Week 6: Linear Regression with Two Regressors

Brandon Stewart[1]

Princeton

October 17, 19, 2016

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

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

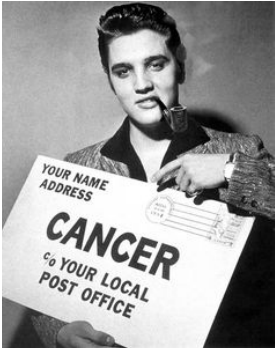
- Last Week
  - mechanics of OLS with one variable
  - properties of OLS
- This Week
  - Monday:
    - adding a second variable
    - new mechanics
  - Wednesday:
    - omitted variable bias
    - multicollinearity
    - interactions
- Next Week
  - multiple regression
- Long Run
  - probability $\rightarrow$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

Questions?

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

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 lowers the death rate by 4 deaths per 1,000 person years.

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

Berkeley gender bias?

- 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?

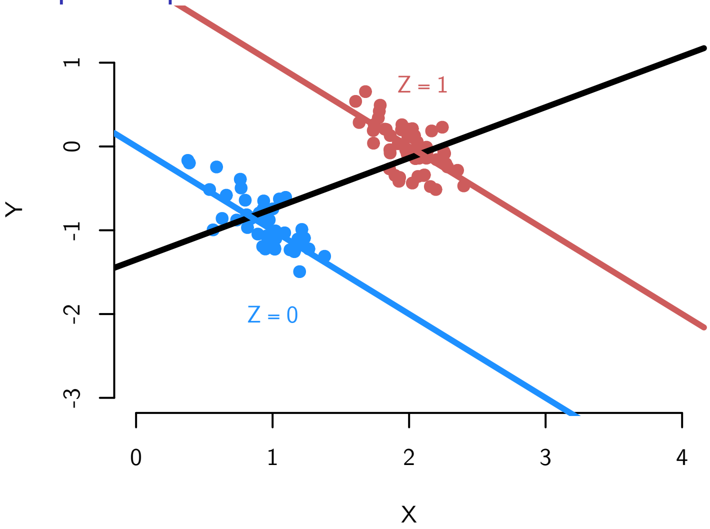
Berkeley gender bias?

- 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? Women apply to more challenging departments.
- Marginal relationships (admissions and gender) $\neq$ conditional relationship given third variable (department)

Simpson's paradox



- Overall a positive relationship between Y_i and X_i here
- But within strata defined by Z_i, the opposite

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.
- Cochran's 1968 study is also a case of Simpson's paradox, he originally sought to compare cigarette to cigar smoking, he found that cigar smokers had higher mortality rates than cigarette smokers, but at *any* age level, cigarette smokers had higher mortality than cigar smokers.

Instance of a more general problem called the ecological inference fallacy

Basic idea

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

$$\mathbb{E}[Y_i|X_i]$$

- For continuous X's, 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

Why control for another variable

- Descriptive
  - get a sense for the relationships in the data.
  - describe more precisely our quantity of interest
- Predictive
  - We can usually make better predictions about the dependent variable with more information on independent variables.
- Causal
  - Block potential **confounding**, which is when X doesn't cause Y, but only appears to because a third variable Z causally affects both of them.
  - X_i: ice cream sales on day i
  - Y_i: drowning deaths on day i
  - Z_i: ??

Regression with Two Explanatory Variables

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

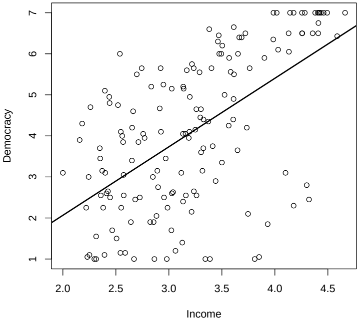
- Variables of interest:
  - Y: Level of democracy, measured as the 10-year average of Freedom House ratings
  - X_1: Country income, measured as log(GDP per capita in \$1000s)
  - X_2: Ethnic heterogeneity (continuous) or British colonial heritage (binary)
- With one predictor we ask: Does income (X_1)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

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

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_1$$

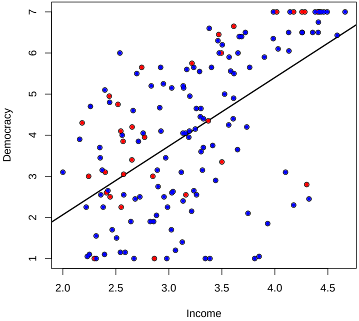
$$\widehat{Demo} = -1.26 + 1.6\, Log(GDP)$$



Interpretation: A one percent increase in GDP is associated with a .016 point increase in 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

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

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 $\widehat{\beta}_0$, $\widehat{\beta}_1$, $\widehat{\beta}_2$:

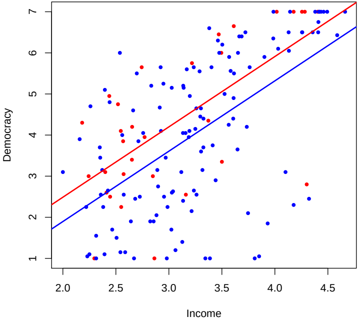
```
Coefficients:
              Estimate
(Intercept)   -1.5060
GDP90LGN       1.7059
BRITCOL        0.5881
```

- Non-British colonies:

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

- Former British colonies:

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

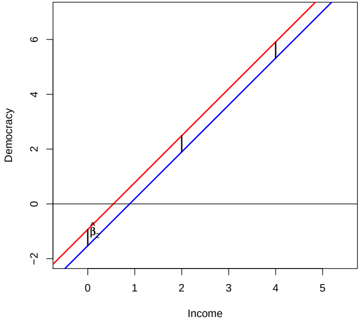


Regression of Democracy on Income

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

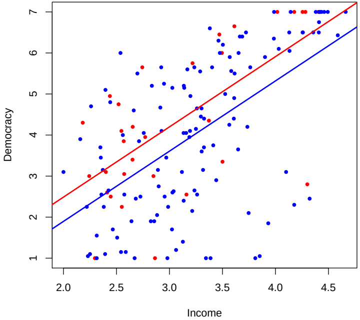
Where do these quantities appear on the graph?

- $\widehat{\beta}_0 = -1.5$ 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



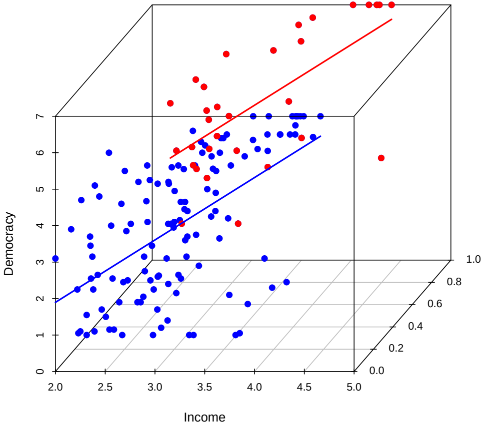
Fitting a regression plane

- 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.



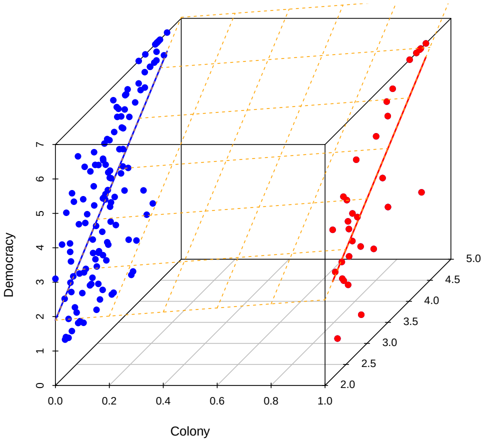
Regression of Democracy on Income

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

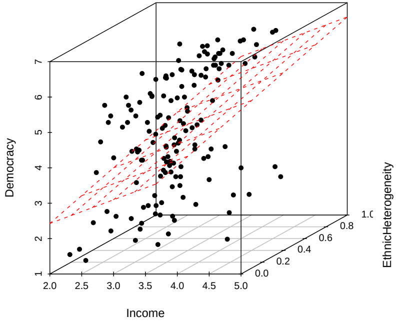
- Since we fit a regression plane to the data whenever we have two explanatory variables, it is easy to move to a case with two continuous explanatory variables.
- For example, we might want to use:
  - X_1 Income and X_2 Ethnic Heterogeneity
  - Y Democracy

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

Regression of Democracy on Income

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



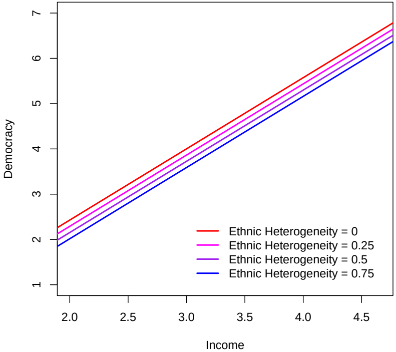
Interpreting a Continuous Covariate

- The coefficient estimates have a similar interpretation in this case as they did in the Income-British Colony example.
- 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 partial effect or ceteris paribus interpretations:

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

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.
- 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.

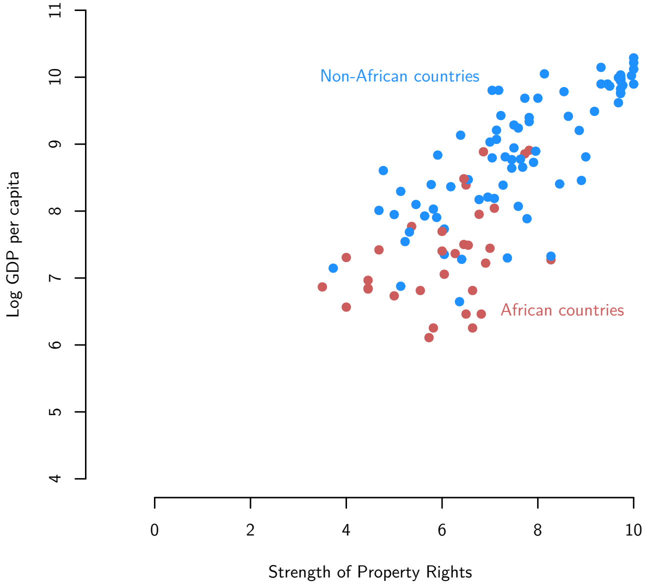


More Complex Predictions

- We can also use the coefficient estimates for more complex predictions that involve changing multiple variables simultaneously.
- 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 $(3.5 - 2.5)\widehat{\beta}_1 + (.06 - .5)\widehat{\beta}_2$

AJR Example



Non-African countries
African countries
Log GDP per capita
Strength of Property Rights
4
5
6
7
8
9
10
11
0
2
4
6
8
10

Basics

- Ye olde model:
$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i$$
- $Z_i = 1$ to indicate that i is an African country
- $Z_i = 0$ to indicate that i is an non-African country
- Concern: AJR might be picking up an "African effect":
  - African countries have low incomes and weak property rights
  - "Control for" country being in Africa or not to remove this
  - Effects are now within Africa or within non-Africa, not between
- New model:
$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{\beta}_2 Z_i$$

AJR model

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.65556    0.31344  18.043  < 2e-16 ***
## avexpr       0.42416    0.03971  10.681  < 2e-16 ***
## africa      -0.87844    0.14707  -5.973 3.03e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6253 on 108 degrees of freedom
##   (52 observations deleted due to missingness)
## Multiple R-squared:  0.7078, Adjusted R-squared:  0.7024
## F-statistic: 130.8 on 2 and 108 DF,  p-value: < 2.2e-16
```

Two lines in one regression

- How can we interpret this model?
- Plug in two possible values for Z_i and rearrange
- When $Z_i = 0$:

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

- When $Z_i = 1$:

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

- Two different intercepts, same slope

Example interpretation of the coefficients

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

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

- In this example, we have:

$$\widehat{Y}_i = 5.656 + 0.424 \times X_i - 0.878 \times Z_i$$

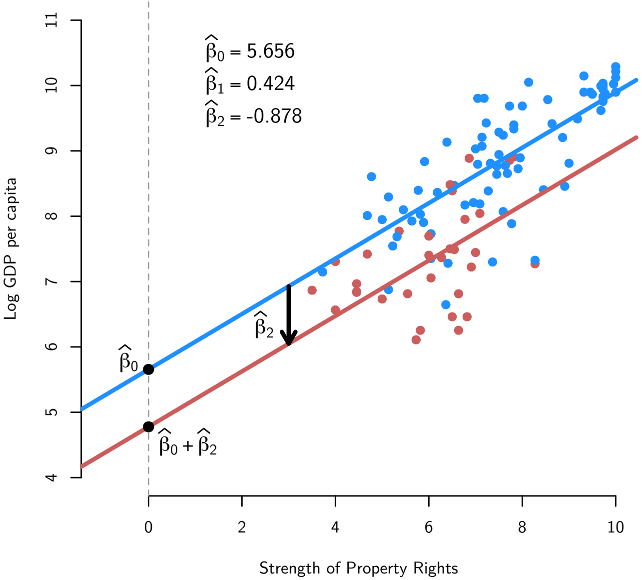
- We can read these as:
  - $\widehat{\beta}_0$: average log income for non-African country ($Z_i = 0$) with property rights measured at 0 is 5.656
  - $\widehat{\beta}_1$: A one-unit increase in property rights is associated with a 0.424 increase in average log incomes for two African countries (or for two non-African countries)
  - $\widehat{\beta}_2$: there is a -0.878 average difference in log income per capita between African and non-African counties **conditional on** property rights

General interpretation of the coefficients

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

- $\widehat{\beta}_0$: 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 conditional on Z_i
- $\widehat{\beta}_2$: average difference in Y_i between $Z_i = 1$ group and $Z_i = 0$ group conditional on X_i

Adding a binary variable, visually



Adding a continuous variable

- Ye olde model:
$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i$$
- Z_i: mean temperature in country i (continuous)
- Concern: geography is confounding the effect
  - geography might affect political institutions
  - geography might affect average incomes (through diseases like malaria)
- New model:
$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{\beta}_2 Z_i$$

AJR model, revisited

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.80627    0.75184   9.053 1.27e-12 ***
## avexpr       0.40568    0.06397   6.342 3.94e-08 ***
## meantemp    -0.06025    0.01940  -3.105  0.00296 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6435 on 57 degrees of freedom
##   (103 observations deleted due to missingness)
## Multiple R-squared:  0.6155, Adjusted R-squared:  0.602
## F-statistic: 45.62 on 2 and 57 DF,  p-value: 1.481e-12
```

Interpretation with a continuous Z

	Intercept for X_i	Slope for X_i
$Z_i = 0\,^\circ\mathrm{C}$	$\widehat{\beta}_0$	$\widehat{\beta}_1$
$Z_i = 21\,^\circ\mathrm{C}$	$\widehat{\beta}_0 + \widehat{\beta}_2 \times 21$	$\widehat{\beta}_1$
$Z_i = 24\,^\circ\mathrm{C}$	$\widehat{\beta}_0 + \widehat{\beta}_2 \times 24$	$\widehat{\beta}_1$
$Z_i = 26\,^\circ\mathrm{C}$	$\widehat{\beta}_0 + \widehat{\beta}_2 \times 26$	$\widehat{\beta}_1$

- In this example we have:

$$\widehat{Y}_i = 6.806 + 0.406 \times X_i + -0.06 \times Z_i$$

- $\widehat{\beta}_0$: average log income for a country with property rights measured at 0 and a mean temperature of 0 is 6.806
- $\widehat{\beta}_1$: A one-unit change in property rights is associated with a 0.406 change in average log incomes conditional on a country's mean temperature
- $\widehat{\beta}_2$: A one-degree increase in mean temperature is associated with a -0.06 change in average log incomes conditional on strength of property rights

General interpretation

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

- The coefficient $\widehat{\beta}_1$ measures how the predicted outcome varies in X_i for a fixed value of Z_i.
- The coefficient $\widehat{\beta}_2$ measures how the predicted outcome varies in Z_i for a fixed value of X_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 some terms from simple linear regression.
- Fitted values for $i = 1, \ldots, n$:

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

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

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

Least squares is still least squares

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

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

- The calculus is the same as last week, with 3 partial derivatives instead of 2
- Let's start with a simple recipe and then rigorously show that it holds

OLS estimator recipe using two steps

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

Regression property rights on mean temperature

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.95678    0.82015  12.140  < 2e-16 ***
## meantemp    -0.14900    0.03469  -4.295 6.73e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.321 on 58 degrees of freedom
##   (103 observations deleted due to missingness)
## Multiple R-squared:  0.2413, Adjusted R-squared:  0.2282
## F-statistic: 18.45 on 1 and 58 DF,  p-value: 6.733e-05
```

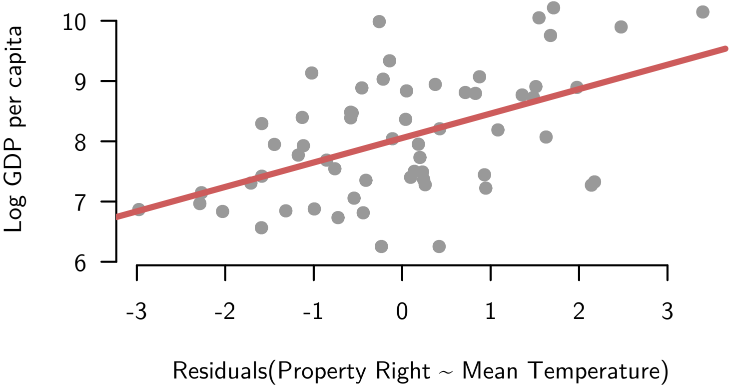
Regression of log income on the residuals

```
## (Intercept)  avexpr.res
##   8.0542783   0.4056757

## (Intercept)      avexpr    meantemp
##  6.80627375  0.40567575 -0.06024937
```

Residual/partial regression plot

Useful for plotting the conditional relationship between property rights and income given temperature:



Deriving the Linear Least Squares Estimator

- In simple regression, we chose $(\widehat{\beta}_0, \widehat{\beta}_1)$ to minimize the sum of the squared residuals
- We use the same principle for picking $(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2)$ for regression with two regressors (x_i and z_i):

$$\begin{aligned}
(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2) &= \underset{\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\beta}_2}{\operatorname{argmin}} \sum_{i=1}^{n} \widehat{u}_i^2 = \underset{\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\beta}_2}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\
&= \underset{\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\beta}_2}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \tilde{\beta}_0 - x_i \tilde{\beta}_1 - z_i \tilde{\beta}_2)^2
\end{aligned}$$

- (The same works more generally for k regressors, but this is done more easily with matrices as we will see next week)

Deriving the Linear Least Squares Estimator

We want to minimize the following quantitity with respect to $(\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\beta}_2)$:

$$S(\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\beta}_2) = \sum_{i=1}^{n}(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i - \tilde{\beta}_2 z_i)^2$$

Plan is conceptually the same as before

1. Take the partial derivatives of S with respect to $\tilde{\beta}_0, \tilde{\beta}_1$ and $\tilde{\beta}_2$.
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 $\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\beta}_2$ and solve for $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ to obtain the OLS estimator.

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 $\hat{\beta}_2$

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

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)
- Wooldridge calls this assumption no perfect collinearity

The OLS Estimator

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:

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

Condition 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 ($\Rightarrow \; Cor(x,z) = 1 \; \Rightarrow \; Var(x)Var(z) = Cov(x,z)^2$)

"Partialling Out" Interpretation of the OLS Estimator

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 is correlation between X and Z. What is our estimator $\hat{\beta}_1$ if $\delta = 0$?

$$r_{xz} = x - \hat{\lambda} = x_i - \bar{x} \quad \text{so} \quad \hat{\beta}_1 = \frac{\sum_i^n \hat{r}_{xz,i}\, y_i}{\sum_i^n \hat{r}_{xz,i}^2} = \frac{\sum_i^n (x_i - \bar{x})\, y_i}{\sum_i^n (x_i - \bar{x})^2}$$

- That is, same as the simple regresson of Y on X alone.

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.

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
$$\mathbb{E}[u_i | X_i, Z_i] = 0$$

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
  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:
  - $X_i = 1$ 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?
- 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:

```
##
## 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:
  - 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

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
$$\mathbb{E}[u_i | X_i, Z_i] = 0$$
5. Homoskedasticity
$$\text{var}[u_i | X_i, Z_i] = \sigma_u^2$$

Inference with two independent variables in large samples

- 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

OLS assumptions for small-sample inference

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
$$\mathbb{E}[u_i|X_i, Z_i] = 0$$
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:

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

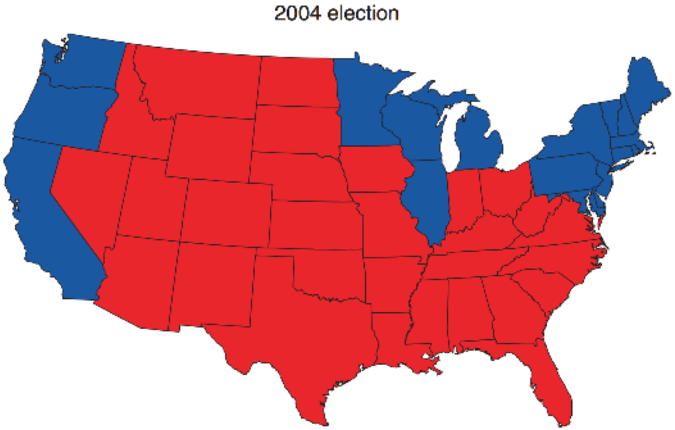
- The same is true for the other coefficient:

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

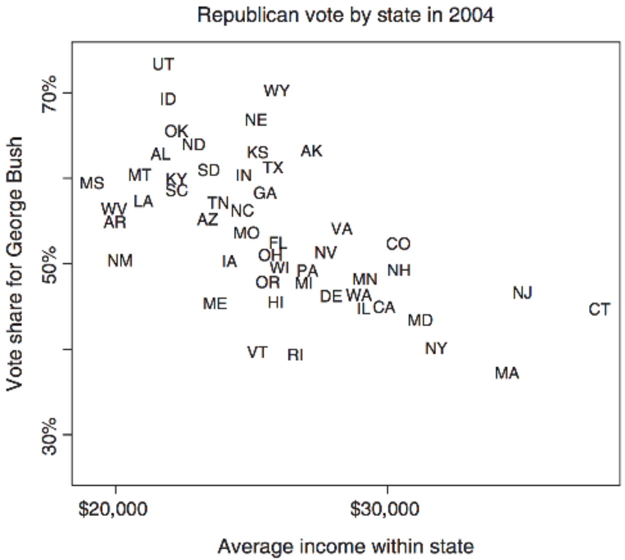
- Why $n-3$?
    - We've estimated another parameter, so we need to take off another degree of freedom.
- $\rightsquigarrow$ small adjustments to the critical values and the t-values for our hypothesis tests and confidence intervals.

Red State Blue State

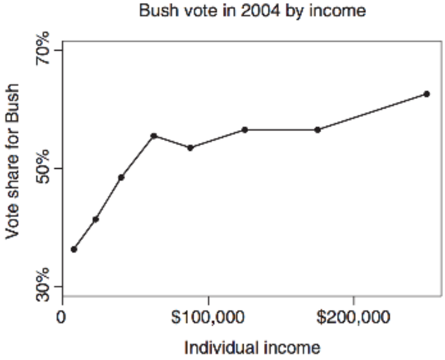
Red and Blue States

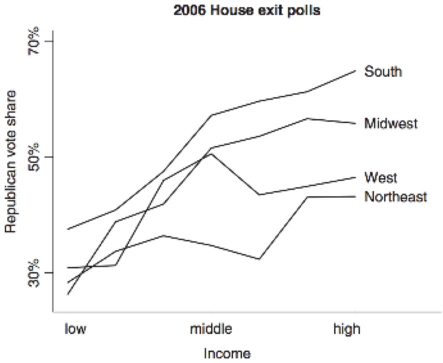


Rich States are More Democratic

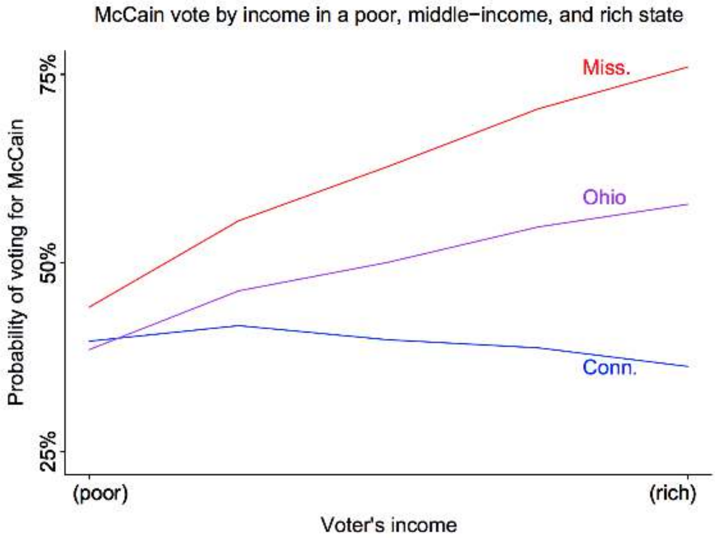


But Rich People are More Republican

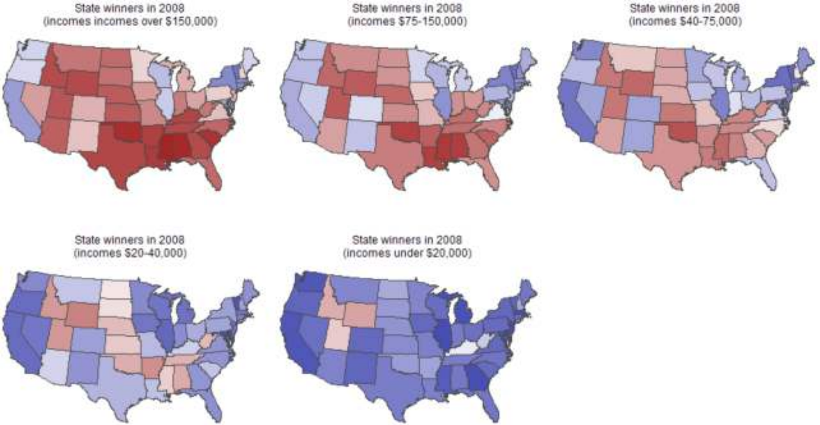




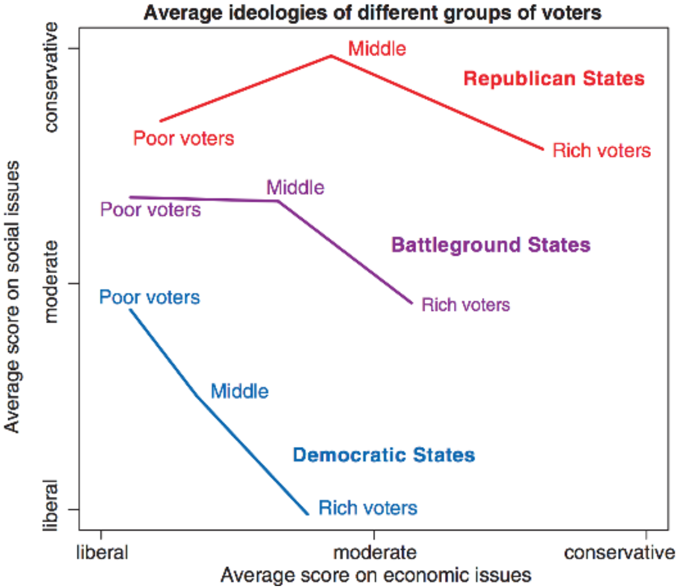
Paradox Resolved



If Only Rich People Voted, it Would Be a Landslide



A Possible Explanation



References

Acemoglu, Daron, Simon Johnson, and James A. Robinson. "The colonial origins of comparative development: An empirical investigation." *American Economic Review*. 91(5). 2001: 1369-1401.

Fish, M. Steven. "Islam and authoritarianism." *World politics* 55(01). 2002: 4-37.

Gelman, Andrew. *Red state, blue state, rich state, poor state: why Americans vote the way they do*. Princeton University Press, 2009.

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

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

Questions?

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:

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

- 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:

$$Y_i = \beta_0 + \beta_1 X_i + u_i^* \qquad u_i^* = \beta_2 Z_i + u_i$$

- OLS estimates from the misspecified model:

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

Omitted Variable Bias: Simple Case

True Population Model:

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

Underspecified Model that we use:

$$\text{Voted Republican} = \tilde{\beta}_0 + \tilde{\beta}_1 \text{Watch Fox News}$$

Q: Which statement is correct?

1. $\beta_1 > \tilde{\beta}_1$
2. $\beta_1 < \tilde{\beta}_1$
3. $\beta_1 = \tilde{\beta}_1$
4. Can't tell

Answer: $\tilde{\beta}_1$ is upward biased since being a strong republican is positively correlated with both watching fox news and voting republican. We have $\beta_1 < \tilde{\beta}_1$.

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:

$$\text{Survival} = \tilde{\beta}_0 + \tilde{\beta}_1 \text{Hospitalized}$$

Q: Which statement is correct?

1. $\beta_1 > \tilde{\beta}_1$
2. $\beta_1 < \tilde{\beta}_1$
3. $\beta_1 = \tilde{\beta}_1$
4. Can't tell

Answer: The negative coefficient $\tilde{\beta}_1$ is downward biased compared to the true β_1 so $\beta_1 > \tilde{\beta}_1$. Being hospitalized is negatively correlated with health, and health is positively correlated with survival.

Omitted Variable Bias: Simple Case

True Population Model:

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

Underspecified Model that we use:

$$\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1$$

We can show that the relationship between $\tilde{\beta}_1$ and $\hat{\beta}_1$ is:

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

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.

Q. When will $\tilde{\beta}_1 = \hat{\beta}_1$?
A. If $\tilde{\delta} = 0$ or $\hat{\beta}_2 = 0$.

Omitted Variable Bias: Simple Case

We take expectations to see what the bias will be:

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

So

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

So the bias depends on the relationship between x_2 and x_1, our $\tilde{\delta}$, and the relationship between x_2 and y, our β_2.

Any variable that is correlated with an included X and the outcome Y is called a confounder.

Omitted Variable Bias: Simple Case

Direction of the bias of $\tilde{\beta}_1$ 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
- If you miss an important confounder, your estimates are biased and inconsistent.
- In the more general case with more than two covariates the bias is more difficult to discern. It depends on all the pairwise correlations.

Including an Irrelevant Variable: Simple Case

True Population Model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u \quad \text{where} \quad \beta_2 = 0$$

and Assumptions I–IV hold.

Overspecified Model that we use:

$$\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + \tilde{\beta}_2 x_2$$

Q: Which statement is correct?

1. $\beta_1 > \tilde{\beta}_1$
2. $\beta_1 < \tilde{\beta}_1$
3. $\beta_1 = \tilde{\beta}_1$
4. Can't tell

Including an Irrelevant Variable: Simple Case

Recall: Given Assumptions I–IV, we have:

$$E[\hat{\beta}_j] = \beta_j$$

for all values of β_j. So, if $\beta_2 = 0$, we get

$$E[\hat{\beta}_0] = \beta_0,\ E[\hat{\beta}_1] = \beta_1,\ E[\hat{\beta}_2] = 0$$

and thus including the irrelevant variable does not generally affect the unbiasedness. The sampling distribution of $\hat{\beta}_2$ will be centered about zero.

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:

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

- 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 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

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:
  - $\widehat{\beta}_1$ 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.
- 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.
- 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.
- 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

Why 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.
    - E.g. does average income vary by region? Are Republicans smarter than Democrats?
- Dummy variables are also used to examine conditional hypothesis via interaction terms
    - E.g. does the effect of education differ by gender?

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:

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

- Hint: Informative variable names help (e.g. call it MAJORITARIAN)

- Let's regress GDP on this dummy variable and a constant:
  $Y = \beta_0 + \beta_1 D + u$

Example: GDP per capita on Electoral System

R Code

```
> summary(lm(REALGDPCAP ~ MAJORITARIAN, data = D))

Call:
lm(formula = REALGDPCAP ~ MAJORITARIAN, data = D)

Residuals:
   Min     1Q Median     3Q    Max
 -5982  -4592  -2112   4293  13685

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    7097.7      763.2    9.30 1.64e-14 ***
MAJORITARIAN  -1053.8     1224.9   -0.86    0.392
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 5504 on 83 degrees of freedom
Multiple R-squared: 0.008838,       Adjusted R-squared: -0.003104
F-statistic: 0.7401 on 1 and 83 DF,  p-value: 0.3921
```

Example: GDP per capita on Electoral System

R Code

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    7097.7      763.2    9.30 1.64e-14 ***
MAJORITARIAN  -1053.8     1224.9   -0.86    0.392
```

R Code

```
> gdp.pro <- D$REALGDPCAP[D$MAJORITARIAN == 0]
> summary(gdp.pro)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1116    2709    5102    7098   10670   20780

> gdp.maj <- D$REALGDPCAP[D$MAJORITARIAN == 1]
> summary(gdp.maj)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  530.2  1431.0  3404.0  6044.0 11770.0 18840.0
```

So this is just like a difference in means two sample t-test!

Example: GDP per capita on Electoral System

R Code

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    7097.7      763.2    9.30 1.64e-14 ***
MAJORITARIAN  -1053.8     1224.9   -0.86    0.392
```

R Code

```
> gdp.pro <- D$REALGDPCAP[D$MAJORITARIAN == 0]
> summary(gdp.pro)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1116    2709    5102    7098   10670   20780

> gdp.maj <- D$REALGDPCAP[D$MAJORITARIAN == 1]
> summary(gdp.maj)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  530.2  1431.0  3404.0  6044.0 11770.0 18840.0
```

So this is just like a difference in means two sample t-test!

Example: GDP per capita on Electoral System

R Code

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    7097.7      763.2    9.30 1.64e-14 ***
MAJORITARIAN  -1053.8     1224.9   -0.86    0.392
```

R Code

```
> gdp.pro <- D$REALGDPCAP[D$MAJORITARIAN == 0]
> summary(gdp.pro)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1116    2709    5102    7098   10670   20780

> gdp.maj <- D$REALGDPCAP[D$MAJORITARIAN == 1]
> summary(gdp.maj)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  530.2  1431.0  3404.0  6044.0 11770.0 18840.0
```

So this is just like a difference in means two sample t-test!

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 violate the no perfect collinearity assumption:

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

- 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 $m = 5$:

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

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

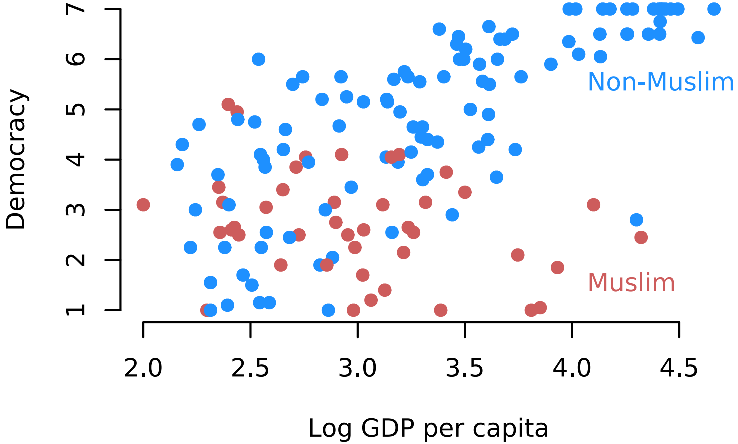
Why Interaction Terms?

- 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 gender?)
  - 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

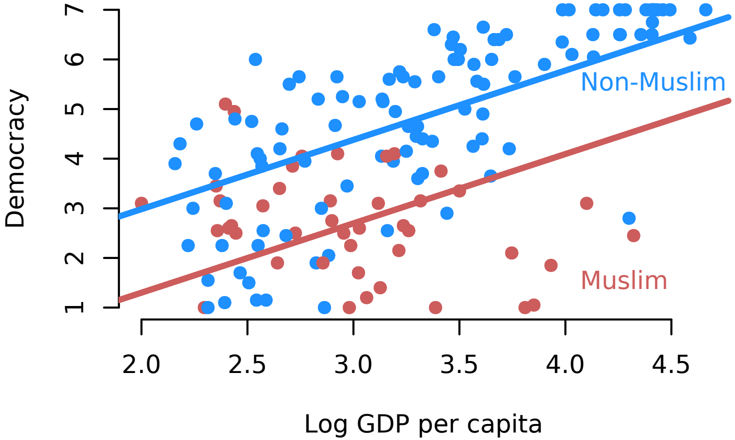
- Data comes from Fish (2002), "Islam and Authoritarianism."
- Basic relationship: does more economic development lead to more democracy?
- We measure economic development with log GDP per capita
- We measure democracy with a Freedom House score, 1 (less free) to 7 (more free)

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



But the regression is a poor fit for Muslim countries

Can we allow for different slopes for each group?

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

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

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

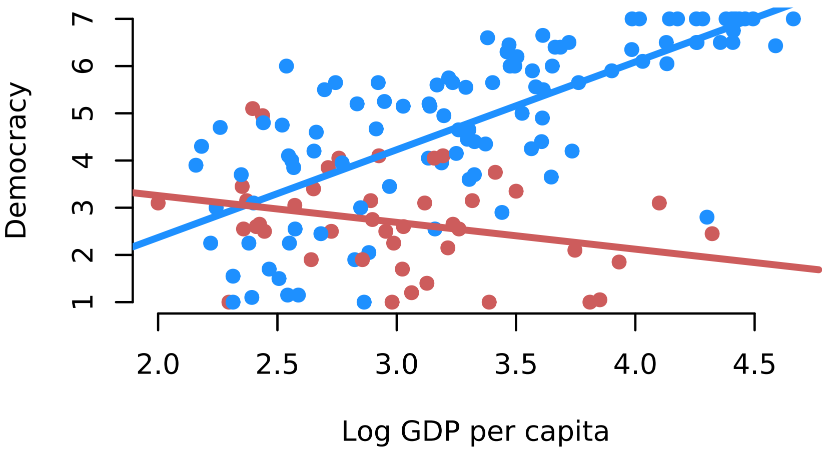
$$\begin{aligned} \widehat{Y}_i &= \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{\beta}_2 Z_i + \widehat{\beta}_3 X_i Z_i \\ &= \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{\beta}_2 \times 0 + \widehat{\beta}_3 X_i \times 0 \\ &= \widehat{\beta}_0 + \widehat{\beta}_1 X_i \end{aligned}$$

- 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 ($Z_i = 1$)	$\widehat{\beta}_0 + \widehat{\beta}_2$	$\widehat{\beta}_1 + \widehat{\beta}_3$

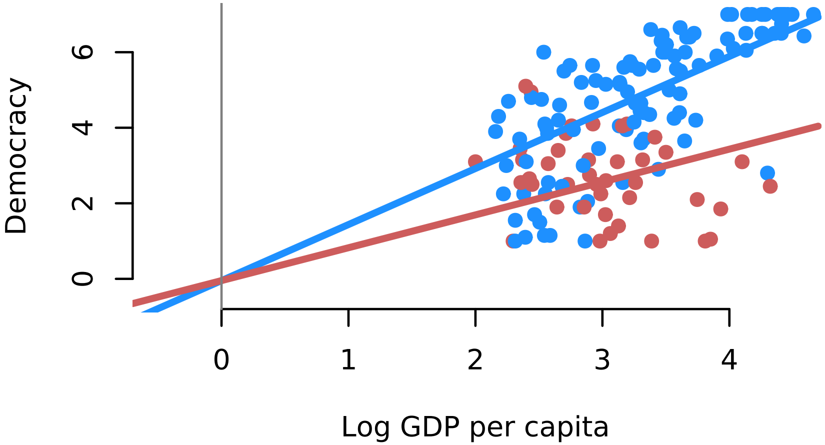


General interpretation of the coefficients

- $\widehat{\beta}_0$: average value of Y_i when both X_i and Z_i are equal to 0
- $\widehat{\beta}_1$: 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

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



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 ($Z_i = 1$)	$\widehat{\beta}_0 + 0$	$\widehat{\beta}_1 + \widehat{\beta}_3$

- Implication: no difference between Muslims and non-Muslims when income is 0
- Distorts slope estimates.
- Very rarely justified.
- Yet for some reason people keep doing it.

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
$Z_i = 0$	$\widehat{\beta}_0$	$\widehat{\beta}_1$
$Z_i = 0.5$	$\widehat{\beta}_0 + \widehat{\beta}_2 \times 0.5$	$\widehat{\beta}_1 + \widehat{\beta}_3 \times 0.5$
$Z_i = 1$	$\widehat{\beta}_0 + \widehat{\beta}_2 \times 1$	$\widehat{\beta}_1 + \widehat{\beta}_3 \times 1$
$Z_i = 5$	$\widehat{\beta}_0 + \widehat{\beta}_2 \times 5$	$\widehat{\beta}_1 + \widehat{\beta}_3 \times 5$

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 $Z_i = 0$.
- 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

Interaction effects are particularly susceptible to model dependence. We are making two assumptions for the estimated effects to be meaningful:

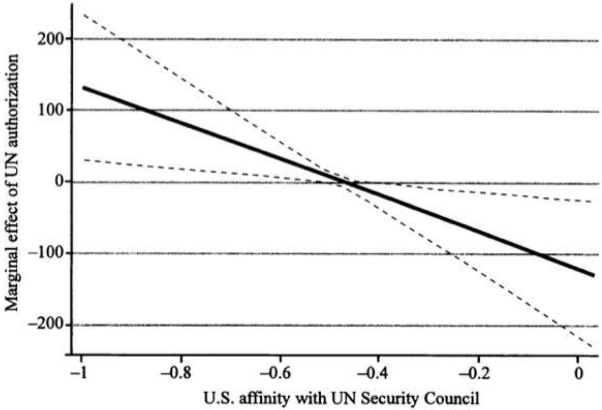
1. Linearity of the interaction effect
2. Common support (variation in X throughout the range of Z)

We will talk about checking these assumptions in a few weeks.

Example: Common Support

Chapman 2009 analysis

example and reanalysis from Hainmueller, Mummolo, Xu 2016

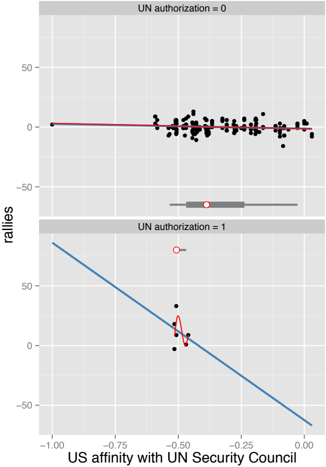


Note: Dashed lines give 95 percent confidence interval.

Example: Common Support

Chapman 2009 analysis

example and reanalysis from Hainmueller, Mummolo, Xu 2016



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* 14 (1): 63-82.

Hainmueller, Mummolo, Xu. 2016. How Much Should We Trust Estimates from Multiplicative Interaction Models? Simple Tools to Improve Empirical Practice. *Working Paper*

Polynomial terms

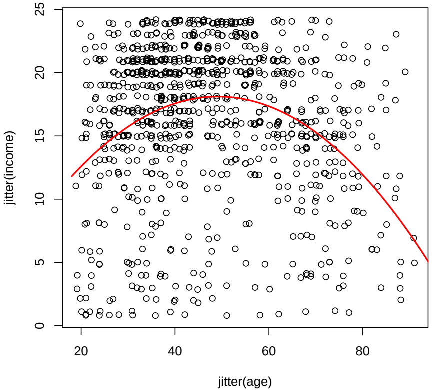
- Polynomial terms are a special case of the continuous variable interactions.
- For example, when $X_1 = X_2$ in the previous interaction model, we get a quadratic:

$$
\begin{aligned}
Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + u \\
Y &= \beta_0 + (\beta_1 + \beta_2) X_1 + \beta_3 X_1 X_1 + u \\
Y &= \beta_0 + \tilde{\beta}_1 X_1 + \tilde{\beta}_2 X_1^2 + u
\end{aligned}
$$

- This is called a second order polynomial in X_1
- A third order polynomial is given by:
  $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_1^3 + u$

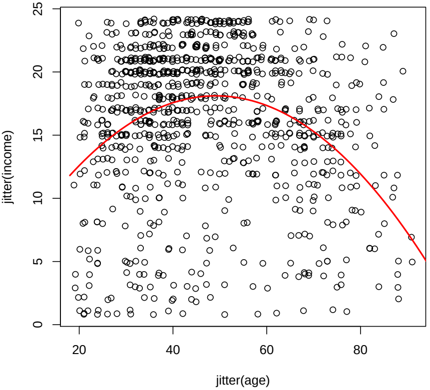
Polynomial Example: Income and Age

- Let's look at data from the U.S. and examine the relationship between Y: income and X: age
- 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: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + u$

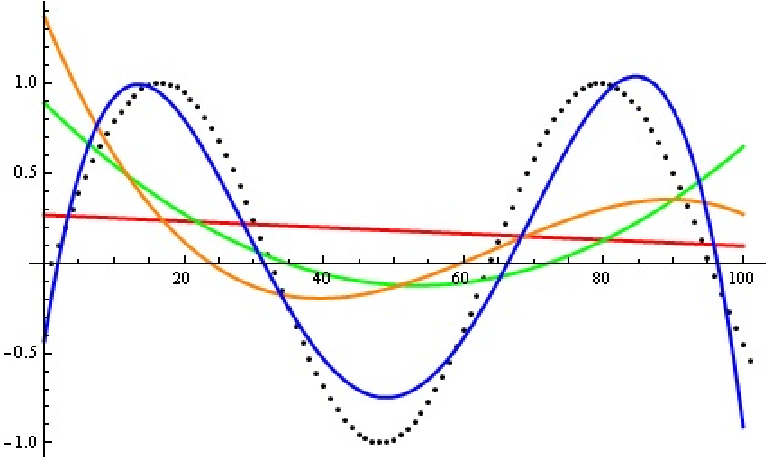


Polynomial Example: Income and Age

- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + u$
- 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

Conclusion

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

- OLS in its full glory
- Reading:
  - Practice up on matrices
  - Fox Chapter 9.1-9.4 (skip 9.1.1-9.1.2) Linear Models in Matrix Form
  - Aronow and Miller 4.1.2-4.1.4 Regression with Matrix Algebra
  - Optional: Fox Chapter 10 Geometry of Regression
  - Optional: Imai Chapter 4.3-4.3.3
  - Optional: Angrist and Pischke Chapter 3.1 Regression Fundamentals

Fun With Interactions

Remember that time I mentioned people doing strange things with interactions?

Brooks and Manza (2006). "Social Policy Responsiveness in Developed Democracies." *American Sociological Review*.

Breznau (2015) "The Missing Main Effect of Welfare State Regimes: A Replication of 'Social Policy Responsiveness in Developed Democracies."' *Sociological Science*.

Original Argument

- Public preferences shape welfare state trajectories over the long term
- Democracy empowers the masses, and that empowerment helps define social outcomes
- Key model is interaction between liberal/non-liberal and public preferences on social spending
- but. . . they leave out a main effect.

Omitted Term

- They omit the marginal term for liberal/non-liberal
- This forces the two regression lines to intersect at public preferences $= 0$.
- They mean center so the 0 represents the average over the entire sample

What Happens?

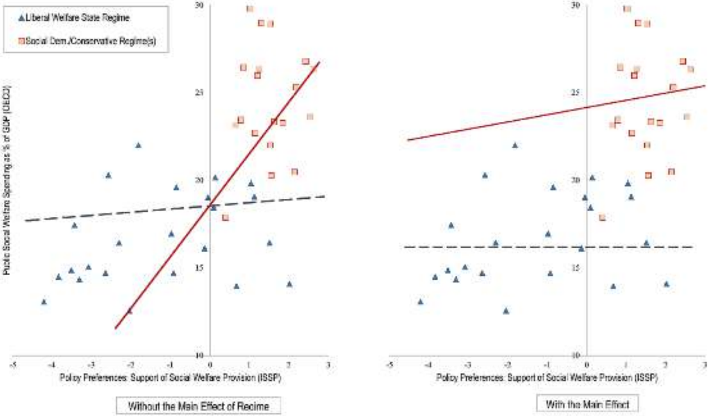


Figure 1: Predicted Regression Lines for the Effect of Policy Preferences on Social Welfare Spending, without and with the Main Effect of Regime

Moral of the Story

Seriously, don't omit lower order terms.

<drops mic>

References

Acemoglu, Daron, Simon Johnson, and James A. Robinson. “The colonial origins of comparative development: An empirical investigation.” *American Economic Review*. 91(5). 2001: 1369-1401.

Fish, M. Steven. "Islam and authoritarianism." *World politics* 55(01). 2002: 4-37.

Gelman, Andrew. *Red state, blue state, rich state, poor state: why Americans vote the way they do*. Princeton University Press, 2009.