Week 6: Linear Regression with Two Regressors

Brandon Stewart[1]

Princeton

October 5–9, 2020

[1] These slides are heavily influenced by Matt Blackwell, Adam Glynn, Jens Hainmueller and Erin Hartman.

Where We've Been and Where We're Going...

Where We've Been and Where We're Going...

- Last Week
  - mechanics of OLS with one variable
  - properties of OLS

Where We've Been and Where We're Going...

- Last Week
  - mechanics of OLS with one variable
  - properties of OLS
- This Week
  - adding a second variable
  - new mechanics
  - omitted variable bias
  - multicollinearity
  - interactions

Where We've Been and Where We're Going...

- Last Week
  - mechanics of OLS with one variable
  - properties of OLS
- This Week
  - adding a second variable
  - new mechanics
  - omitted variable bias
  - multicollinearity
  - interactions
- Next Week
  - multiple regression
- Long Run
  - probability → inference → regression → causal inference

Why Do We Want More Than One Predictor?

Why Do We Want More Than One Predictor?

- Summarize more information for descriptive inference

Why Do We Want More Than One Predictor?

- Summarize more information for descriptive inference
- Improve the fit and predictive power of our model

Why Do We Want More Than One Predictor?

- Summarize more information for descriptive inference
- Improve the fit and predictive power of our model
- Control for confounding factors for causal inference

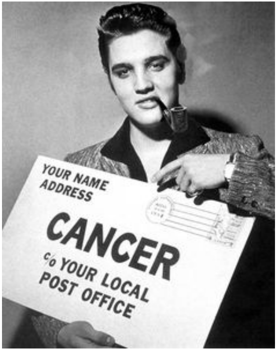
Why Do We Want More Than One Predictor?

- Summarize more information for descriptive inference
- Improve the fit and predictive power of our model
- Control for confounding factors for causal inference
- Model non-linearities (e.g. $Y = \beta_0 + \beta_1 X + \beta_2 X^2$)

Why Do We Want More Than One Predictor?

- Summarize more information for descriptive inference
- Improve the fit and predictive power of our model
- Control for confounding factors for causal inference
- Model non-linearities (e.g. $Y = \beta_0 + \beta_1 X + \beta_2 X^2$)
- Model interactive effects (e.g. $Y = \beta_0 + \beta_1 X + \beta_2 X_2 + \beta_3 X_1 X_2$)

Example 1: Cigarette Smokers and Pipe Smokers



Example 1: Cigarette Smokers and Pipe Smokers

Example 1: Cigarette Smokers and Pipe Smokers

Consider the following example from Cochran (1968). We have a random sample of 20,000 smokers and run a regression using:

- Y: Deaths per 1,000 Person-Years.

Example 1: Cigarette Smokers and Pipe Smokers

Consider the following example from Cochran (1968). We have a random sample of 20,000 smokers and run a regression using:

- Y: Deaths per 1,000 Person-Years.
- X_1: 0 if person is pipe smoker; 1 if person is cigarette smoker

Example 1: Cigarette Smokers and Pipe Smokers

Consider the following example from Cochran (1968). We have a random sample of 20,000 smokers and run a regression using:

- Y: Deaths per 1,000 Person-Years.
- X_1: 0 if person is pipe smoker; 1 if person is cigarette smoker

We fit the regression and find:

$$\widehat{\text{Death Rate}} = 17 - 4\,\text{Cigarette Smoker}$$

Example 1: Cigarette Smokers and Pipe Smokers

Consider the following example from Cochran (1968). We have a random sample of 20,000 smokers and run a regression using:

- Y: Deaths per 1,000 Person-Years.
- X_1: 0 if person is pipe smoker; 1 if person is cigarette smoker

We fit the regression and find:

$$\widehat{\text{Death Rate}} = 17 - 4\,\text{Cigarette Smoker}$$

What do we conclude?

Example 1: Cigarette Smokers and Pipe Smokers

Consider the following example from Cochran (1968). We have a random sample of 20,000 smokers and run a regression using:

- Y: Deaths per 1,000 Person-Years.
- X_1: 0 if person is pipe smoker; 1 if person is cigarette smoker

We fit the regression and find:

$$\widehat{\text{Death Rate}} = 17 - 4\,\text{Cigarette Smoker}$$

What do we conclude?

- The average death rate is 17 deaths per $1,000$ person-years for pipe smokers and 13 (17 - 4) for cigarette smokers.

Example 1: Cigarette Smokers and Pipe Smokers

Consider the following example from Cochran (1968). We have a random sample of 20,000 smokers and run a regression using:

- Y: Deaths per 1,000 Person-Years.
- X_1: 0 if person is pipe smoker; 1 if person is cigarette smoker

We fit the regression and find:

$$\widehat{\text{Death Rate}} = 17 - 4\,\text{Cigarette Smoker}$$

What do we conclude?

- The average death rate is 17 deaths per $1,000$ person-years for pipe smokers and 13 (17 - 4) for cigarette smokers.
- So cigarette smoking appears to lower the death rate by 4 deaths per 1,000 person years (relative to pipe smoking).

Example 1: Cigarette Smokers and Pipe Smokers

Consider the following example from Cochran (1968). We have a random sample of 20,000 smokers and run a regression using:

- Y: Deaths per 1,000 Person-Years.
- X_1: 0 if person is pipe smoker; 1 if person is cigarette smoker

We fit the regression and find:

$$\widehat{\text{Death Rate}} = 17 - 4\,\text{Cigarette Smoker}$$

What do we conclude?

- The average death rate is 17 deaths per $1,000$ person-years for pipe smokers and 13 (17 - 4) for cigarette smokers.
- So cigarette smoking appears to lower the death rate by 4 deaths per 1,000 person years (relative to pipe smoking).

When we "control" for age (in years) we find:

Example 1: Cigarette Smokers and Pipe Smokers

Consider the following example from Cochran (1968). We have a random sample of 20,000 smokers and run a regression using:

- Y: Deaths per 1,000 Person-Years.
- X_1: 0 if person is pipe smoker; 1 if person is cigarette smoker

We fit the regression and find:

$$\widehat{\text{Death Rate}} = 17 - 4\,\text{Cigarette Smoker}$$

What do we conclude?

- The average death rate is 17 deaths per $1,000$ person-years for pipe smokers and 13 (17 - 4) for cigarette smokers.
- So cigarette smoking appears to lower the death rate by 4 deaths per 1,000 person years (relative to pipe smoking).

When we "control" for age (in years) we find:

$$\widehat{\text{Death Rate}} = 14 + 4\,\text{Cigarette Smoker} + 10\,\text{Age}$$

Example 1: Cigarette Smokers and Pipe Smokers

Consider the following example from Cochran (1968). We have a random sample of 20,000 smokers and run a regression using:

- Y: Deaths per 1,000 Person-Years.
- X_1: 0 if person is pipe smoker; 1 if person is cigarette smoker

We fit the regression and find:

$$\widehat{\text{Death Rate}} = 17 - 4\,\text{Cigarette Smoker}$$

What do we conclude?

- The average death rate is 17 deaths per $1,000$ person-years for pipe smokers and 13 (17 - 4) for cigarette smokers.
- So cigarette smoking appears to lower the death rate by 4 deaths per 1,000 person years (relative to pipe smoking).

When we "control" for age (in years) we find:

$$\widehat{\text{Death Rate}} = 14 + 4\,\text{Cigarette Smoker} + 10\,\text{Age}$$

Why did the sign switch?

Example 1: Cigarette Smokers and Pipe Smokers

Consider the following example from Cochran (1968). We have a random sample of 20,000 smokers and run a regression using:

- Y: Deaths per 1,000 Person-Years.
- X_1: 0 if person is pipe smoker; 1 if person is cigarette smoker

We fit the regression and find:

$$\widehat{\text{Death Rate}} = 17 - 4\,\text{Cigarette Smoker}$$

What do we conclude?

- The average death rate is 17 deaths per $1,000$ person-years for pipe smokers and 13 (17 - 4) for cigarette smokers.
- So cigarette smoking appears to lower the death rate by 4 deaths per 1,000 person years (relative to pipe smoking).

When we "control" for age (in years) we find:

$$\widehat{\text{Death Rate}} = 14 + 4\,\text{Cigarette Smoker} + 10\,\text{Age}$$

Why did the sign switch? Which estimate is more useful?

Example 2: Berkeley Graduate Admissions

Example 2: Berkeley Graduate Admissions

Example 2: Berkeley Graduate Admissions

- Graduate admissions data from Berkeley, 1973

Example 2: Berkeley Graduate Admissions

- Graduate admissions data from Berkeley, 1973
- Acceptance rates:

Example 2: Berkeley Graduate Admissions

- Graduate admissions data from Berkeley, 1973
- Acceptance rates:
    - Men: 8442 applicants, 44% admission rate

Example 2: Berkeley Graduate Admissions

- Graduate admissions data from Berkeley, 1973
- Acceptance rates:
  - Men: 8442 applicants, 44% admission rate
  - Women: 4321 applicants, 35% admission rate

Example 2: Berkeley Graduate Admissions

- Graduate admissions data from Berkeley, 1973
- Acceptance rates:
  - Men: 8442 applicants, 44% admission rate
  - Women: 4321 applicants, 35% admission rate
- Evidence of discrimination toward women in admissions?

Example 2: Berkeley Graduate Admissions

- Graduate admissions data from Berkeley, 1973
- Acceptance rates:
  - Men: 8442 applicants, 44% admission rate
  - Women: 4321 applicants, 35% admission rate
- Evidence of discrimination toward women in admissions?
- This is a marginal relationship

Example 2: Berkeley Graduate Admissions

- Graduate admissions data from Berkeley, 1973
- Acceptance rates:
  - Men: 8442 applicants, 44% admission rate
  - Women: 4321 applicants, 35% admission rate
- Evidence of discrimination toward women in admissions?
- This is a marginal relationship
- What about the conditional relationship within departments?

Bias?

- Within departments:

	Men		Women	
Dept	Applied	Admitted	Applied	Admitted
A	825	62%	108	82%

Bias?

- Within departments:

	Men		Women	
Dept	Applied	Admitted	Applied	Admitted
A	825	62%	108	82%

Bias?

- Within departments:

	Men		Women	
Dept	Applied	Admitted	Applied	Admitted
A	825	62%	108	82%
B	560	63%	25	68%

Bias?

- Within departments:

	Men		Women	
Dept	Applied	Admitted	Applied	Admitted
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%

Bias?

- Within departments:

	Men		Women	
Dept	Applied	Admitted	Applied	Admitted
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	373	6%	341	7%

Bias?

- Within departments:

	Men		Women	
Dept	Applied	Admitted	Applied	Admitted
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	373	6%	341	7%

- Within departments, women do somewhat better than men!

Bias?

- Within departments:

	Men		Women	
Dept	Applied	Admitted	Applied	Admitted
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	373	6%	341	7%

- Within departments, women do somewhat better than men!
- How? Overall admission rates are lower for the departments women apply to.

Bias?

- Within departments:

	Men		Women	
Dept	Applied	Admitted	Applied	Admitted
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	373	6%	341	7%

- Within departments, women do somewhat better than men!
- How? Overall admission rates are lower for the departments women apply to.
- Marginal relationships (admissions and gender) $\neq$ conditional relationship given third variable (department)

Sex Bias in Graduate Admissions: Data from Berkeley

Measuring bias is harder than is usually assumed, and the evidence is sometimes contrary to expectation.

P. J. Bickel, E. A. Hammel, J. W. O'Connell

Determining whether discrimination because of sex or ethnic identity is being practiced against persons seeking passage from one social status or locus to another is an important problem in our society today. It is legally important and morally important. It is also often quite difficult. This article is an exploration of some of the issues of measurement and assessment involved in one example of the general problem, by means of which we hope to shed some light on the difficulties. We deceision to admit or to deny admission. The question we wish to pursue is whether the decision to admit or to deny was influenced by the sex of the applicant. We cannot know with any certainty the influences on the evaluators in the Graduate Admissions Office, or on the faculty reviewing committees, or on any other administrative personnel participating in the chain of actions that led to a decision on an individual application. We can, however, say that if the admissions decision and the sex by using a familiar statistic, chi-square. As already noted, we are aware of the pitfalls ahead in this naive approach, but we intend to stumble into every one of them for didactic reasons.

We must first make clear two assumptions that underlie consideration of the data in this contingency table approach. Assumption 1 is that in any given discipline male and female applicants do not differ in respect of their intelligence, skill, qualifications, promise, or other attribute deemed legitimately pertinent to their acceptance as students. It is precisely this assumption that makes the study of "sex bias" meaningful, for if we did not hold it any differences in acceptance of applicants by sex could be attributed to differences in their qualifications, promise as scholars, and so on. Theoretically one could test the assumption, for example, by examining presumably unbiased estimators of academic qualification such as Graduate Record Examination scores, undergraduate grade point averages, and so on. There are, however, enormous practical difficulties in this. We therefore predicate our discussion on the validity of assumption 1.

Bickel, Peter J., Eugene A. Hammel, and J. William O'Connell. "Sex bias in graduate admissions: Data from Berkeley." *Science* 187, no. 4175 (1975): 398-404.

Bias?

Bias?

'If prejudicial treatment is to be minimized, it must first be located accurately.

Bias?

'If prejudicial treatment is to be minimized, it must first be located accurately. We have shown that it is not characteristic of the graduate admissions process here examined...

Bias?

'If prejudicial treatment is to be minimized, it must first be located accurately. We have shown that it is not characteristic of the graduate admissions process here examined... The bias in the aggregated data stems not from any pattern of discrimination on the part of admissions committees, which seem quite fair on the whole,

Bias?

'If prejudicial treatment is to be minimized, it must first be located accurately. We have shown that it is not characteristic of the graduate admissions process here examined... The bias in the aggregated data stems not from any pattern of discrimination on the part of admissions committees, which seem quite fair on the whole,but apparently from prior screening at earlier levels of the educational system.

Bias?

'If prejudicial treatment is to be minimized, it must first be located accurately. We have shown that it is not characteristic of the graduate admissions process here examined. . . The bias in the aggregated data stems not from any pattern of discrimination on the part of admissions committees, which seem quite fair on the whole,but apparently from prior screening at earlier levels of the educational system. Women are shunted by their socialization and education toward fields of graduate study that are generally more crowded, less productive of completed degrees, and less well funded, and that frequently offer poorer professional employment prospects.' (Bickel et al 1975, 403, emphasis mine)

Bias?

'If prejudicial treatment is to be minimized, it must first be located accurately. We have shown that it is not characteristic of the graduate admissions process here examined... The bias in the aggregated data stems not from any pattern of discrimination on the part of admissions committees, which seem quite fair on the whole,but apparently from prior screening at earlier levels of the educational system. Women are shunted by their socialization and education toward fields of graduate study that are generally more crowded, less productive of completed degrees, and less well funded, and that frequently offer poorer professional employment prospects.' (Bickel et al 1975, 403, emphasis mine)

Bias? (a short digression)

Bias? (a short digression)

- Today we are covering the mechanics of how we get these results, but there is an important leap to their meaning for a particular policy argument.

Bias? (a short digression)

- Today we are covering the mechanics of how we get these results, but there is an important leap to their meaning for a particular policy argument.
- Bickel et al conclude that there is no evidence of bias at the admissions committee level.

Bias? (a short digression)

- Today we are covering the mechanics of how we get these results, but there is an important leap to their meaning for a particular policy argument.
- Bickel et al conclude that there is no evidence of bias at the admissions committee level.
- Key assumption: admits are equally qualified.

Bias? (a short digression)

- Today we are covering the mechanics of how we get these results, but there is an important leap to their meaning for a particular policy argument.
- Bickel et al conclude that there is no evidence of bias at the admissions committee level.
- Key assumption: admits are equally qualified.
- If the women are stronger admits (because e.g. a pattern of sexist behavior imposes a high barrier for women to even consider graduate school), we should expect them to be admitted at better than equal rates as men in a discrimination-free environment.

Bias? (a short digression)

- Today we are covering the mechanics of how we get these results, but there is an important leap to their meaning for a particular policy argument.
- Bickel et al conclude that there is no evidence of bias at the admissions committee level.
- Key assumption: admits are equally qualified.
- If the women are stronger admits (because e.g. a pattern of sexist behavior imposes a high barrier for women to even consider graduate school), we should expect them to be admitted at better than equal rates as men in a discrimination-free environment.
- Two general takeaways:

Bias? (a short digression)

- Today we are covering the mechanics of how we get these results, but there is an important leap to their meaning for a particular policy argument.
- Bickel et al conclude that there is no evidence of bias at the admissions committee level.
- Key assumption: admits are equally qualified.
- If the women are stronger admits (because e.g. a pattern of sexist behavior imposes a high barrier for women to even consider graduate school), we should expect them to be admitted at better than equal rates as men in a discrimination-free environment.
- Two general takeaways:
    1. interpreting results requires assumptions about the world

Bias? (a short digression)

- Today we are covering the mechanics of how we get these results, but there is an important leap to their meaning for a particular policy argument.
- Bickel et al conclude that there is no evidence of bias at the admissions committee level.
- Key assumption: admits are equally qualified.
- If the women are stronger admits (because e.g. a pattern of sexist behavior imposes a high barrier for women to even consider graduate school), we should expect them to be admitted at better than equal rates as men in a discrimination-free environment.
- Two general takeaways:
  1. interpreting results requires assumptions about the world
  2. the story of how people select into the group we are studying is important.

Bias? (a short digression)

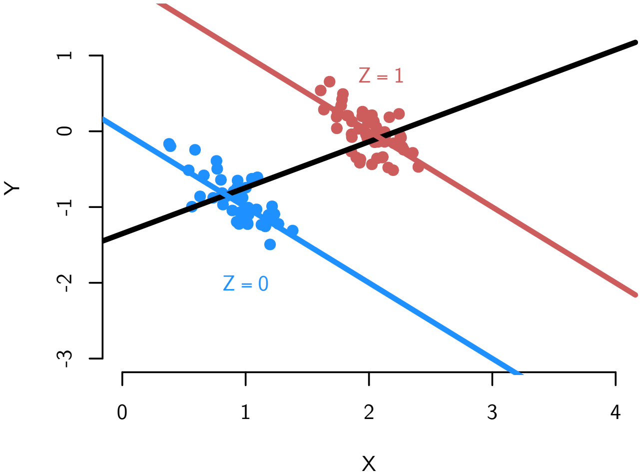
- Today we are covering the mechanics of how we get these results, but there is an important leap to their meaning for a particular policy argument.
- Bickel et al conclude that there is no evidence of bias at the admissions committee level.
- Key assumption: admits are equally qualified.
- If the women are stronger admits (because e.g. a pattern of sexist behavior imposes a high barrier for women to even consider graduate school), we should expect them to be admitted at better than equal rates as men in a discrimination-free environment.
- Two general takeaways:
  1. interpreting results requires assumptions about the world
  2. the story of how people select into the group we are studying is important.
- This general pattern repeats in many debates, often because of the limits of data collection.

Simpson's Paradox

The smoking and gender bias patterns are instances of Simpson's Paradox.

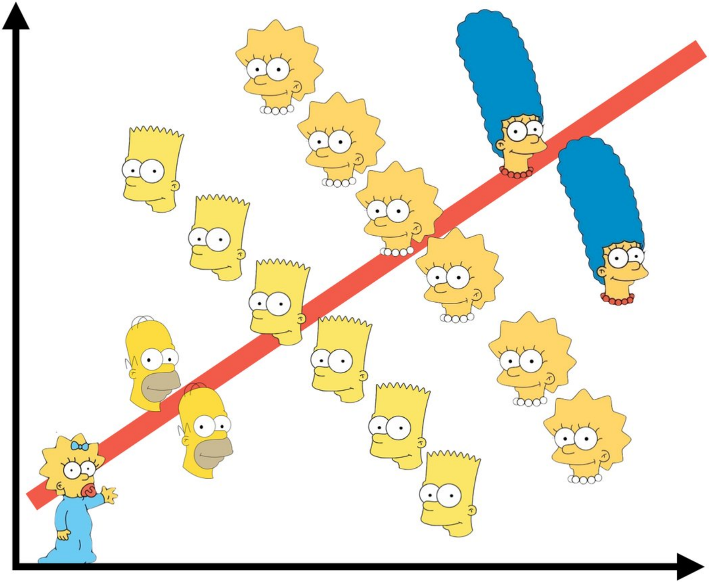
Simpson's Paradox

The smoking and gender bias patterns are instances of Simpson's Paradox.



Core idea: a relationship in one direction between Y_i and X_i but the opposite relationship within strata defined by Z_i.

Simpson's Paradox



Paradoxes Are Rarely A Paradox

Paradoxes Are Rarely A Paradox

- The paradox here is really just a case of imprecision of language.

Paradoxes Are Rarely A Paradox

- The paradox here is really just a case of imprecision of language.
- We observe the statement

$$P(\text{Death}|\text{Cigarette Smoking}) < P(\text{Death}|\text{Pipe Smoking})$$

and translate it to 'cigarette smoking lowers the death rate.'

Paradoxes Are Rarely A Paradox

- The paradox here is really just a case of imprecision of language.
- We observe the statement

$$P(\text{Death}|\text{Cigarette Smoking}) < P(\text{Death}|\text{Pipe Smoking})$$

  and translate it to 'cigarette smoking lowers the death rate.'
- This is in tension with then the subsequent finding

$$P(\text{Death}|\text{Cigarette Smoking}, \text{Age}) > P(\text{Death}|\text{Pipe Smoking}, \text{Age})$$

  which we translate to 'cigarette smoking increases the death rate for each age group.'

Paradoxes Are Rarely A Paradox

- The paradox here is really just a case of imprecision of language.
- We observe the statement

$$P(\text{Death}|\text{Cigarette Smoking}) < P(\text{Death}|\text{Pipe Smoking})$$

  and translate it to 'cigarette smoking lowers the death rate.'
- This is in tension with then the subsequent finding

$$P(\text{Death}|\text{Cigarette Smoking}, \text{Age}) > P(\text{Death}|\text{Pipe Smoking}, \text{Age})$$

  which we translate to 'cigarette smoking increases the death rate for each age group.'
- Both text translations cannot be true, but the math does not imply the causal interpretation given in the text.

Paradoxes Are Rarely A Paradox

- The paradox here is really just a case of imprecision of language.
- We observe the statement

$$P(\text{Death}|\text{Cigarette Smoking}) < P(\text{Death}|\text{Pipe Smoking})$$

  and translate it to 'cigarette smoking lowers the death rate.'
- This is in tension with then the subsequent finding

$$P(\text{Death}|\text{Cigarette Smoking}, \text{Age}) > P(\text{Death}|\text{Pipe Smoking}, \text{Age})$$

  which we translate to 'cigarette smoking increases the death rate for each age group.'
- Both text translations cannot be true, but the math does not imply the causal interpretation given in the text.
- Conditioning is just a way of looking at subgroups—we will see later that this plays a key role in making causal inferences but it requires careful assumptions.

Simpson's Paradox

Simpson's Paradox

- Simpson's paradox arises in many contexts- particularly where there is selection on ability

Simpson's Paradox

- Simpson's paradox arises in many contexts- particularly where there is selection on ability
- It is a particular problem in medical or demographic contexts, e.g. kidney stones, low-birth weight paradox.

Simpson's Paradox

- Simpson's paradox arises in many contexts- particularly where there is selection on ability
- It is a particular problem in medical or demographic contexts, e.g. kidney stones, low-birth weight paradox.
- It isn't clear that one version (the marginal or conditional) is necessarily the right way to examine the data. They just have different meanings.

Simpson's Paradox

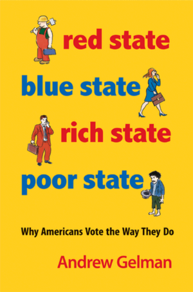
- Simpson's paradox arises in many contexts- particularly where there is selection on ability
- It is a particular problem in medical or demographic contexts, e.g. kidney stones, low-birth weight paradox.
- It isn't clear that one version (the marginal or conditional) is necessarily the right way to examine the data. They just have different meanings.
- This is often an issue of not being clear what we want.

Simpson's Paradox

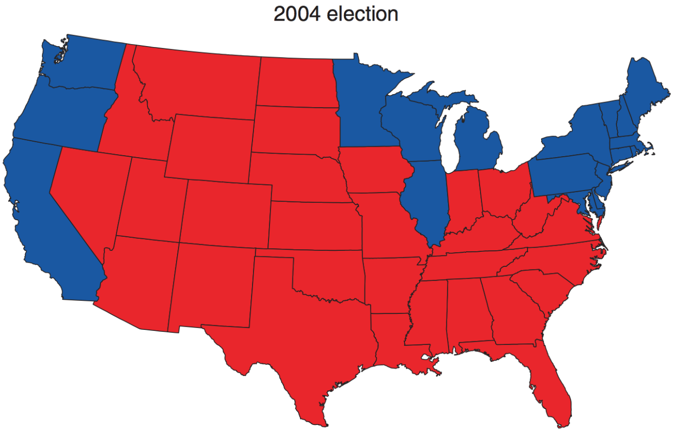
- Simpson's paradox arises in many contexts- particularly where there is selection on ability
- It is a particular problem in medical or demographic contexts, e.g. kidney stones, low-birth weight paradox.
- It isn't clear that one version (the marginal or conditional) is necessarily the right way to examine the data. They just have different meanings.
- This is often an issue of not being clear what we want.

Instance of a more general problem called the ecological inference fallacy.

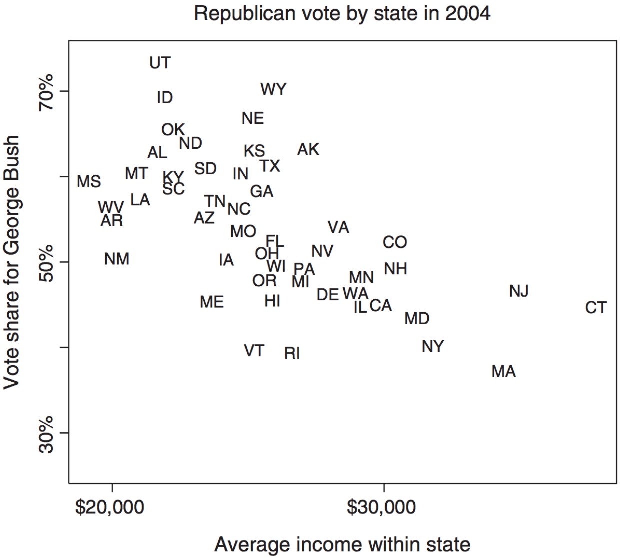
Red State Blue State



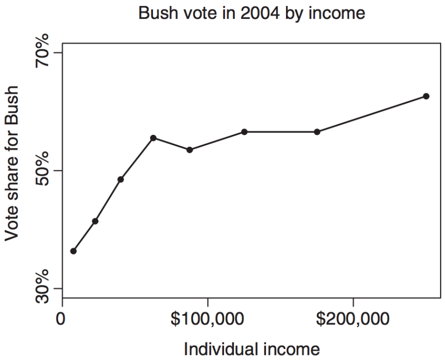
Red and Blue States

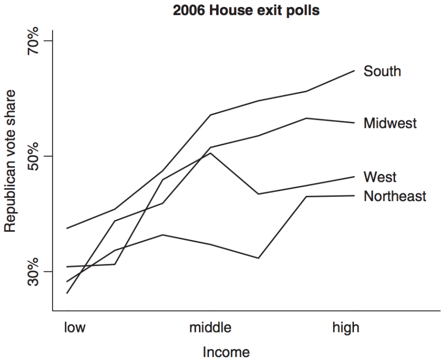


Rich States are More Democratic

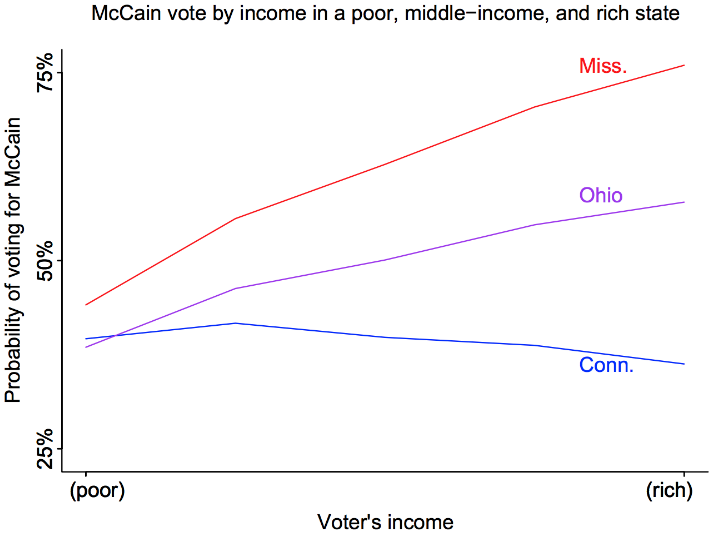


But Rich People are More Republican

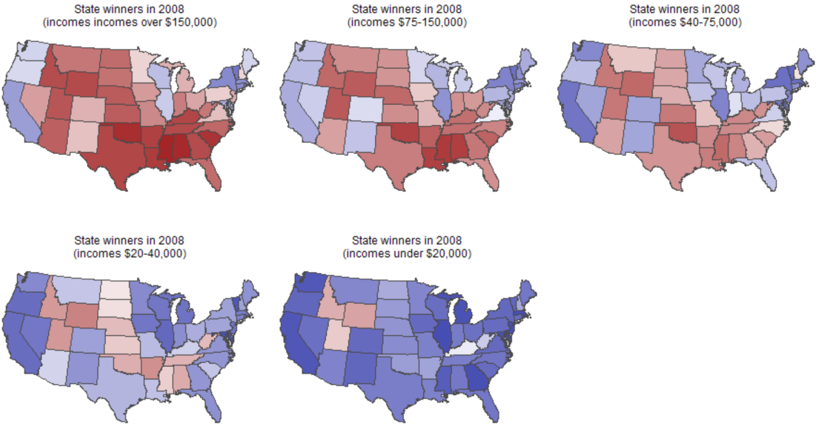




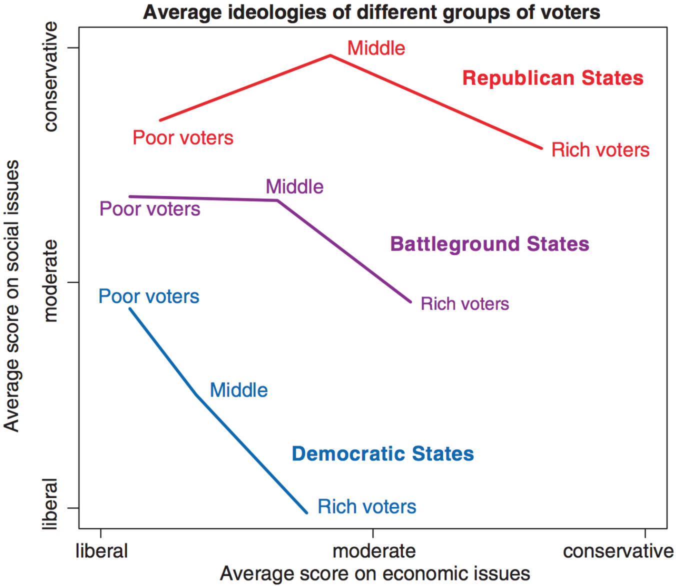
Paradox Resolved



If Only Rich People Voted, it Would Be a Landslide



A Possible Explanation



We Covered

We Covered

- Why controlling for a variable makes a difference

We Covered

- Why controlling for a variable makes a difference
- Simpson's paradox

We Covered

- Why controlling for a variable makes a difference
- Simpson's paradox

Next Time: How to Add a Variable

Where We've Been and Where We're Going...

Where We've Been and Where We're Going...

- Last Week
  - mechanics of OLS with one variable
  - properties of OLS
- This Week
  - adding a second variable
  - new mechanics
  - omitted variable bias
  - multicollinearity
  - interactions
- Next Week
  - multiple regression
- Long Run
  - probability $\rightarrow$ inference $\rightarrow$ regression $\rightarrow$ causal inference

Basic Idea of Two Variable Regressions

- Old goal: estimate the mean of Y as a function of one independent variable, X:

$$E[Y_i|X_i]$$

Basic Idea of Two Variable Regressions

- Old goal: estimate the mean of Y as a function of one independent variable, X:
$$E[Y_i|X_i]$$
- We modeled the CEF/regression function with a line:
$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

Basic Idea of Two Variable Regressions

- Old goal: estimate the mean of Y as a function of one independent variable, X:

$$E[Y_i|X_i]$$

- We modeled the CEF/regression function with a line:

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

- New goal: estimate the relationship of two variables, Y_i and X_i, conditional on a third variable, Z_i:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i$$

Basic Idea of Two Variable Regressions

- Old goal: estimate the mean of Y as a function of one independent variable, X:

$$E[Y_i|X_i]$$

- We modeled the CEF/regression function with a line:

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

- New goal: estimate the relationship of two variables, Y_i and X_i, conditional on a third variable, Z_i:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i$$

- β's are the population parameters we want to estimate

Regression with Two Explanatory Variables

Example: data from Fish (2002) "Islam and Authoritarianism." *World Politics.* 55: 4-37. Data from 157 countries.

Regression with Two Explanatory Variables

Example: data from Fish (2002) "Islam and Authoritarianism." *World Politics.* 55: 4-37. Data from 157 countries.

- Variables of interest:
  - Y: Level of democracy, measured as the 10-year average of Freedom House ratings

Regression with Two Explanatory Variables

Example: data from Fish (2002) "Islam and Authoritarianism." *World Politics.* 55: 4-37. Data from 157 countries.

- Variables of interest:
  - Y: Level of democracy, measured as the 10-year average of Freedom House ratings
  - X_1: Country income, measured as log(GDP per capita in $1000s)

Regression with Two Explanatory Variables

Example: data from Fish (2002) "Islam and Authoritarianism." *World Politics.* 55: 4-37. Data from 157 countries.

- Variables of interest:
  - Y: Level of democracy, measured as the 10-year average of Freedom House ratings
  - X_1: Country income, measured as log(GDP per capita in \$1000s)
  - X_2: Ethnic heterogeneity (continuous) or British colonial heritage (binary)

Regression with Two Explanatory Variables

Example: data from Fish (2002) "Islam and Authoritarianism." *World Politics.* 55: 4-37. Data from 157 countries.

- Variables of interest:
  - Y: Level of democracy, measured as the 10-year average of Freedom House ratings
  - X_1: Country income, measured as log(GDP per capita in \$1000s)
  - X_2: Ethnic heterogeneity (continuous) or British colonial heritage (binary)
- With one predictor we ask: Does income (X_1) predict or explain the level of democracy (Y)?

Regression with Two Explanatory Variables

Example: data from Fish (2002) "Islam and Authoritarianism." *World Politics.* 55: 4-37. Data from 157 countries.

- Variables of interest:
  - Y: Level of democracy, measured as the 10-year average of Freedom House ratings
  - X_1: Country income, measured as log(GDP per capita in \$1000s)
  - X_2: Ethnic heterogeneity (continuous) or British colonial heritage (binary)
- With one predictor we ask: Does income (X_1) predict or explain the level of democracy (Y)?

Regression with Two Explanatory Variables

Example: data from Fish (2002) “Islam and Authoritarianism.” *World Politics.* 55: 4-37. Data from 157 countries.

- Variables of interest:
  - Y: Level of democracy, measured as the 10-year average of Freedom House ratings
  - X_1: Country income, measured as log(GDP per capita in $1000s)
  - X_2: Ethnic heterogeneity (continuous) or British colonial heritage (binary)
- With one predictor we ask: Does income (X_1) predict or explain the level of democracy (Y)?
- With two predictors we ask questions like: Does income (X_1) predict or explain the level of democracy (Y), once we “control” for ethnic heterogeneity or British colonial heritage (X_2)?

Regression with Two Explanatory Variables

Example: data from Fish (2002) "Islam and Authoritarianism." *World Politics.* 55: 4-37. Data from 157 countries.

- Variables of interest:
  - Y: Level of democracy, measured as the 10-year average of Freedom House ratings
  - X_1: Country income, measured as log(GDP per capita in \$1000s)
  - X_2: Ethnic heterogeneity (continuous) or British colonial heritage (binary)
- With one predictor we ask: Does income (X_1) predict or explain the level of democracy (Y)?
- With two predictors we ask questions like: Does income (X_1) predict or explain the level of democracy (Y), once we "control" for ethnic heterogeneity or British colonial heritage (X_2)?
- The rest of this lecture is designed to explain what is meant by "controlling for another variable" with linear regression.

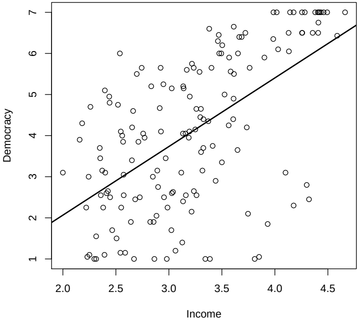
Simple Regression of Democracy on Income

Simple Regression of Democracy on Income

- Let's look at the bivariate regression of Democracy on Income:

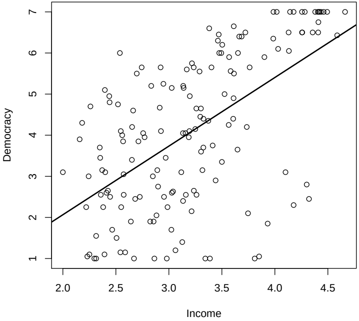
$$\widehat{y_i} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1$$

$$\widehat{Demo} = -1.26 + 1.6\, Log(GDP)$$



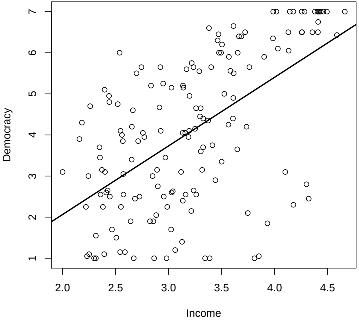
Simple Regression of Democracy on Income

- But we can use more information in our prediction equation.



Simple Regression of Democracy on Income

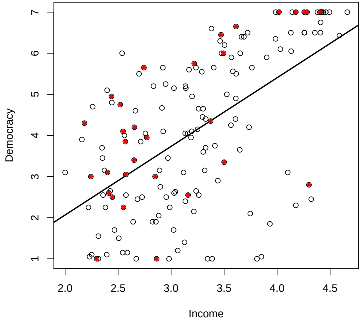
- But we can use more information in our prediction equation.
- For example, some countries were originally British colonies and others were not:



Simple Regression of Democracy on Income

- But we can use more information in our prediction equation.

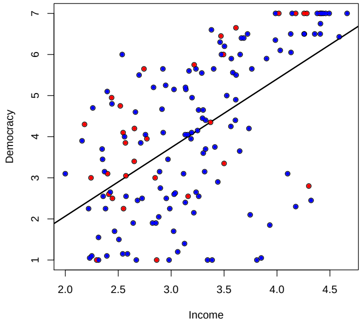
- For example, some countries were originally British colonies and others were not:
  - Former British colonies tend to have higher levels of democracy



Simple Regression of Democracy on Income

- But we can use more information in our prediction equation.

- For example, some countries were originally British colonies and others were not:
  - Former British colonies tend to have higher levels of democracy
  - Non-colony countries tend to have lower levels of democracy



Adding a Covariate

Adding a Covariate

How do we do this?

Adding a Covariate

How do we do this? We can generalize the prediction equation:

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{1i} + \widehat{\beta}_2 x_{2i}$$

This implies that we want to predict y using the information we have about x_1 and x_2, and we are assuming a linear functional form.

Adding a Covariate

How do we do this? We can generalize the prediction equation:

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{1i} + \widehat{\beta}_2 x_{2i}$$

This implies that we want to predict y using the information we have about x_1 and x_2, and we are assuming a linear functional form.

Notice that now we write X_{ji} where:

- $j = 1, ..., k$ is the index for the explanatory variables

Adding a Covariate

How do we do this? We can generalize the prediction equation:

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{1i} + \widehat{\beta}_2 x_{2i}$$

This implies that we want to predict y using the information we have about x_1 and x_2, and we are assuming a linear functional form.

Notice that now we write X_{ji} where:

- $j = 1, ..., k$ is the index for the explanatory variables
- $i = 1, ..., n$ is the index for the observation

Adding a Covariate

How do we do this? We can generalize the prediction equation:

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{1i} + \widehat{\beta}_2 x_{2i}$$

This implies that we want to predict y using the information we have about x_1 and x_2, and we are assuming a linear functional form.

Notice that now we write X_{ji} where:

- $j = 1, ..., k$ is the index for the explanatory variables
- $i = 1, ..., n$ is the index for the observation
- we often omit i to avoid clutter

Adding a Covariate

How do we do this? We can generalize the prediction equation:

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{1i} + \widehat{\beta}_2 x_{2i}$$

This implies that we want to predict y using the information we have about x_1 and x_2, and we are assuming a linear functional form.

Notice that now we write X_{ji} where:

- $j = 1, ..., k$ is the index for the explanatory variables
- $i = 1, ..., n$ is the index for the observation
- we often omit i to avoid clutter

In words:

$$\widehat{Democracy} = \widehat{\beta}_0 + \widehat{\beta}_1 \, Log(GDP) + \widehat{\beta}_2 \, Colony$$

Interpreting a Binary Covariate

Interpreting a Binary Covariate

Assume X_{2i} indicates whether country i used to be a British colony.

When $X_2 = 0$, the model becomes:

Interpreting a Binary Covariate

Assume X_{2i} indicates whether country i used to be a British colony.

When $X_2 = 0$, the model becomes:

$$\begin{aligned}
\widehat{y} &= \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2\, 0 \\
&= \widehat{\beta}_0 + \widehat{\beta}_1 x_1
\end{aligned}$$

Interpreting a Binary Covariate

Assume X_{2i} indicates whether country i used to be a British colony.

When $X_2 = 0$, the model becomes:

$$\begin{aligned}\widehat{y} &= \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2\, 0 \\ &= \widehat{\beta}_0 + \widehat{\beta}_1 x_1\end{aligned}$$

When $X_2 = 1$, the model becomes:

Interpreting a Binary Covariate

Assume X_{2i} indicates whether country i used to be a British colony.

When $X_2 = 0$, the model becomes:

$$\begin{aligned}\widehat{y} &= \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2\, 0 \\ &= \widehat{\beta}_0 + \widehat{\beta}_1 x_1\end{aligned}$$

When $X_2 = 1$, the model becomes:

$$\begin{aligned}\widehat{y} &= \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2\, 1 \\ &= (\widehat{\beta}_0 + \widehat{\beta}_2) + \widehat{\beta}_1 x_1\end{aligned}$$

What does this mean?

Interpreting a Binary Covariate

Assume X_{2i} indicates whether country i used to be a British colony.

When $X_2 = 0$, the model becomes:

$$\begin{aligned}\widehat{y} &= \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2\, 0 \\ &= \widehat{\beta}_0 + \widehat{\beta}_1 x_1\end{aligned}$$

When $X_2 = 1$, the model becomes:

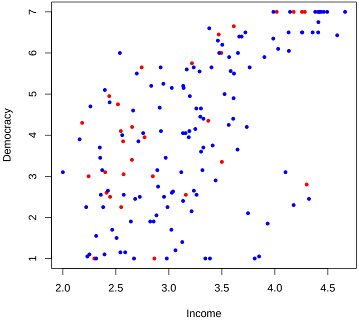
$$\begin{aligned}\widehat{y} &= \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2\, 1 \\ &= (\widehat{\beta}_0 + \widehat{\beta}_2) + \widehat{\beta}_1 x_1\end{aligned}$$

What does this mean? We are fitting two lines with the same slope but different intercepts.

Regression of Democracy on Income

From R, we obtain estimates $\widehat{\beta}_0$, $\widehat{\beta}_1$, $\widehat{\beta}_2$:

```
Coefficients:
             Estimate
(Intercept)   -1.5060
GDP90LGN       1.7059
BRITCOL        0.5881
```



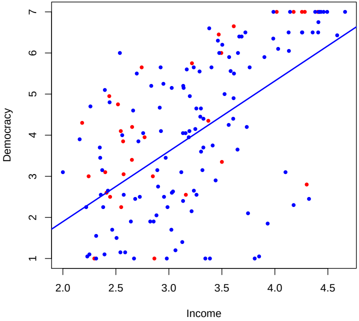
Regression of Democracy on Income

From R, we obtain estimates $\widehat{\beta}_0$, $\widehat{\beta}_1$, $\widehat{\beta}_2$:

```
Coefficients:
             Estimate
(Intercept)  -1.5060
GDP90LGN      1.7059
BRITCOL       0.5881
```

- Non-British colonies:

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1$$
$$\widehat{y} = -1.5 + 1.7\, x_1$$



Regression of Democracy on Income

From R, we obtain estimates
$\widehat{\beta}_0$, $\widehat{\beta}_1$, $\widehat{\beta}_2$:

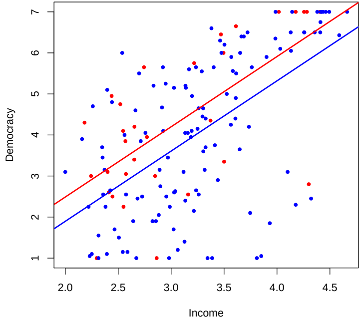
```
Coefficients:
             Estimate
(Intercept)   -1.5060
GDP90LGN       1.7059
BRITCOL        0.5881
```

- Non-British colonies:

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1$$
$$\widehat{y} = -1.5 + 1.7\, x_1$$

- Former British colonies:

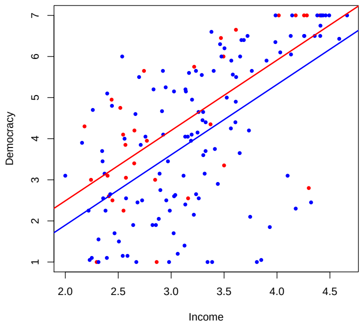
$$\widehat{y} = (\widehat{\beta}_0 + \widehat{\beta}_2) + \widehat{\beta}_1 x_1$$
$$\widehat{y} = -.92 + 1.7\, x_1$$



Regression of Democracy on Income

Our prediction equation is:
$\widehat{y} = -1.5 + 1.7\, x_1 + .58\, x_2$

Where do these quantities appear on the graph?

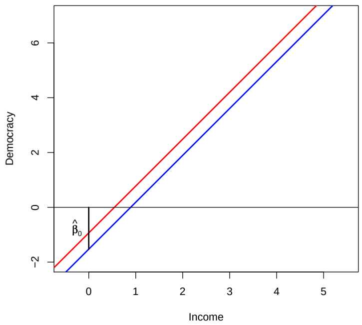


Regression of Democracy on Income

Our prediction equation is:
$\widehat{y} = -1.5 + 1.7\,x_1 + .58\,x_2$

Where do these quantities appear on the graph?

- $\widehat{\beta}_0 = -1.5$ is the intercept for the prediction line for non-British colonies.



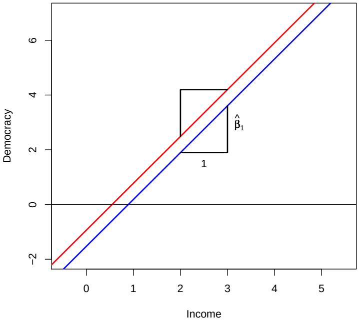
Regression of Democracy on Income

Our prediction equation is:
$\widehat{y} = -1.5 + 1.7\, x_1 + .58\, x_2$

Where do these quantities appear on the graph?

- $\widehat{\beta}_0 = -1.5$ is the intercept for the prediction line for non-British colonies.

- $\widehat{\beta}_1 = 1.7$ is the slope for both lines.

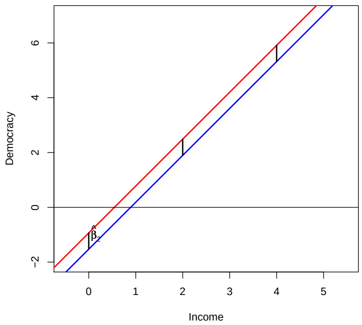


Regression of Democracy on Income

Our prediction equation is:
$\widehat{y} = -1.5 + 1.7\, x_1 + .58\, x_2$

Where do these quantities appear on the graph?

- $\widehat{\beta}_0 = -1.5$ is the intercept for the prediction line for non-British colonies.
- $\widehat{\beta}_1 = 1.7$ is the slope for both lines.
- $\widehat{\beta}_2 = .58$ is the vertical distance between the two lines for Ex-British colonies and non-colonies respectively



Example Interpretation of the Coefficients

- Let's review what we've seen so far:

	Intercept for X_1	Slope for X_1
Non-Colony ($X_2 = 0$)	$\widehat{\beta}_0$	$\widehat{\beta}_1$
Former Colony ($X_2 = 1$)	$\widehat{\beta}_0 + \widehat{\beta}_2$	$\widehat{\beta}_1$

Example Interpretation of the Coefficients

- Let's review what we've seen so far:

	Intercept for X_1	Slope for X_1
Non-Colony ($X_2 = 0$)	$\widehat{\beta}_0$	$\widehat{\beta}_1$
Former Colony ($X_2 = 1$)	$\widehat{\beta}_0 + \widehat{\beta}_2$	$\widehat{\beta}_1$

- In this example, we have:

$$\widehat{Y}_i = -1.5060 + 1.7059 \cdot X_1 + 0.5881 \cdot X_2$$

Example Interpretation of the Coefficients

- Let's review what we've seen so far:

	Intercept for X_1	Slope for X_1
Non-Colony ($X_2 = 0$)	$\widehat{\beta}_0$	$\widehat{\beta}_1$
Former Colony ($X_2 = 1$)	$\widehat{\beta}_0 + \widehat{\beta}_2$	$\widehat{\beta}_1$

- In this example, we have:

$$\widehat{Y}_i = -1.5060 + 1.7059 \cdot X_1 + 0.5881 \cdot X_2$$

- We can read these as:

Example Interpretation of the Coefficients

- Let's review what we've seen so far:

	Intercept for X_1	Slope for X_1
Non-Colony ($X_2 = 0$)	$\widehat{\beta}_0$	$\widehat{\beta}_1$
Former Colony ($X_2 = 1$)	$\widehat{\beta}_0 + \widehat{\beta}_2$	$\widehat{\beta}_1$

- In this example, we have:

$$\widehat{Y}_i = -1.5060 + 1.7059 \cdot X_1 + 0.5881 \cdot X_2$$

- We can read these as:
  - $\widehat{\beta}_0$: average democracy for non-British colony with log income of 0 is -1.5060 (note this is an extrapolation for this data!).

Example Interpretation of the Coefficients

- Let's review what we've seen so far:

	Intercept for X_1	Slope for X_1
Non-Colony ($X_2 = 0$)	$\widehat{\beta}_0$	$\widehat{\beta}_1$
Former Colony ($X_2 = 1$)	$\widehat{\beta}_0 + \widehat{\beta}_2$	$\widehat{\beta}_1$

- In this example, we have:

$$\widehat{Y}_i = -1.5060 + 1.7059 \cdot X_1 + 0.5881 \cdot X_2$$

- We can read these as:
  - $\widehat{\beta}_0$: average democracy for non-British colony with log income of 0 is -1.5060 (note this is an extrapolation for this data!).
  - $\widehat{\beta}_1$: countries with a one unit higher log income have on average a 1.7059 higher democracy score.

Example Interpretation of the Coefficients

- Let's review what we've seen so far:

	Intercept for X_1	Slope for X_1
Non-Colony ($X_2 = 0$)	$\widehat{\beta}_0$	$\widehat{\beta}_1$
Former Colony ($X_2 = 1$)	$\widehat{\beta}_0 + \widehat{\beta}_2$	$\widehat{\beta}_1$

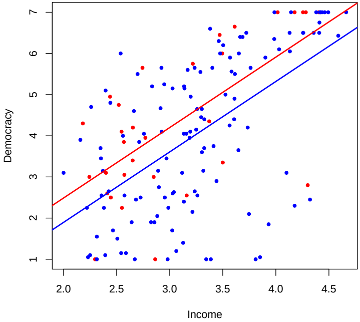
- In this example, we have:

$$\widehat{Y}_i = -1.5060 + 1.7059 \cdot X_1 + 0.5881 \cdot X_2$$

- We can read these as:
  - $\widehat{\beta}_0$: average democracy for non-British colony with log income of 0 is -1.5060 (note this is an extrapolation for this data!).
  - $\widehat{\beta}_1$: countries with a one unit higher log income have on average a 1.7059 higher democracy score.
  - $\widehat{\beta}_2$: former british colonies are predicted to have a 0.5881 higher average democracy score than non-british colonies with the same level of income.

Fitting a regression plane

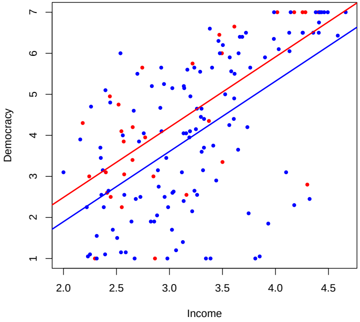
- We have considered an example of multiple regression with one continuous explanatory variable and one binary explanatory variable.



Fitting a regression plane

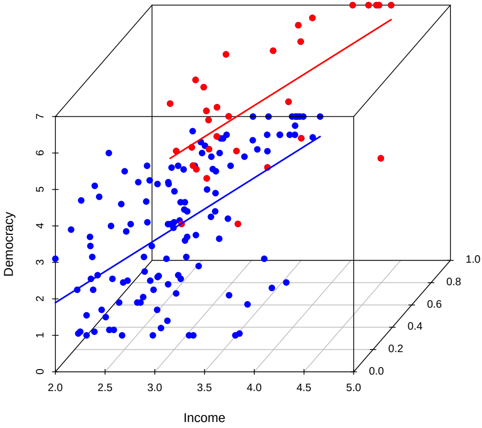
- We have considered an example of multiple regression with one continuous explanatory variable and one binary explanatory variable.

- This is easy to represent graphically in two dimensions because we can use colors to distinguish the two groups in the data.



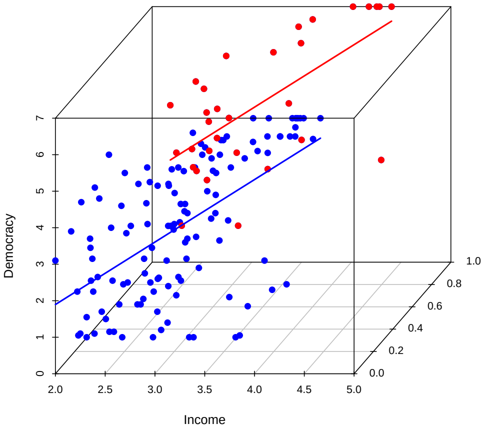
Regression of Democracy on Income

- These observations are actually located in a three-dimensional space.



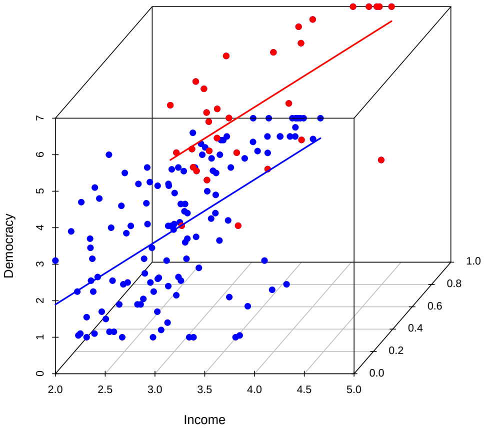
Regression of Democracy on Income

- These observations are actually located in a three-dimensional space.
- We can try to represent this using a 3D scatterplot.



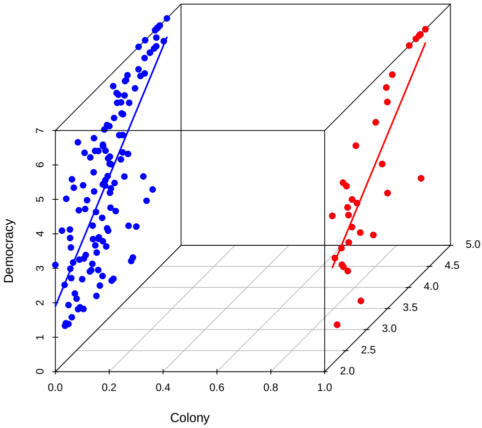
Regression of Democracy on Income

- These observations are actually located in a three-dimensional space.
- We can try to represent this using a 3D scatterplot.
- In this view, we are looking at the data from the Income side; the two regression lines are drawn in the appropriate locations.



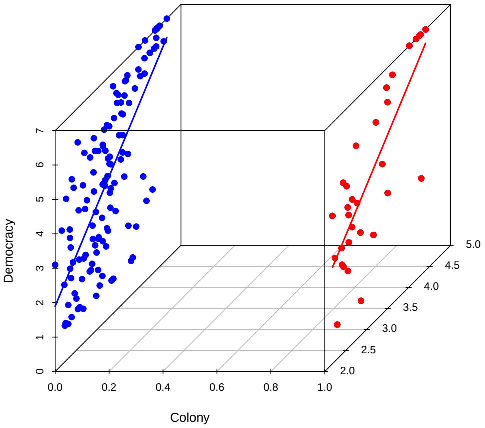
Regression of Democracy on Income

- We can also look at the 3D scatterplot from the British colony side.



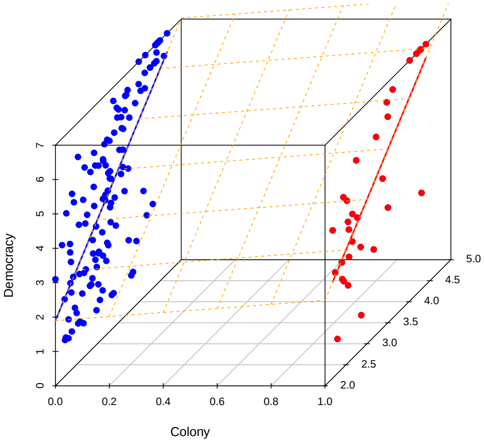
Regression of Democracy on Income

- We can also look at the 3D scatterplot from the British colony side.
- While the British colonial status variable is either 0 or 1, there is nothing in the prediction equation that requires this to be the case.



Regression of Democracy on Income

- We can also look at the 3D scatterplot from the British colony side.
- While the British colonial status variable is either 0 or 1, there is nothing in the prediction equation that requires this to be the case.
- In fact, the prediction equation defines a regression plane that connects the lines when $x_2 = 0$ and $x_2 = 1$.



Regression with two continuous variables

- Since we fit a regression plane to the data whenever we have two explanatory variables, it is easy to move to a case with two continuous explanatory variables.

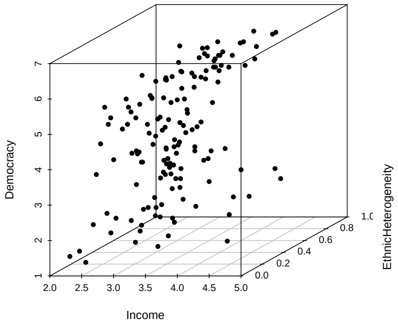
Regression with two continuous variables

- Since we fit a regression plane to the data whenever we have two explanatory variables, it is easy to move to a case with two continuous explanatory variables.
- For example, we might want to use:
  - X_1 Income and X_2 Ethnic Heterogeneity
  - Y Democracy

$$\widehat{\text{Democracy}} = \hat{\beta}_0 + \hat{\beta}_1 \text{Income} + \hat{\beta}_2 \text{Ethnic Heterogeneity}$$

Regression of Democracy on Income

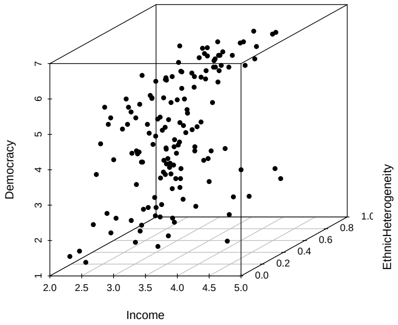
- We can plot the points in a 3D scatterplot.



Regression of Democracy on Income

- We can plot the points in a 3D scatterplot.
- R returns:
  - $\widehat{\beta}_0 = -.71$
  - $\widehat{\beta}_1 = 1.6$ for Income
  - $\widehat{\beta}_2 = -.6$ for Ethnic Heterogeneity

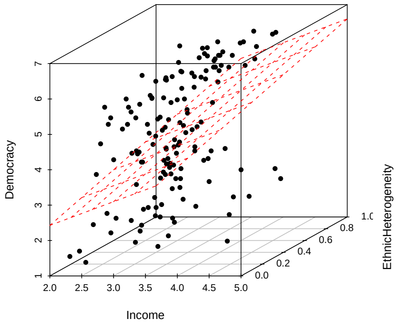
  How does this look graphically?



Regression of Democracy on Income

- We can plot the points in a 3D scatterplot.
- R returns:
  - $\widehat{\beta}_0 = -.71$
  - $\widehat{\beta}_1 = 1.6$ for Income
  - $\widehat{\beta}_2 = -.6$ for Ethnic Heterogeneity

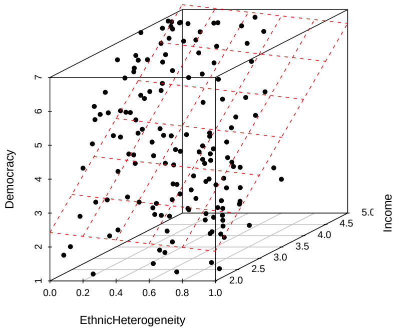
  How does this look graphically?
- These estimates define a regression plane through the data.



Regression of Democracy on Income

- We can plot the points in a 3D scatterplot.
- R returns:
  - $\widehat{\beta}_0 = -.71$
  - $\widehat{\beta}_1 = 1.6$ for Income
  - $\widehat{\beta}_2 = -.6$ for Ethnic Heterogeneity

  How does this look graphically?
- These estimates define a regression plane through the data.



Interpreting a Continuous Covariate

- The coefficient estimates have a similar interpretation in this case as they did in the Income-British Colony example.

Interpreting a Continuous Covariate

- The coefficient estimates have a similar interpretation in this case as they did in the Income-British Colony example.
- For example, $\widehat{\beta}_1 = 1.6$ represents our prediction of the difference in Democracy between two observations that differ by one unit of Income but have the same value of Ethnic Heterogeneity.

Interpreting a Continuous Covariate

- The coefficient estimates have a similar interpretation in this case as they did in the Income-British Colony example.
- For example, $\widehat{\beta}_1 = 1.6$ represents our prediction of the difference in Democracy between two observations that differ by one unit of Income but have the same value of Ethnic Heterogeneity.
- The slope estimates have an interpretation in terms of the partial derivative:

$$\frac{\partial(y = \beta_0 + \beta_1 X_1 + \beta_2 X_2)}{\partial X_1} =$$

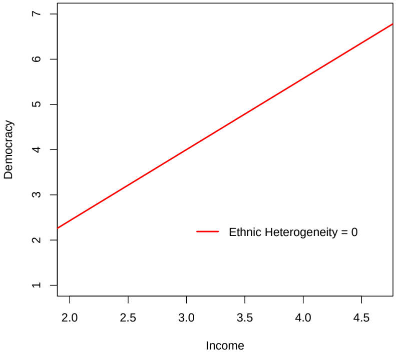
Interpreting a Continuous Covariate

- The coefficient estimates have a similar interpretation in this case as they did in the Income-British Colony example.
- For example, $\widehat{\beta}_1 = 1.6$ represents our prediction of the difference in Democracy between two observations that differ by one unit of Income but have the same value of Ethnic Heterogeneity.
- The slope estimates have an interpretation in terms of the partial derivative:

$$\frac{\partial(y = \beta_0 + \beta_1 X_1 + \beta_2 X_2)}{\partial X_1} = \beta_1$$

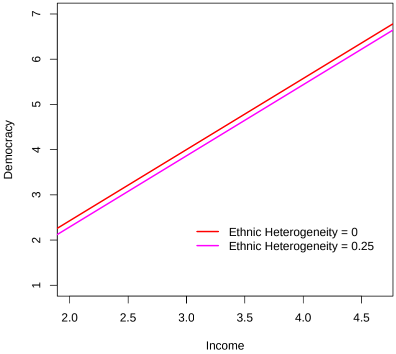
Interpreting a Continuous Covariate

- Again, we can think of this as defining a regression line for the relationship between Democracy and Income at every level of Ethnic Heterogeneity.



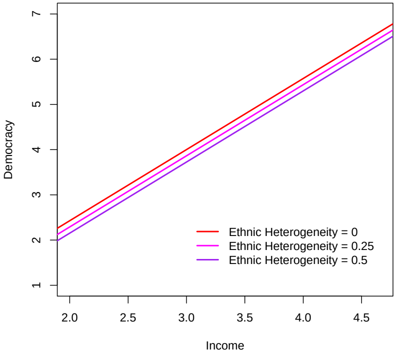
Interpreting a Continuous Covariate

- Again, we can think of this as defining a regression line for the relationship between Democracy and Income at every level of Ethnic Heterogeneity.
- All of these lines are parallel since they have the slope $\widehat{\beta}_1 = 1.6$



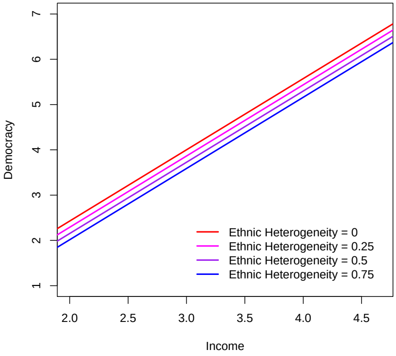
Interpreting a Continuous Covariate

- Again, we can think of this as defining a regression line for the relationship between Democracy and Income at every level of Ethnic Heterogeneity.
- All of these lines are parallel since they have the slope $\widehat{\beta}_1 = 1.6$
- The lines shift up or down based on the value of Ethnic Heterogeneity.



Interpreting a Continuous Covariate

- Again, we can think of this as defining a regression line for the relationship between Democracy and Income at every level of Ethnic Heterogeneity.
- All of these lines are parallel since they have the slope $\widehat{\beta}_1 = 1.6$
- The lines shift up or down based on the value of Ethnic Heterogeneity.



More Complex Predictions

- We can also use the coefficient estimates for more complex predictions that involve changing multiple variables simultaneously.

More Complex Predictions

- We can also use the coefficient estimates for more complex predictions that involve changing multiple variables simultaneously.
- Consider our results for the regression of democracy on X_1 income and X_2 ethnic heterogeneity:
  - $\widehat{\beta}_0 = -.71$
  - $\widehat{\beta}_1 = 1.6$
  - $\widehat{\beta}_2 = -.6$

More Complex Predictions

- We can also use the coefficient estimates for more complex predictions that involve changing multiple variables simultaneously.
- Consider our results for the regression of democracy on X_1 income and X_2 ethnic heterogeneity:
  - $\widehat{\beta}_0 = -.71$
  - $\widehat{\beta}_1 = 1.6$
  - $\widehat{\beta}_2 = -.6$
- What is the predicted difference in democracy between
  - Chile with $X_1 = 3.5$ and $X_2 = .06$?
  - China with $X_1 = 2.5$ and $X_2 = .5$?

More Complex Predictions

- We can also use the coefficient estimates for more complex predictions that involve changing multiple variables simultaneously.
- Consider our results for the regression of democracy on X_1 income and X_2 ethnic heterogeneity:
  - $\widehat{\beta}_0 = -.71$
  - $\widehat{\beta}_1 = 1.6$
  - $\widehat{\beta}_2 = -.6$
- What is the predicted difference in democracy between
  - Chile with $X_1 = 3.5$ and $X_2 = .06$?
  - China with $X_1 = 2.5$ and $X_2 = .5$?
- Predicted democracy is
  - $-.71 + 1.6 \cdot 3.5 - .6 \cdot .06 = 4.8$ for Chile

More Complex Predictions

- We can also use the coefficient estimates for more complex predictions that involve changing multiple variables simultaneously.
- Consider our results for the regression of democracy on X_1 income and X_2 ethnic heterogeneity:
  - $\widehat{\beta}_0 = -.71$
  - $\widehat{\beta}_1 = 1.6$
  - $\widehat{\beta}_2 = -.6$
- What is the predicted difference in democracy between
  - Chile with $X_1 = 3.5$ and $X_2 = .06$?
  - China with $X_1 = 2.5$ and $X_2 = .5$?
- Predicted democracy is
  - $-.71 + 1.6 \cdot 3.5 - .6 \cdot .06 = 4.8$ for Chile
  - $-.71 + 1.6 \cdot 2.5 - .6 \cdot 0.5 = 3$ for China.

More Complex Predictions

- We can also use the coefficient estimates for more complex predictions that involve changing multiple variables simultaneously.
- Consider our results for the regression of democracy on X_1 income and X_2 ethnic heterogeneity:
  - $\widehat{\beta}_0 = -.71$
  - $\widehat{\beta}_1 = 1.6$
  - $\widehat{\beta}_2 = -.6$
- What is the predicted difference in democracy between
  - Chile with $X_1 = 3.5$ and $X_2 = .06$?
  - China with $X_1 = 2.5$ and $X_2 = .5$?
- Predicted democracy is
  - $-.71 + 1.6 \cdot 3.5 - .6 \cdot .06 = 4.8$ for Chile
  - $-.71 + 1.6 \cdot 2.5 - .6 \cdot 0.5 = 3$ for China.

  Predicted difference is thus: 1.8 or $(3.5 - 2.5)\widehat{\beta}_1 + (.06 - .5)\widehat{\beta}_2$

Dummy Variables

Dummy Variables

- A dummy variable (a.k.a. indicator variable, binary variable, etc.) is a variable that is coded 1 or 0 only.

Dummy Variables

- A dummy variable (a.k.a. indicator variable, binary variable, etc.) is a variable that is coded 1 or 0 only.
- We use dummy variables in regression to represent qualitative information through categorical variables such as different subgroups of the sample (e.g. regions, old and young respondents, etc.)

Dummy Variables

- A dummy variable (a.k.a. indicator variable, binary variable, etc.) is a variable that is coded 1 or 0 only.
- We use dummy variables in regression to represent qualitative information through categorical variables such as different subgroups of the sample (e.g. regions, old and young respondents, etc.)
- By including dummy variables into our regression function, we can easily obtain the conditional mean of the outcome variable for each category.

Dummy Variables

- A dummy variable (a.k.a. indicator variable, binary variable, etc.) is a variable that is coded 1 or 0 only.
- We use dummy variables in regression to represent qualitative information through categorical variables such as different subgroups of the sample (e.g. regions, old and young respondents, etc.)
- By including dummy variables into our regression function, we can easily obtain the conditional mean of the outcome variable for each category.
- Dummy variables are also used to examine conditional hypothesis via interaction terms (more in a few videos).

Dummy Variables

- A dummy variable (a.k.a. indicator variable, binary variable, etc.) is a variable that is coded 1 or 0 only.
- We use dummy variables in regression to represent qualitative information through categorical variables such as different subgroups of the sample (e.g. regions, old and young respondents, etc.)
- By including dummy variables into our regression function, we can easily obtain the conditional mean of the outcome variable for each category.
- Dummy variables are also used to examine conditional hypothesis via interaction terms (more in a few videos).
- NB: if you want to sound like a machine learning person you can call it a one-hot encoding.

How Can I Use a Dummy Variable?

- Consider the easiest case with two categories. The type of electoral system of country i is given by:

  $X_i \in \{Proportional, Majoritarian\}$

How Can I Use a Dummy Variable?

- Consider the easiest case with two categories. The type of electoral system of country i is given by:

  $X_i \in \{Proportional, Majoritarian\}$

- For this we use a single dummy variable which is coded like:

$$D_i = \begin{cases} 1 & \text{if country } i \text{ has a Majoritarian Electoral System} \\ 0 & \text{if country } i \text{ has a Proportional Electoral System} \end{cases}$$

Dummy Variables for Multiple Categories

- More generally, let's say X measures which of m categories each unit i belongs to. E.g. the type of electoral system or region of country i is given by:
  - $X_i \in \{Proportional, Majoritarian\}$ so $m = 2$
  - $X_i \in \{Asia, Africa, LatinAmerica, OECD, Transition\}$ so $m = 5$

Dummy Variables for Multiple Categories

- More generally, let's say X measures which of m categories each unit i belongs to. E.g. the type of electoral system or region of country i is given by:
  - $X_i \in \{Proportional, Majoritarian\}$ so $m = 2$
  - $X_i \in \{Asia, Africa, LatinAmerica, OECD, Transition\}$ so $m = 5$
- To incorporate this information into our regression function we usually create $m - 1$ dummy variables, one for each of the $m - 1$ categories.

Dummy Variables for Multiple Categories

- More generally, let's say X measures which of m categories each unit i belongs to. E.g. the type of electoral system or region of country i is given by:
  - $X_i \in \{Proportional, Majoritarian\}$ so $m = 2$
  - $X_i \in \{Asia, Africa, LatinAmerica, OECD, Transition\}$ so $m = 5$
- To incorporate this information into our regression function we usually create $m - 1$ dummy variables, one for each of the $m - 1$ categories.
- Why not all m?

Dummy Variables for Multiple Categories

- More generally, let's say X measures which of m categories each unit i belongs to. E.g. the type of electoral system or region of country i is given by:
  - $X_i \in \{Proportional, Majoritarian\}$ so $m = 2$
  - $X_i \in \{Asia, Africa, LatinAmerica, OECD, Transition\}$ so $m = 5$
- To incorporate this information into our regression function we usually create $m - 1$ dummy variables, one for each of the $m - 1$ categories.
- Why not all m? Including all m category indicators as dummies would be indistinguishable from the intercept (more to come in one video!):

$$D_m = 1 - (D_1 + \cdots + D_{m-1})$$

Dummy Variables for Multiple Categories

- More generally, let's say X measures which of m categories each unit i belongs to. E.g. the type of electoral system or region of country i is given by:
  - $X_i \in \{Proportional, Majoritarian\}$ so $m = 2$
  - $X_i \in \{Asia, Africa, LatinAmerica, OECD, Transition\}$ so $m = 5$
- To incorporate this information into our regression function we usually create $m - 1$ dummy variables, one for each of the $m - 1$ categories.
- Why not all m? Including all m category indicators as dummies would be indistinguishable from the intercept (more to come in one video!):

$$D_m = 1 - (D_1 + \cdots + D_{m-1})$$

- The omitted category is our baseline case (also called a reference category) against which we compare the conditional means of Y for the other $m - 1$ categories.

Example: Regions of the World

- Consider the case of our "polytomous" variable world region with $m = 5$:

  $X_i \in \{Asia, Africa, LatinAmerica, OECD, Transition\}$

Example: Regions of the World

- Consider the case of our "polytomous" variable world region with $m = 5$:

  $X_i \in \{Asia, Africa, LatinAmerica, OECD, Transition\}$
- This five-category classification can be represented in the regression equation by introducing $m - 1 = 4$ dummy regressors:

Category	D_1	D_2	D_3	D_4
Asia	1	0	0	0
Africa	0	1	0	0
LatinAmerica	0	0	1	0
OECD	0	0	0	1
Transition	0	0	0	0

Example: Regions of the World

- Consider the case of our "polytomous" variable world region with $m = 5$:

  $X_i \in \{Asia, Africa, LatinAmerica, OECD, Transition\}$
- This five-category classification can be represented in the regression equation by introducing $m - 1 = 4$ dummy regressors:

Category	D_1	D_2	D_3	D_4
Asia	1	0	0	0
Africa	0	1	0	0
LatinAmerica	0	0	1	0
OECD	0	0	0	1
Transition	0	0	0	0

Our regression equation is:

$$Y = \beta_0 + \beta_1 D_1 + \beta_2 D_2 + \beta_3 D_3 + \beta_4 D_4 + u$$

Example: GDP per capita on Regions

R Code

```
> summary(lm(REALGDPCAP ~ Asia + Africa + LatAmerica + Oecd, data = D))

~~~~~~


Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4452.7      783.4   5.684 2.07e-07 ***
Asia           148.9     1149.8   0.129   0.8973
Africa       -2552.8     1204.5  -2.119   0.0372 *
LatAmerica    -271.3     1007.0  -0.269   0.7883
Oecd          9671.3     1007.0   9.604 5.74e-15 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 3034 on 80 degrees of freedom
Multiple R-squared: 0.7096,       Adjusted R-squared: 0.6951
F-statistic: 48.88 on 4 and 80 DF,  p-value: < 2.2e-16
```

What does β_0 mean?

Example: GDP per capita on Regions

R Code

```
> summary(lm(REALGDPCAP ~ Asia + Africa + LatAmerica + Oecd, data = D))

~~~~~


Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4452.7      783.4   5.684 2.07e-07 ***
Asia           148.9     1149.8   0.129   0.8973
Africa       -2552.8     1204.5  -2.119   0.0372 *
LatAmerica    -271.3     1007.0  -0.269   0.7883
Oecd          9671.3     1007.0   9.604 5.74e-15 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 3034 on 80 degrees of freedom
Multiple R-squared: 0.7096,        Adjusted R-squared: 0.6951
F-statistic: 48.88 on 4 and 80 DF,  p-value: < 2.2e-16
```

What does β_0 mean?

$$\beta_0 = E[GDP|D_j = 0 \text{ for all } j] = E[GDP|Transition]$$

So the mean for the baseline category shows up as the intercept.

Example: GDP per capita on Regions

R Code

```
> summary(lm(REALGDPCAP ~ Asia + Africa + LatAmerica + Oecd, data = D))

~~~~~


Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4452.7      783.4   5.684 2.07e-07 ***
Asia           148.9     1149.8   0.129   0.8973
Africa       -2552.8     1204.5  -2.119   0.0372 *
LatAmerica    -271.3     1007.0  -0.269   0.7883
Oecd          9671.3     1007.0   9.604 5.74e-15 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 3034 on 80 degrees of freedom
Multiple R-squared: 0.7096,        Adjusted R-squared: 0.6951
F-statistic: 48.88 on 4 and 80 DF,  p-value: < 2.2e-16
```

What does β_{Africa} mean?

Example: GDP per capita on Regions

R Code

```
> summary(lm(REALGDPCAP ~ Asia + Africa + LatAmerica + Oecd, data = D))

~~~~~


Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4452.7      783.4   5.684 2.07e-07 ***
Asia           148.9     1149.8   0.129   0.8973
Africa       -2552.8     1204.5  -2.119   0.0372 *
LatAmerica    -271.3     1007.0  -0.269   0.7883
Oecd          9671.3     1007.0   9.604 5.74e-15 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 3034 on 80 degrees of freedom
Multiple R-squared: 0.7096,       Adjusted R-squared: 0.6951
F-statistic: 48.88 on 4 and 80 DF,  p-value: < 2.2e-16
```

What does β_{Africa} mean?

$$\beta_{Africa} = E[GDP|Africa] - E[GDP|Transition]$$

The difference in means between the baseline and that category.

Example: GDP per capita on Regions

R Code

```
> summary(lm(REALGDPCAP ~ Asia + Africa + LatAmerica + Oecd, data = D))

~~~~~


Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4452.7      783.4   5.684 2.07e-07 ***
Asia           148.9     1149.8   0.129   0.8973
Africa       -2552.8     1204.5  -2.119   0.0372 *
LatAmerica    -271.3     1007.0  -0.269   0.7883
Oecd          9671.3     1007.0   9.604 5.74e-15 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 3034 on 80 degrees of freedom
Multiple R-squared: 0.7096,       Adjusted R-squared: 0.6951
F-statistic: 48.88 on 4 and 80 DF,  p-value: < 2.2e-16
```

Do Latin America economies have higher or lower average GDP than Asian economies?

Example: GDP per capita on Regions

R Code

```
> summary(lm(REALGDPCAP ~ Asia + Africa + LatAmerica + Oecd, data = D))

~~~~~


Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4452.7      783.4   5.684 2.07e-07 ***
Asia           148.9     1149.8   0.129   0.8973
Africa       -2552.8     1204.5  -2.119   0.0372 *
LatAmerica    -271.3     1007.0  -0.269   0.7883
Oecd          9671.3     1007.0   9.604 5.74e-15 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 3034 on 80 degrees of freedom
Multiple R-squared: 0.7096,        Adjusted R-squared: 0.6951
F-statistic: 48.88 on 4 and 80 DF,  p-value: < 2.2e-16
```

Do Latin America economies have higher or lower average GDP than Asian economies?
$\beta_{LatAmerica} = E[GDP|LatAmerica] - E[GDP|Transition]$, and
$\beta_{Asia} = E[GDP|Asia] - E[GDP|Transition]$, so

Example: GDP per capita on Regions

R Code

```
> summary(lm(REALGDPCAP ~ Asia + Africa + LatAmerica + Oecd, data = D))

~~~~~


Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4452.7      783.4   5.684 2.07e-07 ***
Asia           148.9     1149.8   0.129   0.8973
Africa       -2552.8     1204.5  -2.119   0.0372 *
LatAmerica    -271.3     1007.0  -0.269   0.7883
Oecd          9671.3     1007.0   9.604 5.74e-15 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 3034 on 80 degrees of freedom
Multiple R-squared: 0.7096,       Adjusted R-squared: 0.6951
F-statistic: 48.88 on 4 and 80 DF,  p-value: < 2.2e-16
```

Do Latin America economies have higher or lower average GDP than Asian economies?
$\beta_{LatAmerica} = E[GDP|LatAmerica] - E[GDP|Transition]$, and
$\beta_{Asia} = E[GDP|Asia] - E[GDP|Transition]$, so

$$\beta_{LatAmerica} - \beta_{Asia} = E[GDP|LatAmerica] - E[GDP|Asia] = -420$$

Dealing with a Categorical Variable in R

- In fact, R automatically expands an m-category variable into an $m-1$ dummy variables:

R Code

```
> head(D$Region)
[1] LatAmerica Oecd       Oecd       LatAmerica Asia       LatAmerica
Levels: Africa Asia LatAmerica Oecd Transition

> summary(lm(REALGDPCAP ~ Region, data = D))

~~~~~
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)        1899.9      914.9   2.077   0.0410 *
RegionAsia         2701.7     1243.0   2.173   0.0327 *
RegionLatAmerica   2281.5     1112.3   2.051   0.0435 *
RegionOecd        12224.2     1112.3  10.990   <2e-16 ***
RegionTransition   2552.8     1204.5   2.119   0.0372 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 3034 on 80 degrees of freedom
Multiple R-squared: 0.7096,       Adjusted R-squared: 0.6951
F-statistic: 48.88 on 4 and 80 DF,  p-value: < 2.2e-16
```

Dealing with a Categorical Variable in R

- You can change the baseline category by the `relevel()` function:

R Code

```
> D$Region <- relevel(D$Region, ref="Transition")
> summary(lm(REALGDPCAP ~ Region, data = D))

~~~~~

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)        4452.7      783.4   5.684 2.07e-07 ***
RegionAfrica      -2552.8     1204.5  -2.119   0.0372 *
RegionAsia          148.9     1149.8   0.129   0.8973
RegionLatAmerica   -271.3     1007.0  -0.269   0.7883
RegionOecd         9671.3     1007.0   9.604 5.74e-15 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 3034 on 80 degrees of freedom
Multiple R-squared: 0.7096,       Adjusted R-squared: 0.6951
F-statistic: 48.88 on 4 and 80 DF,  p-value: < 2.2e-16
```

Saturated Models

Saturated Models

- A model is saturated if there are as many parameters as there are possible combination of the X_i variables.

Saturated Models

- A model is saturated if there are as many parameters as there are possible combination of the X_i variables.
- This happens when we have a dummy variable for every possible configuration of X variables in the data.

Saturated Models

- A model is saturated if there are as many parameters as there are possible combination of the X_i variables.
- This happens when we have a dummy variable for every possible configuration of X variables in the data.
- In this setting, linearity holds by construction because we are estimating a single mean for every combination of X_i variables.

Saturated Model Example

- Two binary variables, X_{1i} for marriage status and X_{2i} for having children.

Saturated Model Example

- Two binary variables, X_{1i} for marriage status and X_{2i} for having children.

Saturated Model Example

- Two binary variables, X_{1i} for marriage status and X_{2i} for having children.
- Four possible values of X_i, four possible values of $\mu(X_i)$:

$$E[Y_i|X_{1i} = 0, X_{2i} = 0]$$
$$E[Y_i|X_{1i} = 1, X_{2i} = 0]$$
$$E[Y_i|X_{1i} = 0, X_{2i} = 1]$$
$$E[Y_i|X_{1i} = 1, X_{2i} = 1]$$

Saturated Model Example

- Two binary variables, X_{1i} for marriage status and X_{2i} for having children.
- Four possible values of X_i, four possible values of $\mu(X_i)$:

$$E[Y_i|X_{1i} = 0, X_{2i} = 0] = \alpha$$
$$E[Y_i|X_{1i} = 1, X_{2i} = 0]$$
$$E[Y_i|X_{1i} = 0, X_{2i} = 1]$$
$$E[Y_i|X_{1i} = 1, X_{2i} = 1]$$

- We can write the CEF as follows:

$$E[Y_i|X_{1i}, X_{2i}] = \alpha + \beta X_{1i} + \gamma X_{2i} + \delta(X_{1i}X_{2i})$$

Saturated Model Example

- Two binary variables, X_{1i} for marriage status and X_{2i} for having children.
- Four possible values of X_i, four possible values of $\mu(X_i)$:

$$
\begin{aligned}
E[Y_i|X_{1i} = 0, X_{2i} = 0] &= \alpha \\
E[Y_i|X_{1i} = 1, X_{2i} = 0] &= \alpha + \beta \\
E[Y_i|X_{1i} = 0, X_{2i} = 1] & \\
E[Y_i|X_{1i} = 1, X_{2i} = 1] &
\end{aligned}
$$

- We can write the CEF as follows:

$$E[Y_i|X_{1i}, X_{2i}] = \alpha + \beta X_{1i} + \gamma X_{2i} + \delta(X_{1i}X_{2i})$$

Saturated Model Example

- Two binary variables, X_{1i} for marriage status and X_{2i} for having children.
- Four possible values of X_i, four possible values of $\mu(X_i)$:

$$
\begin{aligned}
E[Y_i|X_{1i} = 0, X_{2i} = 0] &= \alpha \\
E[Y_i|X_{1i} = 1, X_{2i} = 0] &= \alpha + \beta \\
E[Y_i|X_{1i} = 0, X_{2i} = 1] &= \alpha + \gamma \\
E[Y_i|X_{1i} = 1, X_{2i} = 1] &
\end{aligned}
$$

- We can write the CEF as follows:

$$E[Y_i|X_{1i}, X_{2i}] = \alpha + \beta X_{1i} + \gamma X_{2i} + \delta(X_{1i}X_{2i})$$

Saturated Model Example

- Two binary variables, X_{1i} for marriage status and X_{2i} for having children.
- Four possible values of X_i, four possible values of $\mu(X_i)$:

$$
\begin{aligned}
E[Y_i|X_{1i} = 0, X_{2i} = 0] &= \alpha \\
E[Y_i|X_{1i} = 1, X_{2i} = 0] &= \alpha + \beta \\
E[Y_i|X_{1i} = 0, X_{2i} = 1] &= \alpha + \gamma \\
E[Y_i|X_{1i} = 1, X_{2i} = 1] &= \alpha + \beta + \gamma + \delta
\end{aligned}
$$

- We can write the CEF as follows:

$$E[Y_i|X_{1i}, X_{2i}] = \alpha + \beta X_{1i} + \gamma X_{2i} + \delta(X_{1i}X_{2i})$$

Saturated model example

$$E[Y_i|X_{1i}, X_{2i}] = \alpha + \beta X_{1i} + \gamma X_{2i} + \delta(X_{1i}X_{2i})$$

- Basically, each value of the CEF is being estimated separately.

Saturated model example

$$E[Y_i|X_{1i}, X_{2i}] = \alpha + \beta X_{1i} + \gamma X_{2i} + \delta(X_{1i}X_{2i})$$

- Basically, each value of the CEF is being estimated separately.
  - $\rightsquigarrow$ within-strata estimation.

Saturated model example

$$E[Y_i|X_{1i}, X_{2i}] = \alpha + \beta X_{1i} + \gamma X_{2i} + \delta(X_{1i}X_{2i})$$

- Basically, each value of the CEF is being estimated separately.
  - $\rightsquigarrow$ within-strata estimation.
  - No borrowing of information from across values of X_i.

Saturated model example

$$E[Y_i|X_{1i}, X_{2i}] = \alpha + \beta X_{1i} + \gamma X_{2i} + \delta(X_{1i}X_{2i})$$

- Basically, each value of the CEF is being estimated separately.
  - $\rightsquigarrow$ within-strata estimation.
  - No borrowing of information from across values of X_i.
- Requires a set of dummies for each categorical variable plus all interactions.

Saturated model example

$$E[Y_i|X_{1i}, X_{2i}] = \alpha + \beta X_{1i} + \gamma X_{2i} + \delta(X_{1i}X_{2i})$$

- Basically, each value of the CEF is being estimated separately.
  - $\rightsquigarrow$ within-strata estimation.
  - No borrowing of information from across values of X_i.
- Requires a set of dummies for each categorical variable plus all interactions.
- i.e. a series of dummies for each unique combination of X_i.

Saturated model example

- Ebonya Washington (AER) data from *AER* paper "Female socialization: how daughters affect their legislator fathers"

Saturated model example

- Ebonya Washington (AER) data from *AER* paper "Female socialization: how daughters affect their legislator fathers"
- We'll look at the relationship between voting and number of kids.

Saturated model example

- Ebonya Washington (AER) data from *AER* paper "Female socialization: how daughters affect their legislator fathers"
- We'll look at the relationship between voting and number of kids.

```r
girls <- foreign::read.dta("girls.dta")
head(girls[, c("name", "totchi", "aauw")])
```

```
##                     name totchi aauw
## 1   ABERCROMBIE, NEIL       0  100
## 2   ACKERMAN, GARY L.       3   88
## 3 ADERHOLT, ROBERT B.       0    0
## 4     ALLEN, THOMAS H.      2  100
## 5  ANDREWS, ROBERT E.       2  100
## 6          ARCHER, W.R.     7    0
```

Linear model

```r
summary(lm(aauw ~ totchi, data = girls))
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    61.31       1.81   33.81   <2e-16 ***
## totchi         -5.33       0.62   -8.59   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 42 on 1733 degrees of freedom
##   (5 observations deleted due to missingness)
## Multiple R-squared:  0.0408, Adjusted R-squared:  0.0403
## F-statistic: 73.8 on 1 and 1733 DF,  p-value: <2e-16
```

Saturated model

```r
summary(lm(aauw ~ as.factor(totchi), data = girls))
```

```
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)             56.41       2.76   20.42  < 2e-16 ***
## as.factor(totchi)1       5.45       4.11    1.33   0.1851
## as.factor(totchi)2      -3.80       3.27   -1.16   0.2454
## as.factor(totchi)3     -13.65       3.45   -3.95  8.1e-05 ***
## as.factor(totchi)4     -19.31       4.01   -4.82  1.6e-06 ***
## as.factor(totchi)5     -15.46       4.85   -3.19   0.0015 **
## as.factor(totchi)6     -33.59      10.42   -3.22   0.0013 **
## as.factor(totchi)7     -17.13      11.41   -1.50   0.1336
## as.factor(totchi)8     -55.33      12.28   -4.51  7.0e-06 ***
## as.factor(totchi)9     -50.41      24.08   -2.09   0.0364 *
## as.factor(totchi)10    -53.41      20.90   -2.56   0.0107 *
## as.factor(totchi)12    -56.41      41.53   -1.36   0.1745
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 41 on 1723 degrees of freedom
##   (5 observations deleted due to missingness)
## Multiple R-squared:  0.0506, Adjusted R-squared:  0.0446
## F-statistic: 8.36 on 11 and 1723 DF,  p-value: 1.84e-14
```

Saturated model minus the constant

```r
summary(lm(aauw ~ as.factor(totchi) - 1, data = girls))
```

```
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## as.factor(totchi)0     56.41       2.76   20.42   <2e-16 ***
## as.factor(totchi)1     61.86       3.05   20.31   <2e-16 ***
## as.factor(totchi)2     52.62       1.75   30.13   <2e-16 ***
## as.factor(totchi)3     42.76       2.07   20.62   <2e-16 ***
## as.factor(totchi)4     37.11       2.90   12.79   <2e-16 ***
## as.factor(totchi)5     40.95       3.99   10.27   <2e-16 ***
## as.factor(totchi)6     22.82      10.05    2.27   0.0233 *
## as.factor(totchi)7     39.29      11.07    3.55   0.0004 ***
## as.factor(totchi)8      1.08      11.96    0.09   0.9278
## as.factor(totchi)9      6.00      23.92    0.25   0.8020
## as.factor(totchi)10     3.00      20.72    0.14   0.8849
## as.factor(totchi)12     0.00      41.43    0.00   1.0000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 41 on 1723 degrees of freedom
##   (5 observations deleted due to missingness)
## Multiple R-squared:  0.587,  Adjusted R-squared:  0.584
## F-statistic:  204 on 12 and 1723 DF,  p-value: <2e-16
```

Compare to within-strata means

- The saturated model makes no assumptions about the between-strata relationships.

Compare to within-strata means

- The saturated model makes no assumptions about the between-strata relationships.
- Just calculates within-strata means:

Compare to within-strata means

- The saturated model makes no assumptions about the between-strata relationships.
- Just calculates within-strata means:

```
c1 <- coef(lm(aauw ~ as.factor(totchi) - 1, data = girls))
c2 <- with(girls, tapply(aauw, totchi, mean, na.rm = TRUE))
rbind(c1, c2)
```

```
##     0  1  2  3  4  5  6  7   8 9 10 12
## c1 56 62 53 43 37 41 23 39 1.1 6  3  0
## c2 56 62 53 43 37 41 23 39 1.1 6  3  0
```

We Covered

We Covered

- How to add a binary variable.

We Covered

- How to add a binary variable.
- How to add a continuous variable.

We Covered

- How to add a binary variable.
- How to add a continuous variable.
- Dummy variables and saturated models.

We Covered

- How to add a binary variable.
- How to add a continuous variable.
- Dummy variables and saturated models.

Next Time: Estimation and Inference!

Where We've Been and Where We're Going...

Where We've Been and Where We're Going...

- Last Week
  - mechanics of OLS with one variable
  - properties of OLS
- This Week
  - adding a second variable
  - new mechanics
  - omitted variable bias
  - multicollinearity
  - interactions
- Next Week
  - multiple regression
- Long Run
  - probability → inference → regression → causal inference

Fitted values and Residuals

- Where do we get our hats?

Fitted values and Residuals

- Where do we get our hats?

Fitted values and Residuals

- Where do we get our hats? $\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2$

Fitted values and Residuals

- Where do we get our hats? $\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2$
- To answer this, we first need to redefine and generalize some terms from linear regression with one predictor.

Fitted values and Residuals

- Where do we get our hats? $\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2$
- To answer this, we first need to redefine and generalize some terms from linear regression with one predictor.
- Let's change notation to call our second variable Z_i so its a bit clearer.

Fitted values and Residuals

- Where do we get our hats? $\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2$
- To answer this, we first need to redefine and generalize some terms from linear regression with one predictor.
- Let's change notation to call our second variable Z_i so its a bit clearer.
- Fitted values for $i = 1, \ldots, n$:

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{\beta}_2 Z_i$$

Fitted values and Residuals

- Where do we get our hats? $\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2$
- To answer this, we first need to redefine and generalize some terms from linear regression with one predictor.
- Let's change notation to call our second variable Z_i so its a bit clearer.
- Fitted values for $i = 1, \dots, n$:

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{\beta}_2 Z_i$$

- Residuals for $i = 1, \dots, n$:

$$\widehat{u}_i = Y_i - \widehat{Y}_i$$

Least Squares is Still Least Squares

- How do we estimate $\widehat{\beta}_0$, $\widehat{\beta}_1$, and $\widehat{\beta}_2$?

Least Squares is Still Least Squares

- How do we estimate $\widehat{\beta}_0$, $\widehat{\beta}_1$, and $\widehat{\beta}_2$?
- Minimize the sum of the squared residuals,

$$(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2) = \underset{b_0, b_1, b_2}{\arg\min} \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i - b_2 Z_i)^2$$

Least Squares is Still Least Squares

- How do we estimate $\widehat{\beta}_0$, $\widehat{\beta}_1$, and $\widehat{\beta}_2$?
- Minimize the sum of the squared residuals,

$$(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2) = \underset{b_0, b_1, b_2}{\arg\min} \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i - b_2 Z_i)^2$$

Plan is conceptually the same as before

Least Squares is Still Least Squares

- How do we estimate $\widehat{\beta}_0$, $\widehat{\beta}_1$, and $\widehat{\beta}_2$?
- Minimize the sum of the squared residuals,

$$(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2) = \underset{b_0, b_1, b_2}{\arg\min} \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i - b_2 Z_i)^2$$

Plan is conceptually the same as before

1. Take the partial derivatives of S with respect to b_0, b_1, b_2.

Least Squares is Still Least Squares

- How do we estimate $\widehat{\beta}_0$, $\widehat{\beta}_1$, and $\widehat{\beta}_2$?
- Minimize the sum of the squared residuals,

$$(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2) = \underset{b_0, b_1, b_2}{\arg\min} \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i - b_2 Z_i)^2$$

Plan is conceptually the same as before

1. Take the partial derivatives of S with respect to b_0, b_1, b_2.
2. Set each of the partial derivatives to 0 to obtain the first order conditions.

Least Squares is Still Least Squares

- How do we estimate $\widehat{\beta}_0$, $\widehat{\beta}_1$, and $\widehat{\beta}_2$?
- Minimize the sum of the squared residuals,

$$(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2) = \underset{b_0, b_1, b_2}{\arg\min} \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i - b_2 Z_i)^2$$

Plan is conceptually the same as before

1. Take the partial derivatives of S with respect to b_0, b_1, b_2.
2. Set each of the partial derivatives to 0 to obtain the first order conditions.
3. Substitute $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ for b_0, b_1, b_2 and solve for $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ to obtain the OLS estimator.

Take partial derivatives

$(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2) = \arg\min_{b_0,b_1,b_2} \sum_{i=1}^{n}(Y_i - b_0 - b_1X_i - b_2Z_i)^2$
After some calculus and algebra we can show that:

$$\frac{\partial S}{\partial b_0} = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i - \hat{\beta}_2 z_i)$$

$$\frac{\partial S}{\partial b_1} = \sum_{i=1}^{n} x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i - \hat{\beta}_2 z_i)$$

$$\frac{\partial S}{\partial b_2} = \sum_{i=1}^{n} z_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i - \hat{\beta}_2 z_i)$$

First Order Conditions

Setting the partial derivatives equal to zero leads to a system of 3 linear equations in 3 unknowns: $\hat{\beta}_0, \hat{\beta}_1$ and $\hat{\beta}_2$

$$
\begin{aligned}
\frac{\partial S}{\partial b_0} &= \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i - \hat{\beta}_2 z_i) = 0 \\
\frac{\partial S}{\partial b_1} &= \sum_{i=1}^{n} x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i - \hat{\beta}_2 z_i) = 0 \\
\frac{\partial S}{\partial b_2} &= \sum_{i=1}^{n} z_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i - \hat{\beta}_2 z_i) = 0
\end{aligned}
$$

When will this linear system have a unique solution?

First Order Conditions

Setting the partial derivatives equal to zero leads to a system of 3 linear equations in 3 unknowns: $\hat{\beta}_0, \hat{\beta}_1$ and $\hat{\beta}_2$

$$\begin{aligned}
\frac{\partial S}{\partial b_0} &= \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i - \hat{\beta}_2 z_i) = 0 \\
\frac{\partial S}{\partial b_1} &= \sum_{i=1}^{n} x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i - \hat{\beta}_2 z_i) = 0 \\
\frac{\partial S}{\partial b_2} &= \sum_{i=1}^{n} z_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i - \hat{\beta}_2 z_i) = 0
\end{aligned}$$

When will this linear system have a unique solution?

- More observations than predictors (i.e. $n > 2$)

First Order Conditions

Setting the partial derivatives equal to zero leads to a system of 3 linear equations in 3 unknowns: $\hat{\beta}_0, \hat{\beta}_1$ and $\hat{\beta}_2$

$$\begin{aligned}
\frac{\partial S}{\partial b_0} &= \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i - \hat{\beta}_2 z_i) = 0 \\
\frac{\partial S}{\partial b_1} &= \sum_{i=1}^{n} x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i - \hat{\beta}_2 z_i) = 0 \\
\frac{\partial S}{\partial b_2} &= \sum_{i=1}^{n} z_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i - \hat{\beta}_2 z_i) = 0
\end{aligned}$$

When will this linear system have a unique solution?

- More observations than predictors (i.e. $n > 2$)
- x and z are linearly independent, i.e.,
  - neither x nor z is a constant
  - x is not a linear function of z (or vice versa)

First Order Conditions

Setting the partial derivatives equal to zero leads to a system of 3 linear equations in 3 unknowns: $\hat{\beta}_0, \hat{\beta}_1$ and $\hat{\beta}_2$

$$
\begin{aligned}
\frac{\partial S}{\partial b_0} &= \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i - \hat{\beta}_2 z_i) = 0 \\
\frac{\partial S}{\partial b_1} &= \sum_{i=1}^{n} x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i - \hat{\beta}_2 z_i) = 0 \\
\frac{\partial S}{\partial b_2} &= \sum_{i=1}^{n} z_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i - \hat{\beta}_2 z_i) = 0
\end{aligned}
$$

When will this linear system have a unique solution?

- More observations than predictors (i.e. $n > 2$)
- x and z are linearly independent, i.e.,
  - neither x nor z is a constant
  - x is not a linear function of z (or vice versa)
- Typically called no perfect collinearity

The OLS Estimator

After lots of algebra, the OLS estimator for $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ can be written as

$$\begin{aligned}
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} - \hat{\beta}_2 \bar{z} \\
\hat{\beta}_1 &= \frac{Cov(x,y)Var(z) - Cov(z,y)Cov(x,z)}{Var(x)Var(z) - Cov(x,z)^2} \\
\hat{\beta}_2 &= \frac{Cov(z,y)Var(x) - Cov(x,y)Cov(z,x)}{Var(x)Var(z) - Cov(x,z)^2}
\end{aligned}$$

The OLS Estimator

After lots of algebra, the OLS estimator for $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ can be written as

$$\begin{aligned}
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} - \hat{\beta}_2 \bar{z} \\
\hat{\beta}_1 &= \frac{Cov(x,y)Var(z) - Cov(z,y)Cov(x,z)}{Var(x)Var(z) - Cov(x,z)^2} \\
\hat{\beta}_2 &= \frac{Cov(z,y)Var(x) - Cov(x,y)Cov(z,x)}{Var(x)Var(z) - Cov(x,z)^2}
\end{aligned}$$

For $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ to be well-defined we need:

$$Var(x)Var(z) \neq Cov(x,z)^2$$

This requirement fails if:

The OLS Estimator

After lots of algebra, the OLS estimator for $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ can be written as

$$
\begin{aligned}
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} - \hat{\beta}_2 \bar{z} \\
\hat{\beta}_1 &= \frac{Cov(x,y)Var(z) - Cov(z,y)Cov(x,z)}{Var(x)Var(z) - Cov(x,z)^2} \\
\hat{\beta}_2 &= \frac{Cov(z,y)Var(x) - Cov(x,y)Cov(z,x)}{Var(x)Var(z) - Cov(x,z)^2}
\end{aligned}
$$

For $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ to be well-defined we need:

$$Var(x)Var(z) \neq Cov(x,z)^2$$

This requirement fails if:

1. If x or z is a constant ($\Rightarrow Var(x)Var(z) = Cov(x,z) = 0$)

The OLS Estimator

After lots of algebra, the OLS estimator for $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ can be written as

$$\begin{aligned}
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} - \hat{\beta}_2 \bar{z} \\
\hat{\beta}_1 &= \frac{Cov(x,y)Var(z) - Cov(z,y)Cov(x,z)}{Var(x)Var(z) - Cov(x,z)^2} \\
\hat{\beta}_2 &= \frac{Cov(z,y)Var(x) - Cov(x,y)Cov(z,x)}{Var(x)Var(z) - Cov(x,z)^2}
\end{aligned}$$

For $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ to be well-defined we need:

$$Var(x)Var(z) \neq Cov(x,z)^2$$

This requirement fails if:

1. If x or z is a constant ($\Rightarrow Var(x)Var(z) = Cov(x,z) = 0$)
2. One explanatory variable is an exact linear function of another ($\Rightarrow \; Cor(x,z) = 1 \; \Rightarrow \; Var(x)Var(z) = Cov(x,z)^2$)

OLS assumptions for unbiasedness

OLS assumptions for unbiasedness

- When we have more than one independent variable, we need the following assumptions in order for OLS to be unbiased:

OLS assumptions for unbiasedness

- When we have more than one independent variable, we need the following assumptions in order for OLS to be unbiased:

1. Linearity

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i$$

OLS assumptions for unbiasedness

- When we have more than one independent variable, we need the following assumptions in order for OLS to be unbiased:

1. Linearity
$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i$$
2. Random/iid sample

OLS assumptions for unbiasedness

- When we have more than one independent variable, we need the following assumptions in order for OLS to be unbiased:

1. Linearity
$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i$$
2. Random/iid sample
3. No perfect collinearity

OLS assumptions for unbiasedness

- When we have more than one independent variable, we need the following assumptions in order for OLS to be unbiased:

1. Linearity
$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i$$
2. Random/iid sample
3. No perfect collinearity
4. Zero conditional mean error
$$E[u_i | X_i, Z_i] = 0$$

New assumption

Assumption 3: No perfect collinearity

(1) No explanatory variable is constant in the sample and (2) there are no exactly linear relationships among the explanatory variables.

- Two components

New assumption

Assumption 3: No perfect collinearity

(1) No explanatory variable is constant in the sample and (2) there are no exactly linear relationships among the explanatory variables.

- Two components
    1. Both X_i and Z_i have to vary.

New assumption

Assumption 3: No perfect collinearity

(1) No explanatory variable is constant in the sample and (2) there are no exactly linear relationships among the explanatory variables.

- Two components
  1. Both X_i and Z_i have to vary.
  2. Z_i cannot be a deterministic, linear function of X_i.

New assumption

Assumption 3: No perfect collinearity

(1) No explanatory variable is constant in the sample and (2) there are no exactly linear relationships among the explanatory variables.

- Two components
  1. Both X_i and Z_i have to vary.
  2. Z_i cannot be a deterministic, linear function of X_i.
- Part 2 rules out anything of the form:

$$Z_i = a + bX_i$$

New assumption

Assumption 3: No perfect collinearity

(1) No explanatory variable is constant in the sample and (2) there are no exactly linear relationships among the explanatory variables.

- Two components
  1. Both X_i and Z_i have to vary.
  2. Z_i cannot be a deterministic, linear function of X_i.
- Part 2 rules out anything of the form:

$$Z_i = a + bX_i$$

- Notice how this is linear (equation of a line) and there is no error, so it is deterministic.

New assumption

Assumption 3: No perfect collinearity

(1) No explanatory variable is constant in the sample and (2) there are no exactly linear relationships among the explanatory variables.

- Two components
  1. Both X_i and Z_i have to vary.
  2. Z_i cannot be a deterministic, linear function of X_i.
- Part 2 rules out anything of the form:

$$Z_i = a + bX_i$$

- Notice how this is linear (equation of a line) and there is no error, so it is deterministic.
- What's the correlation between Z_i and X_i? 1!

Perfect collinearity example (I)

- Simple example:

Perfect collinearity example (I)

- Simple example:
  - $X_i = 1$ if a country is not in Africa and 0 otherwise.

Perfect collinearity example (I)

- Simple example:
  - $X_i = 1$ if a country is not in Africa and 0 otherwise.
  - $Z_i = 1$ if a country is in Africa and 0 otherwise.

Perfect collinearity example (I)

- Simple example:
  - $X_i = 1$ if a country is not in Africa and 0 otherwise.
  - $Z_i = 1$ if a country is in Africa and 0 otherwise.
- But, clearly we have the following:

$$Z_i = 1 - X_i$$

Perfect collinearity example (I)

- Simple example:
  - $X_i = 1$ if a country is not in Africa and 0 otherwise.
  - $Z_i = 1$ if a country is in Africa and 0 otherwise.
- But, clearly we have the following:

$$Z_i = 1 - X_i$$

- These two variables are perfectly collinear.

Perfect collinearity example (I)

- Simple example:
  - $X_i = 1$ if a country is not in Africa and 0 otherwise.
  - $Z_i = 1$ if a country is in Africa and 0 otherwise.
- But, clearly we have the following:

$$Z_i = 1 - X_i$$

- These two variables are perfectly collinear.
- What about the following:

Perfect collinearity example (I)

- Simple example:
  - $X_i = 1$ if a country is not in Africa and 0 otherwise.
  - $Z_i = 1$ if a country is in Africa and 0 otherwise.
- But, clearly we have the following:

$$Z_i = 1 - X_i$$

- These two variables are perfectly collinear.
- What about the following:
  - $X_i =$ income

Perfect collinearity example (I)

- Simple example:
  - $X_i = 1$ if a country is not in Africa and 0 otherwise.
  - $Z_i = 1$ if a country is in Africa and 0 otherwise.
- But, clearly we have the following:

$$Z_i = 1 - X_i$$

- These two variables are perfectly collinear.
- What about the following:
  - $X_i =$ income
  - $Z_i = X_i^2$

Perfect collinearity example (I)

- Simple example:
  - $X_i = 1$ if a country is not in Africa and 0 otherwise.
  - $Z_i = 1$ if a country is in Africa and 0 otherwise.
- But, clearly we have the following:

$$Z_i = 1 - X_i$$

- These two variables are perfectly collinear.
- What about the following:
  - $X_i =$ income
  - $Z_i = X_i^2$
- Do we have to worry about collinearity here?

Perfect collinearity example (I)

- Simple example:
  - $X_i = 1$ if a country is not in Africa and 0 otherwise.
  - $Z_i = 1$ if a country is in Africa and 0 otherwise.
- But, clearly we have the following:

$$Z_i = 1 - X_i$$

- These two variables are perfectly collinear.
- What about the following:
  - $X_i = \text{income}$
  - $Z_i = X_i^2$
- Do we have to worry about collinearity here?
- No! Because while Z_i is a deterministic function of X_i, it is not a linear function of X_i.

R and perfect collinearity

- R, and all other packages, will drop one of the variables if there is perfect collinearity:

R and perfect collinearity

- R, and all other packages, will drop one of the variables if there is perfect collinearity:

R and perfect collinearity

- R, and all other packages, will drop one of the variables if there is perfect collinearity:

```
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.71638    0.08991  96.941  < 2e-16 ***
## africa      -1.36119    0.16306  -8.348 4.87e-14 ***
## nonafrica         NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9125 on 146 degrees of freedom
##   (15 observations deleted due to missingness)
## Multiple R-squared:  0.3231, Adjusted R-squared:  0.3184
## F-statistic: 69.68 on 1 and 146 DF,  p-value: 4.87e-14
```

Perfect collinearity example (II)

- Another example:

Perfect collinearity example (II)

- Another example:
  - X_i = mean temperature in Celsius

Perfect collinearity example (II)

- Another example:
  - $X_i =$ mean temperature in Celsius
  - $Z_i = 1.8X_i + 32$ (mean temperature in Fahrenheit)

Perfect collinearity example (II)

- Another example:
  - $X_i =$ mean temperature in Celsius
  - $Z_i = 1.8X_i + 32$ (mean temperature in Fahrenheit)

Perfect collinearity example (II)

- Another example:
  - X_i = mean temperature in Celsius
  - $Z_i = 1.8X_i + 32$ (mean temperature in Fahrenheit)

```
## (Intercept)    meantemp  meantemp.f
##  10.8454999  -0.1206948          NA
```

OLS assumptions for large-sample inference

OLS assumptions for large-sample inference

For large-sample inference and calculating SEs, we need the two-variable version of the Gauss-Markov assumptions:

OLS assumptions for large-sample inference

For large-sample inference and calculating SEs, we need the two-variable version of the Gauss-Markov assumptions:

1. Linearity
$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i$$
2. Random/iid sample
3. No perfect collinearity
4. Zero conditional mean error
$$E[u_i|X_i, Z_i] = 0$$

OLS assumptions for large-sample inference

For large-sample inference and calculating SEs, we need the two-variable version of the Gauss-Markov assumptions:

1. Linearity

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i$$

2. Random/iid sample
3. No perfect collinearity
4. Zero conditional mean error

$$E[u_i|X_i, Z_i] = 0$$

5. Homoskedasticity

$$\text{var}[u_i|X_i, Z_i] = \sigma_u^2$$

OLS assumptions for large-sample inference

- We have our OLS estimate $\widehat{\beta}_1$

OLS assumptions for large-sample inference

- We have our OLS estimate $\widehat{\beta}_1$
- We have an estimate of the standard error for that coefficient, $\widehat{SE}[\widehat{\beta}_1]$.

OLS assumptions for large-sample inference

- We have our OLS estimate $\widehat{\beta}_1$
- We have an estimate of the standard error for that coefficient, $\widehat{SE}[\widehat{\beta}_1]$.
- Under assumption 1-5, in large samples, we'll have the following:

$$\frac{\widehat{\beta}_1 - \beta_1}{\widehat{SE}[\widehat{\beta}_1]} \sim N(0, 1)$$

OLS assumptions for large-sample inference

- We have our OLS estimate $\widehat{\beta}_1$
- We have an estimate of the standard error for that coefficient, $\widehat{SE}[\widehat{\beta}_1]$.
- Under assumption 1-5, in large samples, we'll have the following:

$$\frac{\widehat{\beta}_1 - \beta_1}{\widehat{SE}[\widehat{\beta}_1]} \sim N(0,1)$$

- The same holds for the other coefficient:

$$\frac{\widehat{\beta}_2 - \beta_2}{\widehat{SE}[\widehat{\beta}_2]} \sim N(0,1)$$

OLS assumptions for large-sample inference

- We have our OLS estimate $\widehat{\beta}_1$
- We have an estimate of the standard error for that coefficient, $\widehat{SE}[\widehat{\beta}_1]$.
- Under assumption 1-5, in large samples, we'll have the following:

$$\frac{\widehat{\beta}_1 - \beta_1}{\widehat{SE}[\widehat{\beta}_1]} \sim N(0,1)$$

- The same holds for the other coefficient:

$$\frac{\widehat{\beta}_2 - \beta_2}{\widehat{SE}[\widehat{\beta}_2]} \sim N(0,1)$$

- Inference is exactly the same in large samples!

OLS assumptions for large-sample inference

- We have our OLS estimate $\widehat{\beta}_1$
- We have an estimate of the standard error for that coefficient, $\widehat{SE}[\widehat{\beta}_1]$.
- Under assumption 1-5, in large samples, we'll have the following:

$$\frac{\widehat{\beta}_1 - \beta_1}{\widehat{SE}[\widehat{\beta}_1]} \sim N(0,1)$$

- The same holds for the other coefficient:

$$\frac{\widehat{\beta}_2 - \beta_2}{\widehat{SE}[\widehat{\beta}_2]} \sim N(0,1)$$

- Inference is exactly the same in large samples!
- Hypothesis tests and CIs are good to go

OLS assumptions for large-sample inference

- We have our OLS estimate $\widehat{\beta}_1$
- We have an estimate of the standard error for that coefficient, $\widehat{SE}[\widehat{\beta}_1]$.
- Under assumption 1-5, in large samples, we'll have the following:

$$\frac{\widehat{\beta}_1 - \beta_1}{\widehat{SE}[\widehat{\beta}_1]} \sim N(0,1)$$

- The same holds for the other coefficient:

$$\frac{\widehat{\beta}_2 - \beta_2}{\widehat{SE}[\widehat{\beta}_2]} \sim N(0,1)$$

- Inference is exactly the same in large samples!
- Hypothesis tests and CIs are good to go
- The SE's will change, though

Inference with two independent variables in large samples

Inference with two independent variables in large samples

For small-sample inference, we need the Gauss-Markov plus Normal errors:

Inference with two independent variables in large samples

For small-sample inference, we need the Gauss-Markov plus Normal errors:

1. Linearity
$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i$$
2. Random/iid sample
3. No perfect collinearity
4. Zero conditional mean error
$$E[u_i|X_i, Z_i] = 0$$
5. Homoskedasticity
$$\text{var}[u_i|X_i, Z_i] = \sigma_u^2$$

Inference with two independent variables in large samples

For small-sample inference, we need the Gauss-Markov plus Normal errors:

1. Linearity
$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i$$
2. Random/iid sample
3. No perfect collinearity
4. Zero conditional mean error
$$E[u_i | X_i, Z_i] = 0$$
5. Homoskedasticity
$$\text{var}[u_i | X_i, Z_i] = \sigma_u^2$$
6. Normal conditional errors
$$u_i \sim N(0, \sigma_u^2)$$

Inference with two independent variables in small samples

- Under assumptions 1-6, we have the following small change to our small-n sampling distribution:

$$\frac{\widehat{\beta}_1 - \beta_1}{\widehat{SE}[\widehat{\beta}_1]} \sim t_{n-3}$$

Inference with two independent variables in small samples

- Under assumptions 1-6, we have the following small change to our small-n sampling distribution:

$$\frac{\widehat{\beta}_1 - \beta_1}{\widehat{SE}[\widehat{\beta}_1]} \sim t_{n-3}$$

- The same is true for the other coefficient:

$$\frac{\widehat{\beta}_2 - \beta_2}{\widehat{SE}[\widehat{\beta}_2]} \sim t_{n-3}$$

Inference with two independent variables in small samples

- Under assumptions 1-6, we have the following small change to our small-n sampling distribution:

$$\frac{\widehat{\beta}_1 - \beta_1}{\widehat{SE}[\widehat{\beta}_1]} \sim t_{n-3}$$

- The same is true for the other coefficient:

$$\frac{\widehat{\beta}_2 - \beta_2}{\widehat{SE}[\widehat{\beta}_2]} \sim t_{n-3}$$

- Why $n - 3$?

Inference with two independent variables in small samples

- Under assumptions 1-6, we have the following small change to our small-n sampling distribution:

$$\frac{\widehat{\beta}_1 - \beta_1}{\widehat{SE}[\widehat{\beta}_1]} \sim t_{n-3}$$

- The same is true for the other coefficient:

$$\frac{\widehat{\beta}_2 - \beta_2}{\widehat{SE}[\widehat{\beta}_2]} \sim t_{n-3}$$

- Why $n - 3$?
  - We've estimated another parameter, so we need to take off another degree of freedom.

Inference with two independent variables in small samples

- Under assumptions 1-6, we have the following small change to our small-n sampling distribution:

$$\frac{\widehat{\beta}_1 - \beta_1}{\widehat{SE}[\widehat{\beta}_1]} \sim t_{n-3}$$

- The same is true for the other coefficient:

$$\frac{\widehat{\beta}_2 - \beta_2}{\widehat{SE}[\widehat{\beta}_2]} \sim t_{n-3}$$

- Why $n - 3$?
  - We've estimated another parameter, so we need to take off another degree of freedom.
- $\rightsquigarrow$ small adjustments to the critical values and the t-values for our hypothesis tests and confidence intervals.

Another view of OLS with two predictors

- "Partialling out" OLS recipe:

Another view of OLS with two predictors

- "Partialling out" OLS recipe:
  1. Run regression of X_i on Z_i:

$$\widehat{X}_i = \widehat{\delta}_0 + \widehat{\delta}_1 Z_i$$

Another view of OLS with two predictors

- "Partialling out" OLS recipe:
  1. Run regression of X_i on Z_i:

$$\widehat{X}_i = \widehat{\delta}_0 + \widehat{\delta}_1 Z_i$$

  2. Calculate residuals from this regression:

$$\widehat{r}_{xz,i} = X_i - \widehat{X}_i$$

Another view of OLS with two predictors

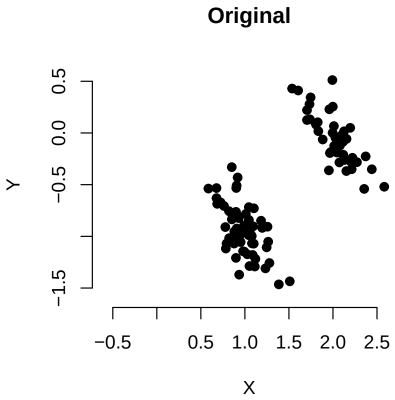
- "Partialling out" OLS recipe:
  1. Run regression of X_i on Z_i:
  $$\widehat{X}_i = \widehat{\delta}_0 + \widehat{\delta}_1 Z_i$$
  2. Calculate residuals from this regression:
  $$\widehat{r}_{xz,i} = X_i - \widehat{X}_i$$
  3. Run a simple regression of Y_i on residuals, $\widehat{r}_{xz,i}$:
  $$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 \widehat{r}_{xz,i}$$

Another view of OLS with two predictors

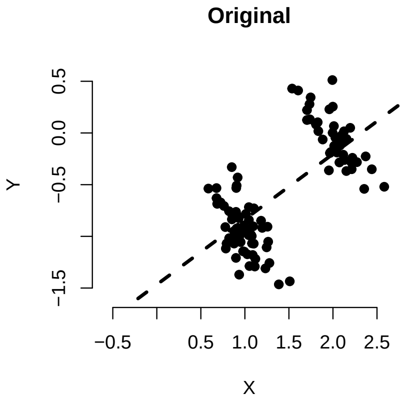
- "Partialling out" OLS recipe:
  1. Run regression of X_i on Z_i:
  $$\widehat{X}_i = \widehat{\delta}_0 + \widehat{\delta}_1 Z_i$$
  2. Calculate residuals from this regression:
  $$\widehat{r}_{xz,i} = X_i - \widehat{X}_i$$
  3. Run a simple regression of Y_i on residuals, $\widehat{r}_{xz,i}$:
  $$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 \widehat{r}_{xz,i}$$

- Estimate of $\widehat{\beta}_1$ will be the same as running:
$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{\beta}_2 Z_i$$

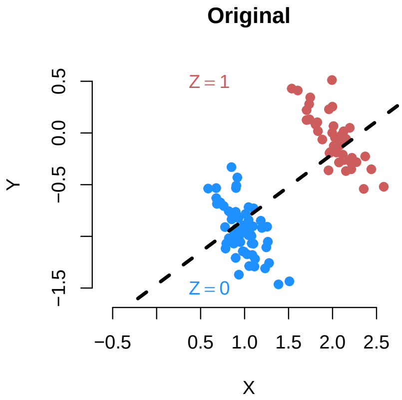
A Visual of Partialling Out



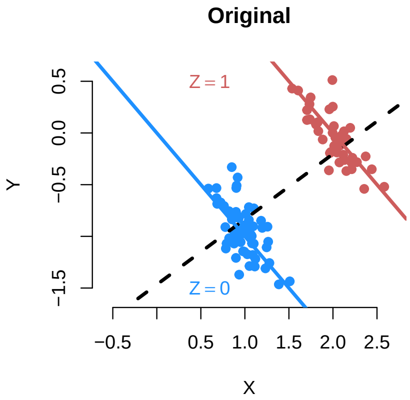
A Visual of Partialling Out



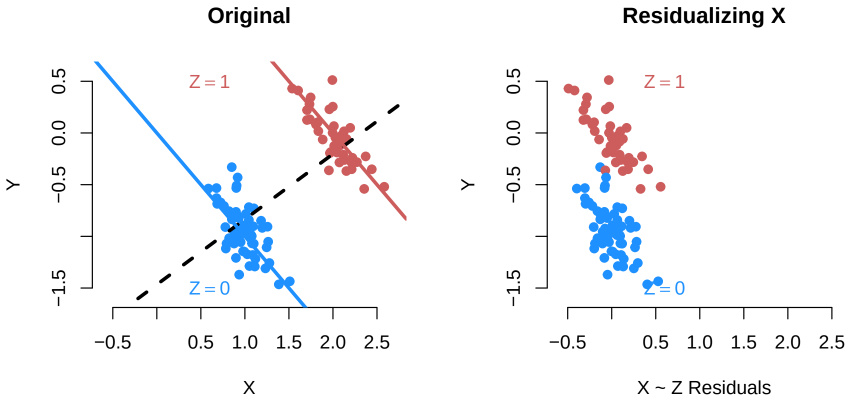
A Visual of Partialling Out



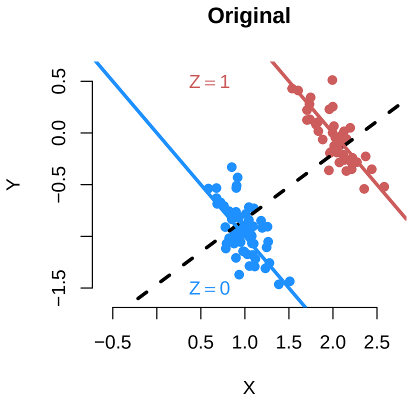
A Visual of Partialling Out

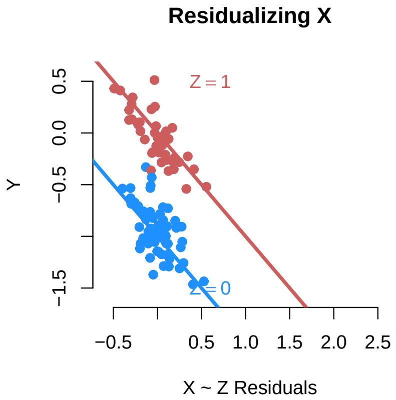


A Visual of Partialling Out

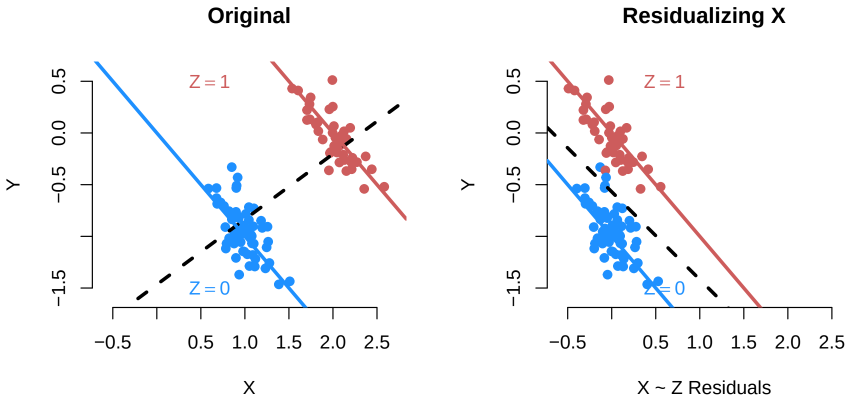


A Visual of Partialling Out

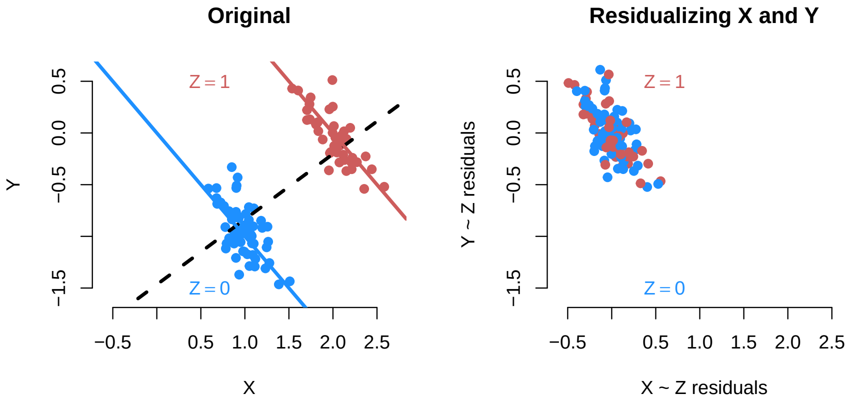




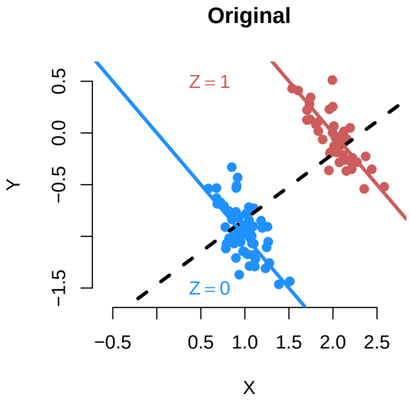
A Visual of Partialling Out

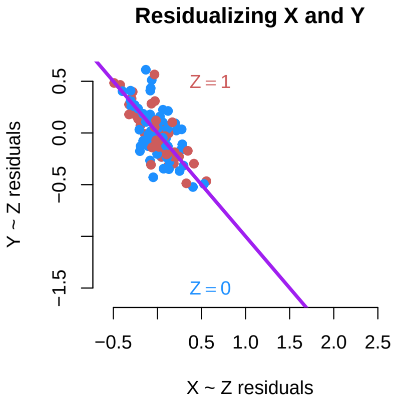


A Visual of Partialling Out

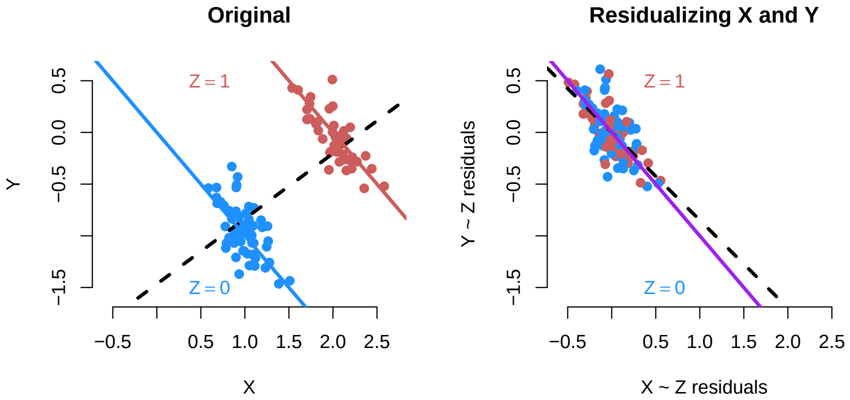


A Visual of Partialling Out





A Visual of Partialling Out



Origin of the Partial Out Recipe

Assume $Y = \beta_0 + \beta_1 X + \beta_2 Z + u$. Another way to write the OLS estimator is:

$$\hat{\beta}_1 = \frac{\sum_i^n \hat{r}_{xz,i}\, y_i}{\sum_i^n \hat{r}_{xz,i}^2}$$

where $\hat{r}_{xz,i}$ are the residuals from the regression of X on Z:

$$X = \lambda + \delta Z + r_{xz}$$

In other words, both of these regressions yield identical estimates $\hat{\beta}_1$:

$$y = \hat{\gamma}_0 + \hat{\beta}_1 \hat{r}_{xz} \quad \text{and} \quad y = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 z$$

Origin of the Partial Out Recipe

Assume $Y = \beta_0 + \beta_1 X + \beta_2 Z + u$. Another way to write the OLS estimator is:

$$\hat{\beta}_1 = \frac{\sum_i^n \hat{r}_{xz,i}\, y_i}{\sum_i^n \hat{r}_{xz,i}^2}$$

where $\hat{r}_{xz,i}$ are the residuals from the regression of X on Z:

$$X = \lambda + \delta Z + r_{xz}$$

In other words, both of these regressions yield identical estimates $\hat{\beta}_1$:

$$y = \hat{\gamma}_0 + \hat{\beta}_1 \hat{r}_{xz} \quad \text{and} \quad y = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 z$$

- δ measures the correlation between X and Z.

Origin of the Partial Out Recipe

Assume $Y = \beta_0 + \beta_1 X + \beta_2 Z + u$. Another way to write the OLS estimator is:

$$\hat{\beta}_1 = \frac{\sum_i^n \hat{r}_{xz,i}\, y_i}{\sum_i^n \hat{r}_{xz,i}^2}$$

where $\hat{r}_{xz,i}$ are the residuals from the regression of X on Z:

$$X = \lambda + \delta Z + r_{xz}$$

In other words, both of these regressions yield identical estimates $\hat{\beta}_1$:

$$y = \hat{\gamma}_0 + \hat{\beta}_1 \hat{r}_{xz} \quad \text{and} \quad y = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 z$$

- δ measures the correlation between X and Z.
- Residuals $\hat{r}_{xz}$ are the part of X that is uncorrelated with Z. Put differently, $\hat{r}_{xz}$ is X, after the effect of Z on X has been partialled out or netted out.

Origin of the Partial Out Recipe

Assume $Y = \beta_0 + \beta_1 X + \beta_2 Z + u$. Another way to write the OLS estimator is:

$$\hat{\beta}_1 = \frac{\sum_i^n \hat{r}_{xz,i}\, y_i}{\sum_i^n \hat{r}_{xz,i}^2}$$

where $\hat{r}_{xz,i}$ are the residuals from the regression of X on Z:

$$X = \lambda + \delta Z + r_{xz}$$

In other words, both of these regressions yield identical estimates $\hat{\beta}_1$:

$$y = \hat{\gamma}_0 + \hat{\beta}_1 \hat{r}_{xz} \quad \text{and} \quad y = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 z$$

- δ measures the correlation between X and Z.
- Residuals $\hat{r}_{xz}$ are the part of X that is uncorrelated with Z. Put differently, $\hat{r}_{xz}$ is X, after the effect of Z on X has been partialled out or netted out.
- Can use same equation with k explanatory variables; $\hat{r}_{xz}$ will then come from a regression of X on all the other explanatory variables.

Why Should We Care About Partialling Out?

- Won't `R` just calculate it for us?

Why Should We Care About Partialling Out?

- Won't `R` just calculate it for us?
- Sure—but the partialling out strategy provides great intuition for what the regression controlling for another variable is doing.

Why Should We Care About Partialling Out?

- Won't `R` just calculate it for us?
- Sure—but the partialling out strategy provides great intuition for what the regression controlling for another variable is doing.
- This set up will also be the basis of diagnostic plots that we will cover in a couple of weeks. It allows us to visualize the conditional relationship.

Why Should We Care About Partialling Out?

- Won't `R` just calculate it for us?
- Sure—but the partialling out strategy provides great intuition for what the regression controlling for another variable is doing.
- This set up will also be the basis of diagnostic plots that we will cover in a couple of weeks. It allows us to visualize the conditional relationship.
- Finally, it forms the foundation of a number of machine learning strategies including double machine learning by breaking down the regression problem.

We Covered

We Covered

- Estimation and inference for the regression model with 2 variables.

We Covered

- Estimation and inference for the regression model with 2 variables.
- Partialling out strategy.

We Covered

- Estimation and inference for the regression model with 2 variables.
- Partialling out strategy.

Next Time: Omitted Variables and Multicollinearity

Where We've Been and Where We're Going...

Where We've Been and Where We're Going...

- Last Week
  - mechanics of OLS with one variable
  - properties of OLS
- This Week
  - adding a second variable
  - new mechanics
  - omitted variable bias
  - multicollinearity
  - interactions
- Next Week
  - multiple regression
- Long Run
  - probability $\rightarrow$ inference $\rightarrow$ regression $\rightarrow$ causal inference

Remember This?

- **Identification / Data Description**
  - Variation in X
- **Unbiasedness / Consistency**
  - Variation in X
  - Random Sampling
  - Linearity in Parameters
  - Zero Conditional Mean
- **Gauss-Markov (BLUE) / Asymptotic Inference (z and χ^2)**
  - Variation in X
  - Random Sampling
  - Linearity in Parameters
  - Zero Conditional Mean
  - Homoskedasticity
- **Classical LM (BUE) / Small-Sample Inference (t and F)**
  - Variation in X
  - Random Sampling
  - Linearity in Parameters
  - Zero Conditional Mean
  - Homoskedasticity
  - Normality of Errors

Unbiasedness revisited

- True model:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i$$

Unbiasedness revisited

- True model:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i$$

- Assumptions 1-4 $\Rightarrow$ we get unbiased estimates of the coefficients

Unbiasedness revisited

- True model:
$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i$$
- Assumptions 1-4 $\Rightarrow$ we get unbiased estimates of the coefficients
- What happens if we ignore the Z_i and just run the simple linear regression with just X_i?

Unbiasedness revisited

- True model:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i$$

- Assumptions 1-4 $\Rightarrow$ we get unbiased estimates of the coefficients
- What happens if we ignore the Z_i and just run the simple linear regression with just X_i?
- Misspecified model:

$$Y_i = \alpha_0 + \alpha_1 X_i + u_i^*$$

Unbiasedness revisited

- True model:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i$$

- Assumptions 1-4 $\Rightarrow$ we get unbiased estimates of the coefficients
- What happens if we ignore the Z_i and just run the simple linear regression with just X_i?
- Misspecified model:

$$Y_i = \alpha_0 + \alpha_1 X_i + u_i^*$$

- $\hat{\alpha}_1$ is the alternative estimator for β_1 when we fail to control for Z_i.

Unbiasedness revisited

- True model:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i$$

- Assumptions 1-4 $\Rightarrow$ we get unbiased estimates of the coefficients
- What happens if we ignore the Z_i and just run the simple linear regression with just X_i?
- Misspecified model:

$$Y_i = \alpha_0 + \alpha_1 X_i + u_i^*$$

- $\hat{\alpha}_1$ is the alternative estimator for β_1 when we fail to control for Z_i.
- OLS estimates from the misspecified model:

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

Omitted Variable Bias: Simple Case

True Population Model:

$$\text{Voted Republican} = \beta_0 + \beta_1 \text{Watch Fox News} + \beta_2 \text{Strong Republican} + u$$

Omitted Variable Bias: Simple Case

True Population Model:

$$\text{Voted Republican} = \beta_0 + \beta_1 \text{Watch Fox News} + \beta_2 \text{Strong Republican} + u$$

Underspecified Model that we use:

$$\text{Voted} \widehat{\text{Republican}} = \hat{\alpha}_0 + \hat{\alpha}_1 \text{Watch Fox News}$$

Omitted Variable Bias: Simple Case

True Population Model:

$$\text{Voted Republican} = \beta_0 + \beta_1 \text{Watch Fox News} + \beta_2 \text{Strong Republican} + u$$

Underspecified Model that we use:

$$\widehat{\text{Voted Republican}} = \hat{\alpha}_0 + \hat{\alpha}_1 \text{Watch Fox News}$$

Expected Behavior: $\hat{\alpha}_1$ is upward biased for β_1 since being a strong Republican is positively correlated with both watching Fox News and voting Republican. We have $E[\hat{\alpha}_1] > \beta_1$.

Omitted Variable Bias: Simple Case

Omitted Variable Bias: Simple Case

True Population Model:

$$\text{Survival} = \beta_0 + \beta_1 \text{Hospitalized} + \beta_2 \text{Health} + u$$

Omitted Variable Bias: Simple Case

True Population Model:

$$\text{Survival} = \beta_0 + \beta_1 \text{Hospitalized} + \beta_2 \text{Health} + u$$

Under-specified Model that we use:

$$\widehat{\text{Survival}} = \hat{\alpha}_0 + \hat{\alpha}_1 \text{Hospitalized}$$

Omitted Variable Bias: Simple Case

True Population Model:

$$\text{Survival} = \beta_0 + \beta_1 \text{Hospitalized} + \beta_2 \text{Health} + u$$

Under-specified Model that we use:

$$\widehat{\text{Survival}} = \hat{\alpha}_0 + \hat{\alpha}_1 \text{Hospitalized}$$

Expected Behavior: The negative coefficient $\hat{\alpha}_1$ is downward biased compared to the true β_1 so $E[\hat{\alpha}_1] < \beta_1$. Being hospitalized is negatively correlated with health, and health is positively correlated with survival.

Omitted Variable Bias: Simple Case

True Population Model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$$

Omitted Variable Bias: Simple Case

True Population Model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$$

Underspecified Model that we use:

$$\hat{y} = \hat{\alpha}_0 + \hat{\alpha}_1 x_1$$

Omitted Variable Bias: Simple Case

True Population Model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$$

Underspecified Model that we use:

$$\hat{y} = \hat{\alpha}_0 + \hat{\alpha}_1 x_1$$

We can show that for the same sample, the relationship between $\hat{\alpha}_1$ and $\hat{\beta}_1$ is:

$$\hat{\alpha}_1 = \hat{\beta}_1 + \hat{\beta}_2 \cdot \tilde{\delta}$$

where:

Omitted Variable Bias: Simple Case

True Population Model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$$

Underspecified Model that we use:

$$\hat{y} = \hat{\alpha}_0 + \hat{\alpha}_1 x_1$$

We can show that for the same sample, the relationship between $\hat{\alpha}_1$ and $\hat{\beta}_1$ is:

$$\hat{\alpha}_1 = \hat{\beta}_1 + \hat{\beta}_2 \cdot \tilde{\delta}$$

where:

- $\tilde{\delta}$ is the slope of a regression of x_2 on x_1. If $\tilde{\delta} > 0$ then $\text{cor}(x_1, x_2) > 0$ and if $\tilde{\delta} < 0$ then $\text{cor}(x_1, x_2) < 0$.

Omitted Variable Bias: Simple Case

True Population Model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$$

Underspecified Model that we use:

$$\hat{y} = \hat{\alpha}_0 + \hat{\alpha}_1 x_1$$

We can show that for the same sample, the relationship between $\hat{\alpha}_1$ and $\hat{\beta}_1$ is:

$$\hat{\alpha}_1 = \hat{\beta}_1 + \hat{\beta}_2 \cdot \tilde{\delta}$$

where:

- $\tilde{\delta}$ is the slope of a regression of x_2 on x_1. If $\tilde{\delta} > 0$ then $\text{cor}(x_1, x_2) > 0$ and if $\tilde{\delta} < 0$ then $\text{cor}(x_1, x_2) < 0$.
- $\hat{\beta}_2$ is from the true regression and measures the relationship between x_2 and y, conditional on x_1.

Omitted Variable Bias: Simple Case

True Population Model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$$

Underspecified Model that we use:

$$\hat{y} = \hat{\alpha}_0 + \hat{\alpha}_1 x_1$$

We can show that for the same sample, the relationship between $\hat{\alpha}_1$ and $\hat{\beta}_1$ is:

$$\hat{\alpha}_1 = \hat{\beta}_1 + \hat{\beta}_2 \cdot \tilde{\delta}$$

where:

- $\tilde{\delta}$ is the slope of a regression of x_2 on x_1. If $\tilde{\delta} > 0$ then $\text{cor}(x_1, x_2) > 0$ and if $\tilde{\delta} < 0$ then $\text{cor}(x_1, x_2) < 0$.
- $\hat{\beta}_2$ is from the true regression and measures the relationship between x_2 and y, conditional on x_1.

$\hat{\alpha}_1 = \hat{\beta}_1$ when $\tilde{\delta} = 0$ or $\hat{\beta}_2 = 0$.

Omitted Variable Bias: Simple Case

We take expectations to see what the bias will be:

$$\begin{aligned}
\hat{\alpha}_1 &= \hat{\beta}_1 + \hat{\beta}_2 \cdot \tilde{\delta} \\
E[\hat{\alpha}_1 \mid X] &=
\end{aligned}$$

Omitted Variable Bias: Simple Case

We take expectations to see what the bias will be:

$$\begin{aligned}
\hat{\alpha}_1 &= \hat{\beta}_1 + \hat{\beta}_2 \cdot \tilde{\delta} \\
E[\hat{\alpha}_1 \mid X] &= E[\hat{\beta}_1 + \hat{\beta}_2 \cdot \tilde{\delta} \mid X] \\
&=
\end{aligned}$$

Omitted Variable Bias: Simple Case

We take expectations to see what the bias will be:

$$\begin{aligned}
\hat{\alpha}_1 &= \hat{\beta}_1 + \hat{\beta}_2 \cdot \tilde{\delta} \\
E[\hat{\alpha}_1 \mid X] &= E[\hat{\beta}_1 + \hat{\beta}_2 \cdot \tilde{\delta} \mid X] \\
&= E[\hat{\beta}_1 \mid X] + E[\hat{\beta}_2 \mid X] \cdot \tilde{\delta} \ (\tilde{\delta} \text{ nonrandom given } x) \\
&=
\end{aligned}$$

Omitted Variable Bias: Simple Case

We take expectations to see what the bias will be:

$$
\begin{aligned}
\hat{\alpha}_1 &= \hat{\beta}_1 + \hat{\beta}_2 \cdot \tilde{\delta} \\
E[\hat{\alpha}_1 \mid X] &= E[\hat{\beta}_1 + \hat{\beta}_2 \cdot \tilde{\delta} \mid X] \\
&= E[\hat{\beta}_1 \mid X] + E[\hat{\beta}_2 \mid X] \cdot \tilde{\delta} \text{ } (\tilde{\delta} \text{ nonrandom given } x) \\
&= \beta_1 + \beta_2 \cdot \tilde{\delta} \text{ (given assumptions 1-4)}
\end{aligned}
$$

Omitted Variable Bias: Simple Case

We take expectations to see what the bias will be:

$$
\begin{aligned}
\hat{\alpha}_1 &= \hat{\beta}_1 + \hat{\beta}_2 \cdot \tilde{\delta} \\
E[\hat{\alpha}_1 \mid X] &= E[\hat{\beta}_1 + \hat{\beta}_2 \cdot \tilde{\delta} \mid X] \\
&= E[\hat{\beta}_1 \mid X] + E[\hat{\beta}_2 \mid X] \cdot \tilde{\delta} \text{ } (\tilde{\delta} \text{ nonrandom given } x) \\
&= \beta_1 + \beta_2 \cdot \tilde{\delta} \text{ (given assumptions 1-4)}
\end{aligned}
$$

So

$$
\text{Bias}[\hat{\alpha}_1 \mid X] = E[\hat{\alpha}_1 \mid X] - \beta_1 = \beta_2 \cdot \tilde{\delta}
$$

Omitted Variable Bias: Simple Case

Formula:

$$\text{Bias}[\hat{\alpha}_1 \mid X] = \beta_2 \cdot \tilde{\delta}$$

Omitted Variable Bias: Simple Case

Formula:

$$\text{Bias}[\hat{\alpha}_1 \mid X] = \beta_2 \cdot \tilde{\delta}$$

Cinelli and Hazlett (2018) describe this as:

impact times its imbalance

Omitted Variable Bias: Simple Case

Formula:

$$\text{Bias}[\hat{\alpha}_1 \mid X] = \beta_2 \cdot \tilde{\delta}$$

Cinelli and Hazlett (2018) describe this as:

impact times its imbalance

- impact is how looking at different subgroups of the unobserved X_2 'impacts' our best linear prediction of the outcome.

Omitted Variable Bias: Simple Case

Formula:

$$\text{Bias}[\hat{\alpha}_1 \mid X] = \beta_2 \cdot \tilde{\delta}$$

Cinelli and Hazlett (2018) describe this as:

impact times its imbalance

- impact is how looking at different subgroups of the unobserved X_2 'impacts' our best linear prediction of the outcome.
- imbalance is how the expectation of the unobserved X_2 varies across levels of X_1.

Omitted Variable Bias: Simple Case

Direction of the bias of $\hat{\alpha}_1$ compared to β_1 is given by:

	$\text{cov}(X_1, X_2) > 0$	$\text{cov}(X_1, X_2) < 0$	$\text{cov}(X_1, X_2) = 0$
$\beta_2 > 0$	Positive bias	Negative Bias	No bias
$\beta_2 < 0$	Negative bias	Positive Bias	No bias
$\beta_2 = 0$	No bias	No bias	No bias

Omitted Variable Bias: Simple Case

Direction of the bias of $\hat{\alpha}_1$ compared to β_1 is given by:

	$\text{cov}(X_1, X_2) > 0$	$\text{cov}(X_1, X_2) < 0$	$\text{cov}(X_1, X_2) = 0$
$\beta_2 > 0$	Positive bias	Negative Bias	No bias
$\beta_2 < 0$	Negative bias	Positive Bias	No bias
$\beta_2 = 0$	No bias	No bias	No bias

Further points:

- Magnitude of the bias matters too

Omitted Variable Bias: Simple Case

Direction of the bias of $\hat{\alpha}_1$ compared to β_1 is given by:

	$\text{cov}(X_1, X_2) > 0$	$\text{cov}(X_1, X_2) < 0$	$\text{cov}(X_1, X_2) = 0$
$\beta_2 > 0$	Positive bias	Negative Bias	No bias
$\beta_2 < 0$	Negative bias	Positive Bias	No bias
$\beta_2 = 0$	No bias	No bias	No bias

Further points:

- Magnitude of the bias matters too
- The Omitted Variable Bias formula provides the foundation for many forms of sensitivity analysis.

Omitted Variable Bias: Simple Case

Direction of the bias of $\hat{\alpha}_1$ compared to β_1 is given by:

	$\text{cov}(X_1, X_2) > 0$	$\text{cov}(X_1, X_2) < 0$	$\text{cov}(X_1, X_2) = 0$
$\beta_2 > 0$	Positive bias	Negative Bias	No bias
$\beta_2 < 0$	Negative bias	Positive Bias	No bias
$\beta_2 = 0$	No bias	No bias	No bias

Further points:

- Magnitude of the bias matters too
- The Omitted Variable Bias formula provides the foundation for many forms of sensitivity analysis.
- In the more general case with more than two covariates the bias is more difficult to discern. It depends on all the pairwise correlations.

Sampling variance for simple linear regression

- Under simple linear regression, we found that the distribution of the slope was the following:

$$\text{var}(\widehat{\beta}_1) = \frac{\sigma_u^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

Sampling variance for simple linear regression

- Under simple linear regression, we found that the distribution of the slope was the following:

$$\text{var}(\widehat{\beta}_1) = \frac{\sigma_u^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

- Factors affecting the standard errors (the square root of these sampling variances):

Sampling variance for simple linear regression

- Under simple linear regression, we found that the distribution of the slope was the following:

$$\text{var}(\widehat{\beta}_1) = \frac{\sigma_u^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

- Factors affecting the standard errors (the square root of these sampling variances):
  - The error variance σ_u^2 (higher conditional variance of Y_i leads to bigger SEs)

Sampling variance for simple linear regression

- Under simple linear regression, we found that the distribution of the slope was the following:

$$\text{var}(\widehat{\beta}_1) = \frac{\sigma_u^2}{\sum_{i=1}^n (X_i - \overline{X})^2}$$

- Factors affecting the standard errors (the square root of these sampling variances):
  - The error variance σ_u^2 (higher conditional variance of Y_i leads to bigger SEs)
  - The total variation in X_i: $\sum_{i=1}^n (X_i - \overline{X})^2$ (lower variation in X_i leads to bigger SEs)

Sampling variation for linear regression with two covariates

- Regression with an additional independent variable:

$$\text{var}(\widehat{\beta}_1) = \frac{\sigma_u^2}{(1 - R_1^2)\sum_{i=1}^n (X_i - \overline{X})^2}$$

Sampling variation for linear regression with two covariates

- Regression with an additional independent variable:

$$\text{var}(\widehat{\beta}_1) = \frac{\sigma_u^2}{(1 - R_1^2)\sum_{i=1}^n (X_i - \overline{X})^2}$$

- Here, R_1^2 is the R^2 from the regression of X_i on Z_i:

$$\widehat{X}_i = \widehat{\delta}_0 + \widehat{\delta}_1 Z_i$$

Sampling variation for linear regression with two covariates

- Regression with an additional independent variable:

$$\text{var}(\widehat{\beta}_1) = \frac{\sigma_u^2}{(1 - R_1^2)\sum_{i=1}^n (X_i - \overline{X})^2}$$

- Here, R_1^2 is the R^2 from the regression of X_i on Z_i:

$$\widehat{X}_i = \widehat{\delta}_0 + \widehat{\delta}_1 Z_i$$

- Factors now affecting the standard errors:

Sampling variation for linear regression with two covariates

- Regression with an additional independent variable:

$$\text{var}(\widehat{\beta}_1) = \frac{\sigma_u^2}{(1 - R_1^2)\sum_{i=1}^n (X_i - \overline{X})^2}$$

- Here, R_1^2 is the R^2 from the regression of X_i on Z_i:

$$\widehat{X}_i = \widehat{\delta}_0 + \widehat{\delta}_1 Z_i$$

- Factors now affecting the standard errors:
  - The error variance (higher conditional variance of Y_i leads to bigger SEs)

Sampling variation for linear regression with two covariates

- Regression with an additional independent variable:

$$\text{var}(\widehat{\beta}_1) = \frac{\sigma_u^2}{(1 - R_1^2)\sum_{i=1}^n (X_i - \overline{X})^2}$$

- Here, R_1^2 is the R^2 from the regression of X_i on Z_i:

$$\widehat{X}_i = \widehat{\delta}_0 + \widehat{\delta}_1 Z_i$$

- Factors now affecting the standard errors:
  - The error variance (higher conditional variance of Y_i leads to bigger SEs)
  - The total variation of X_i (lower variation in X_i leads to bigger SEs)

Sampling variation for linear regression with two covariates

- Regression with an additional independent variable:

$$\text{var}(\widehat{\beta}_1) = \frac{\sigma_u^2}{(1 - R_1^2)\sum_{i=1}^n (X_i - \overline{X})^2}$$

- Here, R_1^2 is the R^2 from the regression of X_i on Z_i:

$$\widehat{X}_i = \widehat{\delta}_0 + \widehat{\delta}_1 Z_i$$

- Factors now affecting the standard errors:
  - The error variance (higher conditional variance of Y_i leads to bigger SEs)
  - The total variation of X_i (lower variation in X_i leads to bigger SEs)
  - The strength of the linear relationship between X_i and Z_i (stronger relationships mean higher R_1^2 and thus bigger SEs)

Sampling variation for linear regression with two covariates

- Regression with an additional independent variable:

$$\text{var}(\widehat{\beta}_1) = \frac{\sigma_u^2}{(1 - R_1^2)\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

- Here, R_1^2 is the R^2 from the regression of X_i on Z_i:

$$\widehat{X}_i = \widehat{\delta}_0 + \widehat{\delta}_1 Z_i$$

- Factors now affecting the standard errors:
  - The error variance (higher conditional variance of Y_i leads to bigger SEs)
  - The total variation of X_i (lower variation in X_i leads to bigger SEs)
  - The strength of the linear relationship between X_i and Z_i (stronger relationships mean higher R_1^2 and thus bigger SEs)
- What happens with perfect collinearity? $R_1^2 = 1$ and the variances are infinite.

Multicollinearity

Multicollinearity

Definition

Multicollinearity is defined to be high, but not perfect, correlation between two independent variables in a regression.

Multicollinearity

Definition

Multicollinearity is defined to be high, but not perfect, correlation between two independent variables in a regression.

- With multicollinearity, we'll have $R_1^2 \approx 1$, but not exactly.

Multicollinearity

Definition

Multicollinearity is defined to be high, but not perfect, correlation between two independent variables in a regression.

- With multicollinearity, we'll have $R_1^2 \approx 1$, but not exactly.
- The stronger the relationship between X_i and Z_i, the closer the R_1^2 will be to 1, and the higher the SEs will be:

$$\text{var}(\widehat{\beta}_1) = \frac{\sigma_u^2}{(1 - R_1^2)\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

Multicollinearity

Definition

Multicollinearity is defined to be high, but not perfect, correlation between two independent variables in a regression.

- With multicollinearity, we'll have $R_1^2 \approx 1$, but not exactly.
- The stronger the relationship between X_i and Z_i, the closer the R_1^2 will be to 1, and the higher the SEs will be:

$$\text{var}(\widehat{\beta}_1) = \frac{\sigma_u^2}{(1 - R_1^2)\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

- Given the symmetry, it will also increase $\text{var}(\widehat{\beta}_2)$ as well.

Intuition for multicollinearity

- Remember the OLS recipe:

Intuition for multicollinearity

- Remember the OLS recipe:
  - $\widehat{\beta}_1$ from regression of Y_i on $\widehat{r}_{xz,i}$

Intuition for multicollinearity

- Remember the OLS recipe:
  - $\widehat{\beta}_1$ from regression of Y_i on $\widehat{r}_{xz,i}$
  - $\widehat{r}_{xz,i}$ are the residuals from the regression of X_i on Z_i

Intuition for multicollinearity

- Remember the OLS recipe:
  - $\widehat{\beta}_1$ from regression of Y_i on $\widehat{r}_{xz,i}$
  - $\widehat{r}_{xz,i}$ are the residuals from the regression of X_i on Z_i
- Estimated coefficient:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} \widehat{r}_{xz,i} Y_i}{\sum_{i=1}^{n} \widehat{r}_{xz,i}^2}$$

Intuition for multicollinearity

- Remember the OLS recipe:
  - $\widehat{\beta}_1$ from regression of Y_i on $\widehat{r}_{xz,i}$
  - $\widehat{r}_{xz,i}$ are the residuals from the regression of X_i on Z_i
- Estimated coefficient:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} \widehat{r}_{xz,i} Y_i}{\sum_{i=1}^{n} \widehat{r}_{xz,i}^2}$$

- When Z_i and X_i have a strong relationship, then the residuals will have low variation

Intuition for multicollinearity

- Remember the OLS recipe:
  - $\widehat{\beta}_1$ from regression of Y_i on $\widehat{r}_{xz,i}$
  - $\widehat{r}_{xz,i}$ are the residuals from the regression of X_i on Z_i
- Estimated coefficient:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n \widehat{r}_{xz,i} Y_i}{\sum_{i=1}^n \widehat{r}_{xz,i}^2}$$

- When Z_i and X_i have a strong relationship, then the residuals will have low variation
- We explain away a lot of the variation in X_i through Z_i.

Intuition for multicollinearity

- Remember the OLS recipe:
  - $\widehat{\beta}_1$ from regression of Y_i on $\widehat{r}_{xz,i}$
  - $\widehat{r}_{xz,i}$ are the residuals from the regression of X_i on Z_i
- Estimated coefficient:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} \widehat{r}_{xz,i} Y_i}{\sum_{i=1}^{n} \widehat{r}_{xz,i}^2}$$

- When Z_i and X_i have a strong relationship, then the residuals will have low variation
- We explain away a lot of the variation in X_i through Z_i.
- Low variation in an independent variable (here, $\widehat{r}_{xz,i}$) $\rightsquigarrow$ high SEs

Intuition for multicollinearity

- Remember the OLS recipe:
  - $\widehat{\beta}_1$ from regression of Y_i on $\widehat{r}_{xz,i}$
  - $\widehat{r}_{xz,i}$ are the residuals from the regression of X_i on Z_i
- Estimated coefficient:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n \widehat{r}_{xz,i} Y_i}{\sum_{i=1}^n \widehat{r}_{xz,i}^2}$$

- When Z_i and X_i have a strong relationship, then the residuals will have low variation
- We explain away a lot of the variation in X_i through Z_i.
- Low variation in an independent variable (here, $\widehat{r}_{xz,i}$) $\rightsquigarrow$ high SEs
- Basically, there is less residual variation left in X_i after "partialling out" the effect of Z_i

Effects of multicollinearity

- No effect on the bias of OLS.

Effects of multicollinearity

- No effect on the bias of OLS.
- Only increases the standard errors.

Effects of multicollinearity

- No effect on the bias of OLS.
- Only increases the standard errors.
- Really just a sample size problem:

Effects of multicollinearity

- No effect on the bias of OLS.
- Only increases the standard errors.
- Really just a sample size problem:
  - If X_i and Z_i are extremely highly correlated, you're going to need a much bigger sample to accurately differentiate between their effects.

Effects of multicollinearity

- No effect on the bias of OLS.
- Only increases the standard errors.
- Really just a sample size problem:
  - If X_i and Z_i are extremely highly correlated, you're going to need a much bigger sample to accurately differentiate between their effects.

Effects of multicollinearity

- No effect on the bias of OLS.
- Only increases the standard errors.
- Really just a sample size problem:
  - If X_i and Z_i are extremely highly correlated, you're going to need a much bigger sample to accurately differentiate between their effects.

How Do We Detect Multicollinearity?

- The best practice is to directly compute $\text{Cor}(X_1, X_2)$ before running your regression.

How Do We Detect Multicollinearity?

- The best practice is to directly compute $\text{Cor}(X_1, X_2)$ before running your regression.
- But you might (and probably will) forget to do so. Even then, you can detect multicollinearity from your regression result:

How Do We Detect Multicollinearity?

- The best practice is to directly compute $\text{Cor}(X_1, X_2)$ before running your regression.
- But you might (and probably will) forget to do so. Even then, you can detect multicollinearity from your regression result:
  - Large changes in the estimated regression coefficients when a predictor variable is added or deleted

How Do We Detect Multicollinearity?

- The best practice is to directly compute $\text{Cor}(X_1, X_2)$ before running your regression.
- But you might (and probably will) forget to do so. Even then, you can detect multicollinearity from your regression result:
  - Large changes in the estimated regression coefficients when a predictor variable is added or deleted
  - Lack of statistical significance despite high R^2

How Do We Detect Multicollinearity?

- The best practice is to directly compute $\text{Cor}(X_1, X_2)$ before running your regression.
- But you might (and probably will) forget to do so. Even then, you can detect multicollinearity from your regression result:
  - Large changes in the estimated regression coefficients when a predictor variable is added or deleted
  - Lack of statistical significance despite high R^2
  - Estimated regression coefficients have an opposite sign from predicted

How Do We Detect Multicollinearity?

- The best practice is to directly compute $\text{Cor}(X_1, X_2)$ before running your regression.
- But you might (and probably will) forget to do so. Even then, you can detect multicollinearity from your regression result:
  - Large changes in the estimated regression coefficients when a predictor variable is added or deleted
  - Lack of statistical significance despite high R^2
  - Estimated regression coefficients have an opposite sign from predicted
- A more formal indicator is the variance inflation factor (VIF):

$$VIF(\beta_j) = \frac{1}{1 - R_j^2}$$

which measures how much $V[\hat{\beta}_j \mid X]$ is inflated compared to a (hypothetical) uncorrelated data. (where R_j^2 is the coefficient of determination from the partialing out equation)

How Do We Detect Multicollinearity?

- The best practice is to directly compute $\text{Cor}(X_1, X_2)$ before running your regression.
- But you might (and probably will) forget to do so. Even then, you can detect multicollinearity from your regression result:
  - Large changes in the estimated regression coefficients when a predictor variable is added or deleted
  - Lack of statistical significance despite high R^2
  - Estimated regression coefficients have an opposite sign from predicted
- A more formal indicator is the variance inflation factor (VIF):

$$VIF(\beta_j) = \frac{1}{1 - R_j^2}$$

which measures how much $V[\hat{\beta}_j \mid X]$ is inflated compared to a (hypothetical) uncorrelated data. (where R_j^2 is the coefficient of determination from the partialing out equation)
In R, `vif()` in the `car` package.

So How Should I Think about Multicollinearity?

- Multicollinearity does NOT lead to bias; estimates will be unbiased and consistent.

So How Should I Think about Multicollinearity?

- Multicollinearity does NOT lead to bias; estimates will be unbiased and consistent.
- Multicollinearity should in fact be seen as a problem of micronumerosity, or "too little data." You can't ask the OLS estimator to distinguish the partial effects of X_1 and X_2 if they are essentially the same.

So How Should I Think about Multicollinearity?

- Multicollinearity does NOT lead to bias; estimates will be unbiased and consistent.
- Multicollinearity should in fact be seen as a problem of micronumerosity, or "too little data." You can't ask the OLS estimator to distinguish the partial effects of X_1 and X_2 if they are essentially the same.
- If X_1 and X_2 are almost the same, why would you want a unique β_1 and a unique β_2? Think about how you would interpret that?

So How Should I Think about Multicollinearity?

- Multicollinearity does NOT lead to bias; estimates will be unbiased and consistent.
- Multicollinearity should in fact be seen as a problem of micronumerosity, or "too little data." You can't ask the OLS estimator to distinguish the partial effects of X_1 and X_2 if they are essentially the same.
- If X_1 and X_2 are almost the same, why would you want a unique β_1 and a unique β_2? Think about how you would interpret that?
- Relax, you got way more important things to worry about!

So How Should I Think about Multicollinearity?

- Multicollinearity does NOT lead to bias; estimates will be unbiased and consistent.
- Multicollinearity should in fact be seen as a problem of micronumerosity, or "too little data." You can't ask the OLS estimator to distinguish the partial effects of X_1 and X_2 if they are essentially the same.
- If X_1 and X_2 are almost the same, why would you want a unique β_1 and a unique β_2? Think about how you would interpret that?
- Relax, you got way more important things to worry about!
- If possible, get more data

So How Should I Think about Multicollinearity?

- Multicollinearity does NOT lead to bias; estimates will be unbiased and consistent.
- Multicollinearity should in fact be seen as a problem of micronumerosity, or "too little data." You can't ask the OLS estimator to distinguish the partial effects of X_1 and X_2 if they are essentially the same.
- If X_1 and X_2 are almost the same, why would you want a unique β_1 and a unique β_2? Think about how you would interpret that?
- Relax, you got way more important things to worry about!
- If possible, get more data
- Drop one of the variables, or combine them

So How Should I Think about Multicollinearity?

- Multicollinearity does NOT lead to bias; estimates will be unbiased and consistent.
- Multicollinearity should in fact be seen as a problem of micronumerosity, or "too little data." You can't ask the OLS estimator to distinguish the partial effects of X_1 and X_2 if they are essentially the same.
- If X_1 and X_2 are almost the same, why would you want a unique β_1 and a unique β_2? Think about how you would interpret that?
- Relax, you got way more important things to worry about!
- If possible, get more data
- Drop one of the variables, or combine them
- Or maybe linear regression is not the right tool

We Covered

We Covered

- Two challenges:

We Covered

- Two challenges:
  - omitted variable bias

We Covered

- Two challenges:
  - omitted variable bias
  - multicollinearity

We Covered

- Two challenges:
  - omitted variable bias
  - multicollinearity

Next Time: Interactions

Where We've Been and Where We're Going...

Where We've Been and Where We're Going...

- Last Week
  - mechanics of OLS with one variable
  - properties of OLS
- This Week
  - adding a second variable
  - new mechanics
  - omitted variable bias
  - multicollinearity
  - interactions
- Next Week
  - multiple regression
- Long Run
  - probability $\rightarrow$ inference $\rightarrow$ regression $\rightarrow$ causal inference

Why Interaction Terms?

Why Interaction Terms?

- Interaction terms will allow you to let the slope on one variable vary as a function of another variable

Why Interaction Terms?

- Interaction terms will allow you to let the slope on one variable vary as a function of another variable
- Interaction terms are central in regression analysis to:
    - Model and test conditional hypothesis (do the returns to education vary by race?)

Why Interaction Terms?

- Interaction terms will allow you to let the slope on one variable vary as a function of another variable
- Interaction terms are central in regression analysis to:
    - Model and test conditional hypothesis (do the returns to education vary by race?)
    - Make model of the conditional expectation function more realistic by letting coefficients vary across subgroups

Why Interaction Terms?

- Interaction terms will allow you to let the slope on one variable vary as a function of another variable
- Interaction terms are central in regression analysis to:
  - Model and test conditional hypothesis (do the returns to education vary by race?)
  - Make model of the conditional expectation function more realistic by letting coefficients vary across subgroups
- We can interact:
  - two or more dummy variables

Why Interaction Terms?

- Interaction terms will allow you to let the slope on one variable vary as a function of another variable
- Interaction terms are central in regression analysis to:
  - Model and test conditional hypothesis (do the returns to education vary by race?)
  - Make model of the conditional expectation function more realistic by letting coefficients vary across subgroups
- We can interact:
  - two or more dummy variables
  - dummy variables and continuous variables

Why Interaction Terms?

- Interaction terms will allow you to let the slope on one variable vary as a function of another variable
- Interaction terms are central in regression analysis to:
  - Model and test conditional hypothesis (do the returns to education vary by race?)
  - Make model of the conditional expectation function more realistic by letting coefficients vary across subgroups
- We can interact:
  - two or more dummy variables
  - dummy variables and continuous variables
  - two or more continuous variables

Why Interaction Terms?

- Interaction terms will allow you to let the slope on one variable vary as a function of another variable
- Interaction terms are central in regression analysis to:
  - Model and test conditional hypothesis (do the returns to education vary by race?)
  - Make model of the conditional expectation function more realistic by letting coefficients vary across subgroups
- We can interact:
  - two or more dummy variables
  - dummy variables and continuous variables
  - two or more continuous variables
- Interactions often confuses researchers and mistakes in use and interpretation occur frequently (even in top journals)

Return to the Fish Example

- Data comes from Fish (2002), "Islam and Authoritarianism."

Return to the Fish Example

- Data comes from Fish (2002), "Islam and Authoritarianism."
- Basic relationship: does more economic development lead to more democracy?

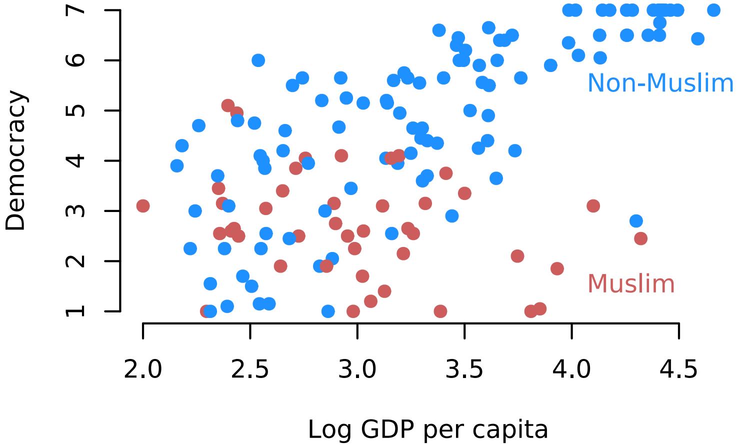
Return to the Fish Example

- Data comes from Fish (2002), "Islam and Authoritarianism."
- Basic relationship: does more economic development lead to more democracy?
- We measure economic development with log GDP per capita

Return to the Fish Example

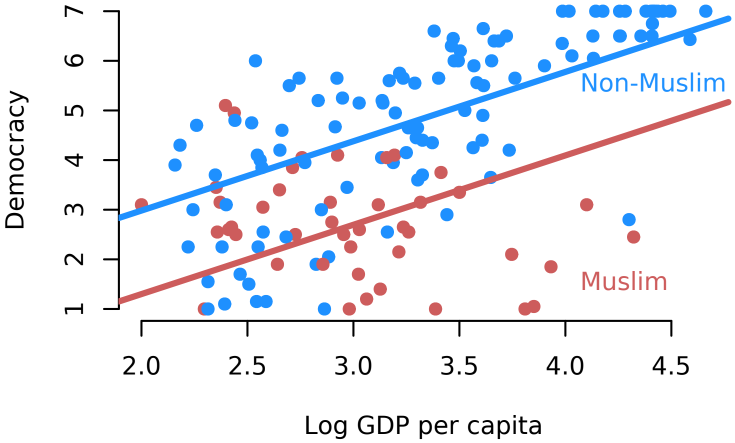
- Data comes from Fish (2002), "Islam and Authoritarianism."
- Basic relationship: does more economic development lead to more democracy?
- We measure economic development with log GDP per capita
- We measure democracy with a Freedom House score, 1 (less free) to 7 (more free)

Let's See the Data

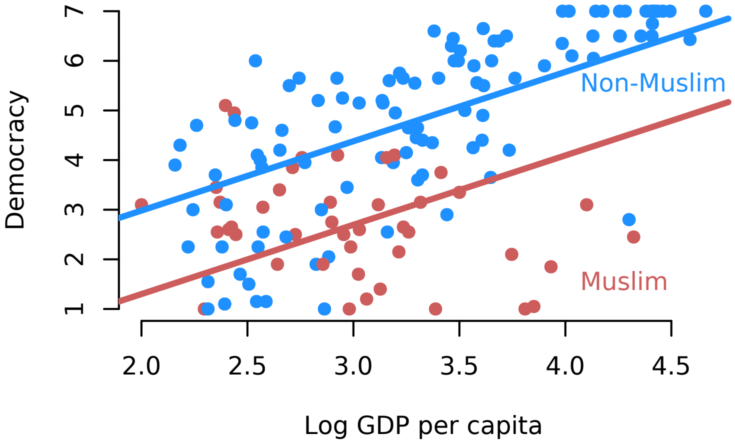


Fish argues that Muslim countries are less likely to be democratic no matter their economic development

Controlling for Religion Additively

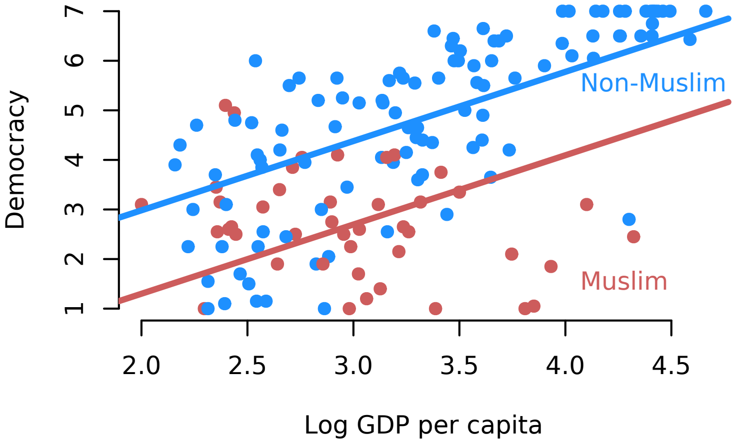


Controlling for Religion Additively



But the regression is a poor fit for Muslim countries

Controlling for Religion Additively



But the regression is a poor fit for Muslim countries

Can we allow for different slopes for each group?

Interactions with a Binary Variable

Interactions with a Binary Variable

- Let Z_i be binary

Interactions with a Binary Variable

- Let Z_i be binary
- In this case, $Z_i = 1$ for the country being Muslim

Interactions with a Binary Variable

- Let Z_i be binary
- In this case, $Z_i = 1$ for the country being Muslim
- We can add another covariate to the baseline model that allows the effect of income to vary by Muslim status.

Interactions with a Binary Variable

- Let Z_i be binary
- In this case, $Z_i = 1$ for the country being Muslim
- We can add another covariate to the baseline model that allows the effect of income to vary by Muslim status.
- This covariate is called an interaction term and it is the product of the two marginal variables of interest: $income_i \times muslim_i$

Interactions with a Binary Variable

- Let Z_i be binary
- In this case, $Z_i = 1$ for the country being Muslim
- We can add another covariate to the baseline model that allows the effect of income to vary by Muslim status.
- This covariate is called an interaction term and it is the product of the two marginal variables of interest: $income_i \times muslim_i$
- Here is the model with the interaction term:

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{\beta}_2 Z_i + \widehat{\beta}_3 X_i Z_i$$

Two Lines in One Regression

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{\beta}_2 Z_i + \widehat{\beta}_3 X_i Z_i$$

Two Lines in One Regression

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{\beta}_2 Z_i + \widehat{\beta}_3 X_i Z_i$$

- How can we interpret this model?

Two Lines in One Regression

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{\beta}_2 Z_i + \widehat{\beta}_3 X_i Z_i$$

- How can we interpret this model?
- We can plug in the two possible values of Z_i

Two Lines in One Regression

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{\beta}_2 Z_i + \widehat{\beta}_3 X_i Z_i$$

- How can we interpret this model?
- We can plug in the two possible values of Z_i
- When $Z_i = 0$:
$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{\beta}_2 Z_i + \widehat{\beta}_3 X_i Z_i$$

Two Lines in One Regression

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{\beta}_2 Z_i + \widehat{\beta}_3 X_i Z_i$$

- How can we interpret this model?
- We can plug in the two possible values of Z_i
- When $Z_i = 0$:

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{\beta}_2 Z_i + \widehat{\beta}_3 X_i Z_i$$

Two Lines in One Regression

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{\beta}_2 Z_i + \widehat{\beta}_3 X_i Z_i$$

- How can we interpret this model?
- We can plug in the two possible values of Z_i
- When $Z_i = 0$:

$$\begin{aligned}\widehat{Y}_i &= \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{\beta}_2 Z_i + \widehat{\beta}_3 X_i Z_i \\ &= \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{\beta}_2 \times 0 + \widehat{\beta}_3 X_i \times 0\end{aligned}$$

Two Lines in One Regression

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{\beta}_2 Z_i + \widehat{\beta}_3 X_i Z_i$$

- How can we interpret this model?
- We can plug in the two possible values of Z_i
- When $Z_i = 0$:

$$\begin{aligned}\widehat{Y}_i &= \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{\beta}_2 Z_i + \widehat{\beta}_3 X_i Z_i \\ &= \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{\beta}_2 \times 0 + \widehat{\beta}_3 X_i \times 0 \\ &= \widehat{\beta}_0 + \widehat{\beta}_1 X_i\end{aligned}$$

Two Lines in One Regression

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{\beta}_2 Z_i + \widehat{\beta}_3 X_i Z_i$$

- How can we interpret this model?
- We can plug in the two possible values of Z_i
- When $Z_i = 0$:

$$\begin{aligned}\widehat{Y}_i &= \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{\beta}_2 Z_i + \widehat{\beta}_3 X_i Z_i \\ &= \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{\beta}_2 \times 0 + \widehat{\beta}_3 X_i \times 0 \\ &= \widehat{\beta}_0 + \widehat{\beta}_1 X_i\end{aligned}$$

- When $Z_i = 1$:

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{\beta}_2 Z_i + \widehat{\beta}_3 X_i Z_i$$

Two Lines in One Regression

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{\beta}_2 Z_i + \widehat{\beta}_3 X_i Z_i$$

- How can we interpret this model?
- We can plug in the two possible values of Z_i
- When $Z_i = 0$:

$$\begin{aligned}\widehat{Y}_i &= \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{\beta}_2 Z_i + \widehat{\beta}_3 X_i Z_i \\ &= \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{\beta}_2 \times 0 + \widehat{\beta}_3 X_i \times 0 \\ &= \widehat{\beta}_0 + \widehat{\beta}_1 X_i\end{aligned}$$

- When $Z_i = 1$:

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{\beta}_2 Z_i + \widehat{\beta}_3 X_i Z_i$$

Two Lines in One Regression

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{\beta}_2 Z_i + \widehat{\beta}_3 X_i Z_i$$

- How can we interpret this model?
- We can plug in the two possible values of Z_i
- When $Z_i = 0$:

$$\begin{aligned}
\widehat{Y}_i &= \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{\beta}_2 Z_i + \widehat{\beta}_3 X_i Z_i \\
&= \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{\beta}_2 \times 0 + \widehat{\beta}_3 X_i \times 0 \\
&= \widehat{\beta}_0 + \widehat{\beta}_1 X_i
\end{aligned}$$

- When $Z_i = 1$:

$$\begin{aligned}
\widehat{Y}_i &= \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{\beta}_2 Z_i + \widehat{\beta}_3 X_i Z_i \\
&= \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{\beta}_2 \times 1 + \widehat{\beta}_3 X_i \times 1
\end{aligned}$$

Two Lines in One Regression

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{\beta}_2 Z_i + \widehat{\beta}_3 X_i Z_i$$

- How can we interpret this model?
- We can plug in the two possible values of Z_i
- When $Z_i = 0$:

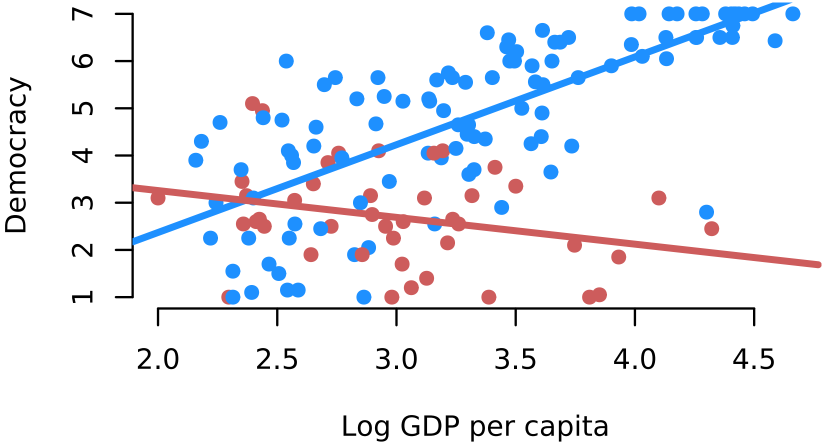
$$\begin{aligned}\widehat{Y}_i &= \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{\beta}_2 Z_i + \widehat{\beta}_3 X_i Z_i \\ &= \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{\beta}_2 \times 0 + \widehat{\beta}_3 X_i \times 0 \\ &= \widehat{\beta}_0 + \widehat{\beta}_1 X_i\end{aligned}$$

- When $Z_i = 1$:

$$\begin{aligned}\widehat{Y}_i &= \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{\beta}_2 Z_i + \widehat{\beta}_3 X_i Z_i \\ &= \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{\beta}_2 \times 1 + \widehat{\beta}_3 X_i \times 1 \\ &= (\widehat{\beta}_0 + \widehat{\beta}_2) + (\widehat{\beta}_1 + \widehat{\beta}_3) X_i\end{aligned}$$

Example Interpretation of the Coefficients

	Intercept for X_i	Slope for X_i
Non-Muslim country ($Z_i = 0$)	$\widehat{\beta}_0$	$\widehat{\beta}_1$
Muslim country ($Z_i = 1$)	$\widehat{\beta}_0 + \widehat{\beta}_2$	$\widehat{\beta}_1 + \widehat{\beta}_3$



General Interpretation of the Coefficients

- $\widehat{\beta}_0$: average value of Y_i when both X_i and Z_i are equal to 0

General Interpretation of the Coefficients

- $\widehat{\beta}_0$: average value of Y_i when both X_i and Z_i are equal to 0
- $\widehat{\beta}_1$: a one-unit change in X_i is associated with a $\widehat{\beta}_1$-unit change in Y_i when $Z_i = 0$

General Interpretation of the Coefficients

- $\widehat{\beta}_0$: average value of Y_i when both X_i and Z_i are equal to 0
- $\widehat{\beta}_1$: a one-unit change in X_i is associated with a $\widehat{\beta}_1$-unit change in Y_i when $Z_i = 0$
- $\widehat{\beta}_2$: average difference in Y_i between $Z_i = 1$ group and $Z_i = 0$ group when $X_i = 0$

General Interpretation of the Coefficients

- $\widehat{\beta}_0$: average value of Y_i when both X_i and Z_i are equal to 0
- $\widehat{\beta}_1$: a one-unit change in X_i is associated with a $\widehat{\beta}_1$-unit change in Y_i when $Z_i = 0$
- $\widehat{\beta}_2$: average difference in Y_i between $Z_i = 1$ group and $Z_i = 0$ group when $X_i = 0$
- $\widehat{\beta}_3$: change in the effect of X_i on Y_i between $Z_i = 1$ group and $Z_i = 0$

Lower Order Terms

Lower Order Terms

- Principle of Marginality: Always include the marginal effects (sometimes called the lower order terms)

Lower Order Terms

- Principle of Marginality: Always include the marginal effects (sometimes called the lower order terms)
- Imagine we omitted the lower order term for `muslim`:

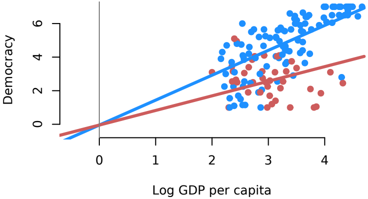
$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + 0 \times Z_i + \widehat{\beta}_3 X_i Z_i$$

Omitting Lower Order Terms

Omitting Lower Order Terms

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + 0 \times Z_i + \widehat{\beta}_3 X_i Z_i$$

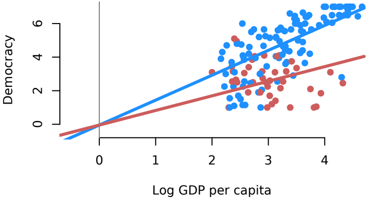
	Intercept for X_i	Slope for X_i
Non-Muslim country ($Z_i = 0$)	$\widehat{\beta}_0$	$\widehat{\beta}_1$
Muslim country ($Z_i = 1$)	$\widehat{\beta}_0 + 0$	$\widehat{\beta}_1 + \widehat{\beta}_3$



Omitting Lower Order Terms

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + 0 \times Z_i + \widehat{\beta}_3 X_i Z_i$$

	Intercept for X_i	Slope for X_i
Non-Muslim country ($Z_i = 0$)	$\widehat{\beta}_0$	$\widehat{\beta}_1$
Muslim country ($Z_i = 1$)	$\widehat{\beta}_0 + 0$	$\widehat{\beta}_1 + \widehat{\beta}_3$

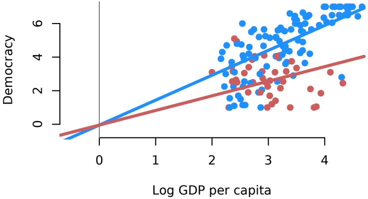


- Model assumption: no difference between Muslim and non-Muslim countries when income is 0

Omitting Lower Order Terms

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + 0 \times Z_i + \widehat{\beta}_3 X_i Z_i$$

	Intercept for X_i	Slope for X_i
Non-Muslim country ($Z_i = 0$)	$\widehat{\beta}_0$	$\widehat{\beta}_1$
Muslim country ($Z_i = 1$)	$\widehat{\beta}_0 + 0$	$\widehat{\beta}_1 + \widehat{\beta}_3$

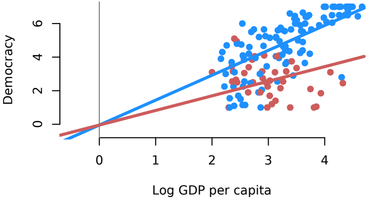


- Model assumption: no difference between Muslim and non-Muslim countries when income is 0
- Distorts slope estimates.

Omitting Lower Order Terms

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + 0 \times Z_i + \widehat{\beta}_3 X_i Z_i$$

	Intercept for X_i	Slope for X_i
Non-Muslim country ($Z_i = 0$)	$\widehat{\beta}_0$	$\widehat{\beta}_1$
Muslim country ($Z_i = 1$)	$\widehat{\beta}_0 + 0$	$\widehat{\beta}_1 + \widehat{\beta}_3$



- Model assumption: no difference between Muslim and non-Muslim countries when income is 0
- Distorts slope estimates.
- Very rarely justified, but for some reason, people keep doing it (as you will see in your problem set).

Interactions with Two Continuous Variables

- Now let Z_i be continuous

Interactions with Two Continuous Variables

- Now let Z_i be continuous
- Z_i is the percent growth in GDP per capita from 1975 to 1998

Interactions with Two Continuous Variables

- Now let Z_i be continuous
- Z_i is the percent growth in GDP per capita from 1975 to 1998
- Is the effect of economic development for rapidly developing countries higher or lower than for stagnant economies?

Interactions with Two Continuous Variables

- Now let Z_i be continuous
- Z_i is the percent growth in GDP per capita from 1975 to 1998
- Is the effect of economic development for rapidly developing countries higher or lower than for stagnant economies?
- We can still define the interaction:

$$income_i \times growth_i$$

Interactions with Two Continuous Variables

- Now let Z_i be continuous
- Z_i is the percent growth in GDP per capita from 1975 to 1998
- Is the effect of economic development for rapidly developing countries higher or lower than for stagnant economies?
- We can still define the interaction:

$$income_i \times growth_i$$

- And include it in the regression:

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{\beta}_2 Z_i + \widehat{\beta}_3 X_i Z_i$$

Interpretation

- With a continuous Z_i, we can have more than two values that it can take on:

	Intercept for X_i	Slope for X_i
$Z_i = 0$	$\widehat{\beta}_0$	$\widehat{\beta}_1$

Interpretation

- With a continuous Z_i, we can have more than two values that it can take on:

	Intercept for X_i	Slope for X_i
$Z_i = 0$	$\widehat{\beta}_0$	$\widehat{\beta}_1$

Interpretation

- With a continuous Z_i, we can have more than two values that it can take on:

	Intercept for X_i	Slope for X_i
$Z_i = 0$	$\widehat{\beta}_0$	$\widehat{\beta}_1$
$Z_i = 0.5$	$\widehat{\beta}_0 + \widehat{\beta}_2 \times 0.5$	$\widehat{\beta}_1 + \widehat{\beta}_3 \times 0.5$

Interpretation

- With a continuous Z_i, we can have more than two values that it can take on:

	Intercept for X_i	Slope for X_i
$Z_i = 0$	$\widehat{\beta}_0$	$\widehat{\beta}_1$
$Z_i = 0.5$	$\widehat{\beta}_0 + \widehat{\beta}_2 \times 0.5$	$\widehat{\beta}_1 + \widehat{\beta}_3 \times 0.5$
$Z_i = 1$	$\widehat{\beta}_0 + \widehat{\beta}_2 \times 1$	$\widehat{\beta}_1 + \widehat{\beta}_3 \times 1$

Interpretation

- With a continuous Z_i, we can have more than two values that it can take on:

	Intercept for X_i	Slope for X_i
$Z_i = 0$	$\widehat{\beta}_0$	$\widehat{\beta}_1$
$Z_i = 0.5$	$\widehat{\beta}_0 + \widehat{\beta}_2 \times 0.5$	$\widehat{\beta}_1 + \widehat{\beta}_3 \times 0.5$
$Z_i = 1$	$\widehat{\beta}_0 + \widehat{\beta}_2 \times 1$	$\widehat{\beta}_1 + \widehat{\beta}_3 \times 1$
$Z_i = 5$	$\widehat{\beta}_0 + \widehat{\beta}_2 \times 5$	$\widehat{\beta}_1 + \widehat{\beta}_3 \times 5$

General Interpretation

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{\beta}_2 Z_i + \widehat{\beta}_3 X_i Z_i$$

- The coefficient $\widehat{\beta}_1$ measures how the predicted outcome varies in X_i when $Z_i = 0$.

General Interpretation

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{\beta}_2 Z_i + \widehat{\beta}_3 X_i Z_i$$

- The coefficient $\widehat{\beta}_1$ measures how the predicted outcome varies in X_i when $Z_i = 0$.
- The coefficient $\widehat{\beta}_2$ measures how the predicted outcome varies in Z_i when $X_i = 0$

General Interpretation

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{\beta}_2 Z_i + \widehat{\beta}_3 X_i Z_i$$

- The coefficient $\widehat{\beta}_1$ measures how the predicted outcome varies in X_i when $Z_i = 0$.
- The coefficient $\widehat{\beta}_2$ measures how the predicted outcome varies in Z_i when $X_i = 0$
- The coefficient $\widehat{\beta}_3$ is the change in the effect of X_i given a one-unit change in Z_i:

$$\frac{\partial E[Y_i|X_i, Z_i]}{\partial X_i} = \beta_1 + \beta_3 Z_i$$

General Interpretation

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{\beta}_2 Z_i + \widehat{\beta}_3 X_i Z_i$$

- The coefficient $\widehat{\beta}_1$ measures how the predicted outcome varies in X_i when $Z_i = 0$.
- The coefficient $\widehat{\beta}_2$ measures how the predicted outcome varies in Z_i when $X_i = 0$
- The coefficient $\widehat{\beta}_3$ is the change in the effect of X_i given a one-unit change in Z_i:
$$\frac{\partial E[Y_i|X_i, Z_i]}{\partial X_i} = \beta_1 + \beta_3 Z_i$$
- The coefficient $\widehat{\beta}_3$ is the change in the effect of Z_i given a one-unit change in X_i:
$$\frac{\partial E[Y_i|X_i, Z_i]}{\partial Z_i} = \beta_2 + \beta_3 X_i$$

Additional Assumptions

Additional Assumptions

Interaction effects are particularly susceptible to model dependence. We are making two assumptions for the estimated effects to be meaningful:

Additional Assumptions

Interaction effects are particularly susceptible to model dependence. We are making two assumptions for the estimated effects to be meaningful:

1. Linearity of the interaction effect

Additional Assumptions

Interaction effects are particularly susceptible to model dependence. We are making two assumptions for the estimated effects to be meaningful:

1. Linearity of the interaction effect
2. Common support (variation in X throughout the range of Z)

Additional Assumptions

Interaction effects are particularly susceptible to model dependence. We are making two assumptions for the estimated effects to be meaningful:

1. Linearity of the interaction effect
2. Common support (variation in X throughout the range of Z)

We will talk about checking these assumptions in a few weeks.

PA

How Much Should We Trust Estimates from Multiplicative Interaction Models? Simple Tools to Improve Empirical Practice

Jens Hainmueller[1], Jonathan Mummolo[2] and Yiqing Xu[3]

[1] *Professor of Political Science, Stanford University, Department of Political Science, Stanford, CA 94305, USA. Email: jhain@stanford.edu*
[2] *Assistant Professor of Politics and Public Affairs, Princeton University, Department of Politics, Woodrow Wilson School of Public and International Affairs, Princeton, NJ 08544, USA. Email: jmummolo@princeton.edu*
[3] *Assistant Professor of Political Science, University of California, San Diego, Department of Political Science, La Jolla, CA 92093, USA. Email: yiqingxu@ucsd.edu*

Abstract

Multiplicative interaction models are widely used in social science to examine whether the relationship between an outcome and an independent variable changes with a moderating variable. Current empirical practice tends to overlook two important problems. First, these models assume a linear interaction effect that changes at a constant rate with the moderator. Second, estimates of the conditional effects of the independent variable can be misleading if there is a lack of common support of the moderator. Replicating 46 interaction effects from 22 recent publications in five top political science journals, we find that these core assumptions often fail in practice, suggesting that a large portion of findings across all political science subfields based on interaction models are fragile and model dependent. We propose a checklist of simple diagnostics to assess the validity of these assumptions and offer flexible estimation strategies that allow for nonlinear interaction effects and safeguard against excessive extrapolation. These statistical routines are available in both R and STATA.

Keywords: misspecification, linear regression, local regression, interaction models, marginal effects

Example: Common Support

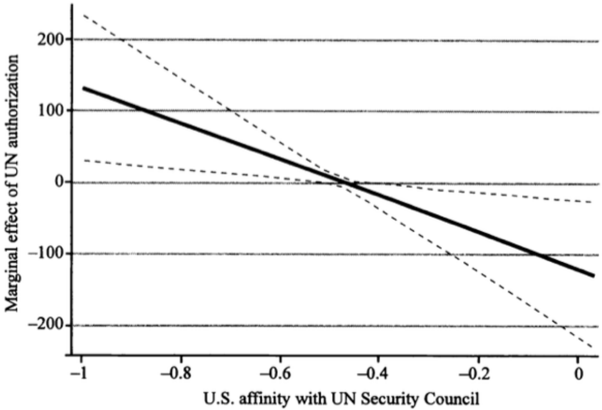
Chapman 2009 analysis
example and reanalysis from Hainmueller, Mummolo, Xu 2019

"The interaction term shows a strong negative and statically significant coefficient; suggesting that when UN authorization occurs and the interaction term is 'switched on' positive movement in the similarity score (towards more similar) reduces rallies. Rallies with UN authorization are only larger than average when the pivotal member is ideologically distant from the United States. This provides strong support for the informational rationale for IO legitimacy… Clearly, the effect of authorization on rallies decreases as similarity increases: foreign policy actions that receive authorization from a less conservative institution receive similar rallies to those that do not receive authorization from an IO."

Example: Common Support

Chapman 2009 analysis

example and reanalysis from Hainmueller, Mummolo, Xu 2019

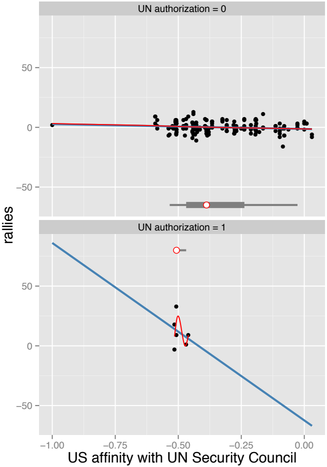


Note: Dashed lines give 95 percent confidence interval.

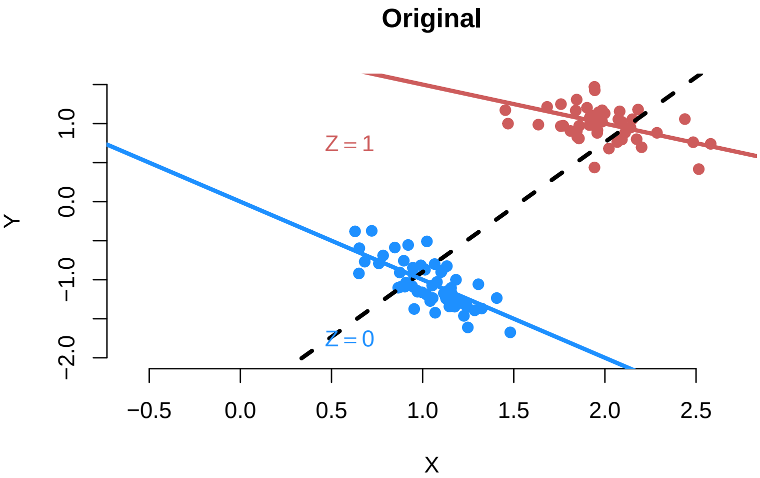
Example: Common Support

Chapman 2009 analysis

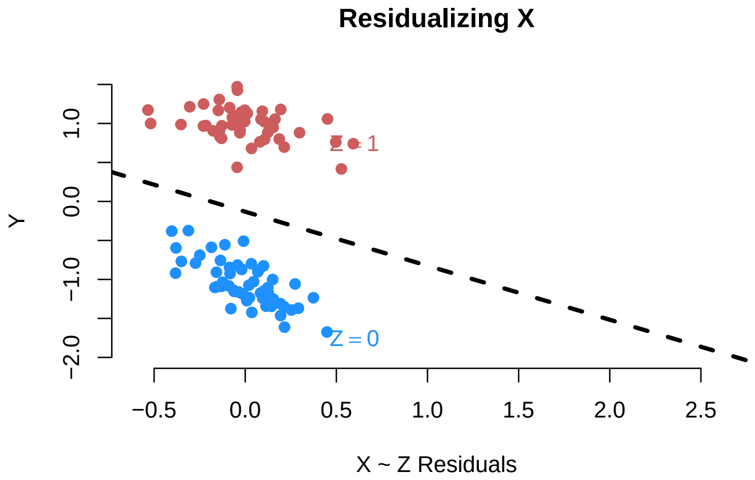
example and reanalysis from Hainmueller, Mummolo, Xu 2019



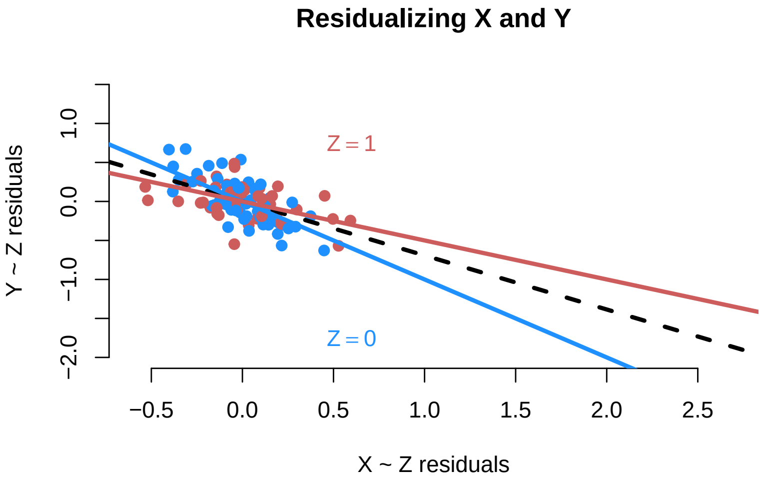
What Happens Without Interactions



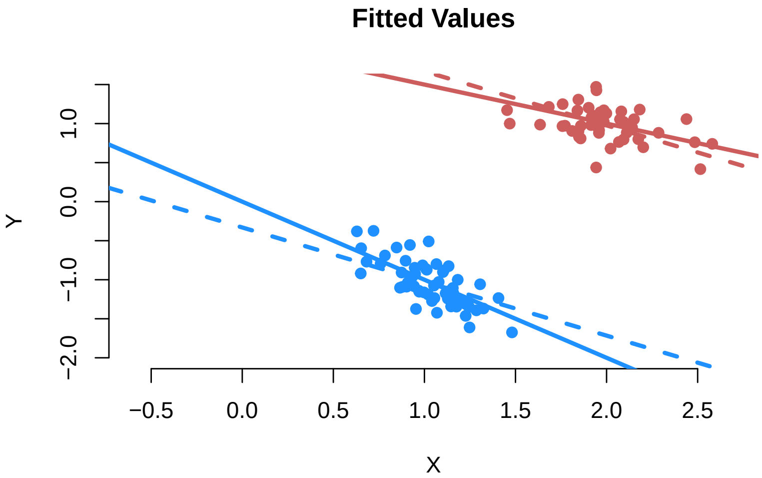
What Happens Without Interactions



What Happens Without Interactions



What Happens Without Interactions



Summary for Interactions

Summary for Interactions

- Do not omit lower order terms (unless you have a strong theory that tells you so) because this usually imposes unrealistic restrictions.

Summary for Interactions

- Do not omit lower order terms (unless you have a strong theory that tells you so) because this usually imposes unrealistic restrictions.
- Do not interpret the coefficients on the lower terms as marginal effects (they give the marginal effect only for the case where the other variable is equal to zero)

Summary for Interactions

- Do not omit lower order terms (unless you have a strong theory that tells you so) because this usually imposes unrealistic restrictions.
- Do not interpret the coefficients on the lower terms as marginal effects (they give the marginal effect only for the case where the other variable is equal to zero)
- Produce tables or figures that summarize the conditional marginal effects of the variable of interest at plausible different levels of the other variable; use correct formula to compute variance for these conditional effects (sum of coefficients)

Summary for Interactions

- Do not omit lower order terms (unless you have a strong theory that tells you so) because this usually imposes unrealistic restrictions.
- Do not interpret the coefficients on the lower terms as marginal effects (they give the marginal effect only for the case where the other variable is equal to zero)
- Produce tables or figures that summarize the conditional marginal effects of the variable of interest at plausible different levels of the other variable; use correct formula to compute variance for these conditional effects (sum of coefficients)
- In simple cases the p-value on the interaction term can be used as a test against the null of no interaction, but significant tests for the lower order terms rarely make sense.

Summary for Interactions

- Do not omit lower order terms (unless you have a strong theory that tells you so) because this usually imposes unrealistic restrictions.
- Do not interpret the coefficients on the lower terms as marginal effects (they give the marginal effect only for the case where the other variable is equal to zero)
- Produce tables or figures that summarize the conditional marginal effects of the variable of interest at plausible different levels of the other variable; use correct formula to compute variance for these conditional effects (sum of coefficients)
- In simple cases the p-value on the interaction term can be used as a test against the null of no interaction, but significant tests for the lower order terms rarely make sense.

Further Reading: Brambor, Clark, and Golder. 2006. Understanding Interaction Models: Improving Empirical Analyses. *Political Analysis*.

Hainmueller, Mummolo, Xu. 2019. How Much Should We Trust Estimates from Multiplicative Interaction Models? Simple Tools to Improve Empirical Practice. *Political Analysis*.

Polynomial Terms

Polynomial Terms

- Polynomial terms are a special case of the continuous variable interactions.

Polynomial Terms

- Polynomial terms are a special case of the continuous variable interactions.
- For example, when $X_1 = X_2$ in the previous interaction model, we get a quadratic:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + u$$
$$Y = \beta_0 + (\beta_1 + \beta_2) X_1 + \beta_3 X_1 X_1 + u$$
$$Y = \beta_0 + \tilde{\beta}_1 X_1 + \tilde{\beta}_2 X_1^2 + u$$

- This is called a second order polynomial in X_1

Polynomial Terms

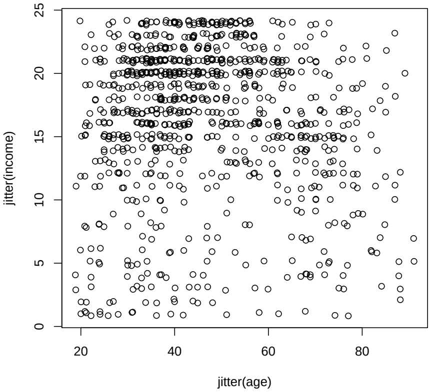
- Polynomial terms are a special case of the continuous variable interactions.
- For example, when $X_1 = X_2$ in the previous interaction model, we get a quadratic:

$$
\begin{aligned}
Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + u \\
Y &= \beta_0 + (\beta_1 + \beta_2) X_1 + \beta_3 X_1 X_1 + u \\
Y &= \beta_0 + \tilde{\beta}_1 X_1 + \tilde{\beta}_2 X_1^2 + u
\end{aligned}
$$

- This is called a second order polynomial in X_1
- A third order polynomial is given by: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_1^3 + u$

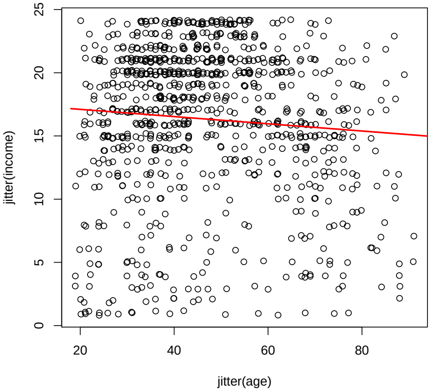
Polynomial Example: Income and Age

- Let's look at data from the U.S. and examine the relationship between Y: income and X: age



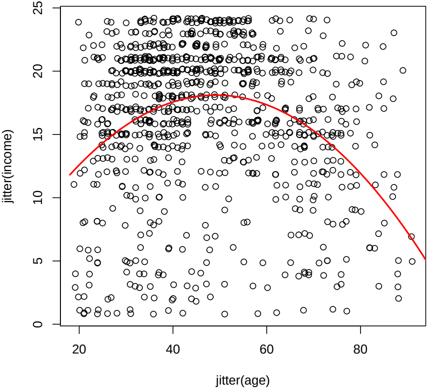
Polynomial Example: Income and Age

- Let's look at data from the U.S. and examine the relationship between Y: income and X: age
- We see that a simple linear specification does not fit the data very well: $Y = \beta_0 + \beta_1 X_1 + u$



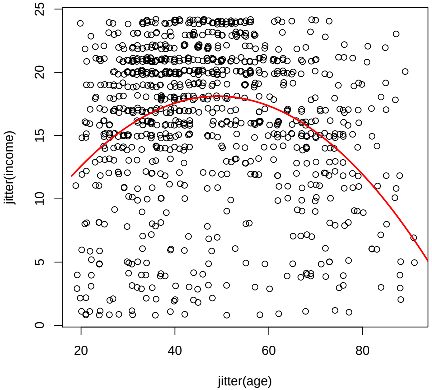
Polynomial Example: Income and Age

- Let's look at data from the U.S. and examine the relationship between Y: income and X: age
- We see that a simple linear specification does not fit the data very well:
  $Y = \beta_0 + \beta_1 X_1 + u$
- A second order polynomial in age fits the data a lot better:
  $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + u$



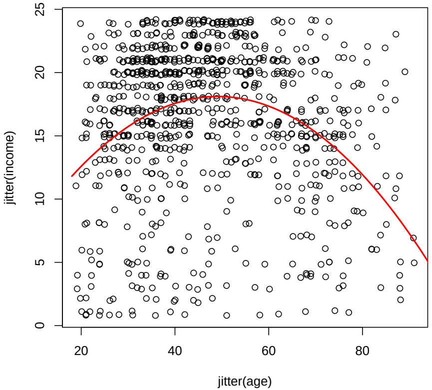
Polynomial Example: Income and Age

- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + u$



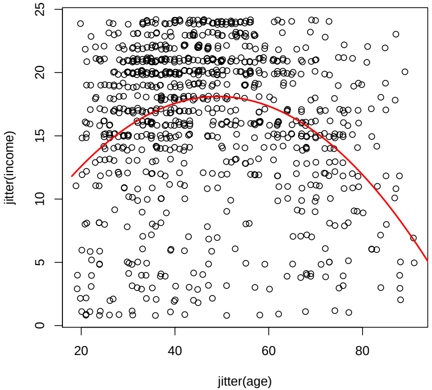
Polynomial Example: Income and Age

- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + u$
- Is β_1 the marginal effect of age on income?



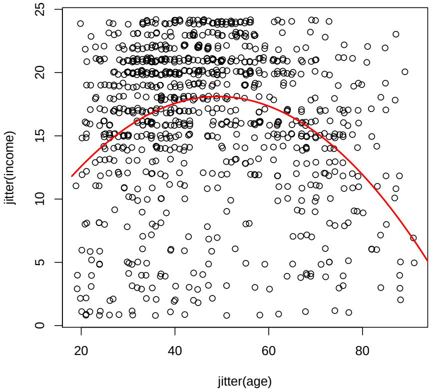
Polynomial Example: Income and Age

- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + u$
- Is β_1 the marginal effect of age on income?
- No! The marginal effect of age depends on the level of age: $\frac{\partial Y}{\partial X_1} = \widehat{\beta}_1 + 2\,\widehat{\beta}_2\,X_1$
  Here the effect of age changes monotonically from positive to negative with income.



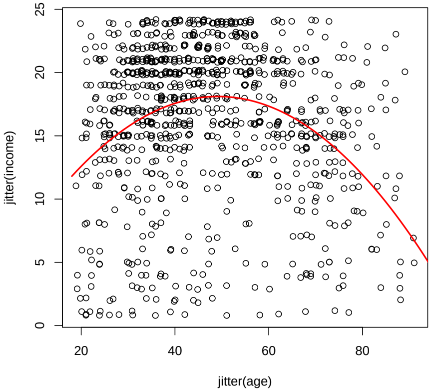
Polynomial Example: Income and Age

- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + u$
- Is β_1 the marginal effect of age on income?
- No! The marginal effect of age depends on the level of age: $\frac{\partial Y}{\partial X_1} = \widehat{\beta}_1 + 2\,\widehat{\beta}_2\,X_1$
  Here the effect of age changes monotonically from positive to negative with income.
- If $\beta_2 > 0$ we get a U-shape, and if $\beta_2 < 0$ we get an inverted U-shape.

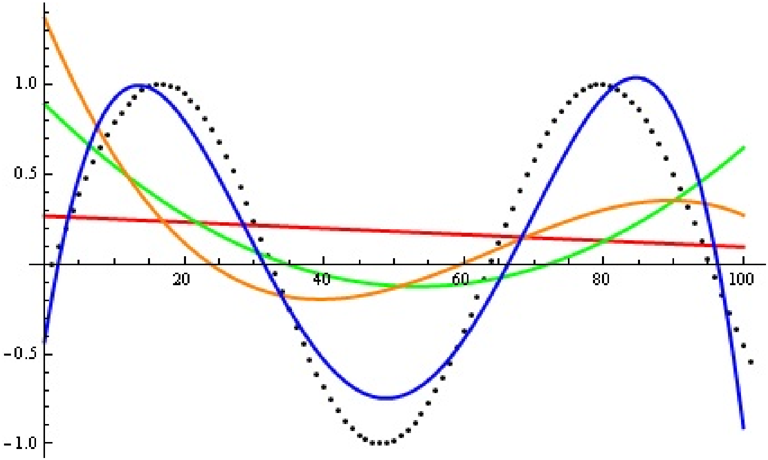


Polynomial Example: Income and Age

- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + u$
- Is β_1 the marginal effect of age on income?
- No! The marginal effect of age depends on the level of age: $\frac{\partial Y}{\partial X_1} = \widehat{\beta}_1 + 2\,\widehat{\beta}_2\,X_1$ Here the effect of age changes monotonically from positive to negative with income.
- If $\beta_2 > 0$ we get a U-shape, and if $\beta_2 < 0$ we get an inverted U-shape.
- Maximum/Minimum occurs at $|\frac{\beta_1}{2\beta_2}|$. Here turning point is at $X_1 = 50$.



Higher Order Polynomials



Approximating data generated with a sine function. Red line is a first degree polynomial, green line is second degree, orange line is third degree and blue is fourth degree

Complex Parameter Interpretation

We can mix and match these model specifications

Complex Parameter Interpretation

We can mix and match these model specifications

- Interactions with higher order terms

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \beta_3 X_i^2 + \beta_4 X_i * Z_i + \beta_5 X_i^2 * Z_i + u$$

Complex Parameter Interpretation

We can mix and match these model specifications

- Interactions with higher order terms

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \beta_3 X_i^2 + \beta_4 X_i * Z_i + \beta_5 X_i^2 * Z_i + u$$

- World is your oyster!

Complex Parameter Interpretation

We can mix and match these model specifications

- Interactions with higher order terms

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \beta_3 X_i^2 + \beta_4 X_i * Z_i + \beta_5 X_i^2 * Z_i + u$$

- World is your oyster!
- But interpretation gets difficult.

Complex Parameter Interpretation

We can mix and match these model specifications

- Interactions with higher order terms

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \beta_3 X_i^2 + \beta_4 X_i * Z_i + \beta_5 X_i^2 * Z_i + u$$

- World is your oyster!
- But interpretation gets difficult.
- We can take partial derivatives as a way to get a sense of the shape of the estimated CEF.

Complex Parameter Interpretation

We can mix and match these model specifications

- Interactions with higher order terms

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \beta_3 X_i^2 + \beta_4 X_i * Z_i + \beta_5 X_i^2 * Z_i + u$$

- World is your oyster!
- But interpretation gets difficult.
- We can take partial derivatives as a way to get a sense of the shape of the estimated CEF.
- An easier approach can be to make predictions and then average over the data

$$\frac{1}{n}\sum_{i}^{n} \widehat{E}[Y|X = x_i + 1, Z = z_i] - \widehat{E}[Y|X = x_i, Z = z_i]$$

Complex Parameter Interpretation

We can mix and match these model specifications

- Interactions with higher order terms

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \beta_3 X_i^2 + \beta_4 X_i * Z_i + \beta_5 X_i^2 * Z_i + u$$

- World is your oyster!
- But interpretation gets difficult.
- We can take partial derivatives as a way to get a sense of the shape of the estimated CEF.
- An easier approach can be to make predictions and then average over the data

$$\frac{1}{n} \sum_i^n \widehat{E}[Y|X = x_i + 1, Z = z_i] - \widehat{E}[Y|X = x_i, Z = z_i]$$

- Getting uncertainty on this quantity is a great use case for the bootstrap!

This Week in Review

In this brave new world with 2 independent variables:

This Week in Review

In this brave new world with 2 independent variables:

1. β's have slightly different interpretations

This Week in Review

In this brave new world with 2 independent variables:

1. β's have slightly different interpretations
2. OLS still minimizing the sum of the squared residuals

This Week in Review

In this brave new world with 2 independent variables:

1. β's have slightly different interpretations
2. OLS still minimizing the sum of the squared residuals
3. Small adjustments to OLS assumptions and inference

This Week in Review

In this brave new world with 2 independent variables:

1. β's have slightly different interpretations
2. OLS still minimizing the sum of the squared residuals
3. Small adjustments to OLS assumptions and inference
4. Adding or omitting variables in a regression can affect the bias and the variance of OLS

This Week in Review

In this brave new world with 2 independent variables:

1. β's have slightly different interpretations
2. OLS still minimizing the sum of the squared residuals
3. Small adjustments to OLS assumptions and inference
4. Adding or omitting variables in a regression can affect the bias and the variance of OLS
5. We can optionally consider interactions, but must take care to interpret them correctly

This Week in Review

In this brave new world with 2 independent variables:

1. β's have slightly different interpretations
2. OLS still minimizing the sum of the squared residuals
3. Small adjustments to OLS assumptions and inference
4. Adding or omitting variables in a regression can affect the bias and the variance of OLS
5. We can optionally consider interactions, but must take care to interpret them correctly

Next week: Linear Regression in its Full Glory!