Lecture3: Testing Bivariate Relationships

Dr. Rachel Blum

October 31, 2019

Housekeeping

1. Today we are going to make R less confusing, among other things.
2. The answer key to problem set 1 is on canvas.
3. Problem set 2 and all necessary materials are on canvas. PS2 is due on Tuesday.
4. I will also have PS1 graded by then.
5. Make sure to complete the required readings!

Table of contents

Big Question: How do we know what we know?

Challenges to assessing cause and effect

- Randomness
- Endogeneity
- Measurement error
- Complexity of human behavior

Today: the Gauss-Markov theorem

We will focus on conditions under which OLS is the ...

- B est (most consistent)
- L inear (and sometimes non-linear)
- U nbiased (most accurate)
- E stimator of β

The Core Model: Overview

Quantifying relationships between two variables

Correlations tell us about the association between two variables (X, Y), but we may want to know more about how **changes in one variable X correspond to changes in another variable Y.**

Quantifying relationships between two variables

Behold, **OLS**: ordinary least squares regression.
OLS lets us quantify the relationship between X and Y to assess:

1. Whether the relationship occurs by chance, and
2. What we expect Y will be for any given value of X.

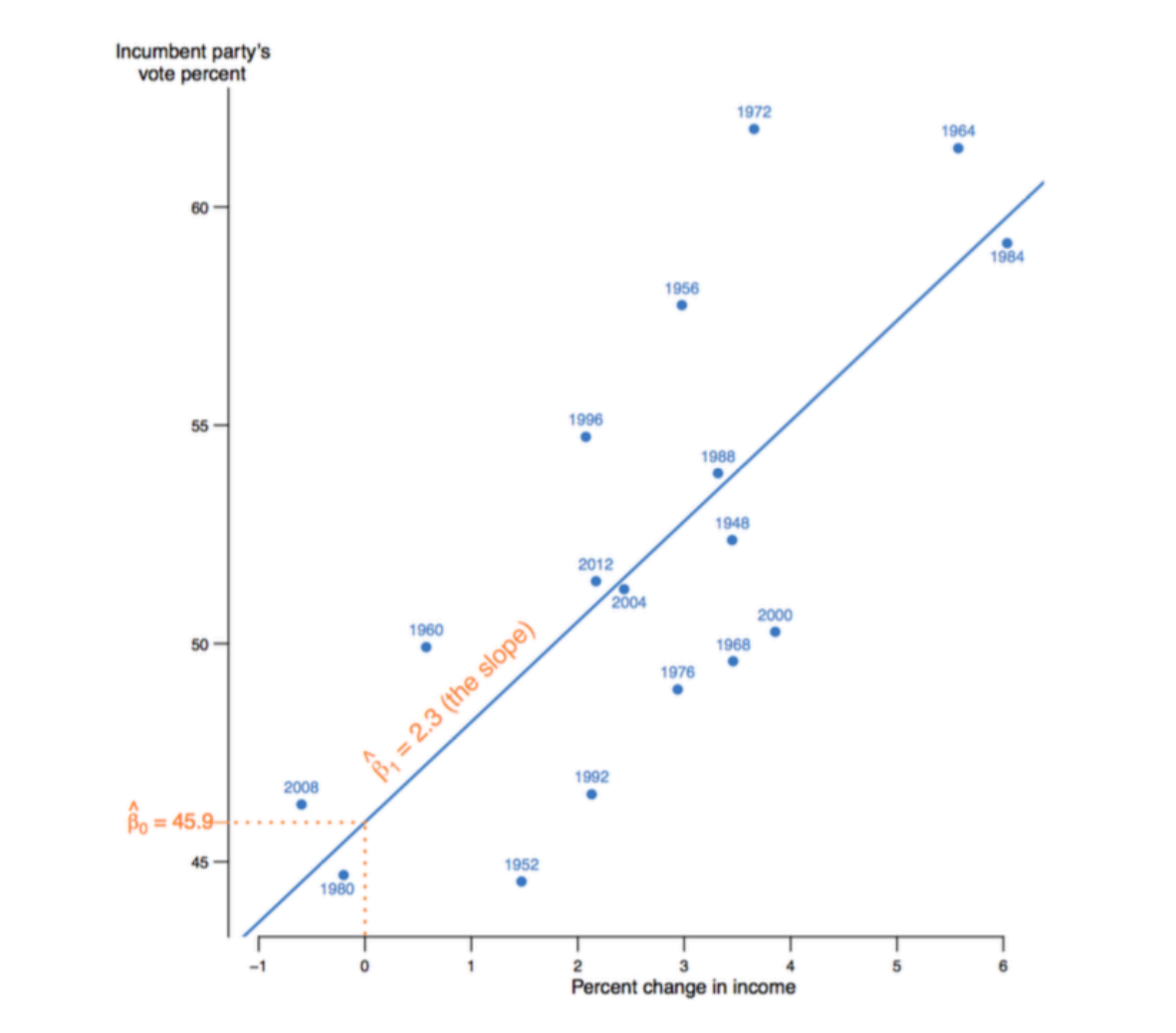
Called *linear* regression because we are estimating a *line* that best characterizes the relationship between X and Y.

The core model

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- Y The **dependent variable**, or the outcome of interest
- X The **independent variable**, or a possible cause
- ϵ The **error term**, or everything we haven't measured in our model
- β_0 The **intercept**, or the value of Y when X is zero
- β_1 The **slope**, or how much change in Y is expected if X changes by one unit

Application: presidential elections



The core model revisited

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

The core model revisited

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

Adding in the parameters and terms from our application:

$$\textit{Vote share}_i = \beta_0 + \beta_1 \textit{Income change}_i + \epsilon_i$$

- β_0 Intercept; expected vote share when income change is zero
- β_1 Slope; expected change in vote share for one-unit increase in income change

The model we actually estimate

Since we don't know the true (population) values of β_0 and β_1, we use data to estimate them as $\hat{\beta}_0$ and $\hat{\beta}_1$.

We estimate each using the concepts of **fitted values**:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

and **residuals**:

$$\begin{aligned} \hat{\epsilon}_i &= Y_i - \hat{Y}_i \\ &= Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i \end{aligned}$$

Estimating β_0 and β_1

OLS identifies values of $\hat{\beta}_0$ and $\hat{\beta}_1$ that define a line that minimizes the **sum of squared residuals** (SSR)

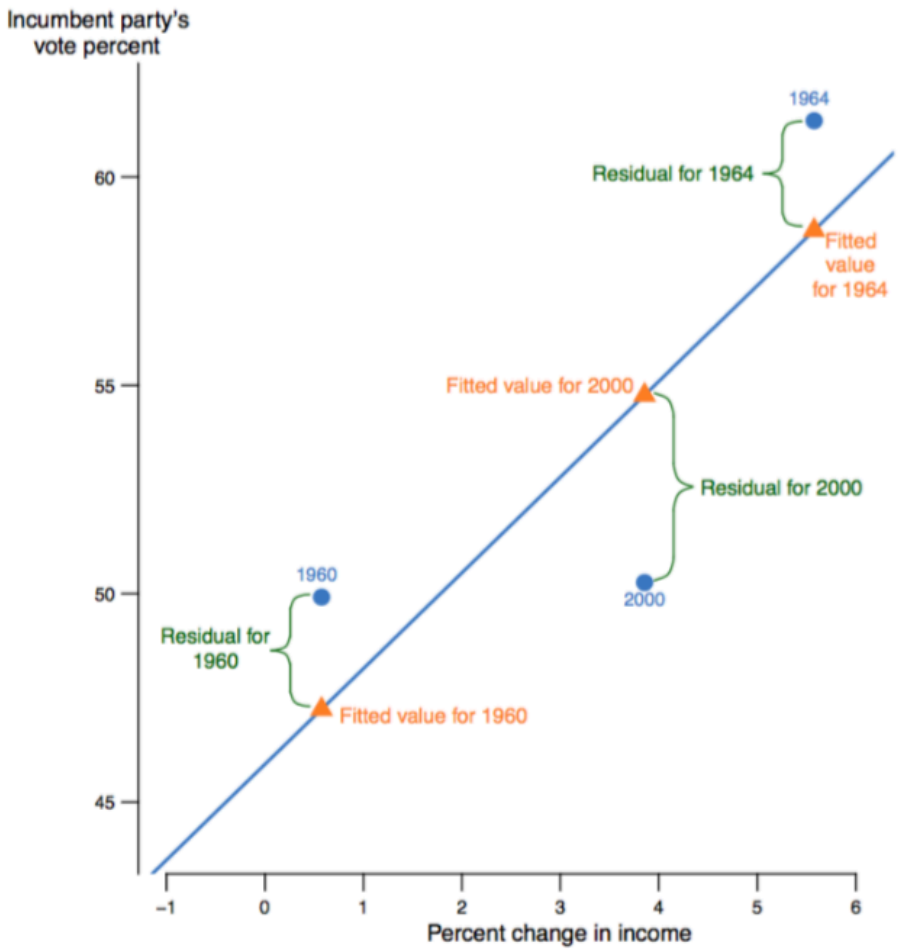
$$\sum_{i=1}^{N}(\hat{\epsilon}_i)^2 = \sum_{i=1}^{N}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

Through some calculus to carry out the minimization of SSR, we get the **OLS estimate of** $\hat{\beta}_1$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{N}(X_i - \bar{X})^2}$$

The **OLS estimate of** $\hat{\beta}_0$ is relatively easy once we have $\hat{\beta}_1$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$


Incumbent party's vote percent
60
55
50
45
1964
Residual for 1964
Fitted value for 1964
Fitted value for 2000
Residual for 2000
2000
1960
Residual for 1960
Fitted value for 1960
−1
0
1
2
3
4
5
6
Percent change in income

Calculating the line of best fit using regression in R

$$ols1 = lm(formula = vote \ \ rdi4, data = PresVote)$$
$$ols1\$residualsplot(vote \ \ rdi4, data = PresVote)$$
$$abline(ols1)$$

Alas, OLS is not a magic formula.

Challenge 1: Regression and bias

Endogeneity

1. Independent variable (X) is **endogenous** if correlated with error term in the model (ϵ) (mutual causality, reverse causality, etc.)
2. Independent variable is **exogenous** if it is not associated with factors captured in the error term.
3. Error term is *unobservable*, so hard to know if an independent variable X is endogenous or exogenous.
4. Difficult to *assess causality* for endogenous independent variables.

Bias in Statistics

Definition: A **biased** coefficient estimate will systematically be higher or lower than the true value.

Bias in Statistics

Definition: A **biased** coefficient estimate will systematically be higher or lower than the true value.

- An estimator is **unbiased** if the expected value equals the true value:

$$E[\hat{\beta}_1] = \beta_1 + \text{corr}(X, \epsilon)\frac{\sigma_\epsilon}{\sigma_X}$$

- $\hat{\beta}_1$ is unbiased if the error is uncorrelated with X – i.e., if X is **exogenous**

OLS & *UNBIASED* ESTIMATES

The distribution of an unbiased estimator is centered around the true value of the parameter. In the case of regression, this parameter is typically β.

OLS & *UNBIASED* ESTIMATES

The distribution of an unbiased estimator is centered around the true value of the parameter. In the case of regression, this parameter is typically β.

On one condition...

> The OLS estimator $\hat{\beta}_1$ is an unbiased estimator of β_1 **if** X and ϵ are not correlated.

Challenge 2: Inconsistency

3 FACTORS INFLUENCE THE ESTIMATED VARIANCE OF $\hat{\beta}_1$

1. Model fit, as represented by the variance of the regression, $\hat{\sigma}^2$.
2. Sample size. The more observations, the lower will be the variance of $\hat{\beta}_1$.
3. Variation in the independent variable, X. The more that X varies, the lower will be the variance of $\hat{\beta}_1$.

Let's look at goodness of fit

Definition: How well the model fits the data.

Let's look at goodness of fit

Definition: How well the model fits the data.

There are at least four ways to assess goodness of fit:

1. Variance of regression, $\hat{\sigma}^2$
2. Standard error of regression, $\hat{\sigma}$
3. Plotting model fit: residuals, fitted values, and diagnostics.
4. Square of correlation between fitted and observed values of Y: R^2

A bit on R^2

Formally, R^2 is calculated as $\text{corr}(\hat{Y}_i, Y_i)^2$.

(It can also be interpreted as the proportion of the variance in Y that is explained by the model.)

- Higher values of R^2 can sometimes be interpreted as indicating a "better fit".
- IMPORTANT: Statistically speaking, a high R^2 is **neither necessary nor sufficient** for an analysis to be useful.

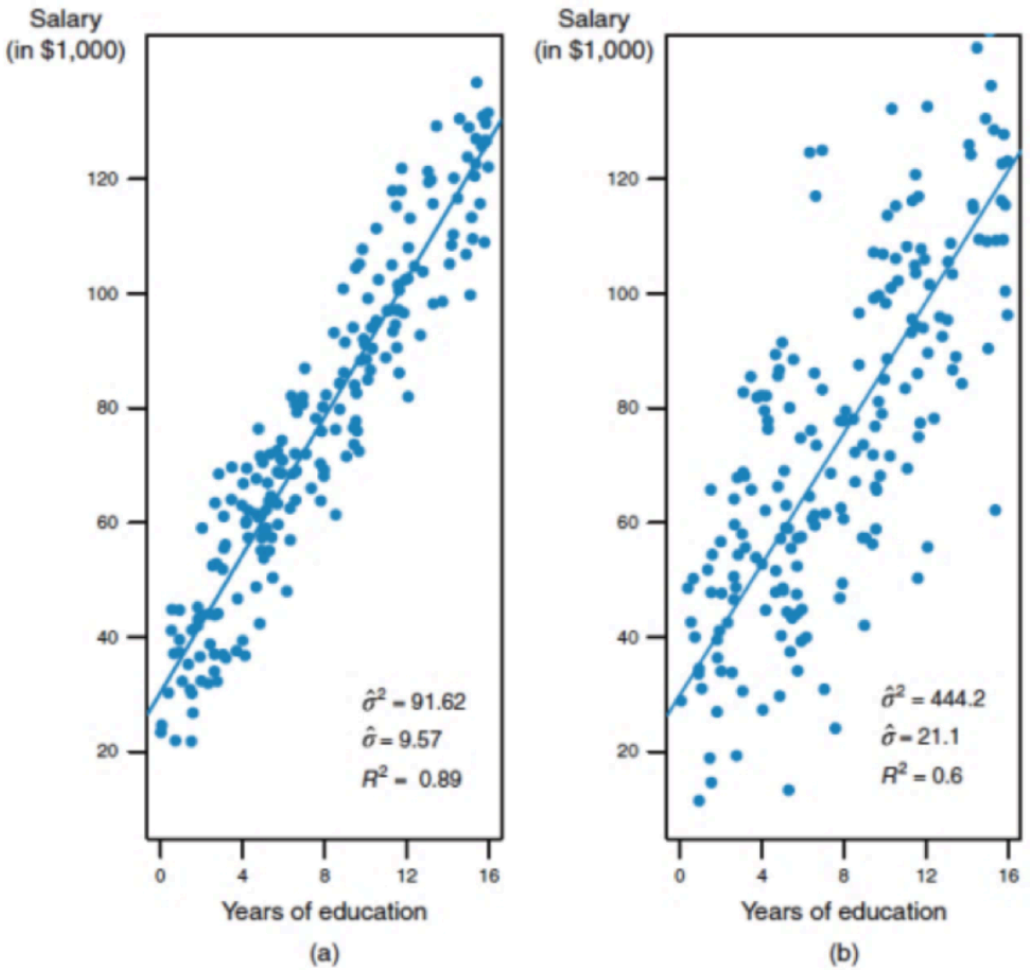
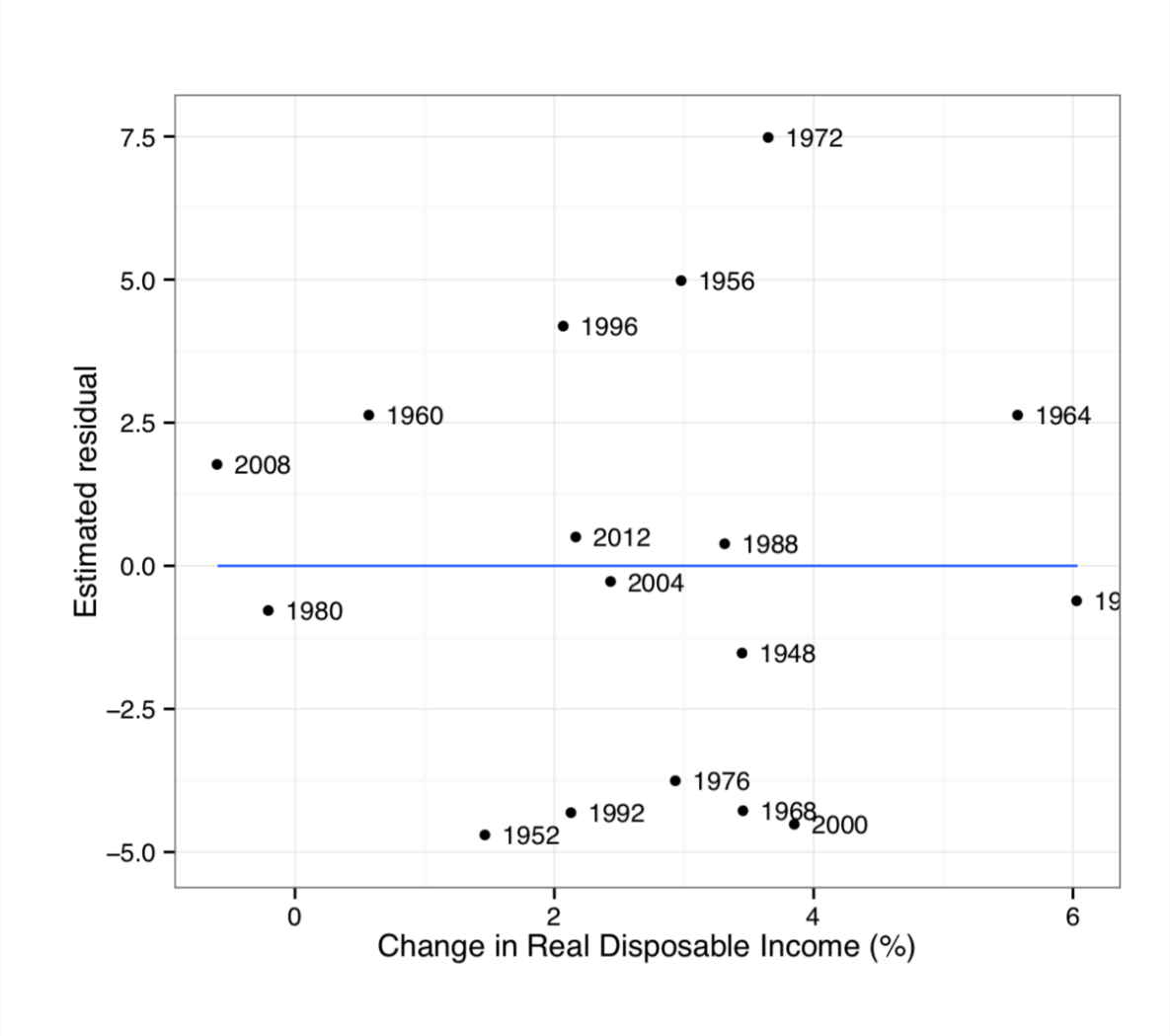
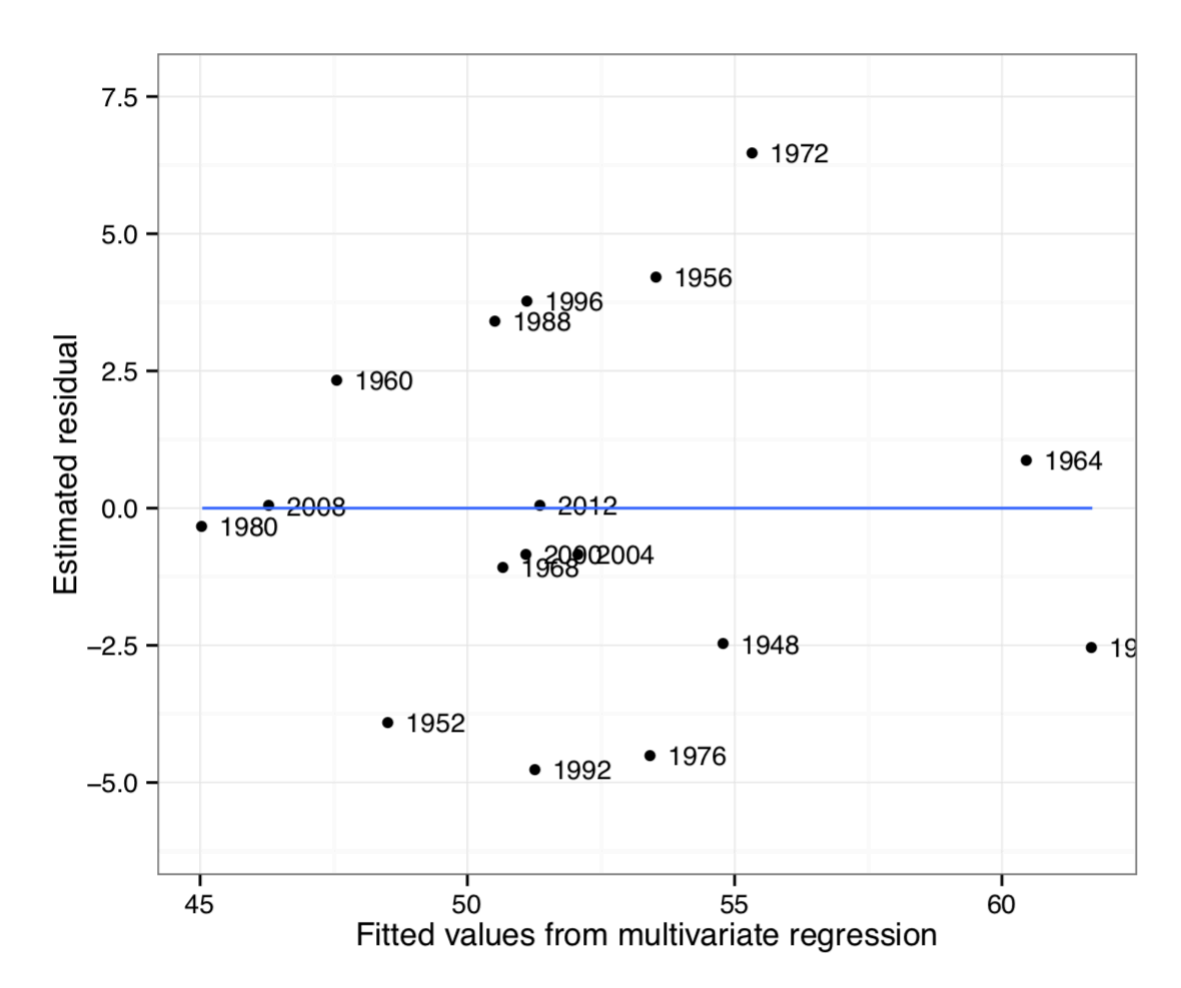


FIGURE 3.9: Plots with Different Goodness of Fit

Looking at the residuals vs. fitted plot for PresVote



What if we controlled for reelection?



What does this figure have to do with consistency?

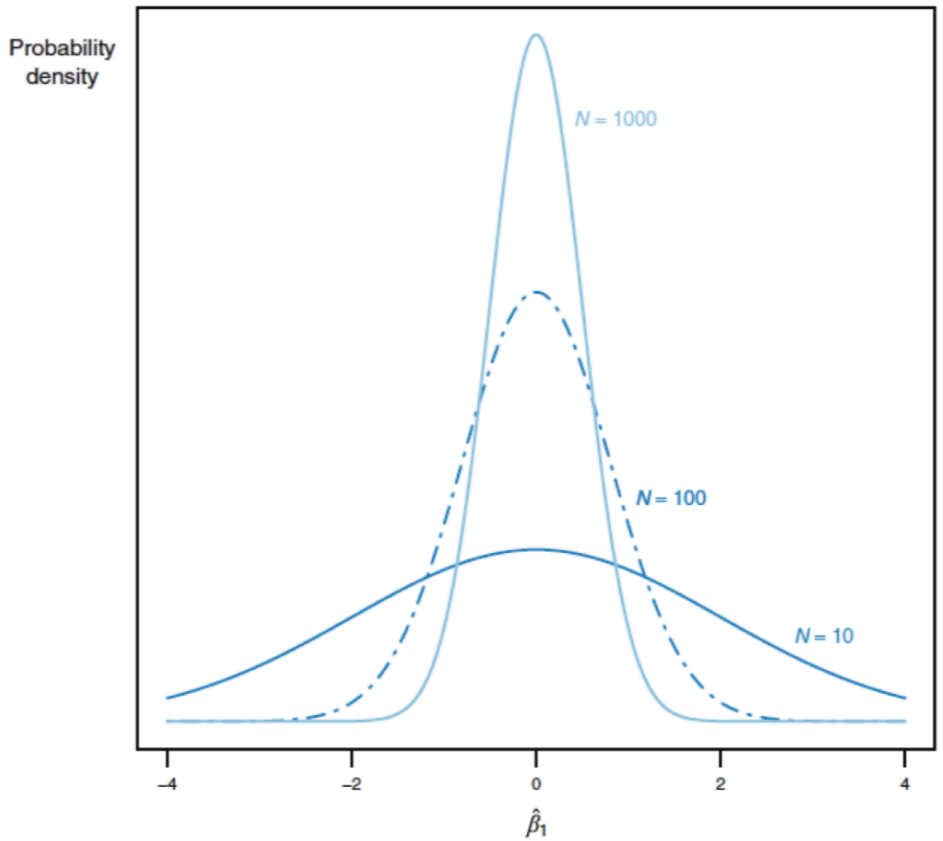


FIGURE 3.8: Distributions of $\hat{\beta}_1$ for Different Sample Sizes

Consistency and Probability Limits

PLIM

plim, or the probability limit, is the value to which a distribution converges as the sample size gets very large (how large is very large?)

Example: the coin toss

- What is the probability limit of a coin toss, if repeated enough times?
- How many times does it need to be repeated?

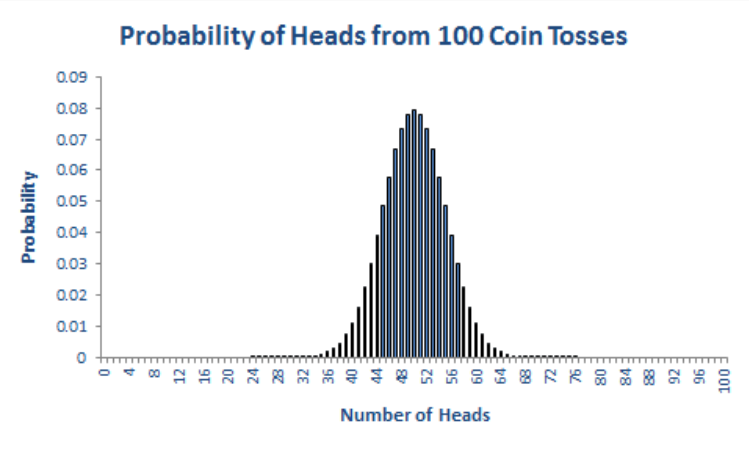
Example: the coin toss

- What is the probability limit of a coin toss, if repeated enough times?

  0.5

- How many times does it need to be repeated?

  >= 100



Challenge 3: Errors with errors

Overview: Rejecting the null hypothesis

Type I errors occur when we reject a null hypothesis that is in fact true. Most common.

Type II errors occur when we fail to reject a null hypothesis that is in fact false. Common with small sample sizes, or when we have a high threshold of statistical significance (e.g., $\alpha = 0.001$)

$\hat{\beta}_1$ and Standard Errors

One of our primary statistics for measuring the uncertainty of our estimates is $se(\hat{\beta}_1)$. This tells us how wide the distribution of $\hat{\beta}_1$ will be under the null hypothesis.

Table 4.2: Effect of Income Changes on Presidential Elections

Variable	Coefficient	Standard error
Change in income	2.29	0.52
Constant	45.91	1.69

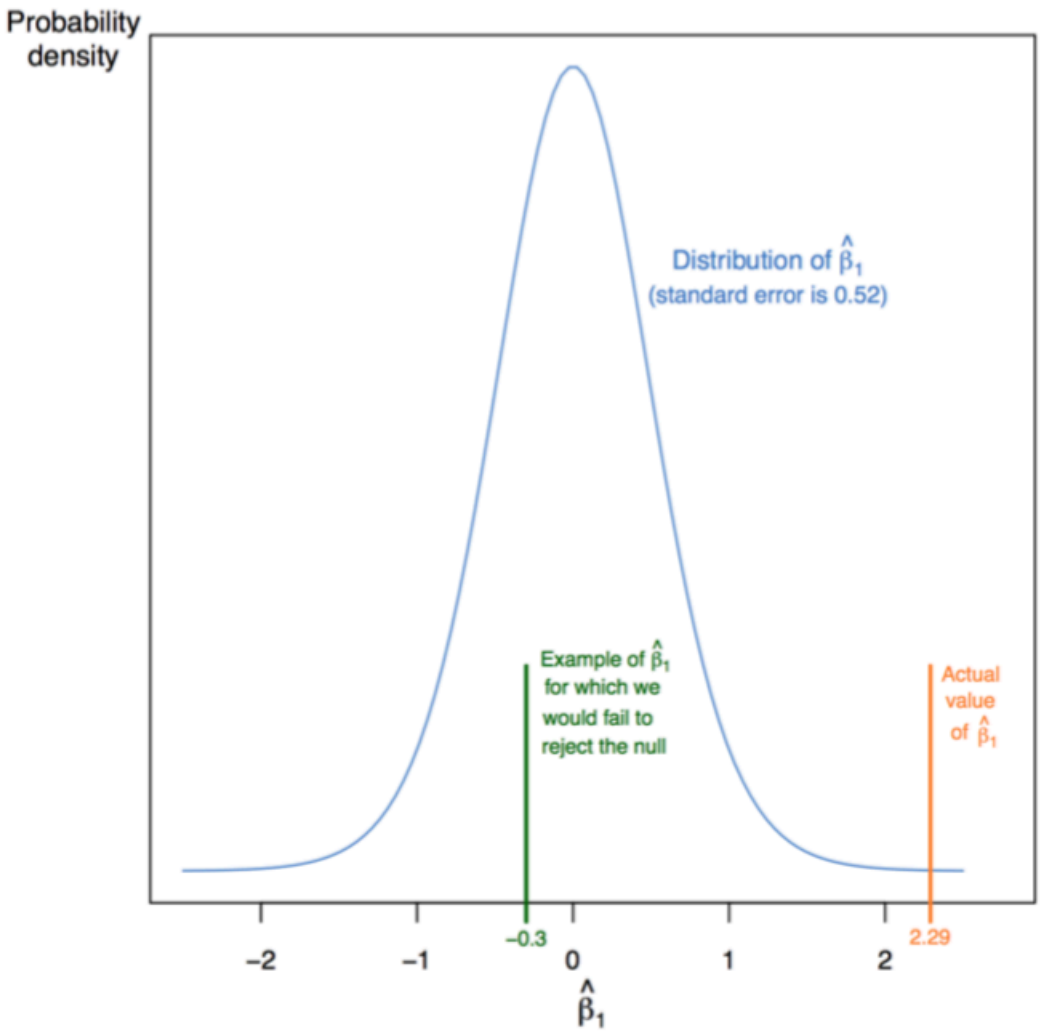
N = 17

Normal distribution of coefficient estimates

The distribution of $\hat{\beta}_1$ will be distributed normally if either:

1. Sample size is large. **Central Limit Theorem** tells us sufficient number of independent draws from any distribution will be normally distributed.
2. Errors are **normally distributed**. True even with a small sample size. *When might errors not be normally distributed?*

Distribution of $\hat{\beta}_1$ under the null hypothesis



$\hat{\beta}_1$ and Standard Errors

Our real concern isn't the standard error itself, but how it relates to the $\hat{\beta}_1$.

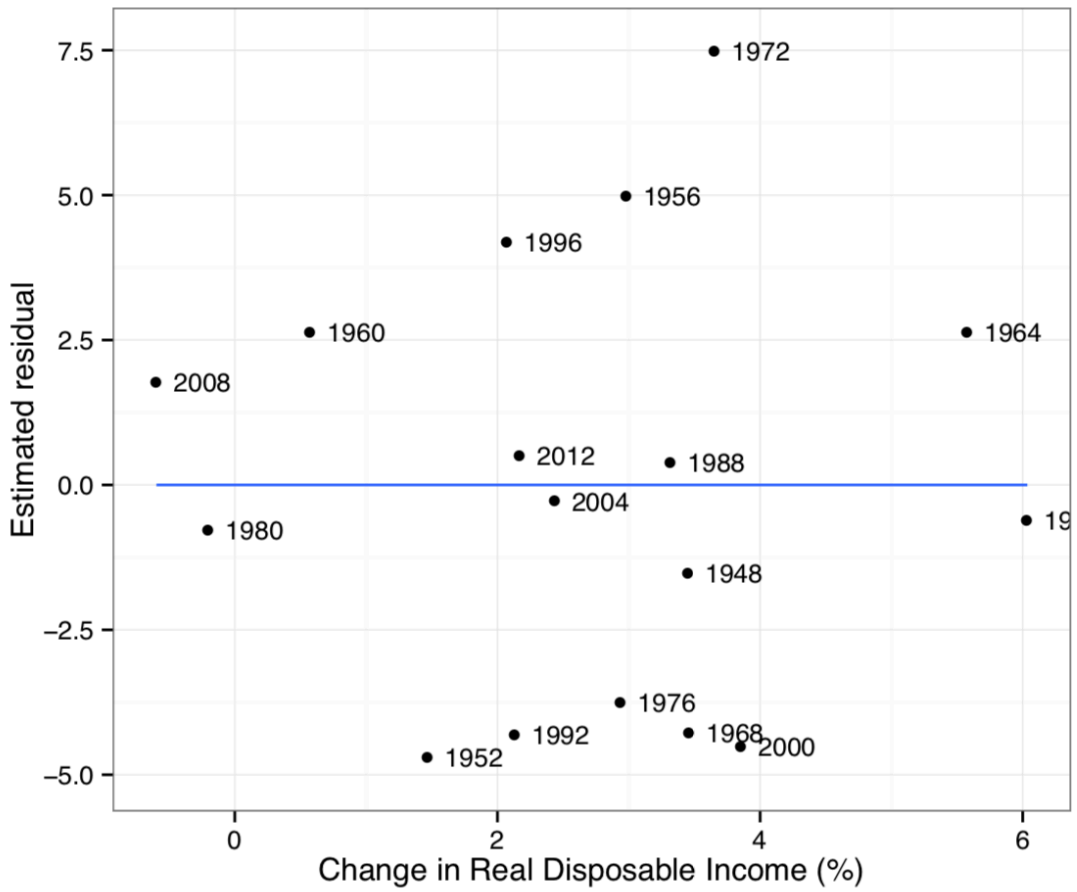
One way to characterize this is the ratio $\frac{\hat{\beta}_1}{se(\hat{\beta}_1)}$

- This ratio (the ***t*-statistic**) tells us how different our coefficient is from 0 in standard error units.
- When not dealing with $H_0 : \beta_1 = 0$, it tells us how far the coefficient is from the null hypothesis value of β

Say what...? 3 Definitions

1. Errors are **homoskedastic** if they have the same variance (this is good—it means errors are just random noise.
2. Errors are **heteroskedastic** if they have different variance (this is v bad).
3. Errors are **autocorrelated** if the error from one observation is correlated with the error of another (also v bad).

Homoskedasticity = errors evenly distributed



Homoskedasticity = errors evenly distributed

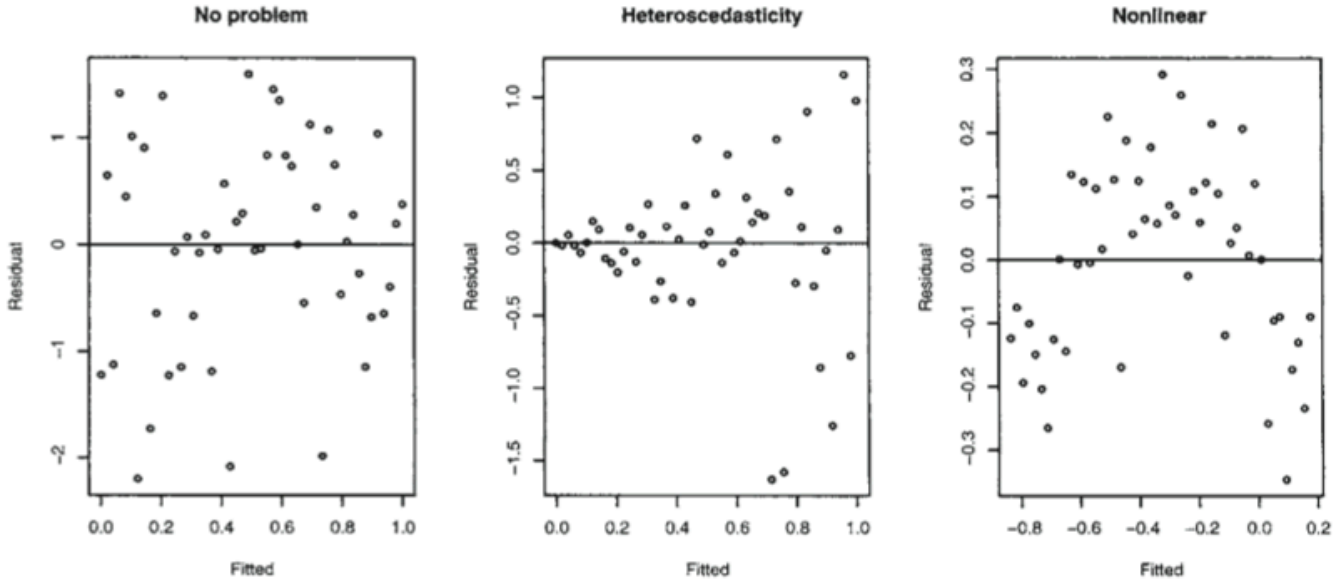


Figure 4.1 *Residuals vs. fitted plots—the first suggests no change to the current model while the second shows nonconstant variance and the third indicates some nonlinearity, which should prompt some change in the structural form of the model.*

One more issue: autocorrelation

If errors are **autocorrelated**, then knowing the error of observation 2 would provide information about the likely error of observation 3.

Checking in

$\Rightarrow$ Violating the homoskedasticity or no-autocorrelation assumptions/conditions does not cause $\hat{\beta}_1$ to be biased.

Why?

$\Rightarrow$ It also doesn't cause $\hat{\beta}_1$ to be inconsistent.

Why?

Bivariate description and regression in R after the break.