and median.

Univariate Activity Solution



Mean as a balance point



Choosing a descriptive statistic

How might we choose between these descriptive statistics?

1. brief: at most a few numbers
2. includes main points:
   - sufficient to describe main characteristics of the data
   - minimizes the information lost by reducing the data

Choosing a descriptive statistic

How might we choose between these descriptive statistics?

1. brief: at most a few numbers
2. includes main points:
   - sufficient to describe main characteristics of the data
   - minimizes the information lost by reducing the data

Choosing a descriptive statistic

How might we choose between these descriptive statistics?

1. brief: at most a few numbers
2. includes main points:
   - sufficient to describe main characteristics of the data
   - minimizes the information lost by reducing the data

A Least Squares Summary

Suppose we want to pick a single number (m) for the middle of the data that summarizes all the values for y, by minimizing the sum of squared residuals (i.e., least squares).

$$SSR(\tilde{m}) = \sum_{i=1}^{N} (y_i - \tilde{m})^2.$$

Activity: Using calculus, derive the least squares estimator

1. Calculate the derivative of SSR with respect to $\tilde{m}$
2. Set the derivative equal to 0
3. Solve for m

A Least Squares Summary

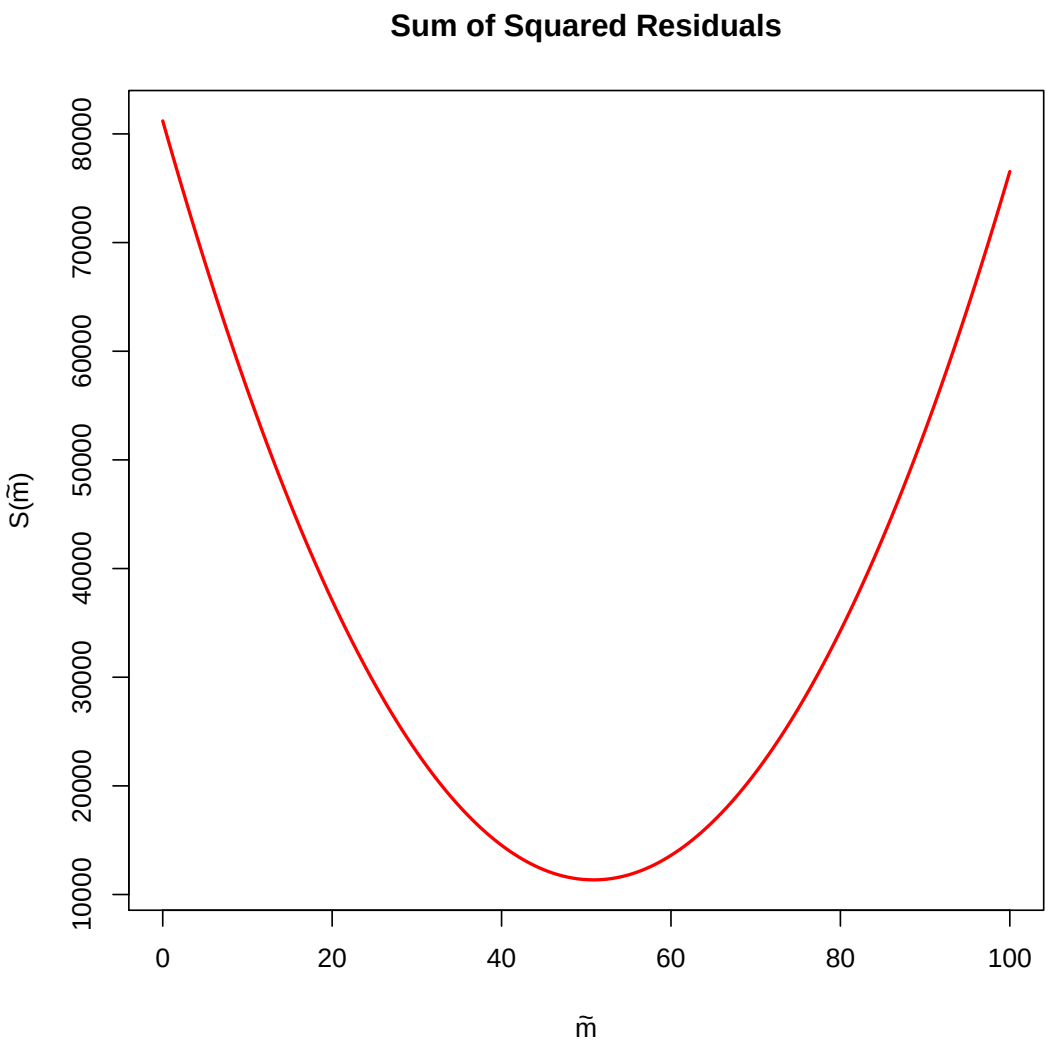
Suppose we want to pick a single number (m) for the middle of the data that summarizes all the values for y, by minimizing the sum of squared residuals (i.e., least squares).

$$SSR(\tilde{m}) = \sum_{i=1}^{N} (y_i - \tilde{m})^2.$$

Activity: Using calculus, derive the least squares estimator

1. Calculate the derivative of SSR with respect to $\tilde{m}$
2. Set the derivative equal to 0
3. Solve for m

The Objective Function for SSR CLlib

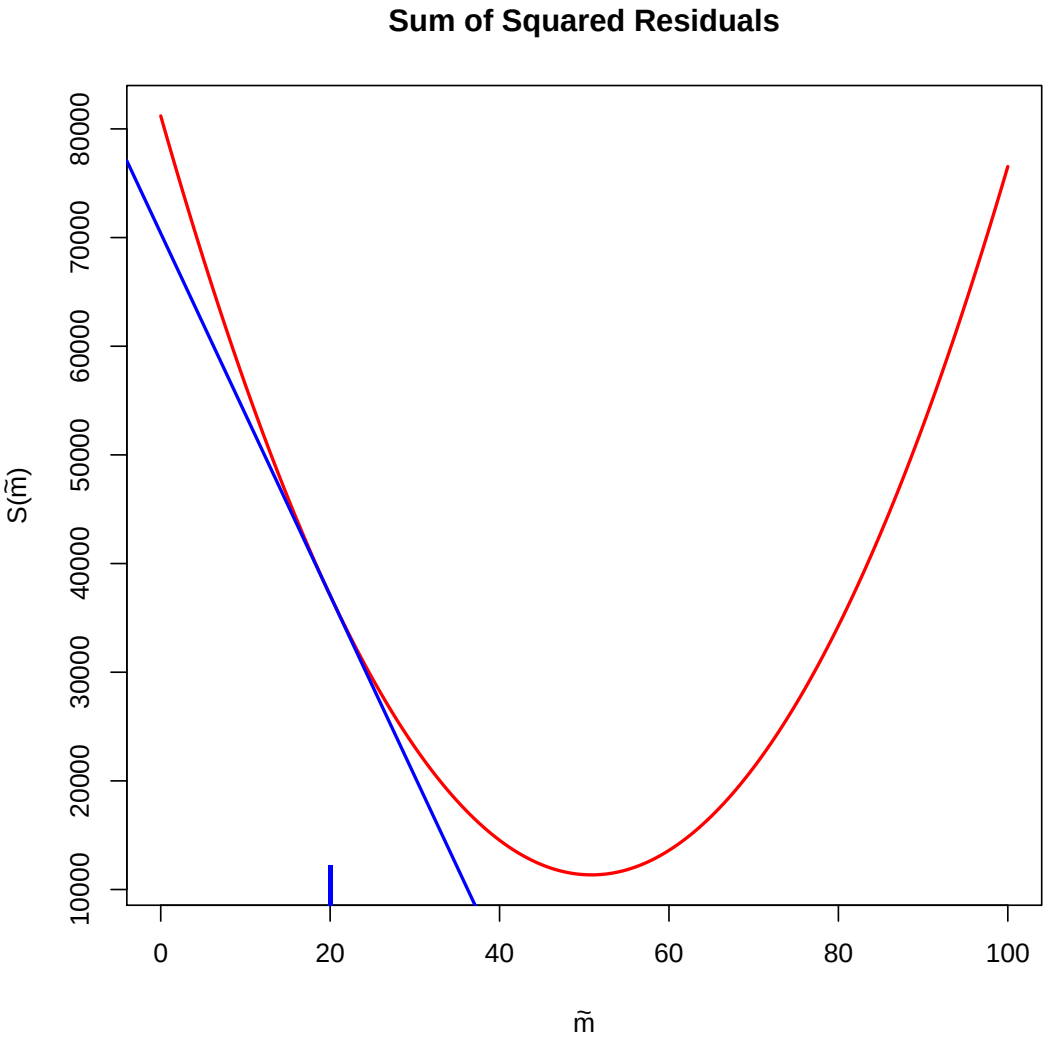


1.

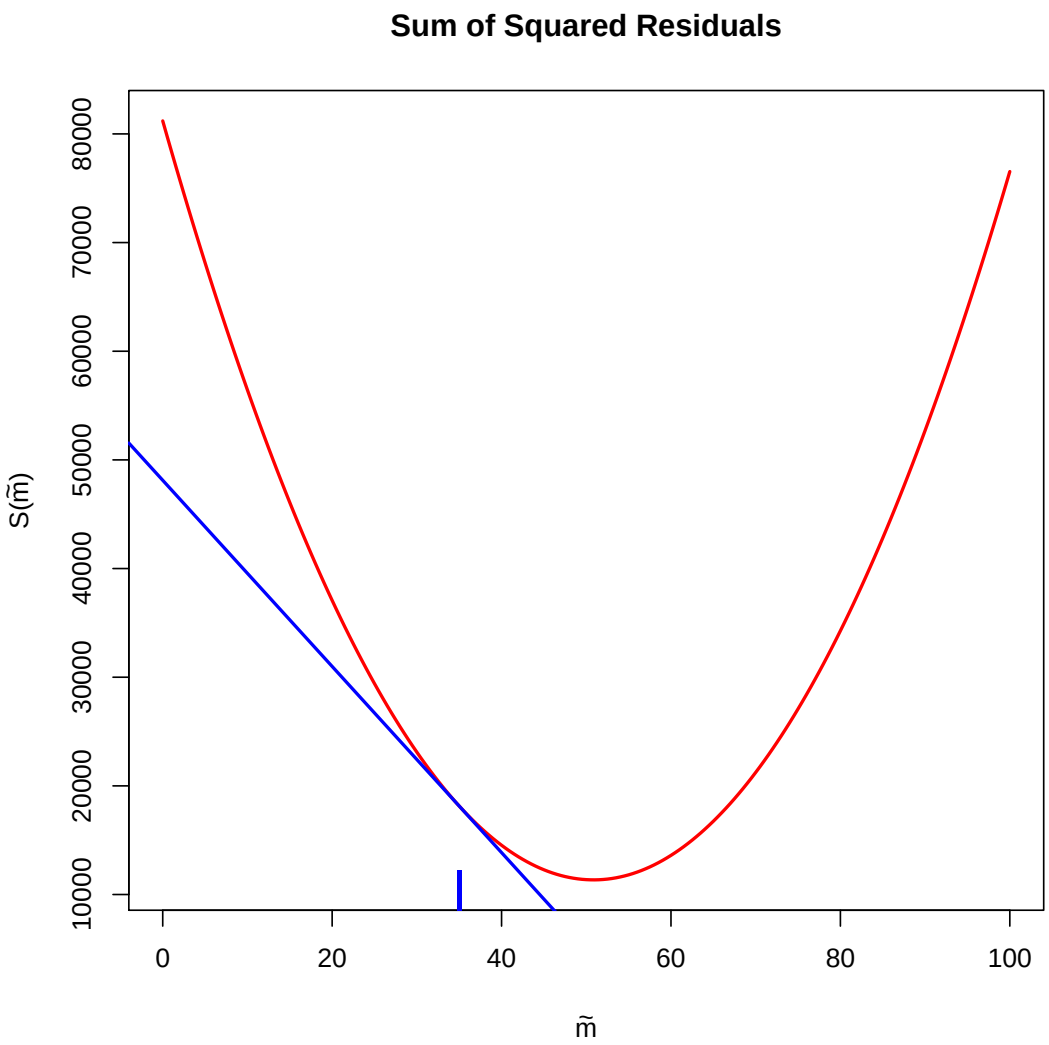
$$
\begin{aligned}
SSR(\tilde{m}) &= \sum_{i=1}^{N}(y_i - \tilde{m})^2 \\
&= \sum_{i=1}^{N}(y_i^2 - 2y_i\tilde{m} + \tilde{m}^2)
\end{aligned}
$$

$$
\frac{\partial SSR(\tilde{m})}{\partial \tilde{m}} = \sum_{i=1}^{N}(-2y_i + 2\tilde{m})
$$

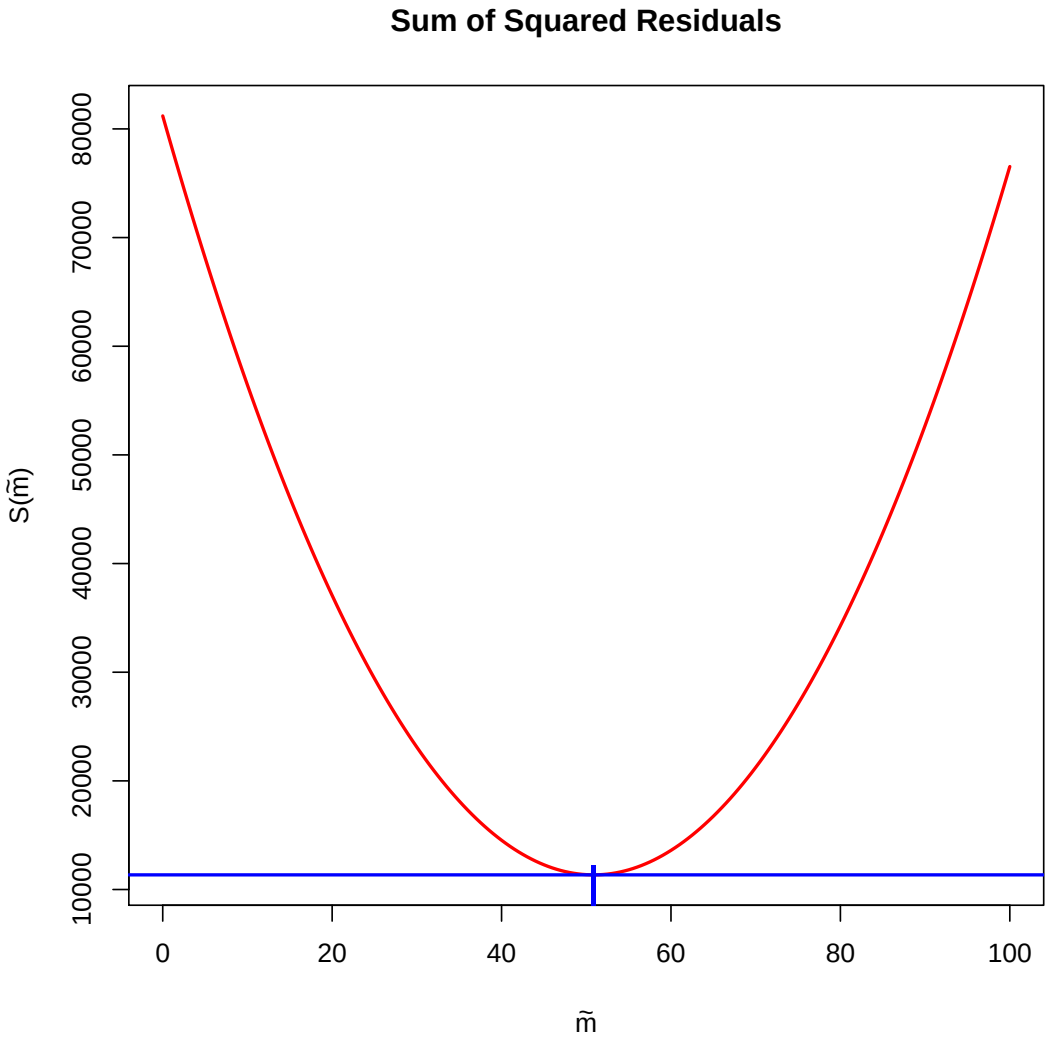
The Slope of the Tangent Line for SSR CLlib



The Slope of the Tangent Line for SSR CLlib



The Slope of the Tangent Line for SSR CLlib



1.

$$SSR(\tilde{m}) = \sum_{i=1}^{N}(y_i - \tilde{m})^2$$
$$= \sum_{i=1}^{N}(y_i^2 - 2y_i\tilde{m} + \tilde{m}^2)$$
$$\frac{\partial S(m)}{\partial \tilde{m}} = \sum_{i=1}^{N}(-2y_i + 2\tilde{m})$$

2.

$$0 = \sum_{i=1}^{N}(-2y_i + 2m)$$

3.

$$m = \frac{1}{N}\sum_{i=1}^{N} y_i$$

Therefore, the mean is a least squares summary.

A Least Absolute Value Summary

Suppose we want to pick a single number (m) for the middle of the data that summarizes all the values for y, by minimizing the sum of absolute deviations.

$$SAV(\tilde{m}) = \sum_{i=1}^{N} |y_i - \tilde{m}|.$$

- Why is this harder to solve?
- Any guesses?

A Least Absolute Value Summary

Suppose we want to pick a single number (m) for the middle of the data that summarizes all the values for y, by minimizing the sum of absolute deviations.

$$SAV(\tilde{m}) = \sum_{i=1}^{N} |y_i - \tilde{m}|.$$

- Why is this harder to solve?
- Any guesses?

Why we focus on means

- Means often provide a “good” summary of the data (and it is relatively easy to tell when they provide bad summaries).
- The properties of means are relatively easy to describe.
- Sufficient in a sense (in a few slides).
- Randomized experiments only identify average causal effects (as do regression and matching when the appropriate assumptions hold).

Why we focus on means

- Means often provide a “good” summary of the data (and it is relatively easy to tell when they provide bad summaries).
- The properties of means are relatively easy to describe.
- Sufficient in a sense (in a few slides).
- Randomized experiments only identify average causal effects (as do regression and matching when the appropriate assumptions hold).

Why we focus on means

- Means often provide a “good” summary of the data (and it is relatively easy to tell when they provide bad summaries).
- The properties of means are relatively easy to describe.
- Sufficient in a sense (in a few slides).
- Randomized experiments only identify average causal effects (as do regression and matching when the appropriate assumptions hold).

Why we focus on means

- Means often provide a “good” summary of the data (and it is relatively easy to tell when they provide bad summaries).
- The properties of means are relatively easy to describe.
- Sufficient in a sense (in a few slides).
- Randomized experiments only identify average causal effects (as do regression and matching when the appropriate assumptions hold).

Variance and Standard Deviation

- Mean is the least squares predictor in terms of SSR.
- How small is "least" (i.e., how much spread around the mean)?
- Variance and Standard Deviation are useful transformations of SSR for the mean.

Population Variance

$$S^2 = \frac{\sum_{i=1}^{N}(y_i - \bar{y})^2}{N-1} = \frac{SSR}{N-1}$$

Population Standard Deviation

$$S = \sqrt{S^2}$$

Variance and Standard Deviation

- Mean is the least squares predictor in terms of SSR.
- How small is "least" (i.e., how much spread around the mean)?
- Variance and Standard Deviation are useful transformations of SSR for the mean.

Population Variance

$$S^2 = \frac{\sum_{i=1}^{N}(y_i - \bar{y})^2}{N-1} = \frac{SSR}{N-1}$$

Population Standard Deviation

$$S = \sqrt{S^2}$$

Variance and Standard Deviation

- Mean is the least squares predictor in terms of SSR.
- How small is "least" (i.e., how much spread around the mean)?
- Variance and Standard Deviation are useful transformations of SSR for the mean.

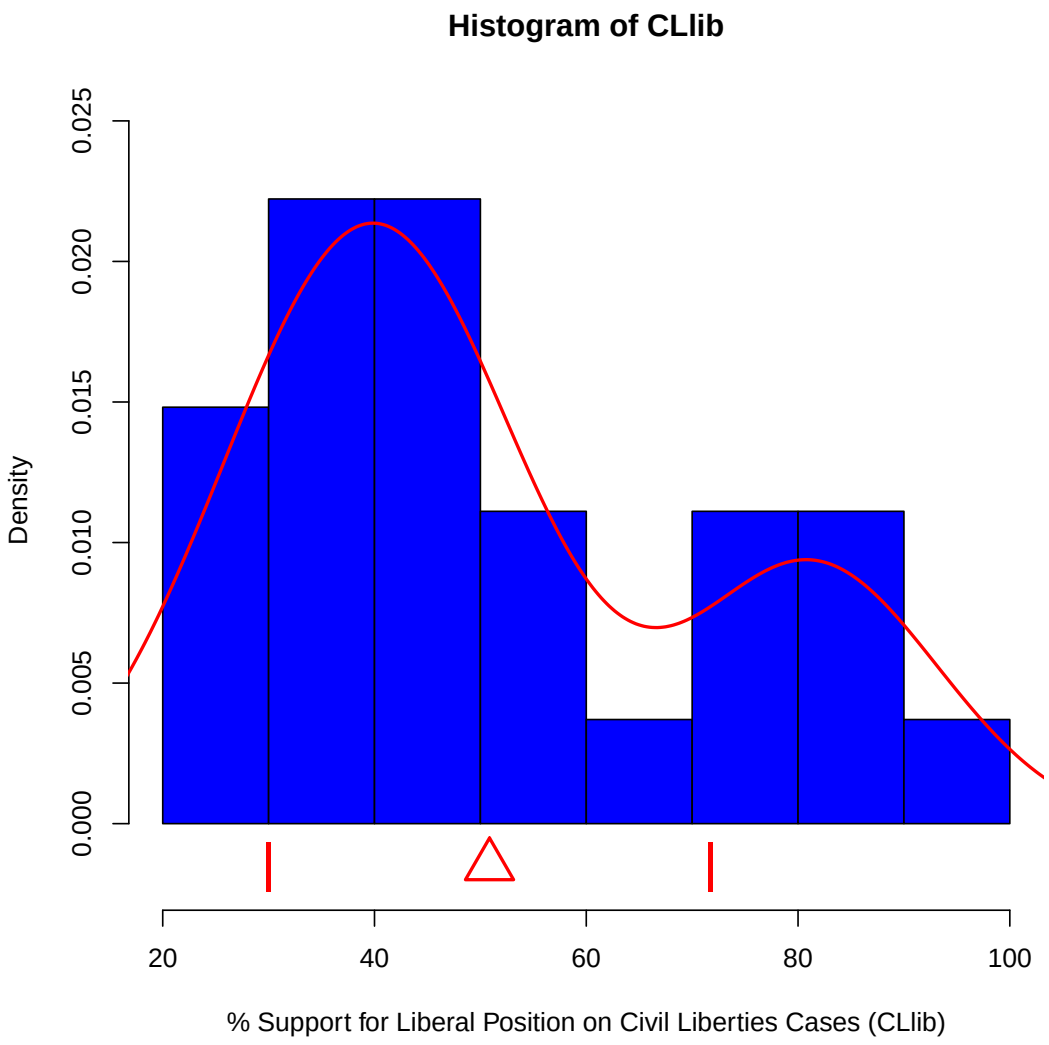
Population Variance

$$S^2 = \frac{\sum_{i=1}^{N}(y_i - \bar{y})^2}{N-1} = \frac{SSR}{N-1}$$

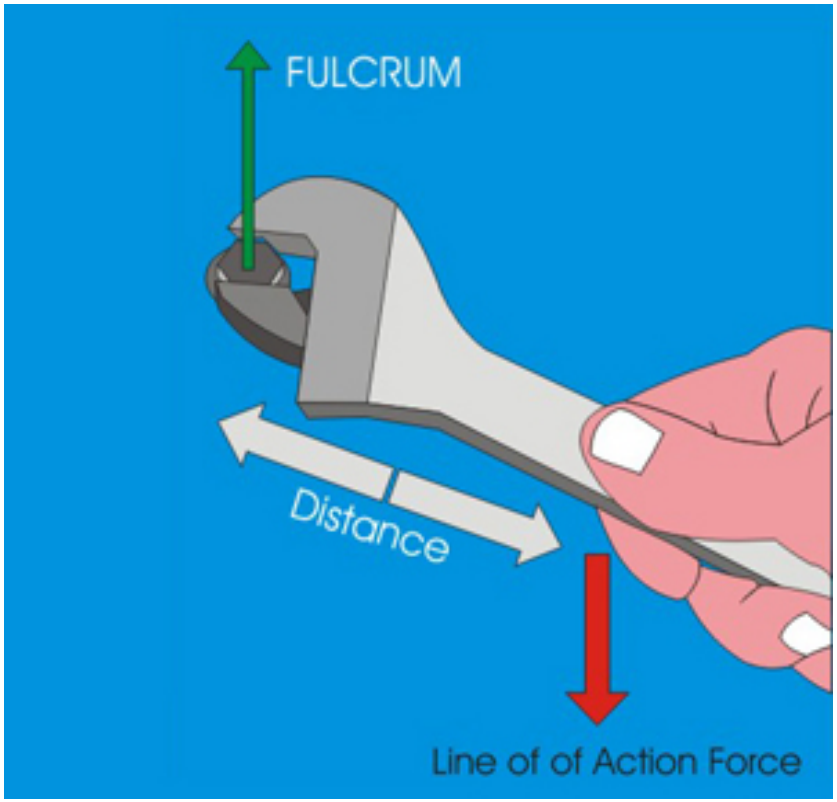
Population Standard Deviation

$$S = \sqrt{S^2}$$

Mean and Standard Deviation



Variance as turning force



Why variance and standard deviation?

- Natural measures of accuracy for means.
- Means and variances are sufficient for normally distributed data.
- For nonnormal data, means and variances provide much information via Chebyshev's inequality and other concentration inequalities.

Why variance and standard deviation?

- Natural measures of accuracy for means.
- Means and variances are sufficient for normally distributed data.
- For nonnormal data, means and variances provide much information via Chebyshev's inequality and other concentration inequalities.

Why variance and standard deviation?

- Natural measures of accuracy for means.
- Means and variances are sufficient for normally distributed data.
- For nonnormal data, means and variances provide much information via Chebyshev's inequality and other concentration inequalities.

Outline

Univariate Descriptive Statistics

Descriptive Regression with a Coarse Covariate

Descriptive Linear Regression

Conditional Means

- In this class, we focus on a single "outcome" variable.
- All additional variables adjust how we consider the outcome, often by conditioning.
- Typically, we consider the mean of y conditional on values of other variables.

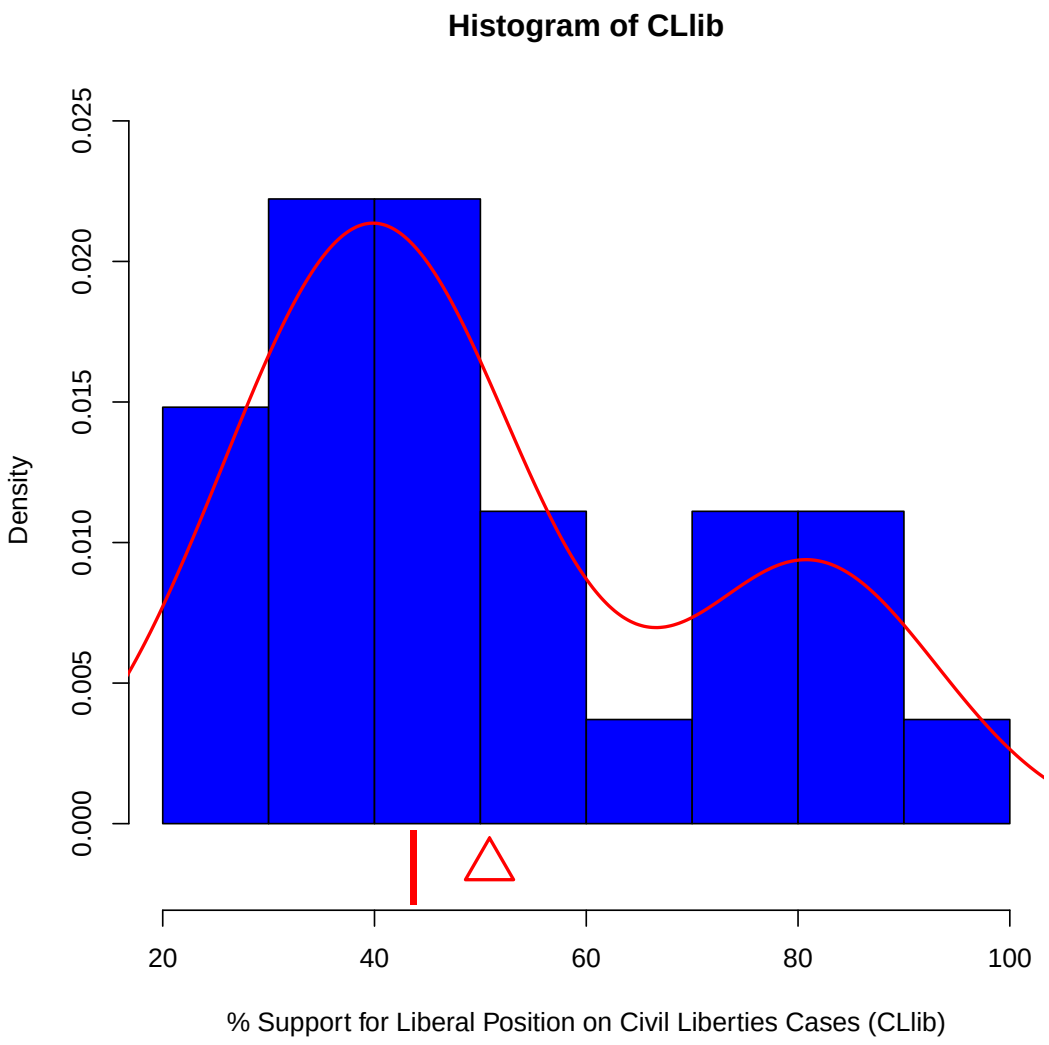
Binary Outcome with Binary Covariate

Table: Population

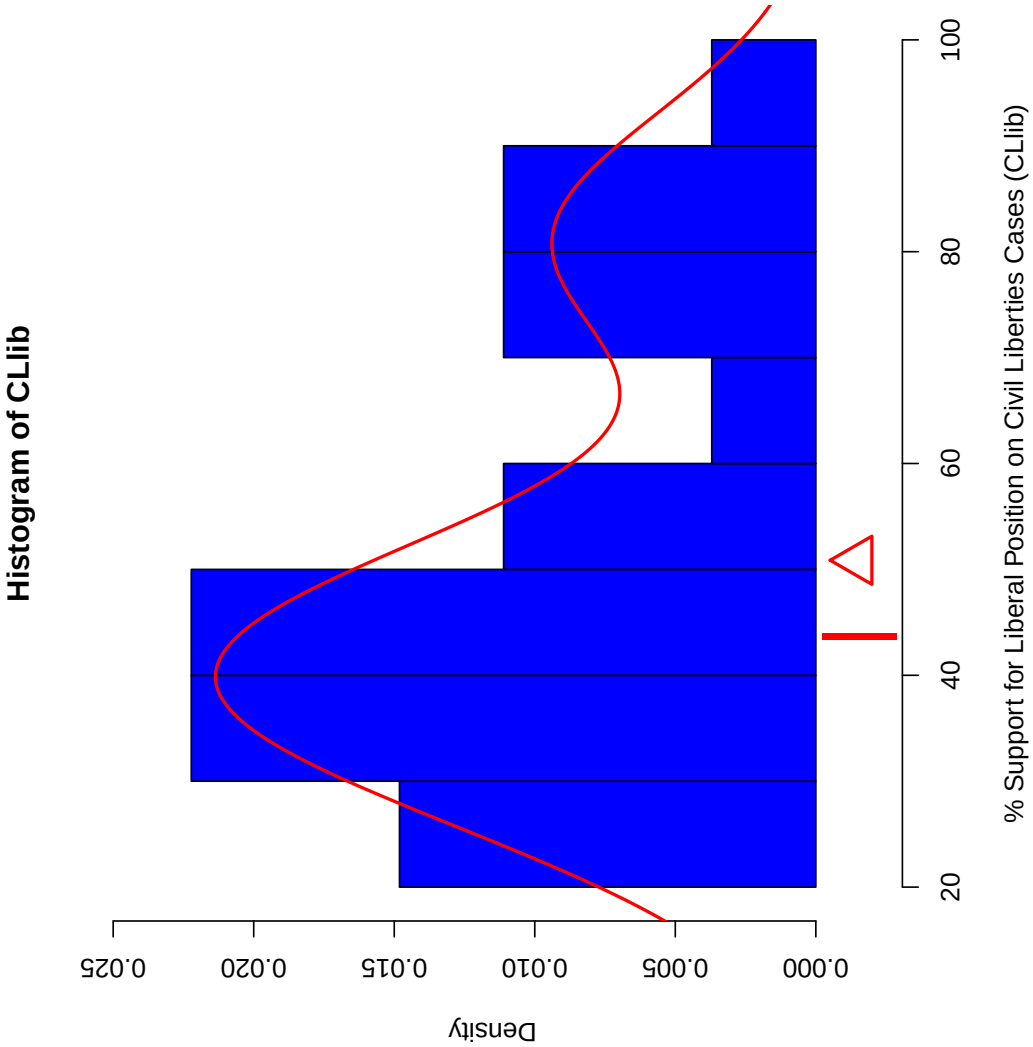
i	1	2	3	4	...	625	mean
y_i	1	1	0	1	...	0	.68
x_i	1	0	1	0	...	0	.5

- Instead of one proportion, we now have two conditional proportions (one for $x = 0$ and one for $x = 1$).
- In regression, we focus on conditional means/proportions (not joint proportions).

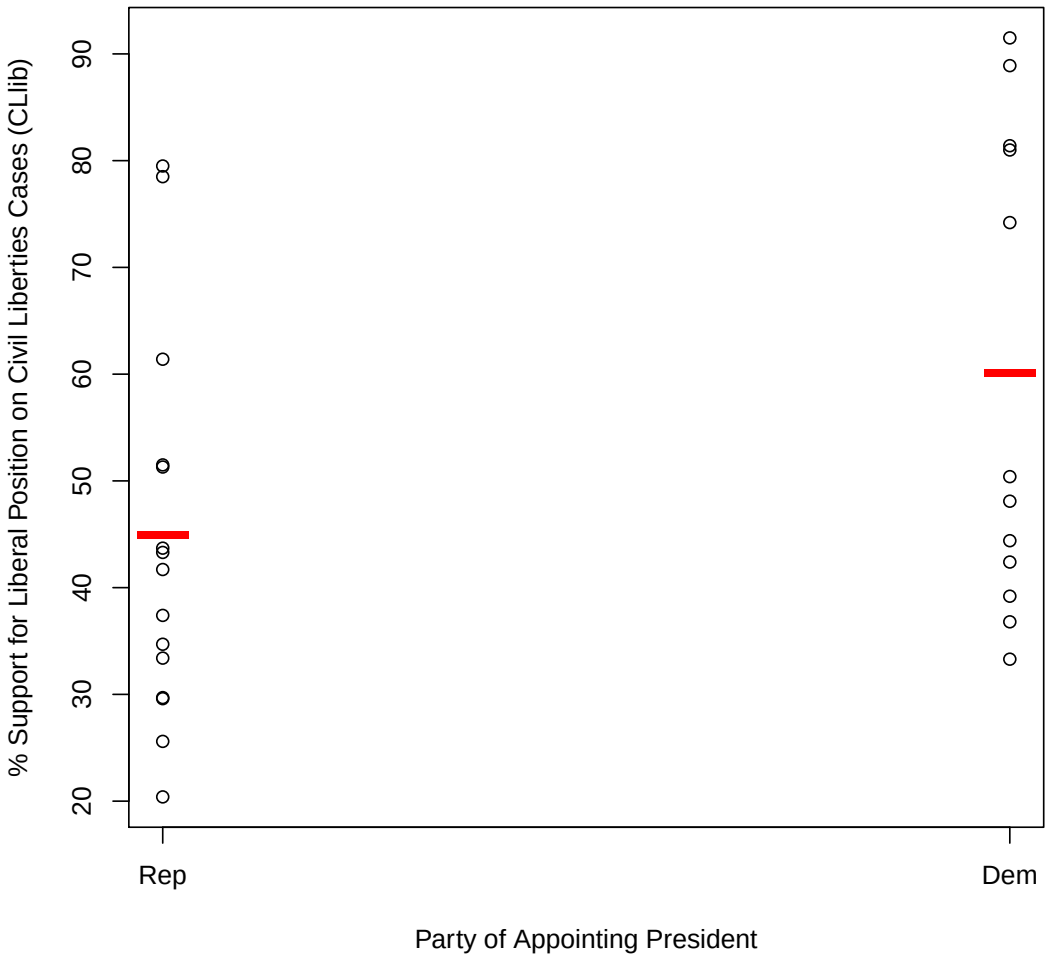
From Univariate to Regression



From Univariate to Regression



Mean CLlib Conditional on Party



Coarse and Fine-grained Covariates

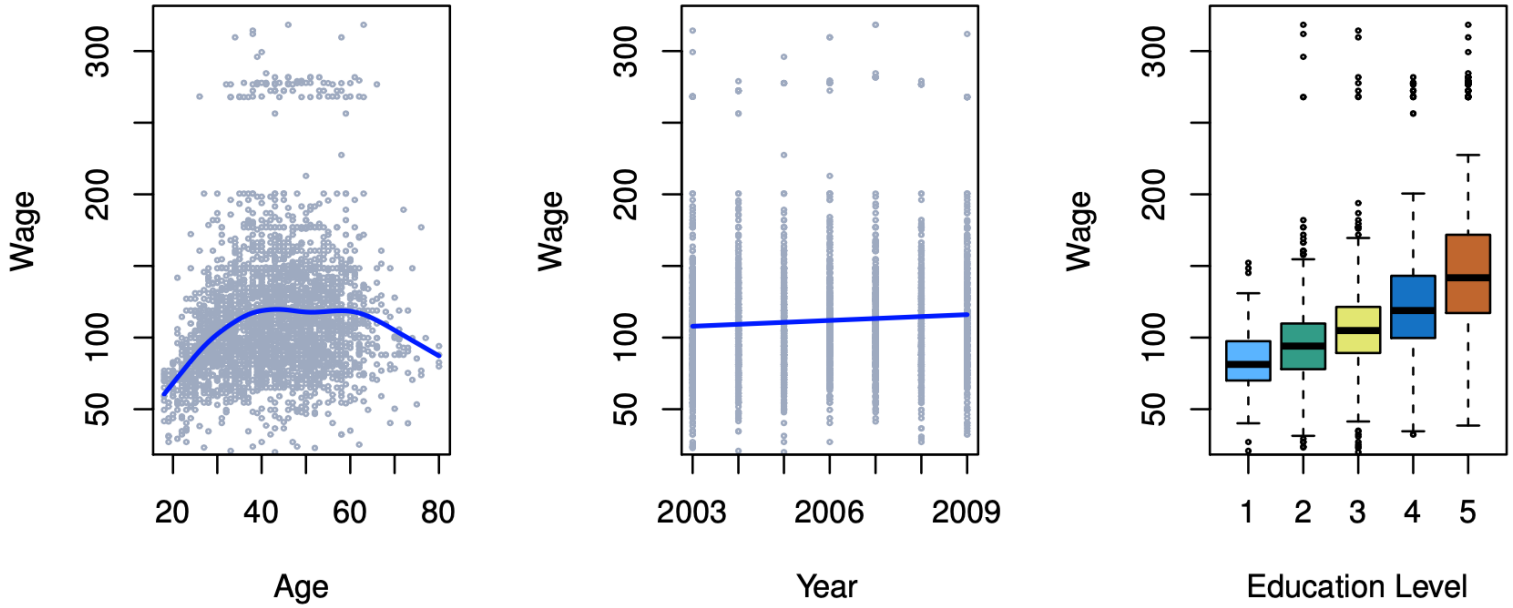


FIGURE 1.1. `Wage` *data, which contains income survey information for men from the central Atlantic region of the United States.* Left: `wage` *as a function of* `age`*. On average,* `wage` *increases with* `age` *until about* 60 *years of age, at which point it begins to decline.* Center: `wage` *as a function of* `year`*. There is a slow but steady increase of approximately* \$10,000 *in the average* `wage` *between* 2003 *and* 2009*.* Right: *Boxplots displaying* `wage` *as a function of* `education`*, with* 1 *indicating the lowest level (no high school diploma) and* 5 *the highest level (an advanced graduate degree). On average,* `wage` *increases with the level of education.*

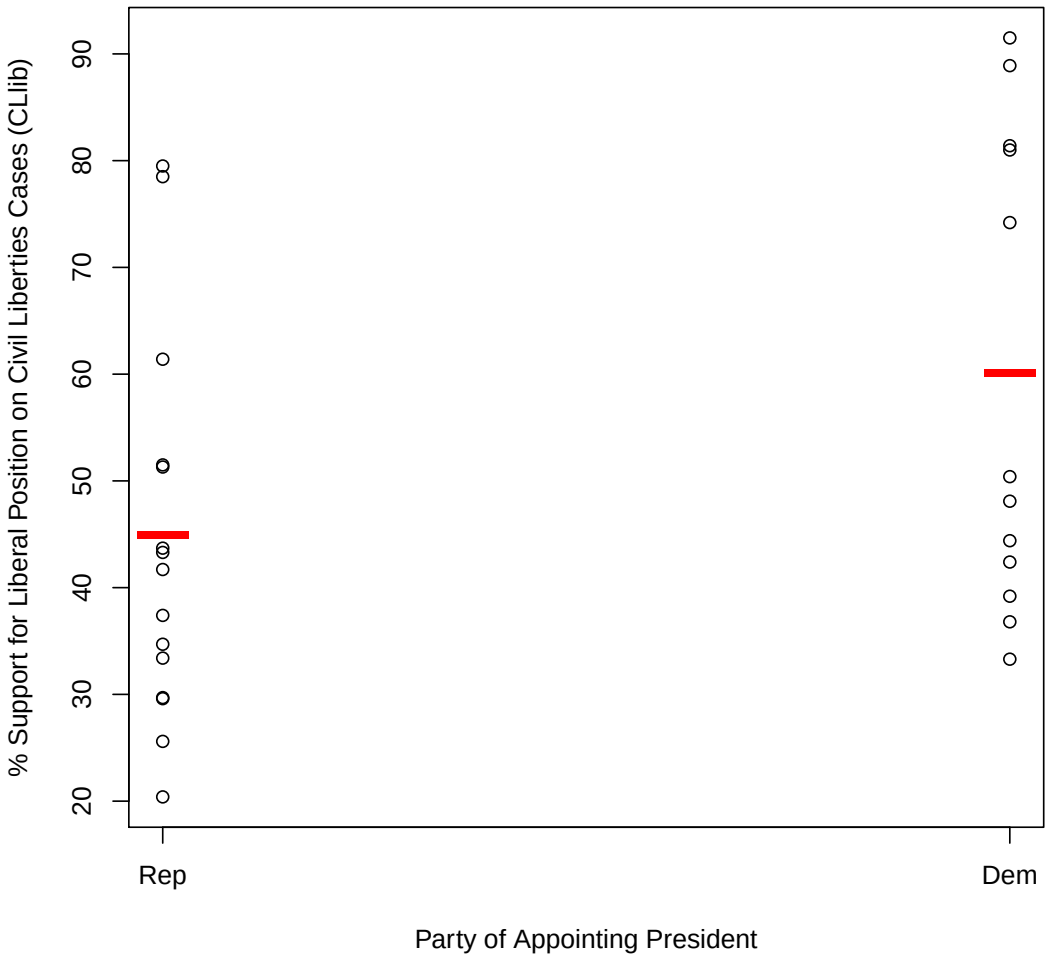
Outline

Univariate Descriptive Statistics

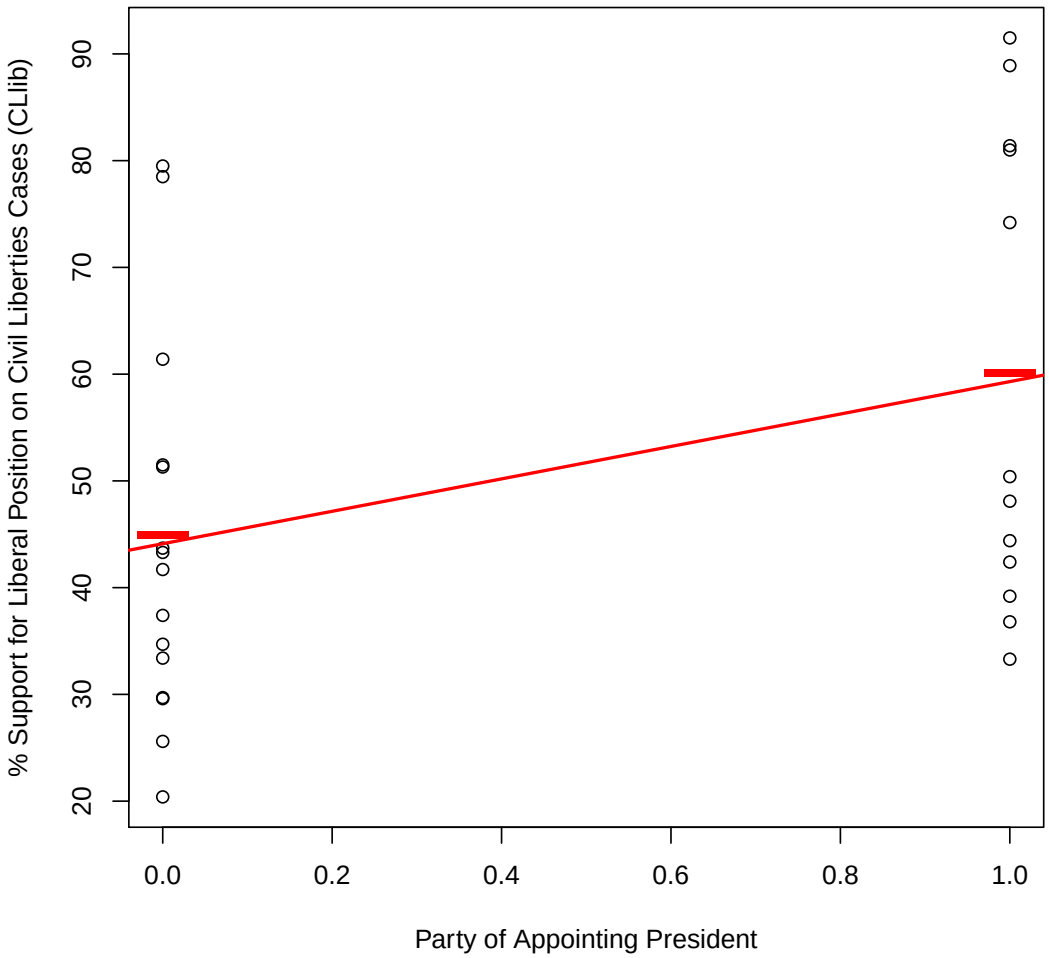
Descriptive Regression with a Coarse Covariate

Descriptive Linear Regression

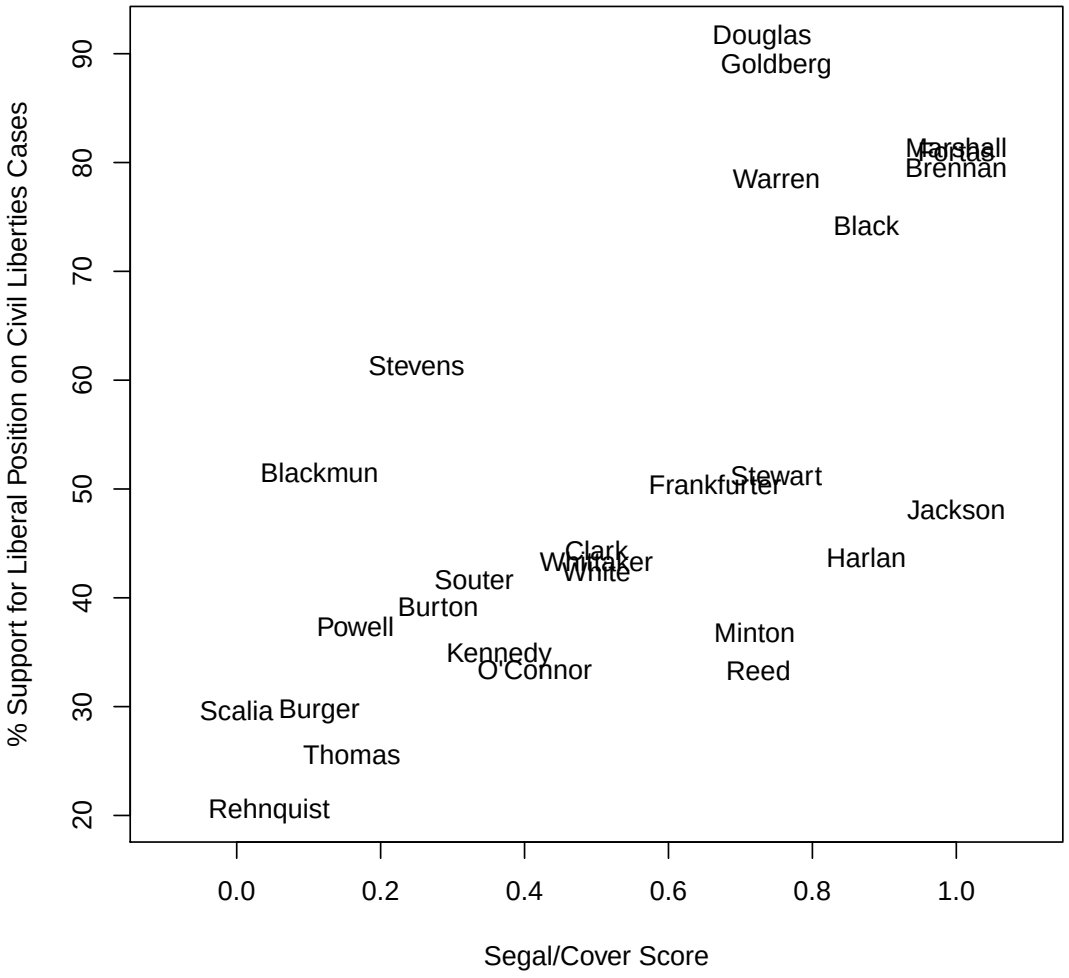
Mean CLlib Conditional on Party



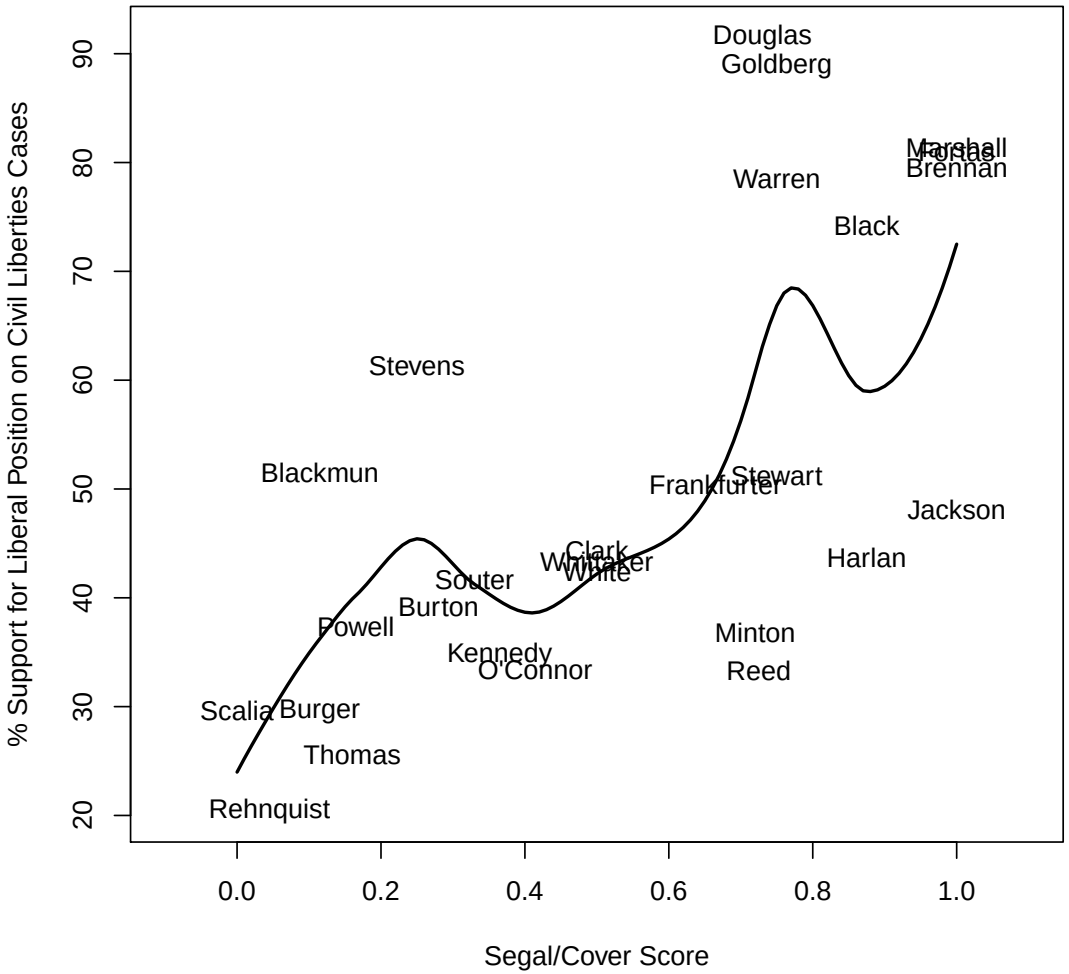
Linear Conditional Mean Function (LCMF)



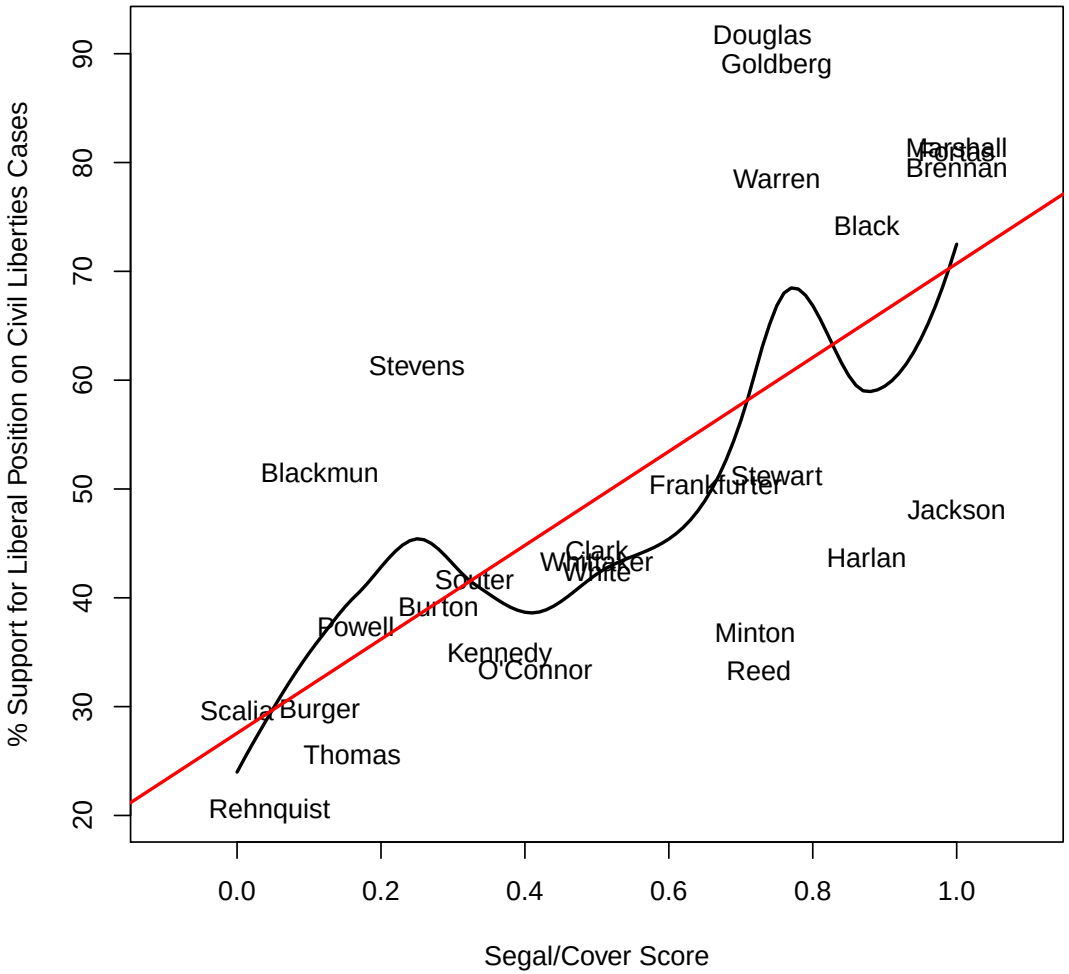
What about a continuous explanatory variable?



What about a continuous explanatory variable?

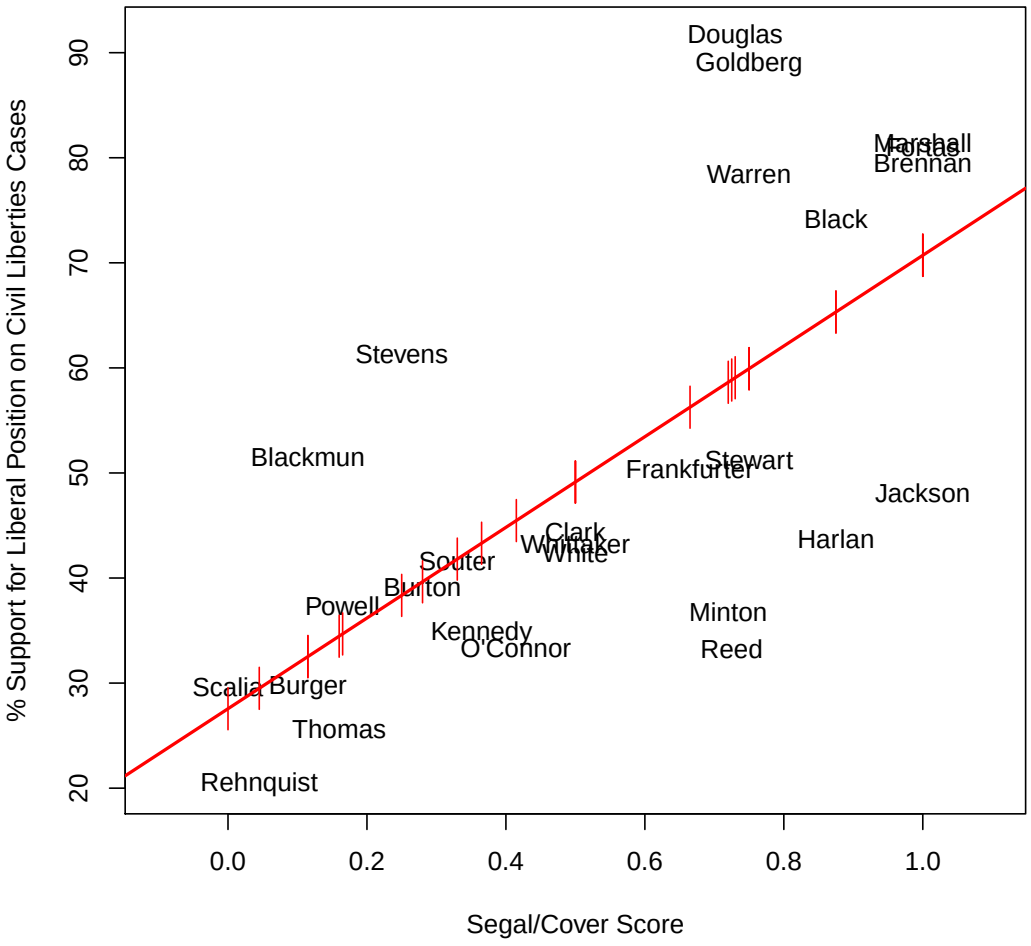


What about a continuous explanatory variable?



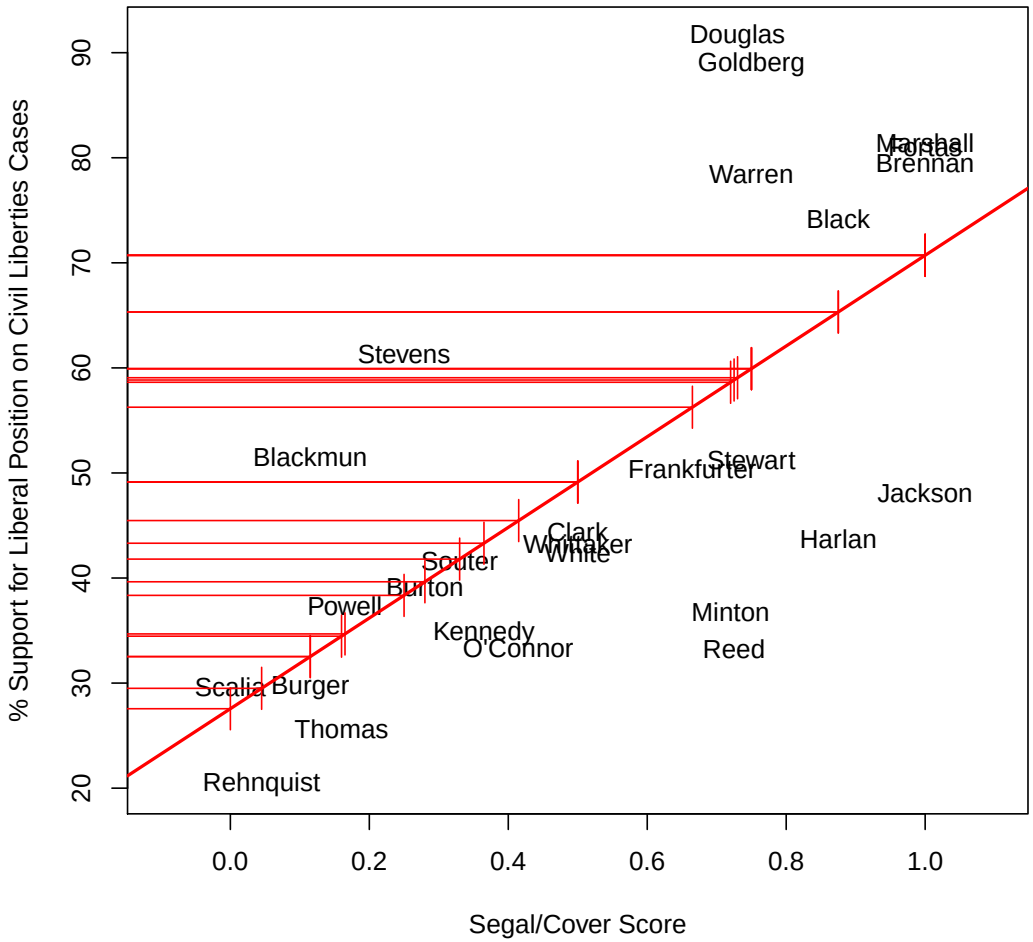
Fitted Values and Residuals

- Fitted Values from LCMF: $\hat{y}_i = B_0 + B_1 x_i$ for $i = 1, ..., N$
- Residuals from Fitted Values: $e_i = y_i - \hat{y}_i$ for $i = 1, ..., N$



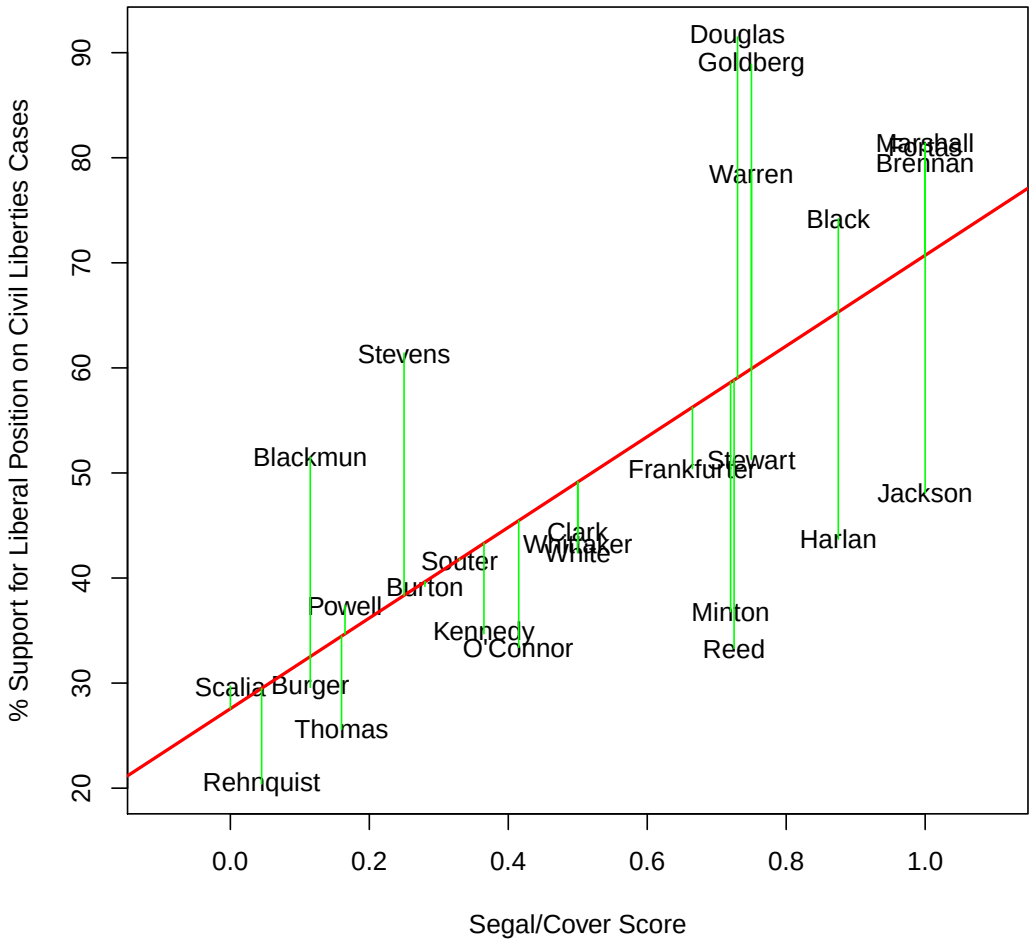
Fitted Values and Residuals

- Fitted Values from LCMF: $\hat{y}_i = B_0 + B_1 x_i$ for $i = 1, ..., N$
- Residuals from Fitted Values: $e_i = y_i - \hat{y}_i$ for $i = 1, ..., N$



Fitted Values and Residuals

- Fitted Values from LCMF: $\hat{y}_i = B_0 + B_1 x_i$ for $i = 1, ..., N$
- Residuals from Fitted Values: $e_i = y_i - \hat{y}_i$ for $i = 1, ..., N$

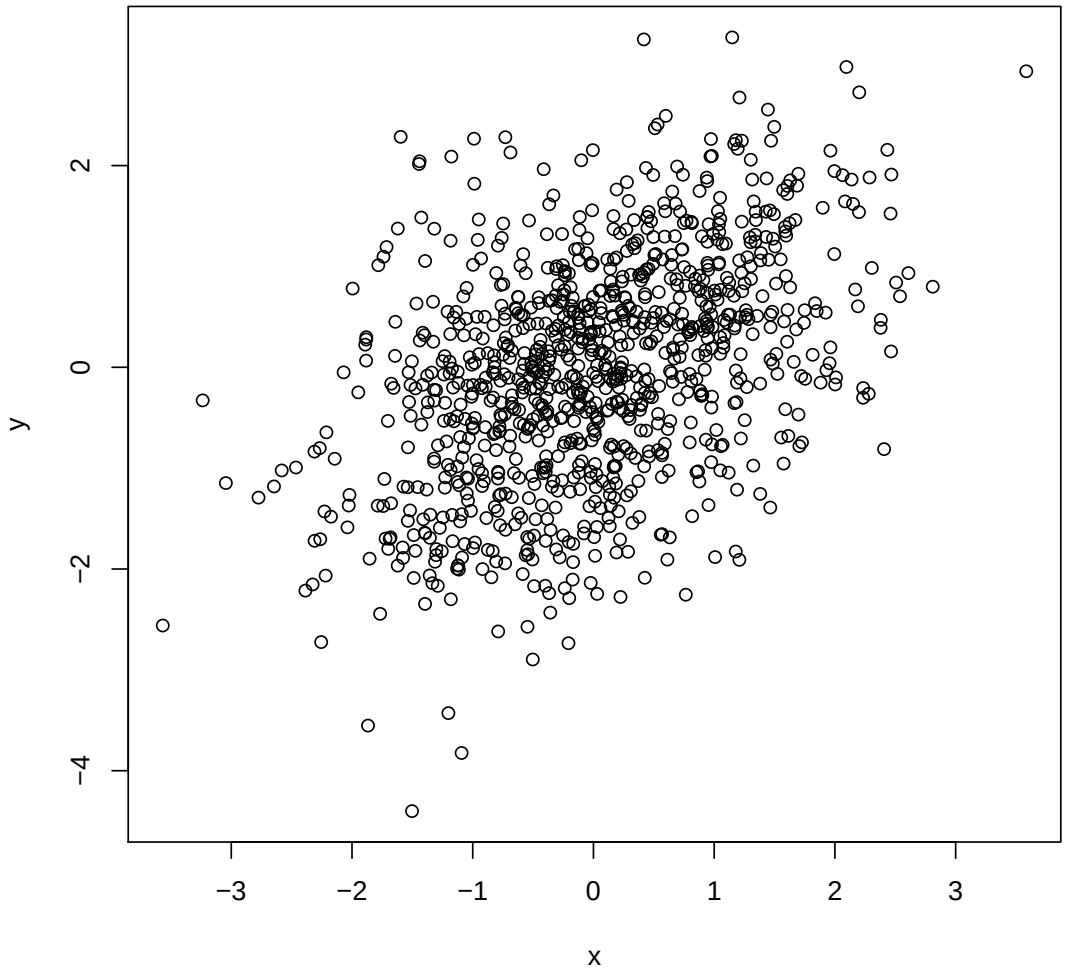


Minimizing the Residuals

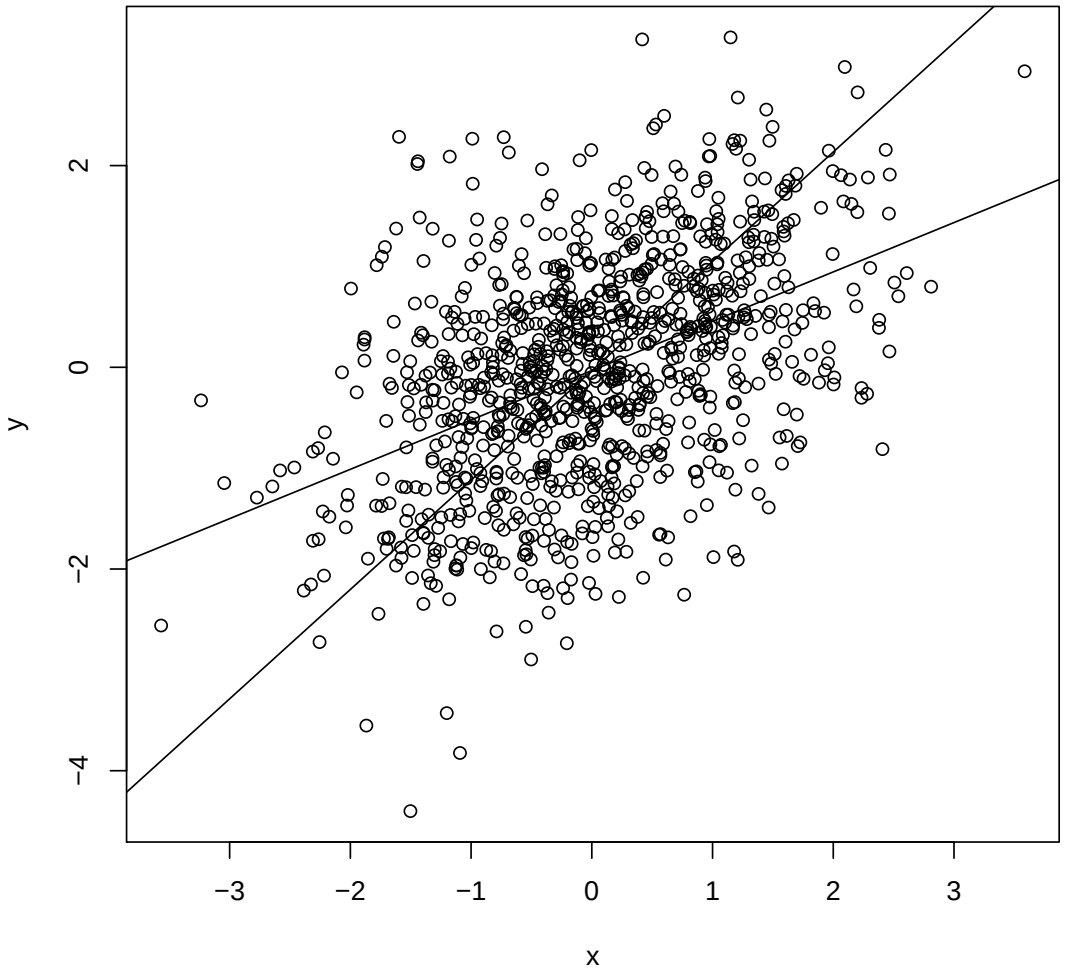
- Let $\tilde{B}_0$ and $\tilde{B}_1$ be possible values for B_0 and B_1 respectively,
- *Least Squares (LS) regression*:

$$(B_0, B_1) = \underset{\tilde{B}_0, \tilde{B}_1}{\operatorname{argmin}} \sum_{i=1}^{N} (y_i - \tilde{B}_0 - x_i \tilde{B}_1)^2$$

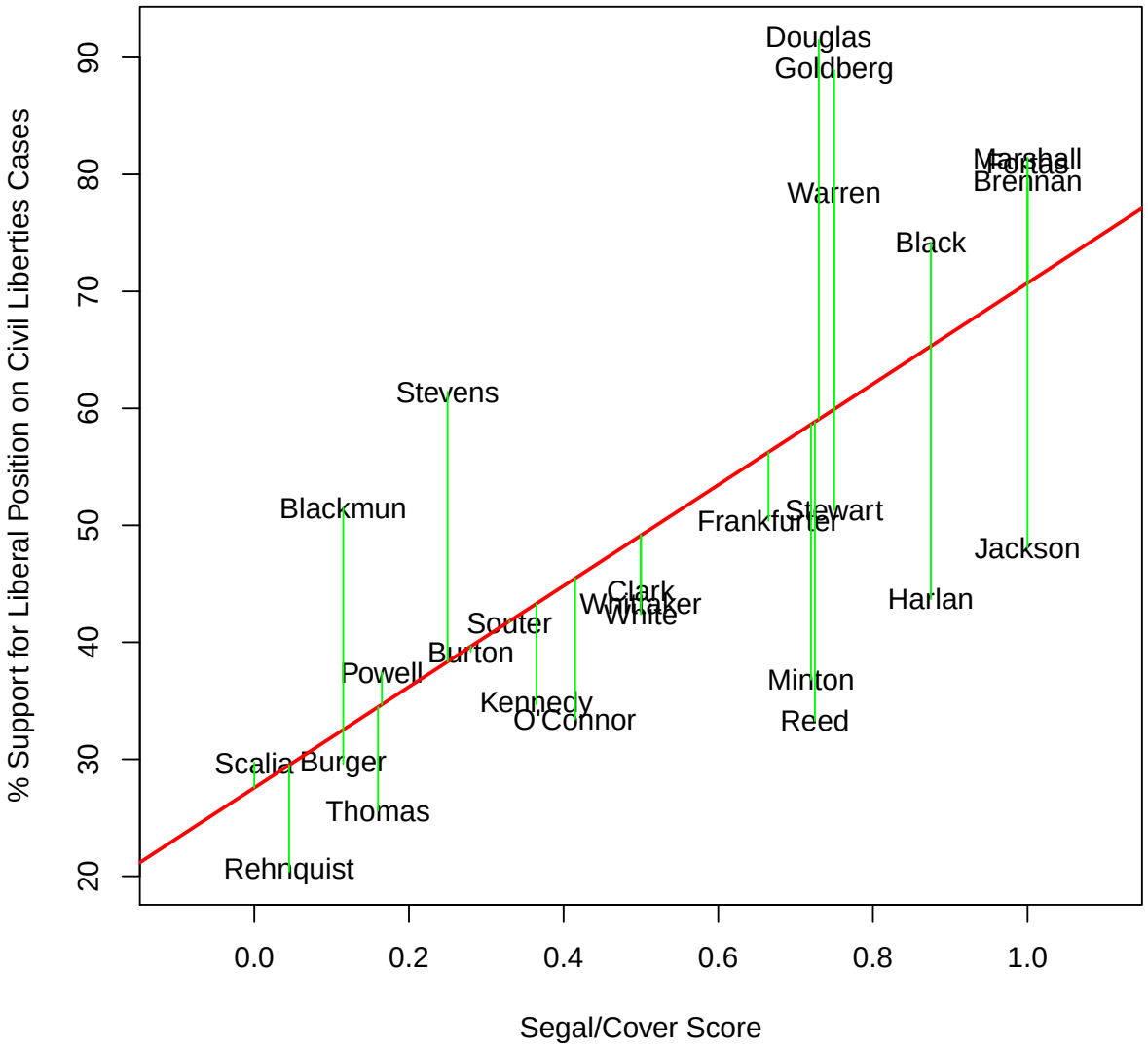
Simple Regression Activity



Simple Regression Activity



Choosing Slope and Intercept to Minimize SSR



Deriving the Linear Least Squares Estimator

Let $\tilde{B}_0$ and $\tilde{B}_1$ be possible values for B_0 and B_1 respectively, and

$$S(\tilde{B}_0, \tilde{B}_1) = \sum_{i=1}^{N} (y_i - \tilde{B}_0 - x_i \tilde{B}_1)^2.$$

1. Take partial derivatives of S with respect to $\tilde{B}_0$ and $\tilde{B}_1$.
2. Set each of the partial derivatives to 0
3. Substitute B_0 and B_1 for $\tilde{B}_0$ and $\tilde{B}_1$ and solve for B_0 and B_1

Useful Results

The partial derivatives imply that

$$\sum_{i=1}^{N} x_i e_i = 0$$

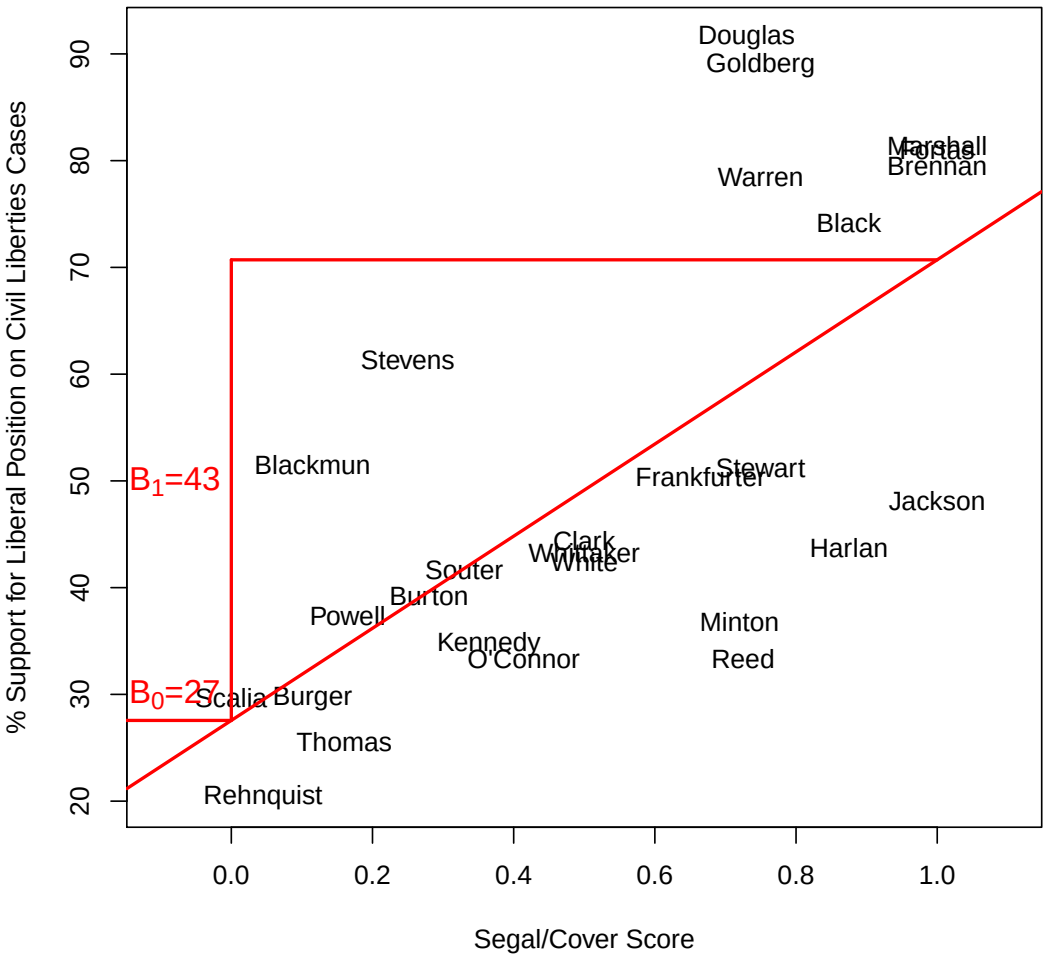
$$\sum_{i=1}^{N} \hat{y}_i e_i = 0$$

and the LS LCMF:

$$B_0 = \bar{y} - B_1 \bar{x}$$

$$B_1 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

Descriptive Interpretation



Variance and Regression

- Again, we want to assess the "leastness" of least squares (i.e., the spread of the data around the line).
- The standard error of the regression, $\sqrt{\frac{\sum_{i=1}^{N} e_i^2}{N-2}}$, is analogous to the standard deviation.
- There are other approaches to this question: correlation and coefficient of determination.

Covariance

The population variances for X and Y can be written as:

$$S_x^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}$$
$$= \frac{\sum_{i=1}^{N}(x_i - \bar{x})(x_i - \bar{x})}{N-1}$$

$$S_y^2 = \frac{\sum_{i=1}^{N}(y_i - \bar{y})(y_i - \bar{y})}{N-1}$$

Covariance

The population variances for X and Y can be written as:

$$S_x^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}$$

$$= \frac{\sum_{i=1}^{N}(x_i - \bar{x})(x_i - \bar{x})}{N-1}$$

$$S_y^2 = \frac{\sum_{i=1}^{N}(y_i - \bar{y})(y_i - \bar{y})}{N-1}$$

Covariance

The population variances for X and Y can be written as:

$$S_x^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}$$
$$= \frac{\sum_{i=1}^{N}(x_i - \bar{x})(x_i - \bar{x})}{N-1}$$

$$S_y^2 = \frac{\sum_{i=1}^{N}(y_i - \bar{y})(y_i - \bar{y})}{N-1}$$

Covariance

The population variances for X and Y can be written as:

$$S_x^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}$$
$$= \frac{\sum_{i=1}^{N}(x_i - \bar{x})(x_i - \bar{x})}{N-1}$$

$$S_y^2 = \frac{\sum_{i=1}^{N}(y_i - \bar{y})(y_i - \bar{y})}{N-1}$$

What if we defined a similar quantity involving both X and Y?

Covariance

The population variances for X and Y can be written as:

$$S_x^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}$$
$$= \frac{\sum_{i=1}^{N}(x_i - \bar{x})(x_i - \bar{x})}{N-1}$$

$$S_y^2 = \frac{\sum_{i=1}^{N}(y_i - \bar{y})(y_i - \bar{y})}{N-1}$$

What if we defined a similar quantity involving both X and Y?

$$S_{xy} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{N-1}$$

Covariance

The population variances for X and Y can be written as:

$$S_x^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}$$
$$= \frac{\sum_{i=1}^{N}(x_i - \bar{x})(x_i - \bar{x})}{N-1}$$

$$S_y^2 = \frac{\sum_{i=1}^{N}(y_i - \bar{y})(y_i - \bar{y})}{N-1}$$

What if we defined a similar quantity involving both X and Y?

$$S_{xy} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{N-1}$$

This quantity is the covariance.

Correlation

The covariance does not have units that are easy to interpret, so we can scale it by the standard deviations of X and Y to obtain the correlation:

$$R = \frac{S_{xy}}{S_x S_y} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2}}$$

The correlation is always in the interval:

$$-1 \leq R \leq 1$$

Correlation

The covariance does not have units that are easy to interpret, so we can scale it by the standard deviations of X and Y to obtain the correlation:

$$R = \frac{S_{xy}}{S_x S_y} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2}}$$

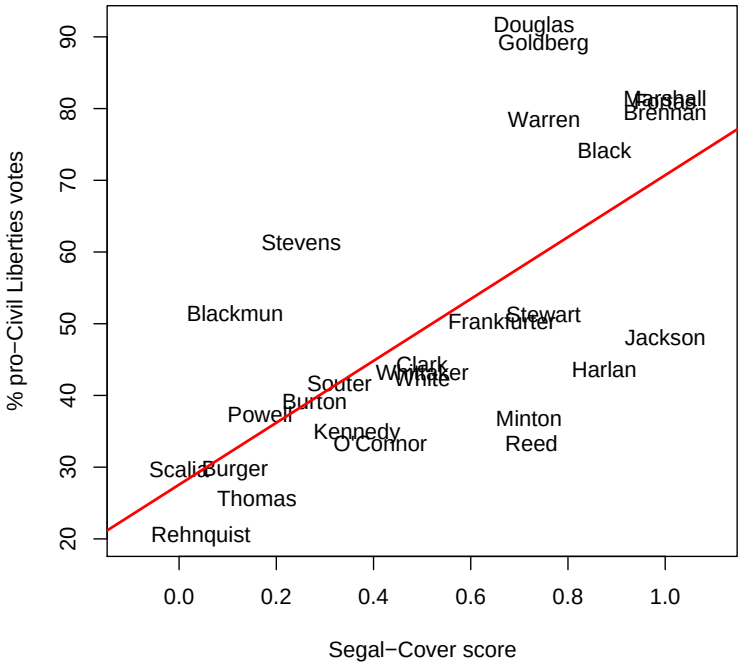
The correlation is always in the interval:

$$-1 \leq R \leq 1$$

Analysis of Variance for the Regression

How well do the Segal-Cover scores predict support for civil liberties in relation to the average support for civil liberties?

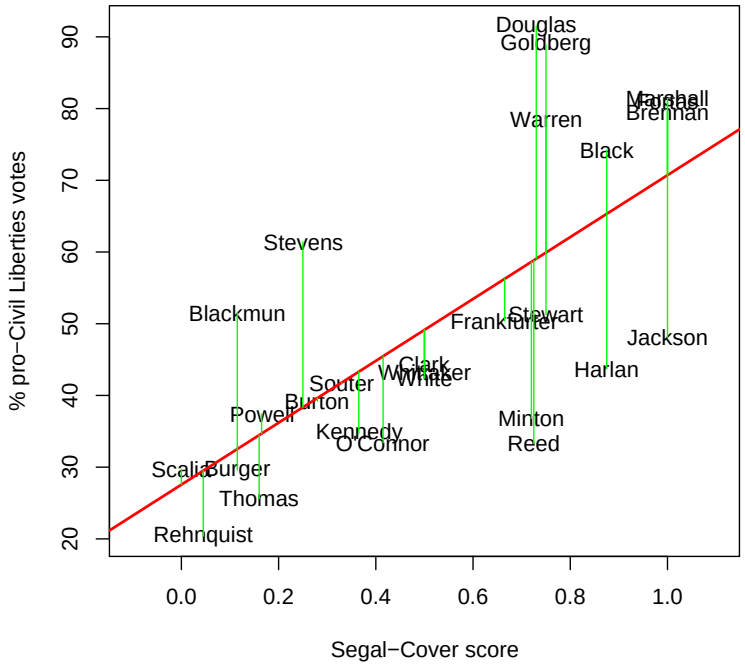
- $e_i = y_i - \widehat{y}_i$
- $y_i - \bar{y}$
- $\widehat{y}_i - \bar{y}$
- $(y_i - \bar{y}) = e_i + (\widehat{y}_i - \bar{y})$



Analysis of Variance for the Regression

How well do the Segal-Cover scores predict support for civil liberties in relation to the average support for civil liberties?

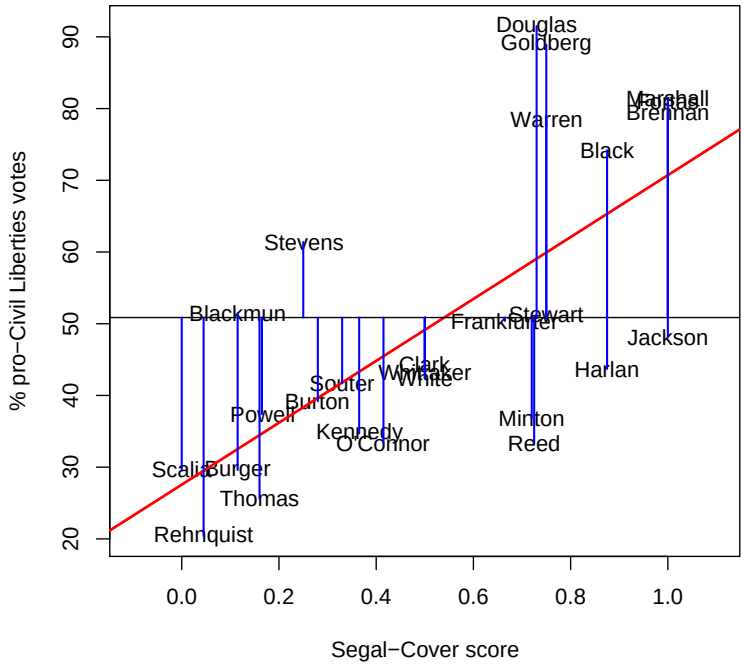
- $e_i = y_i - \widehat{y}_i$
- $y_i - \bar{y}$
- $\widehat{y}_i - \bar{y}$
- $(y_i - \bar{y}) = e_i + (\widehat{y}_i - \bar{y})$



Analysis of Variance for the Regression

How well do the Segal-Cover scores predict support for civil liberties in relation to the average support for civil liberties?

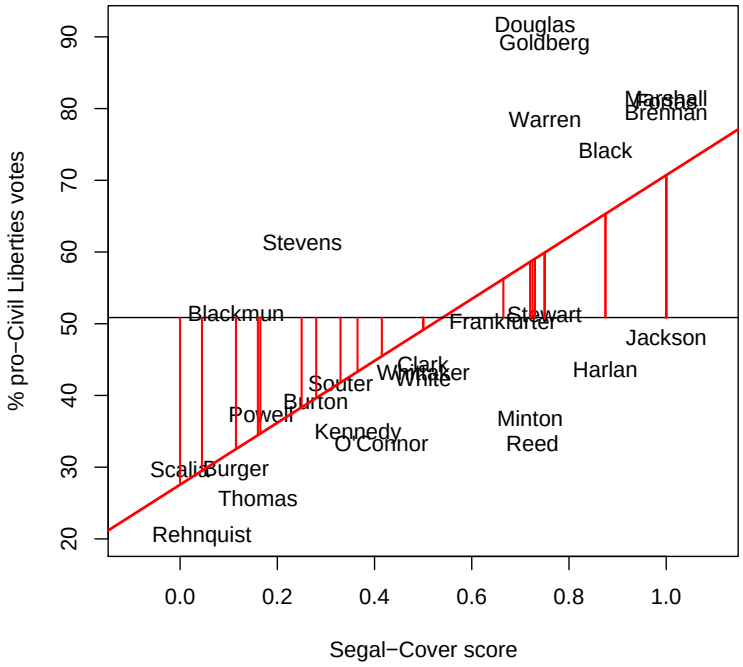
- $e_i = y_i - \widehat{y}_i$
- $y_i - \bar{y}$
- $\widehat{y}_i - \bar{y}$
- $(y_i - \bar{y}) = e_i + (\widehat{y}_i - \bar{y})$



Analysis of Variance for the Regression

How well do the Segal-Cover scores predict support for civil liberties in relation to the average support for civil liberties?

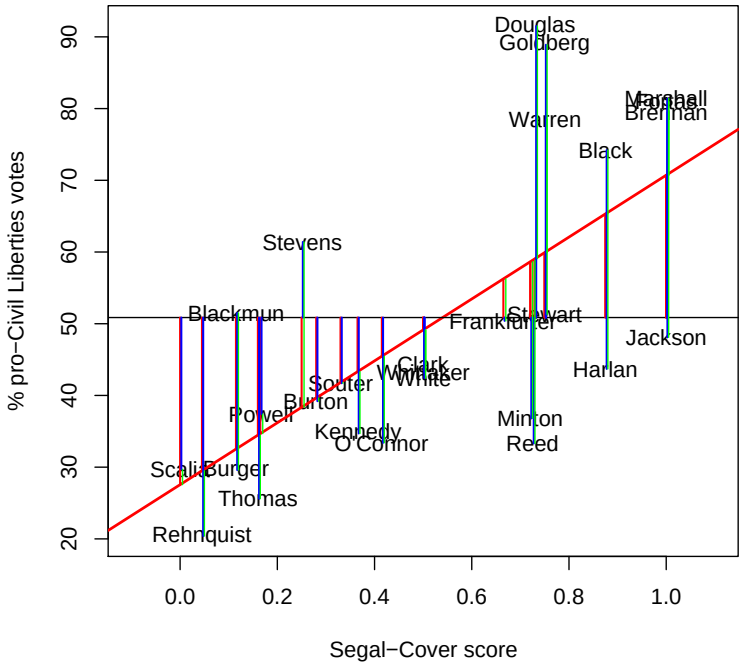
- $e_i = y_i - \widehat{y}_i$
- $y_i - \bar{y}$
- $\widehat{y}_i - \bar{y}$
- $(y_i - \bar{y}) = e_i + (\widehat{y}_i - \bar{y})$



Analysis of Variance for the Regression

How well do the Segal-Cover scores predict support for civil liberties in relation to the average support for civil liberties?

- $e_i = y_i - \widehat{y}_i$
- $y_i - \bar{y}$
- $\widehat{y}_i - \bar{y}$
- $(y_i - \bar{y}) = e_i + (\widehat{y}_i - \bar{y})$



$$\sum_{i=1}^{N}(y_i - \bar{y})^2 = \sum_{i=1}^{N}\{e_i + (\widehat{y}_i - \bar{y})\}^2$$

$$
\begin{aligned}
\sum_{i=1}^{N}(y_i - \bar{y})^2 &= \sum_{i=1}^{N}\{e_i + (\widehat{y}_i - \bar{y})\}^2 \\
&= \sum_{i=1}^{N}\{e_i^2 + 2e_i(\widehat{y}_i - \bar{y}) + (\widehat{y}_i - \bar{y})^2\}
\end{aligned}
$$

$$
\begin{aligned}
\sum_{i=1}^{N}(y_i - \bar{y})^2 &= \sum_{i=1}^{N}\{e_i + (\widehat{y}_i - \bar{y})\}^2 \\
&= \sum_{i=1}^{N}\{e_i^2 + 2e_i(\widehat{y}_i - \bar{y}) + (\widehat{y}_i - \bar{y})^2\} \\
&= \sum_{i=1}^{N} e_i^2 + 2\sum_{i=1}^{N} e_i(\widehat{y}_i - \bar{y}) + \sum_{i=1}^{N}(\widehat{y}_i - \bar{y})^2
\end{aligned}
$$

$$
\begin{aligned}
\sum_{i=1}^{N}(y_i - \bar{y})^2 &= \sum_{i=1}^{N}\{e_i + (\widehat{y}_i - \bar{y})\}^2 \\
&= \sum_{i=1}^{N}\{e_i^2 + 2e_i(\widehat{y}_i - \bar{y}) + (\widehat{y}_i - \bar{y})^2\} \\
&= \sum_{i=1}^{N} e_i^2 + 2\sum_{i=1}^{N} e_i(\widehat{y}_i - \bar{y}) + \sum_{i=1}^{N}(\widehat{y}_i - \bar{y})^2 \\
&= \sum_{i=1}^{N} e_i^2 + \sum_{i=1}^{N}(\widehat{y}_i - \bar{y})^2
\end{aligned}
$$

$$
\begin{aligned}
\sum_{i=1}^{N}(y_i - \bar{y})^2 &= \sum_{i=1}^{N}\{e_i + (\widehat{y}_i - \bar{y})\}^2 \\
&= \sum_{i=1}^{N}\{e_i^2 + 2e_i(\widehat{y}_i - \bar{y}) + (\widehat{y}_i - \bar{y})^2\} \\
&= \sum_{i=1}^{N} e_i^2 + 2\sum_{i=1}^{N} e_i(\widehat{y}_i - \bar{y}) + \sum_{i=1}^{N}(\widehat{y}_i - \bar{y})^2 \\
&= \sum_{i=1}^{N} e_i^2 + \sum_{i=1}^{N}(\widehat{y}_i - \bar{y})^2 \\
TSS &= SSR + RegSS
\end{aligned}
$$

$$
\begin{aligned}
\sum_{i=1}^{N}(y_i - \bar{y})^2 &= \sum_{i=1}^{N}\{e_i + (\widehat{y}_i - \bar{y})\}^2 \\
&= \sum_{i=1}^{N}\{e_i^2 + 2e_i(\widehat{y}_i - \bar{y}) + (\widehat{y}_i - \bar{y})^2\} \\
&= \sum_{i=1}^{N} e_i^2 + 2\sum_{i=1}^{N} e_i(\widehat{y}_i - \bar{y}) + \sum_{i=1}^{N}(\widehat{y}_i - \bar{y})^2 \\
&= \sum_{i=1}^{N} e_i^2 + \sum_{i=1}^{N}(\widehat{y}_i - \bar{y})^2 \\
TSS &= SSR + RegSS
\end{aligned}
$$

$$
\begin{aligned}
\sum_{i=1}^{N}(y_i - \bar{y})^2 &= \sum_{i=1}^{N}\{e_i + (\widehat{y}_i - \bar{y})\}^2 \\
&= \sum_{i=1}^{N}\{e_i^2 + 2e_i(\widehat{y}_i - \bar{y}) + (\widehat{y}_i - \bar{y})^2\} \\
&= \sum_{i=1}^{N} e_i^2 + 2\sum_{i=1}^{N} e_i(\widehat{y}_i - \bar{y}) + \sum_{i=1}^{N}(\widehat{y}_i - \bar{y})^2 \\
&= \sum_{i=1}^{N} e_i^2 + \sum_{i=1}^{N}(\widehat{y}_i - \bar{y})^2 \\
TSS &= SSR + RegSS
\end{aligned}
$$

$$
\begin{aligned}
\sum_{i=1}^{N}(y_i - \bar{y})^2 &= \sum_{i=1}^{N}\{e_i + (\widehat{y}_i - \bar{y})\}^2 \\
&= \sum_{i=1}^{N}\{e_i^2 + 2e_i(\widehat{y}_i - \bar{y}) + (\widehat{y}_i - \bar{y})^2\} \\
&= \sum_{i=1}^{N} e_i^2 + 2\sum_{i=1}^{N} e_i(\widehat{y}_i - \bar{y}) + \sum_{i=1}^{N}(\widehat{y}_i - \bar{y})^2 \\
&= \sum_{i=1}^{N} e_i^2 + \sum_{i=1}^{N}(\widehat{y}_i - \bar{y})^2 \\
TSS &= SSR + RegSS
\end{aligned}
$$

Coefficient of Determination

We can divide each side by the TSS:

$$\frac{SSR}{TSS} + \frac{RegSS}{TSS} = \frac{TSS}{TSS}$$

$$\frac{SSR}{TSS} + \frac{RegSS}{TSS} = 1$$

$$\frac{RegSS}{TSS} = 1 - \frac{SSR}{TSS} = R^2$$

R^2 is a measure of how much of the variation in Y is accounted for by X.

Coefficient of Determination

We can divide each side by the TSS:

$$\frac{SSR}{TSS} + \frac{RegSS}{TSS} = \frac{TSS}{TSS}$$

$$\frac{SSR}{TSS} + \frac{RegSS}{TSS} = 1$$

$$\frac{RegSS}{TSS} = 1 - \frac{SSR}{TSS} = R^2$$

R^2 is a measure of how much of the variation in Y is accounted for by X.

Coefficient of Determination

We can divide each side by the TSS:

$$\frac{SSR}{TSS} + \frac{RegSS}{TSS} = \frac{TSS}{TSS}$$

$$\frac{SSR}{TSS} + \frac{RegSS}{TSS} = 1$$

$$\frac{RegSS}{TSS} = 1 - \frac{SSR}{TSS} = R^2$$

R^2 is a measure of how much of the variation in Y is accounted for by X.

Coefficient of Determination

We can divide each side by the TSS:

$$\frac{SSR}{TSS} + \frac{RegSS}{TSS} = \frac{TSS}{TSS}$$

$$\frac{SSR}{TSS} + \frac{RegSS}{TSS} = 1$$

$$\frac{RegSS}{TSS} = 1 - \frac{SSR}{TSS} = R^2$$

R^2 is a measure of how much of the variation in Y is accounted for by X.

Usefull Identities

$$R^2 = (R)^2$$

$$B_1 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

Usefull Identities

$$R^2 = (R)^2$$

$$B_1 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

Usefull Identities

$$R^2 = (R)^2$$

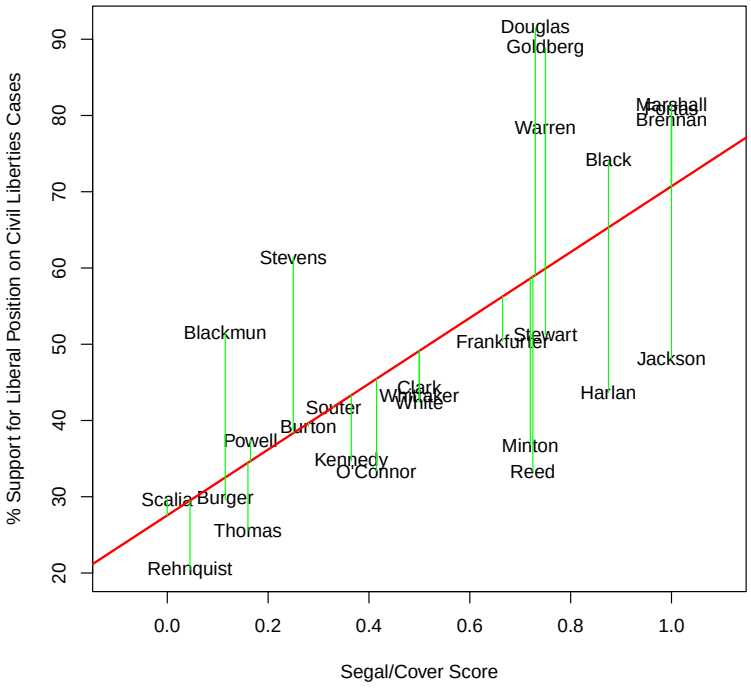
$$\begin{aligned} B_1 &= \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2} \\ &= \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 \sum_{i=1}^{N} n(y_i - \bar{y})^2}} \cdot \frac{\sqrt{\frac{\sum_{i=1}^{N}(y_i - \bar{y})^2}{N-1}}}{\sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}}} \end{aligned}$$

Usefull Identities

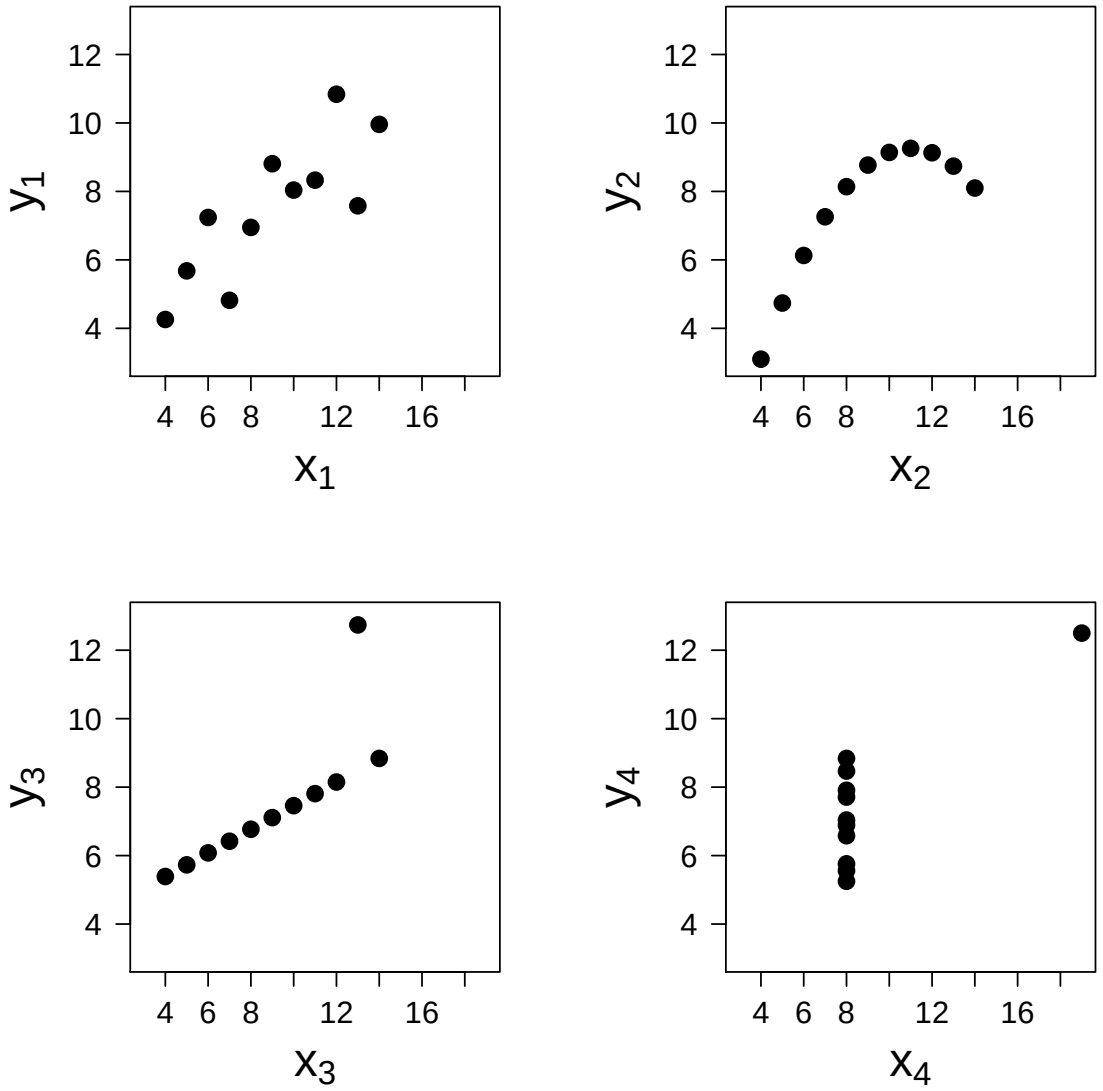
$$R^2 = (R)^2$$

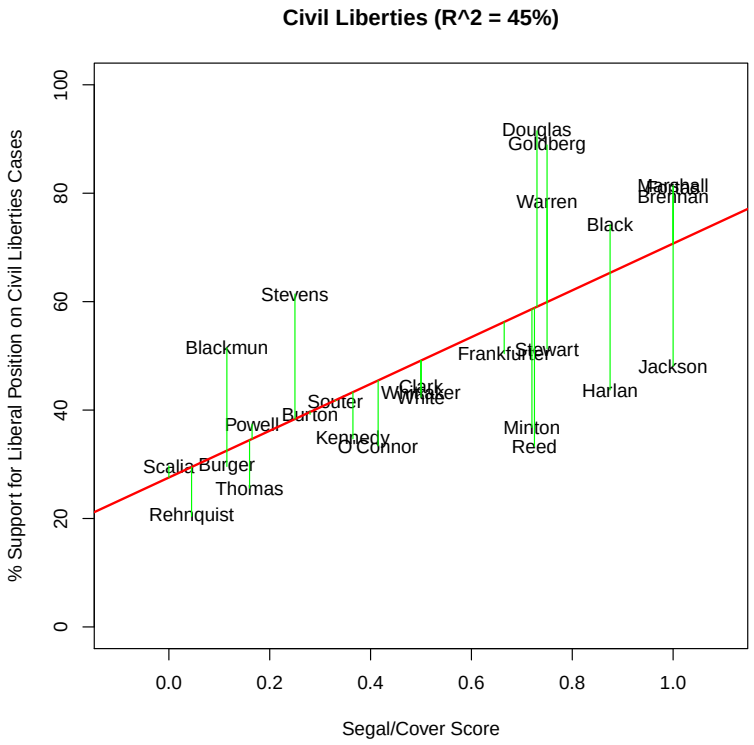
$$\begin{aligned}
B_1 &= \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2} \\
&= \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 \sum_{i=1}^{N} n(y_i - \bar{y})^2}} \cdot \frac{\sqrt{\frac{\sum_{i=1}^{N}(y_i - \bar{y})^2}{N-1}}}{\sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}}} \\
&= R\frac{S_y}{S_x}
\end{aligned}$$

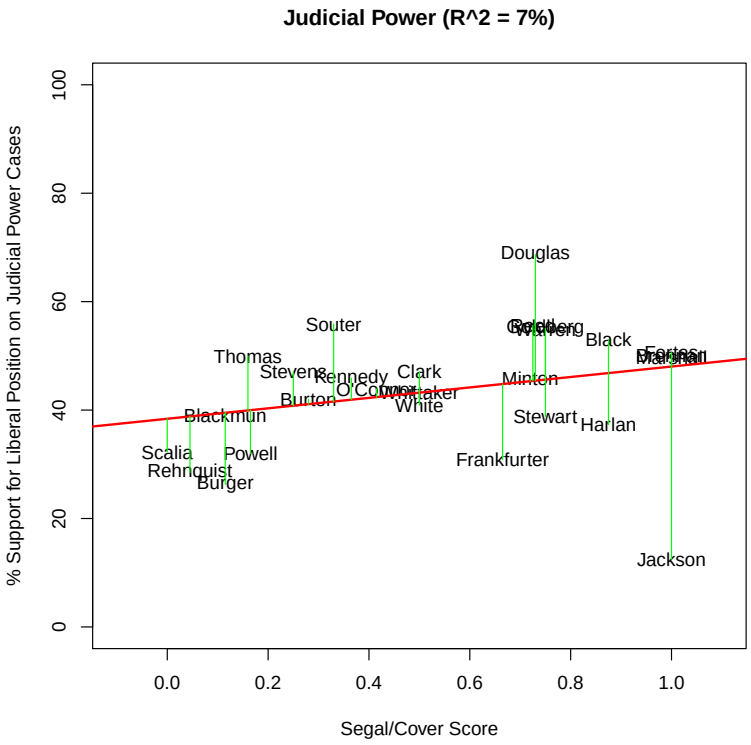
- $R^2 = 0.45$
- $R = \sqrt{R^2} = 0.67$
- $S_x = 0.33$
- $S_y = 20.9$
- $B_1 = R\frac{S_y}{S_x} = 43.15$



Anscombe's Quartet




Civil Liberties (R^2 = 45%)
% Support for Liberal Position on Civil Liberties Cases
Segal/Cover Score
0
20
40
60
80
100
0.0
0.2
0.4
0.6
0.8
1.0
Douglas
Goldberg
Warren
Black
Brennan
Stevens
Blackmun
Frankfurter
Stewart
Jackson
Harlan
Souter
Burton
Powell
Kennedy
O'Connor
Minton
Reed
Scalia
Burger
Thomas
Rehnquist


Judicial Power (R^2 = 7%)
% Support for Liberal Position on Judicial Power Cases
Segal/Cover Score
0
20
40
60
80
100
0.0
0.2
0.4
0.6
0.8
1.0
Douglas
Souter
Black
Thomas
Stevens
Kennedy
Clark
Minton
Burton
White
Blackmun
Stewart
Harlan
Scalia
Powell
Rehnquist
Burger
Frankfurter
Jackson

Tuesday Summary

- We will focus on conditional means as a good conditional summary of a single outcome variable.
- Conditional means are straightforward for a coarse covariate.
- Linear regression provides a parsimonious summary for a fine-grained covariate.

On Thursday

- Binary outcomes with a fine-grained covariate
- Flexible (i.e., ML) approaches