Introduction to Empirical Methods

Lecture 3: Introduction to Multiple Regression

Quasi-Experiment Review

- Experiments ← Random Assignment
  - Treatment Group and Control Group
    - Both groups should have similar characteristics
- The typical Political Science study does not have random assignment
  - Example: Does religion influence approval of the President?
    - People are not randomly assigned to a religion.
  - Random assignment may not be ethical.

Random Assignment Cures Many Problems

- Since the groups are the same, changes are the result of the treatment.
  - Internally valid
- Without random assignment, we need to statistically control for potential confounding variables.

Why Multiple Regression?

- Goal: can we infer that our independent variable causes our dependent variable?
  - We might have multiple hypotheses.
- Regression models only show correlation.
  - We must still infer causation.
- To infer causation, we must rule out alternative explanations.
- Spuriousness
  - Is there another factor that you're not considering?
  - $z \rightarrow x, z \rightarrow y$.

The Multiple Regression Model

- $Y = \alpha + \beta_1 X + \beta_2 Z$
  - Y =Dependent Variable.
  - X =Independent Variable.
  - Z =Controlling for Spuriousness.
- Including Z in the model allows us to examine the effect of X holding Z constant.
  - Therefore, we'll know how X influences Y without worrying about Spuriousness.
    - Don't be confused by the notation...
    - X and Z are both independent variables.
- Instead of a line through points, it's drawing a plane through the points in multiple dimensions.

Bivariate Regression Model vs. Multiple Regression Model

- The extension from the bivariate model to the multiple regression model is simple.
- Bivariate Regression
  - $Y = \alpha + \beta X$
- Multiple Regression
  - $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 ... + \beta_k X_k$
- Each β tells us the partial effect of each different independent variable, holding the others constant.
- Perform separate t-tests to test for statistical significance of each independent variable.

Comparing the Two Models

- Bivariate Regression
  - $Y = \alpha + \beta X$
- Multiple Regression with One Control Variable
  - $Y = \alpha + \beta_1 X_1 + \beta_2 X_2$
- The key to note is that the β in the bivariate regression model will be different from β_1 in the multiple regression model.
  - It will be very different if the relationship between X_1 and Y is spurious.

How Does the Model "Control" for Z?

- Z is related to both X and Y.
  - Z explains some of the variation in X.
  - Z explains some of the variation in Y.
- The formula for β_1 does not use all of the variation in X and Y.
  - The formula for β_1 uses the variation in X and Y that cannot be explained by Z.

What is the Model Telling Us?

- $Y = \alpha + \beta_1 X + \beta_2 Z$
  - β_1 =The effect of X on Y controlling for Z.
  - β_2 =The effect of Z on Y controlling for X.
- With more than two independent variables, each β is the effect of that particular independent variable holding all other independent variables constant.

The Idea in Abstract

- What caused someone to vote for President Obama?
  - Income, partisanship, ideology.
- $Vote = \alpha + \beta_1 Income + \beta_2 Partisanship + \beta_3 Ideology$

The Example Continued

- Imagine you want to know the effect of income on the vote.
- Ideology and partisanship will affect the vote, but they are also related to income.
- $Vote = \alpha + \beta_1 Income + \beta_2 Partisanship + \beta_3 Ideology$
  - β_1 would tell you the effect of income on the vote, *all else equal*: that is, only income varies, and we hold everyone's ideology and partisanship constant. This allows us to compare two people with identical ideology and partisanship who differ only in income.

The Bivariate Regression

- Dependent Variable
  - 1=Voted for Obama; 0=Voted for McCain
- Independent Variable
  - 0–4=Increasing Income Categories

Bivariate Regression Results

Variable	Estimate	Standard Error	T-Ratio
Income	-0.080	0.010	-7.94
Constant	0.767	0.017	43.82

- Predicted Values: $Y = \alpha + \beta X$
  - If a person is in the lowest income category, there is a 76.7% probability they voted for Obama.
  - If a person is in the highest income category, there is a 44.7% probability they voted for Obama.
    - That would be a difference of 32 percentage points.

Multiple Regression Model

- Same variables as the bivariate model.
- Additional Independent Variables (denoted Z above).
  - Partisanship:
    - -1=Republican
    - 0=Independent
    - 1=Democrat
  - Ideology:
    - 1=Very Conservative to 7=Very Liberal

Multiple Regression Results

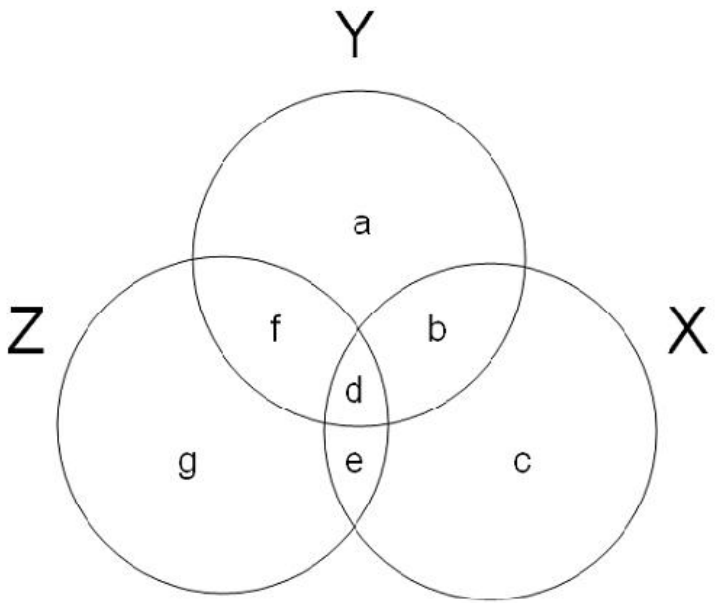
Variable	Estimate	Standard Error	T-Ratio
Income	-0.016	0.008	-1.95
Partisanship	0.338	0.015	22.81
Ideology	0.071	0.008	9.28
Constant	0.302	0.032	9.43

- If a person is in the lowest income category, there is a 63.54% probability they voted for Obama–holding other variables at their means.
- If a person is in the highest income category, there is a 56.97% probability they voted for Obama–holding other variables at their means.
  - That would be a difference of 6.57 percentage points.

What Changed?

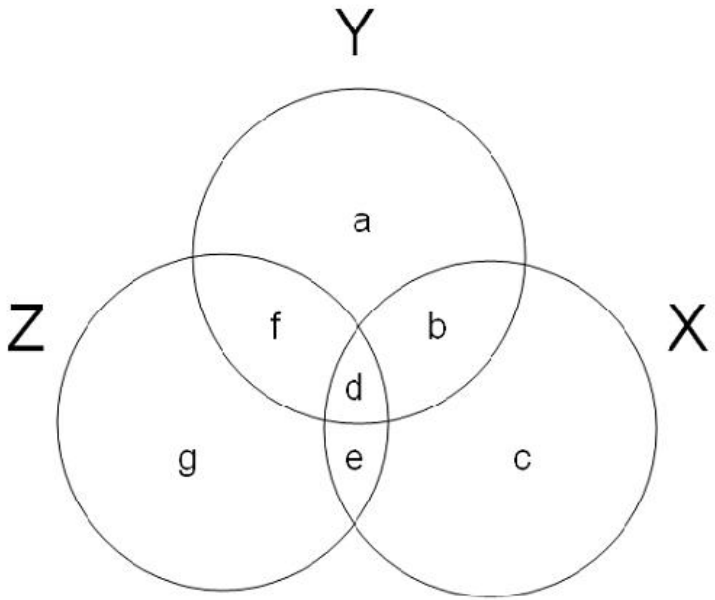
- Income is related to partisanship and ideology:
  - In the first model, the coefficient for income was capturing the effect of income and some of the effect of partisanship and ideology.
- The bivariate regression model has **omitted variable bias**.
  - That is, our estimate of β_1 is incorrect (b/c it includes the impact of the Z variables as well as X).
  - It likely overestimates the effect of the independent variable on the dependent variable.

Another Way of Looking at It



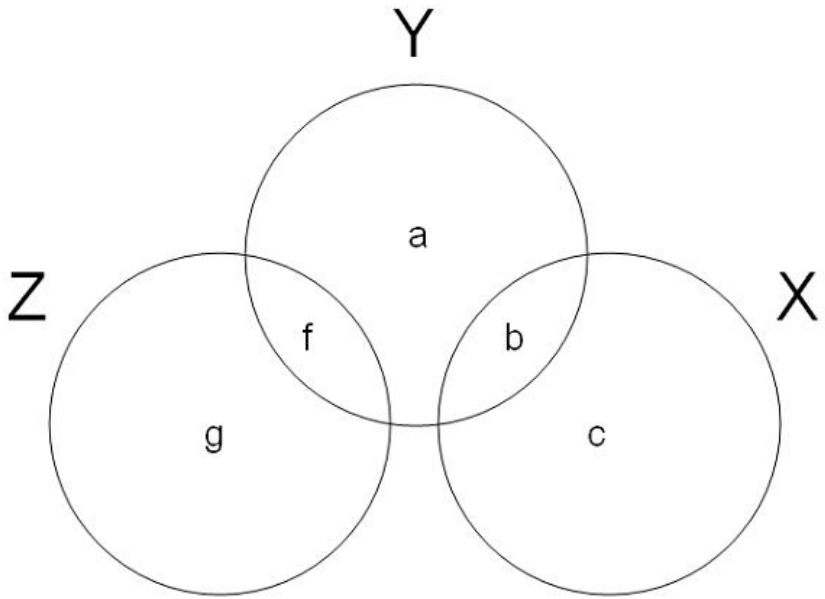
- Y is related to X and Z.
- X and Z are also related.

Another Way of Looking at It



- If we exclude Z from the model, the effect of X on Y is $d + b$
  - But this is a mistake because d is the portion of the variance in Y shared by both X and Z.

What if X and Z are Not Related?



- If we exclude Z from the model, the effect of X on Y is b
- If we include Z in the model, the effect of X on Y is b
  - If X and Z are not correlated, then it does not matter if we include Z.
    - This is rarely the case.

Differentiation

- Discrete versus Instantaneous Change
- Secants versus Tangents
- Limits
- Formal Definition: $lim_{h \to 0} \frac{f(x+h)-f(x)}{h}$
- Notation: $f'(x)$, $\frac{df}{dx}$, $\frac{dy}{dx}$
- Interpretation
- Partial Derivatives

Matrices

- Notation
- Utility for Multiple Regression
- Definitions (Diagonal, Identity, Symmetric, Transpose)
- Addition, Scalar Multiplication
- Matrix multiplication
- Determinant
- Inverse
- Use in OLS

Are these Models any Good?

- Bivariate Regression.
  - $Vote = \alpha + \beta_1 Income$
    - $R^2 = 0.04$
    - root MSE= 0.46
- Multiple Regression.
  - $Vote = \alpha + \beta_1 Income + \beta_2 Partisanship + \beta_3 Ideology$
    - $R^2 = 0.55$
    - root MSE= 0.33
- What does this tell us about our models?

A Real Article

- Gartner, Scott. 2004. "Making the International Local: The Terrorist Attack on the U.S.S. Cole, Local Casualties, and Media Coverage." *Political Communication*, 21:139-159.
- Data
  - Analysis of the websites of 24 newspapers.
- Dependent Variable
  - The number of times the newspaper covered the U.S.S. Cole attack from October 31 to December 24.
- Independent Variables
  - Casualties
    - A count of how many people from the newspaper's community were killed.
  - National
    - A dummy variable indicating that the newspaper is a "national newspaper" like *The New York Times* or the *Washington Post*.

The Results

	Model 1		Model 2	
Variable	Estimate	Standard Error	Estimate	Standard Error
Casualties	5.656	0.783	4.593	0.874
National	7.107	1.814		
Constant	9.393	1.103	11.330	1.179
R^2	.592		.433	
root MSE	4.236		4.880	
N	24		24	

- How many degrees of freedom in model 1?
- How many degrees of freedom in model 2?

The Results

	Model 1		Model 2	
Variable	Estimate	Standard Error	Estimate	Standard Error
Casualties	5.656	0.783	4.593	0.874
National	7.107	1.814		
Constant	9.393	1.103	11.330	1.179
R^2	.592		.433	
root MSE	4.236		4.880	
N	24		24	

- Which variables are statistically significant in model 1? Critical value is 2.08.
- Is casualties statistically significant in model 2? Critical value is 2.07.

The Results

	Model 1		Model 2	
Variable	Estimate	Standard Error	Estimate	Standard Error
Casualties	5.656	0.783	4.593	0.874
National	7.107	1.814		
Constant	9.393	1.103	11.330	1.179
R^2	.592		.433	
root MSE	4.236		4.880	
N	24		24	

- According to model 1, about how many stories be written on the Cole by a national newspaper without any casualties in the area?

The Results

	Model 1		Model 2	
Variable	Estimate	Standard Error	Estimate	Standard Error
Casualties	5.656	0.783	4.593	0.874
National	7.107	1.814		
Constant	9.393	1.103	11.330	1.179
R^2	.592		.433	
root MSE	4.236		4.880	
N	24		24	

- What do the differences in the R^2 values tell us?
- What do the differences in the root MSE values tell us?