ECON B225: ECONOMIC DEVELOPMENT

Sebastian Anti
Assistant Prof. of Economics, Bryn Mawr College

Admin

- **Module II: Econometrics Primer**
  - Goal: To familiarize you with empirical research common in development economics for you to critically examine research in the discipline
- **Module III is coming up**
  - Reading research papers, understanding their contributions and limitations
  - Article briefs: 1 page summaries of papers, their main learnings and limitations
    - Work with a partner that I'll assign randomly
    - Due 24 hours before the class session so your peers can review them

Intro to Econometric Research

Summary

- **The Solow Model and the Australian Model provide frameworks for thinking about economic growth in the short term and the long term**
- **BUT… they leave a lot unanswered. Examples**
  - What is A (TFP) really composed of in the Solow Model?
  - How do changes in components of A actually affect development measures?
  - Does natural resource wealth actually lead to economic instability?

Solow Model in Particular

- **The Solow Model in particular generates a lot of speculation**
  - Anything that plausibly affects TFP in any way becomes a potential question
- **Examples:**
  - Does political corruption cause slower growth? If so how much?
  - Does investment in education cause growth? If so how much?
  - Does trade liberalization cause growth? If so how much?

Research Questions

- **Notice that each of these questions has a particular form:**
  - Does X (example: investment in education) cause Y (example: change in student test scores)? If so how much?
- **<u>Answering these types of questions is the goal of empirical research in economics</u>**
- **Providing answers to these questions reveals and quantifies the various channels through which development can occur**

What is Econometrics?

- **Broadly Speaking: Econometrics is the methodology in economics for finding support for economic theories about how the world works**
  - It is a "way of knowing"
  - Effort to measure causal effects similar to lab experiments in the natural sciences <u>but in a social science setting</u>

- **More specifically, econometrics is the use of data for:**
  - Quantitatively measure the size of economic relationships and economic phenomena
  - Test economic theories
  - Predict economic outcomes (not the focus in this class)

The Third Module of the Class will Expose you to Research providing Answers to these Questions

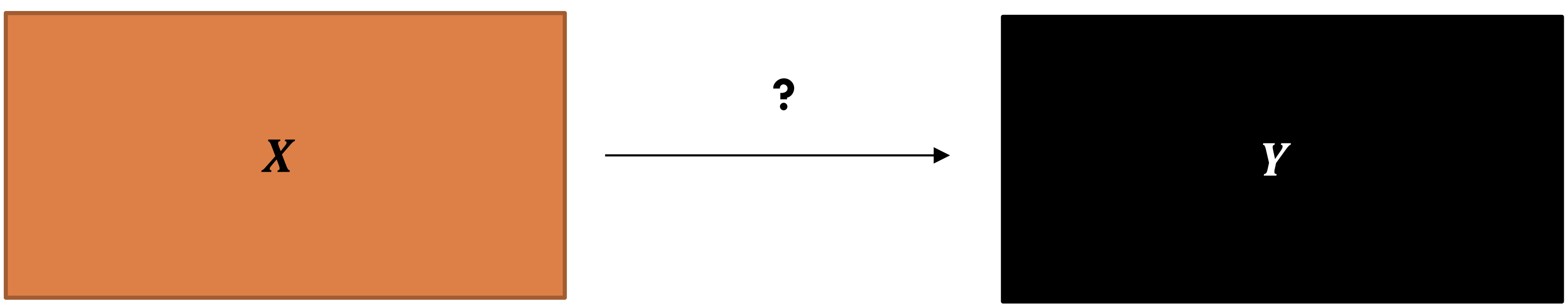
- **First we need to get a grounding in econometric methods**
- **Goal:**
  - Understand basics of econometric methods
  - Understand how to read econometric research and interpret findings
  - Understand the main arguments for causal inference of an effect of X on Y
  - Understand the assumptions that econometricians are making to make causal statements

Remember

- **Econometrics can be tough!**
- **Much of this will be unfamiliar to you**
- **Remember, on the final exam you get one sheet of paper with anything you want to write on it (front and back). So you will not need to memorize all this information**
  - Focus on the intuition and the “rules” or valid applications of these techniques. When you read papers later in the class, you will want to ask whether authors follow these rules and if they don’t what it might mean for their findings

Goal of Econometrics

- **All econometrics papers answer the same type of question: <u>Does X cause Y? If so, how much?</u>**
- **You can represent this in a simple flow chart called a <u>Directed Acyclic Graph (DAG)</u>**



Goal of Econometrics

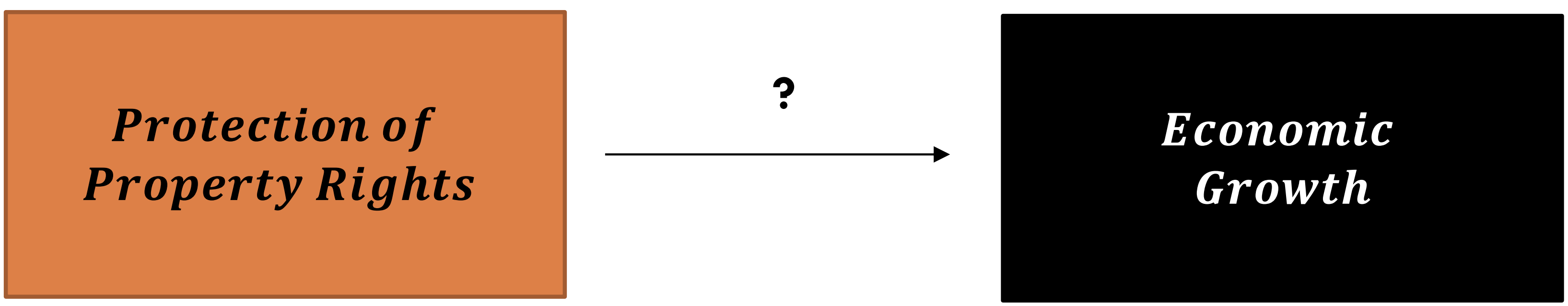
- **Another way to phrase this is: if X changes by some amount, how much does that change Y?**
- **For an example, we're going to work with the question from a famous paper from AJR (2001). Here is the citation:**

Acemoglu, Daron, Simon Johnson, and James A. Robinson. 2001. "The Colonial Origins of Comparative Development: An Empirical Investigation." *American Economic Review*, 91 (5): 1369-1401

AJR (2001)

- **Research Question:**
  - Does the strength of a country's protection of property rights affect economic growth? If so, how much?

- **Representing this in a DAG:**



AJR Hypothesis

- **What do you think the relationship is?**
  - If protection of property rights goes up, would you expect GDP per capita to go up, all else constant?
    - Then you hypothesize the relationship is positive
  - If protection of property rights goes up, would you expect GDP per capita to go down, all else constant?
    - Then you hypothesize the relationship is negative
  - If you think there are reasons for both a negative and positive effect, then you want to ask "which one dominates (is stronger)?"
    - This helps you to know "the net effect"

The Hypothesized "True Relationship"

- **Your hypothesis implies an anticipated relationship between the two variables**
- **Some notation:**
  - Define the outcome variable (Y) as $GDPperCap_c$ as a measure of country c's GDP per capita measured at some period in time
  - Define the independent variable (X) as $PropProtect_c$ a measure of a country's level of property rights protection measured at some period in time

The Hypothesized "True Relationship"

- Using this notation, our hypothesis suggests that we think that

$$GDPperCap_c = f(PropProtect_c) \tag{1}$$

- Let's assume we think that this relationship is positive, and let's also assume for simplicity that it is linear. If we do, we can express $f(PropProtect_c)$ as follows,

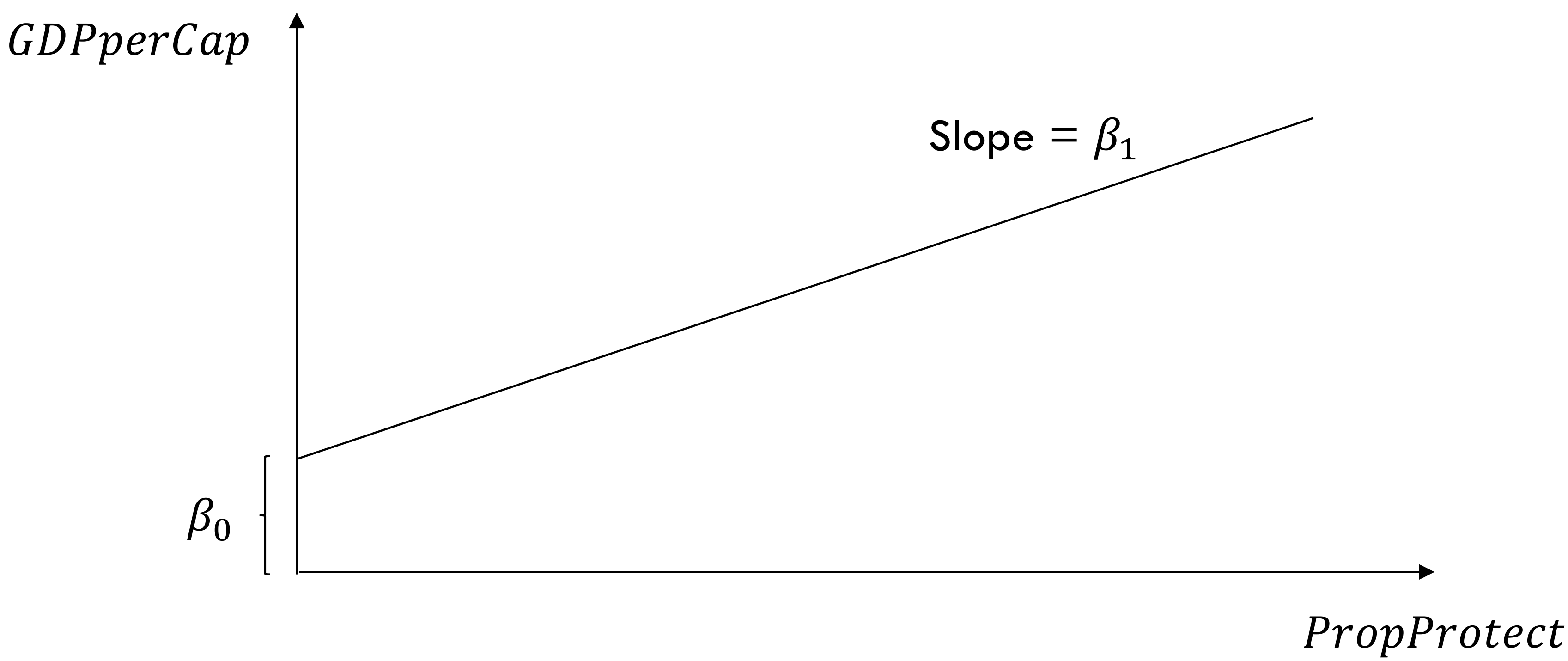
$$GDPperCap_c = f(PropProtect_c) = \beta_0 + \beta_1 PropProtect_c \tag{2}$$

The Hypothesized "True Relationship"

$$GDPperCap_c = f(PropProtect_c) = \beta_0 + \beta_1 PropProtect_c \qquad (2)$$

- Where, β_0 is a constant and β_1 is a slope coefficient

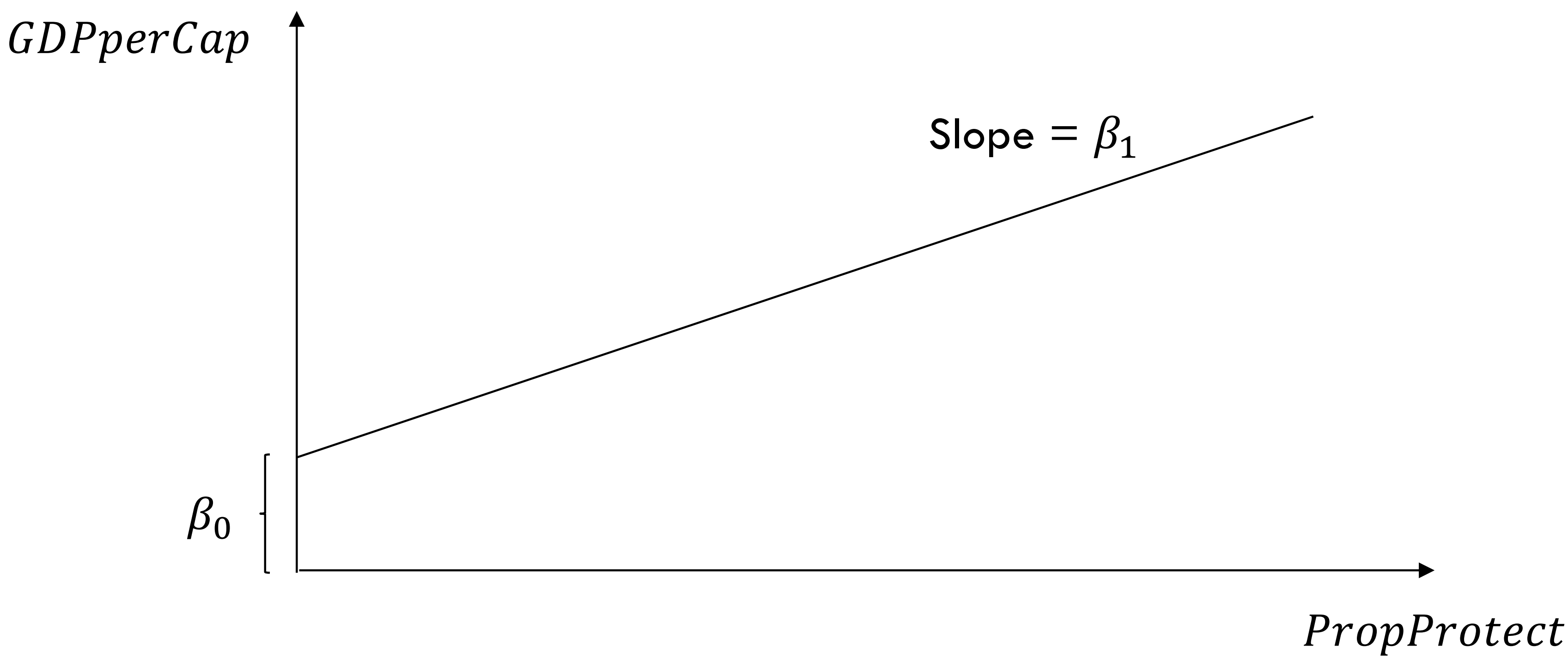
- **This suggests we think that the relationship looks something like:**



The Hypothesized "True Relationship"

- **Econometrics answers the question of how does the protection of property rights affect growth by using data to find evidence for this line and measuring it if it exists**
  - To measure the line, we just need to know β_0 and β_1

$$GDPperCap_c = f(PropProtect_c) = \beta_0 + \beta_1 PropProtect_c \tag{2}$$



Measuring the Relationship

Country (c)	$GDPperCap1995_c$	$PropProtection_c$
Angola	7.771	5.364
Argentina	9.133	6.386
Austria	9.898	9.318
Burkina Faso	6.846	4.455
Bangladesh	6.877	5.136
The Bahamas	9.285	7.5
And so on…	…	…

- **To measure the relationship, you would go out and collect data on GDP per capita and and a measure of a country's protection of property rights**
- **This is an excerpt of the Data from AJR (2001)**

A Note on the Measures

- **For a measure of a country’s property rights protection, AJR uses an index of the level of risk for the government expropriation of property**
- **For a measure of a country’s GDP per capita they use the natural log of GDP per capita in 1995**
  - They use the natural log for a couple reasons I’ll discuss later

Some Quick Math

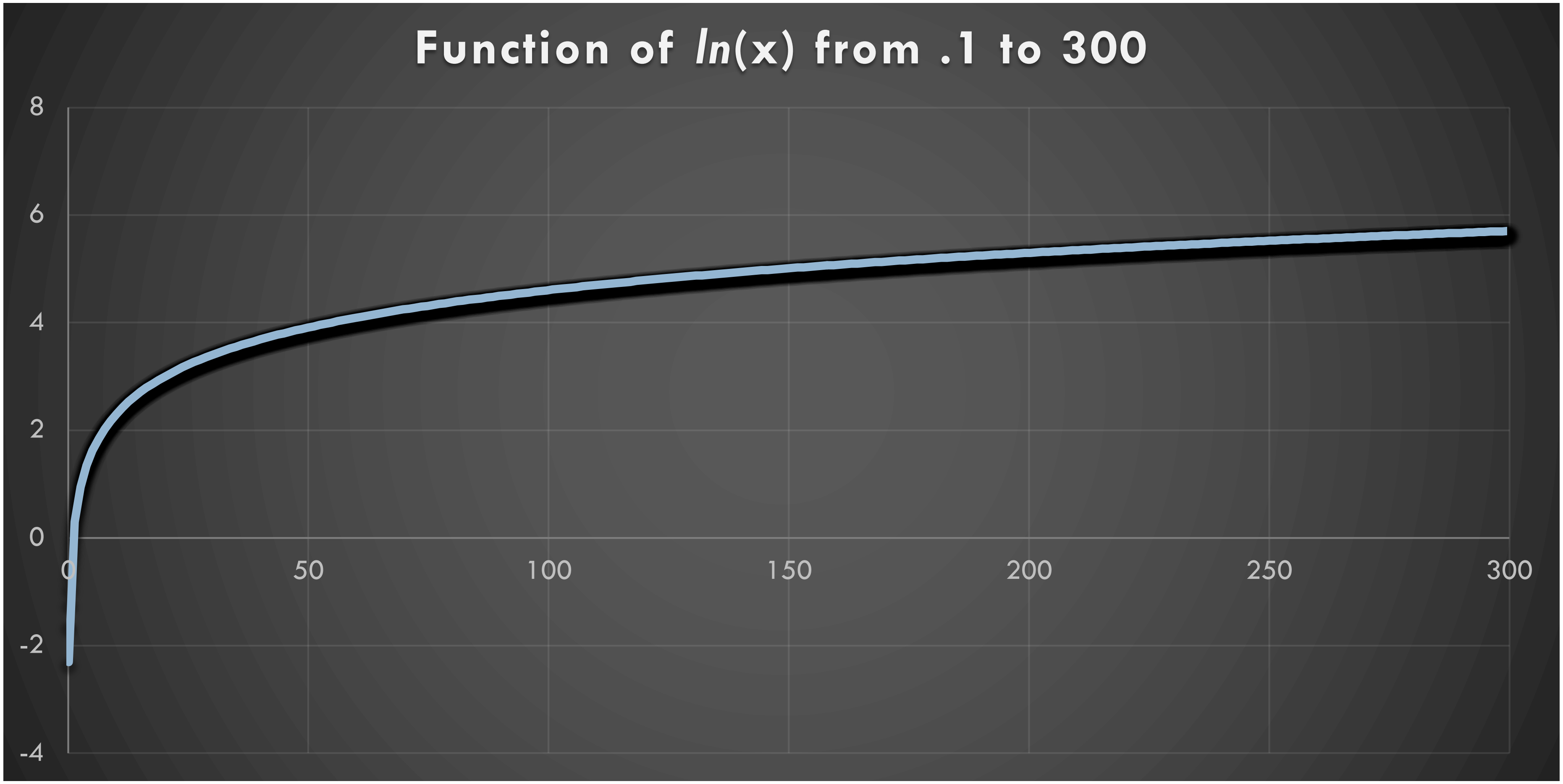
- **A logarithmic transformation of base b of variable X results in the power b would need to be raised to in order to produce X, so**

$$log_b(X) = y, where\ b^y = X \qquad (3)$$

  - Usually we use a base of $e \approx 2.718$ (also known as “Euler’s Number”). This is also known as the “natural log” and is abbreviated $ln()$
  - This is a “monotonic” transformation in that it preserves the order of X

The Logarithmic Function

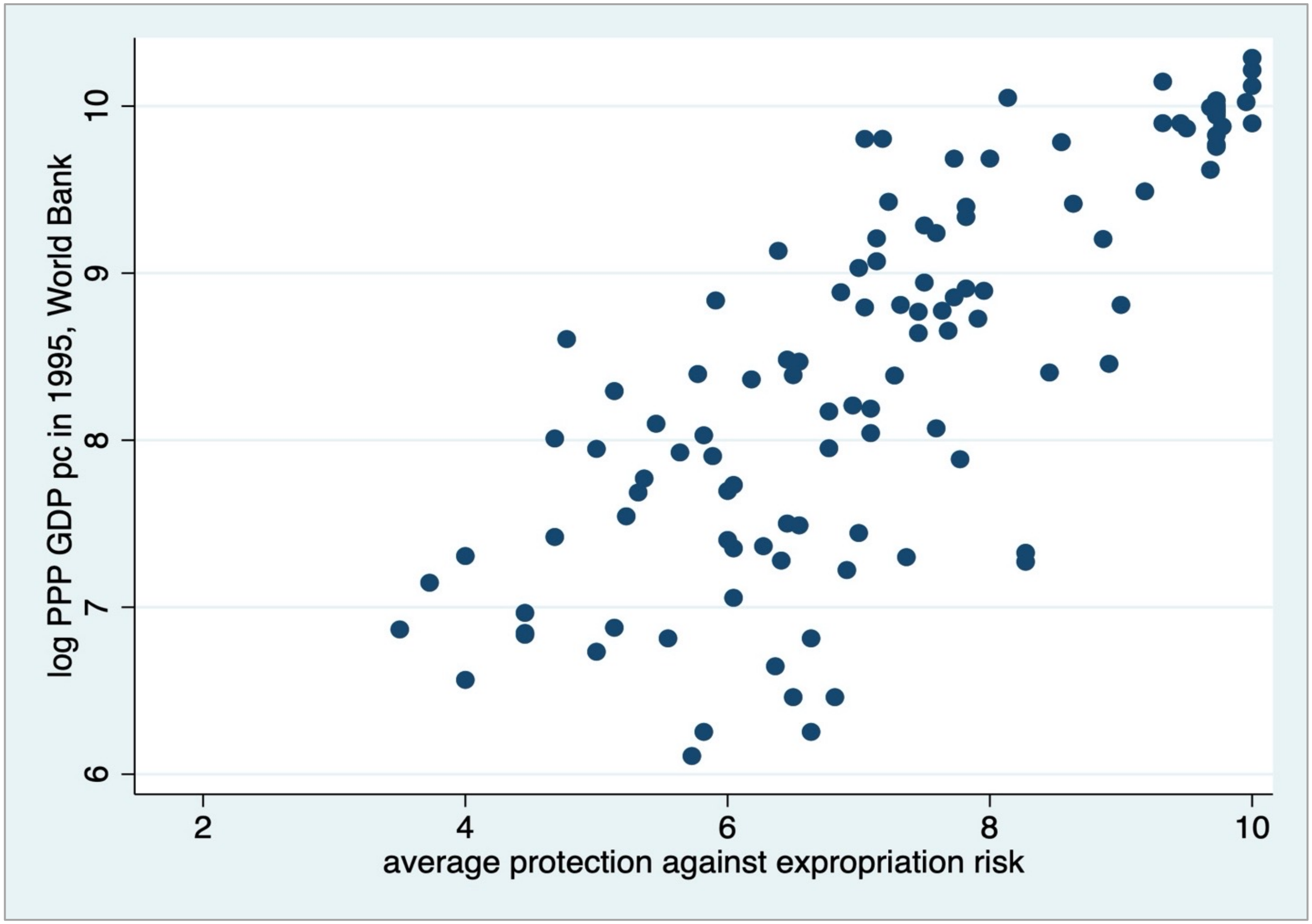
- $f(x) = ln(x)$ **has the following shape:**



Building our Intuition: The Scatter Plot

- Once you have the data, you can simply plot them in a scatter plot
- This provides a sense of how (log of) GDP per capita varies, or changes, as the level of expropriation risk changes
- Here's the scatter plot of the AJR (2001) data (next slide)

Building our Intuition: The Scatter Plot



- Notice how the data seem to conform to a <u>true</u> underlying trend that is unobserved

What you Should Notice

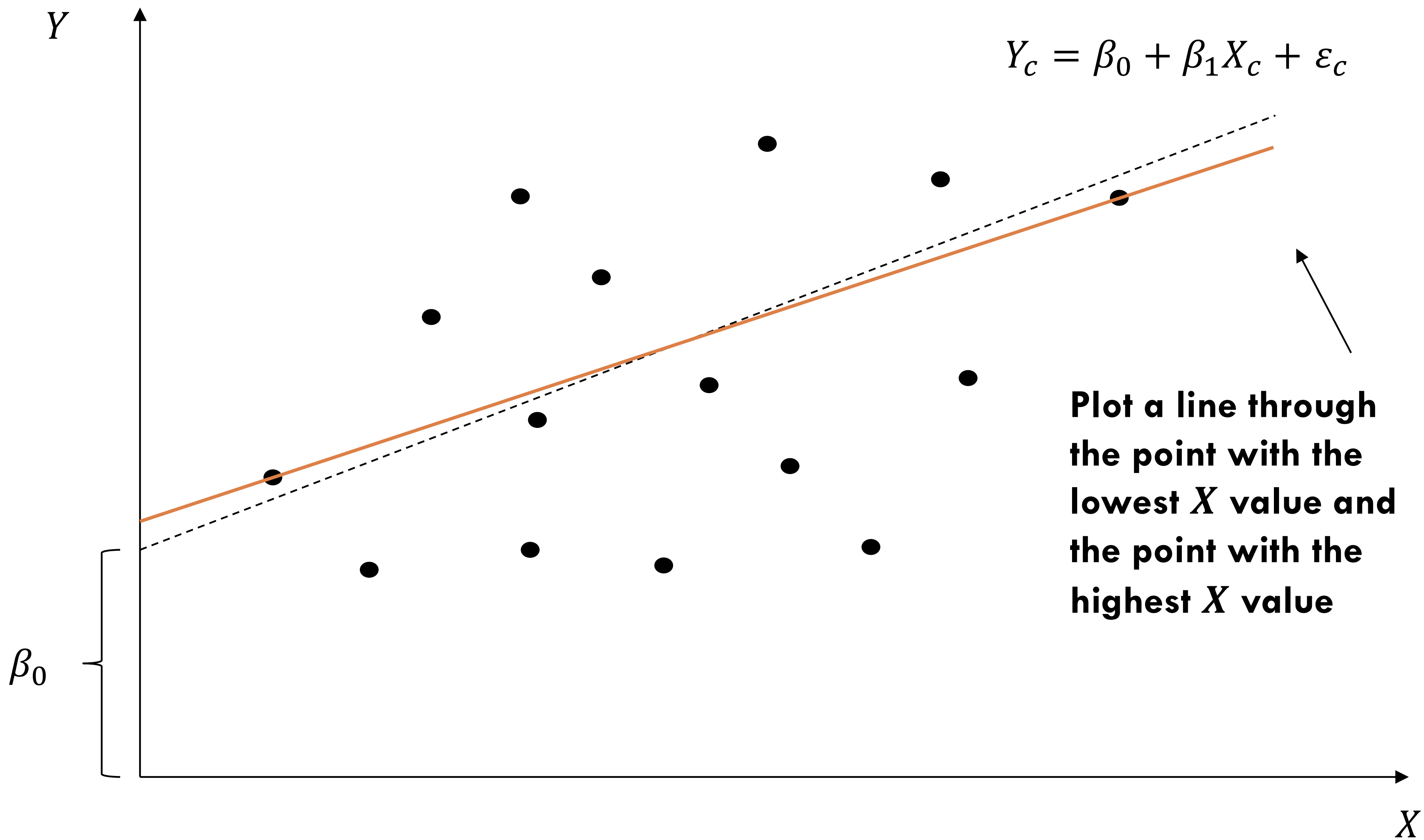
- **The data suggest a positive <u>true relationship that we can't actually see</u>**
- **But there is a lot of "noise" in the data**
  - In other words, the points don't fall exactly on what looks to be the true general trend underlying the data
- **This noise around the true relationship is called the "true error" and it is accounted for in the model with a ε_c, so what we want to estimate is the following:**

$$GDPperCap_c = \beta_0 + \beta_1 PropProtect_c + \varepsilon_c \qquad (4)$$

Estimating the Line of Best Fit

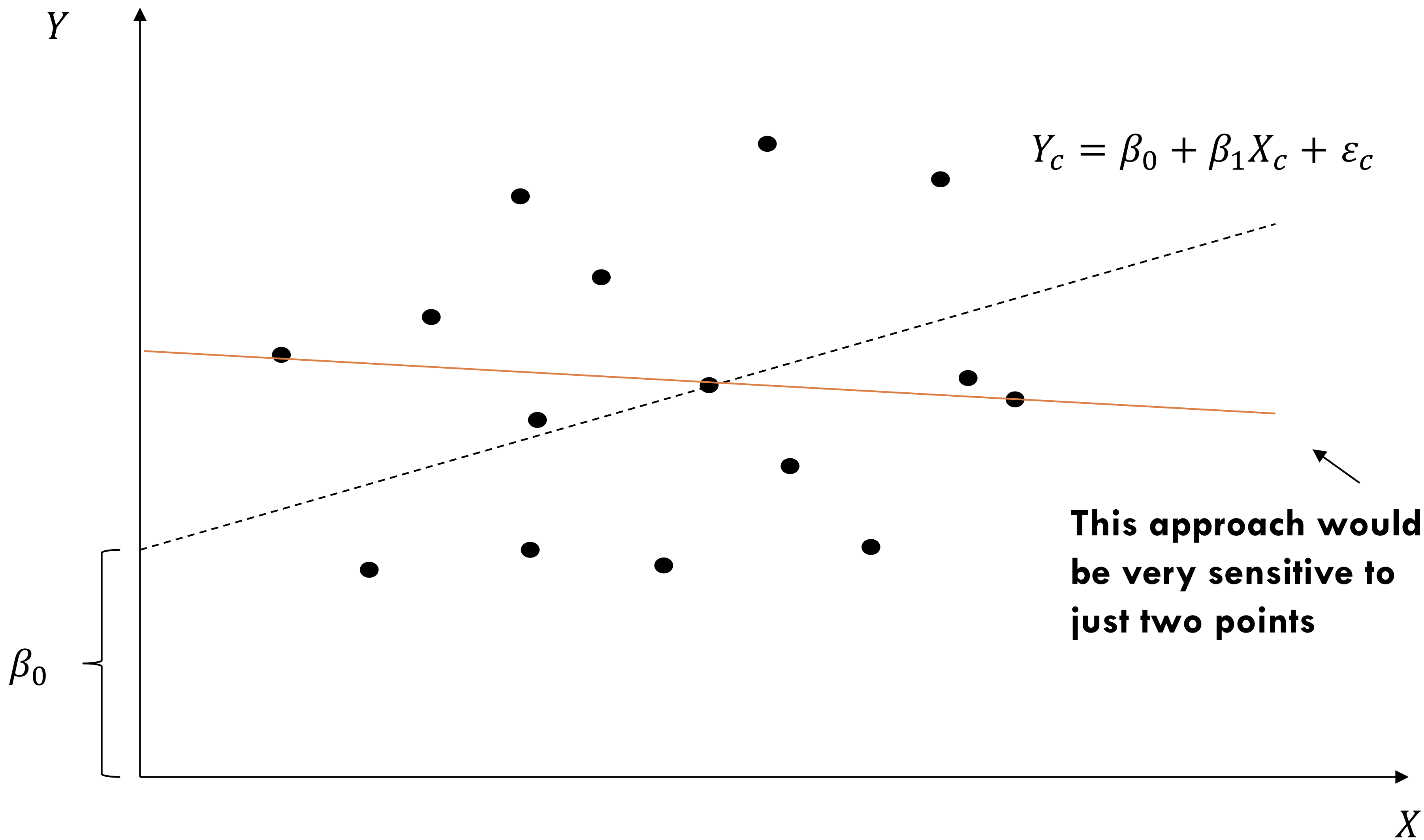
- To actually estimate β_0 and β_1 we need an algorithm called an "estimator" to plot a line through the middle of the "cloud" of data points
- Any ideas?

A Simple Idea for an "Estimator"

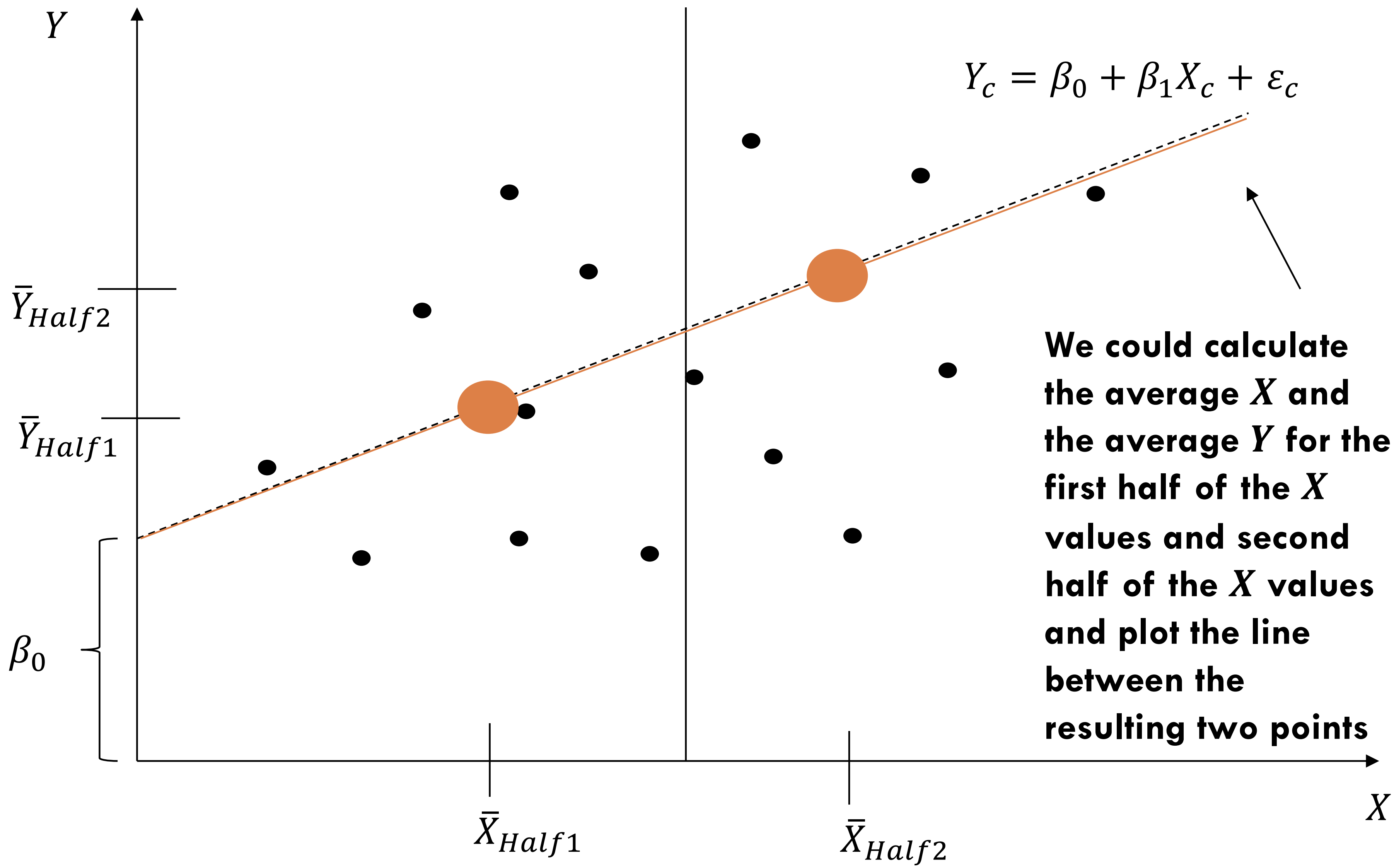


Plot a line through the point with the lowest X value and the point with the highest X value

A "Bad Property"



Another Example of an "Estimator"

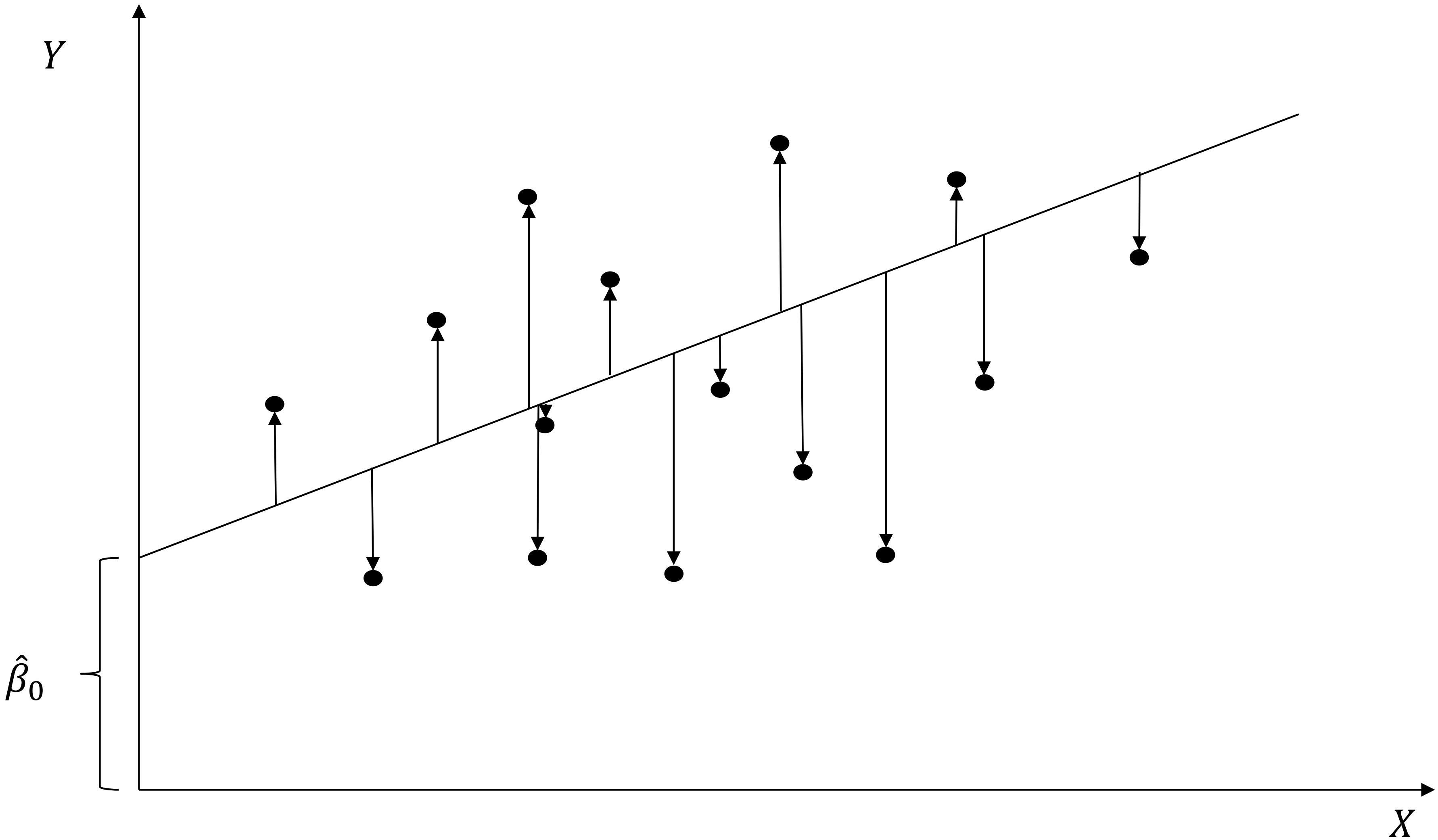


We could calculate the average X and the average Y for the first half of the X values and second half of the X values and plot the line between the resulting two points

The Consensus Estimator in Econometrics: Ordinary Least Squares (OLS)

- **The estimator that most of econometrics uses is the Ordinary Least Squares (OLS) estimator**
- **The basic idea: find the line that "minimizes the sum the squared differences between each data point and the line"**
  - These "differences between each data point and the estimated line" are called "<u>residuals</u>"

Illustrated Residuals



The Main Estimator in Econometrics: Ordinary Least Squares (OLS)

- **The estimator that most of econometrics uses is the Ordinary Least Squares (OLS) estimator**
- **The basic idea: find the line that "<u>minimizes</u> the sum the squared differences between each data point and the line"**
  - These "differences between each data point and the line" are called "<u>residuals</u>"
- **Notice the line is the result of a <u>minimization problem</u>, so calculus allows us to find the equation of the line very easily**

The OLS Formulas

- The OLS formulas for the estimates of the slope coefficient and the constant are

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{N}(X_i - \bar{X})^2} = \frac{\sum_{i=1}^{N}(X_i - \bar{X})Y_i}{\sum_{i=1}^{N}(X_i - \bar{X})X_i} = \frac{Cov(Y,X)}{Var(X)} \quad (4)$$

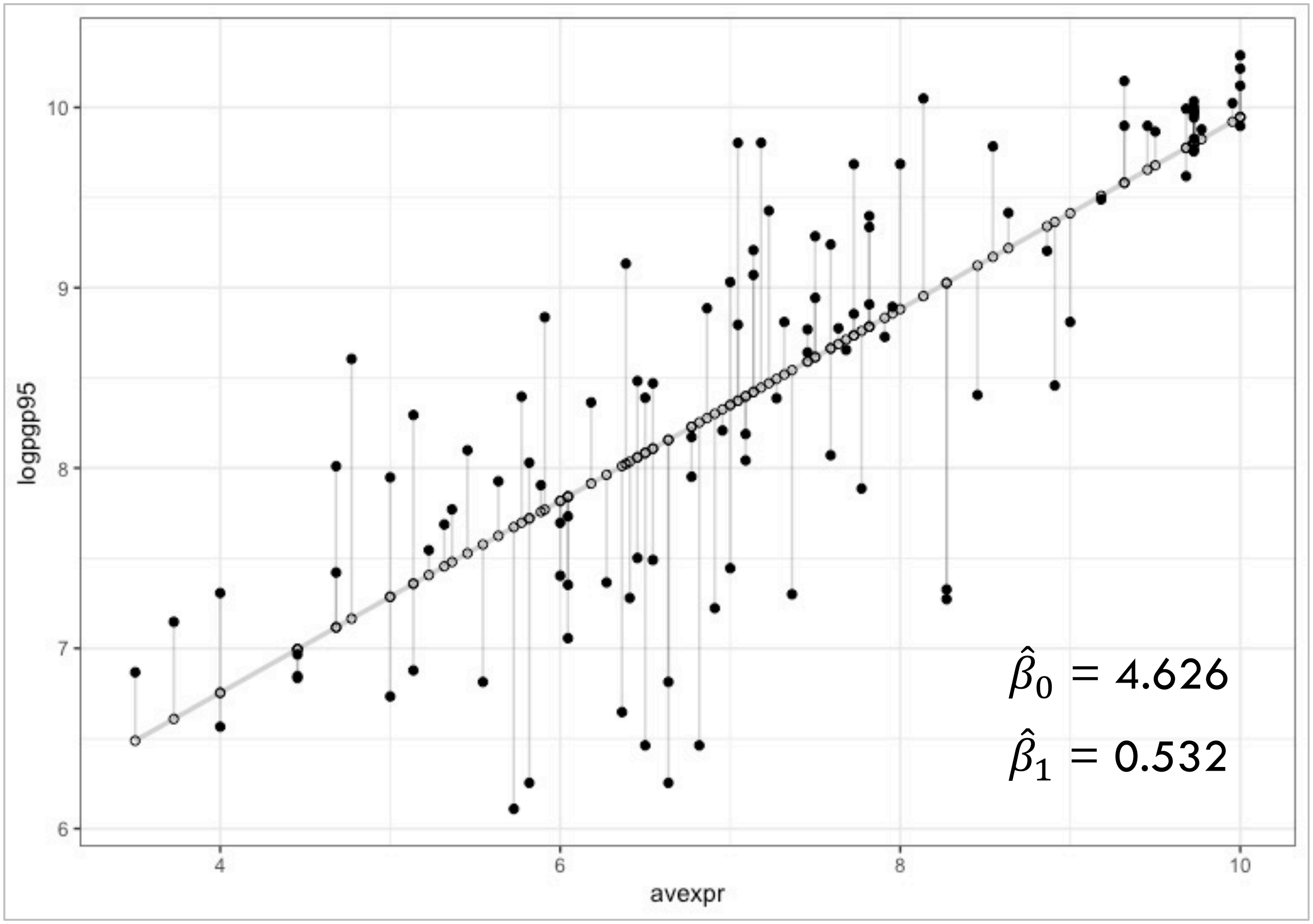
$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X} \quad (5)$$

- The hats on the betas indicate that these are <u>estimates of the true parameters of the true function</u>
- <u>You will not need to use these in this class ever.</u> You just need to see what they are

Why OLS?

- **OLS is favored among economists for several reasons**
  - It's intuitive and transparent
  - It has good statistical properties under certain assumptions, particularly
    - Unbiased
    - Precise
    - Efficient
  - Take ECON B253 if you want to know more about these ideas

Getting Back to AJR (2001)



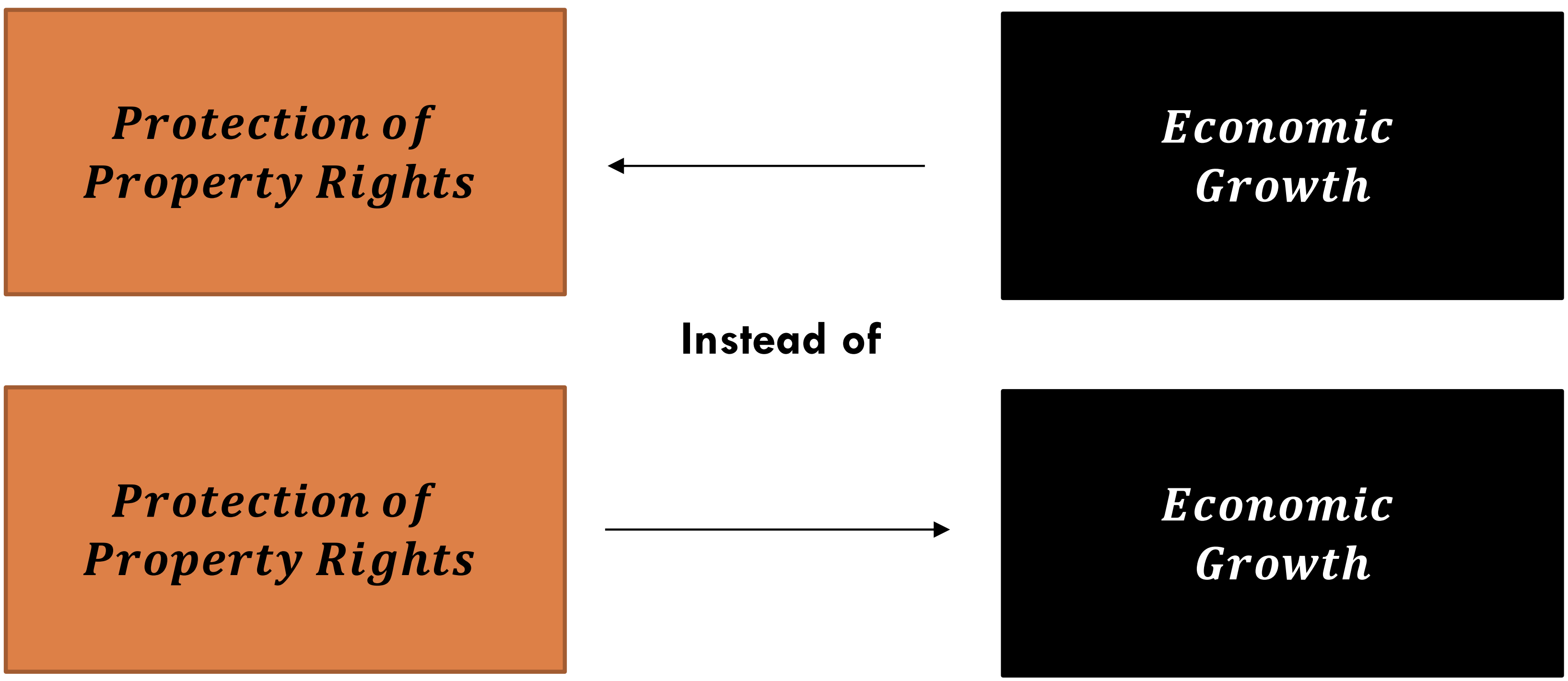
- So if we run OLS and plot the line of best fit with the residuals, here's what it looks like

So Are We All Done?

- **Okay so, for a one unit increase in protection against expropriation, GDP per capita increases by about 53 percent on average… All done right?**
- **No. To make a causal statement like this, we need to assume two things:**
  1. **No reverse causality:** We need to assume that the relationship doesn't go in the other direction. Do you think it is plausible that GDP per capita could have an effect of a country's protection of property rights? If yes, then there is a worry about this
  2. **The error term, ε_c, is independent of our independent variable of interest:** In other words, we need to assume that other things affecting GDP per capita have no relationship with a country's protection of property rights. If you think there is, then we cannot interpret this estimate in a causal way

Reverse Causality

- This is just the idea that rather than the X variable affecting Y, really it's the Y variable that affects X
- In AJR's context, it means the DAG looks like:



Reverse Causality

- **There is no empirical test for this. Whether it is a worry or not is up to you as a reader to decide based on:**
  - Context of the study and the data
  - Economic theory
  - Intuition about how the world works
  - “Social Science Logic”
- **There is no formula here. You need to think critically when reading any paper about whether this is a worry**

The error term, ε_c, is <u>not</u> independent of our independent variable of interest

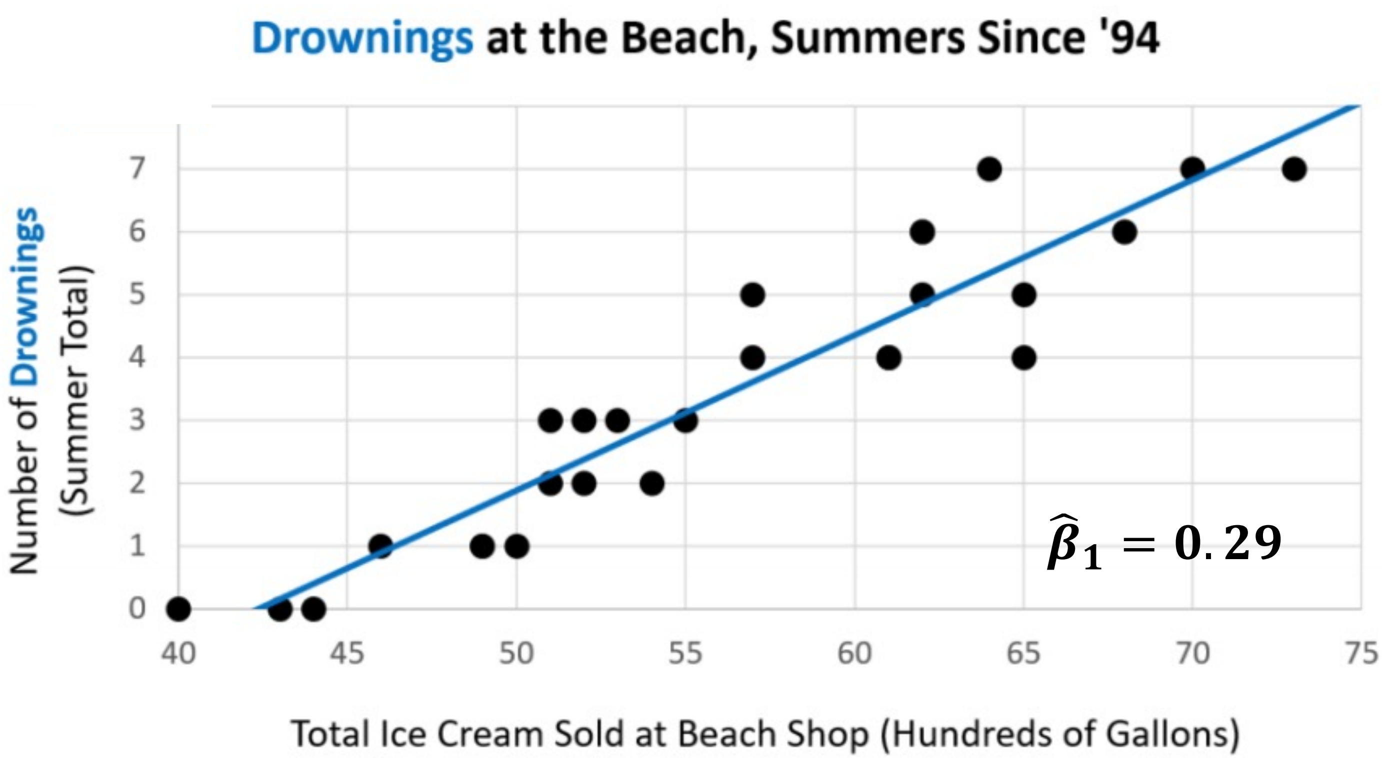
- **This second assumption is usually a lot harder for students to wrap their mind around**
  - What it is saying, roughly, is that we need to assume that there isn't anything else causing Y to change that is also causing X to change
- **If there is, then we say that our estimate of the slope is "endogenous" or "biased" or "misidentified"**
- **Let's consider an extreme, but illustrative example…**

Simple Example of Endogeneity

Summer (c)	$DrowningDeaths_c$	$IceCreamSales_c$
1999	6	62
2000	4	57
2001	0	44
2002	4	62
2003	1	46
2004	3	54
And so on…	…	…

- Suppose we measured the relationship between **ice cream sales** and **drowning deaths** for summers in North America and found the following…

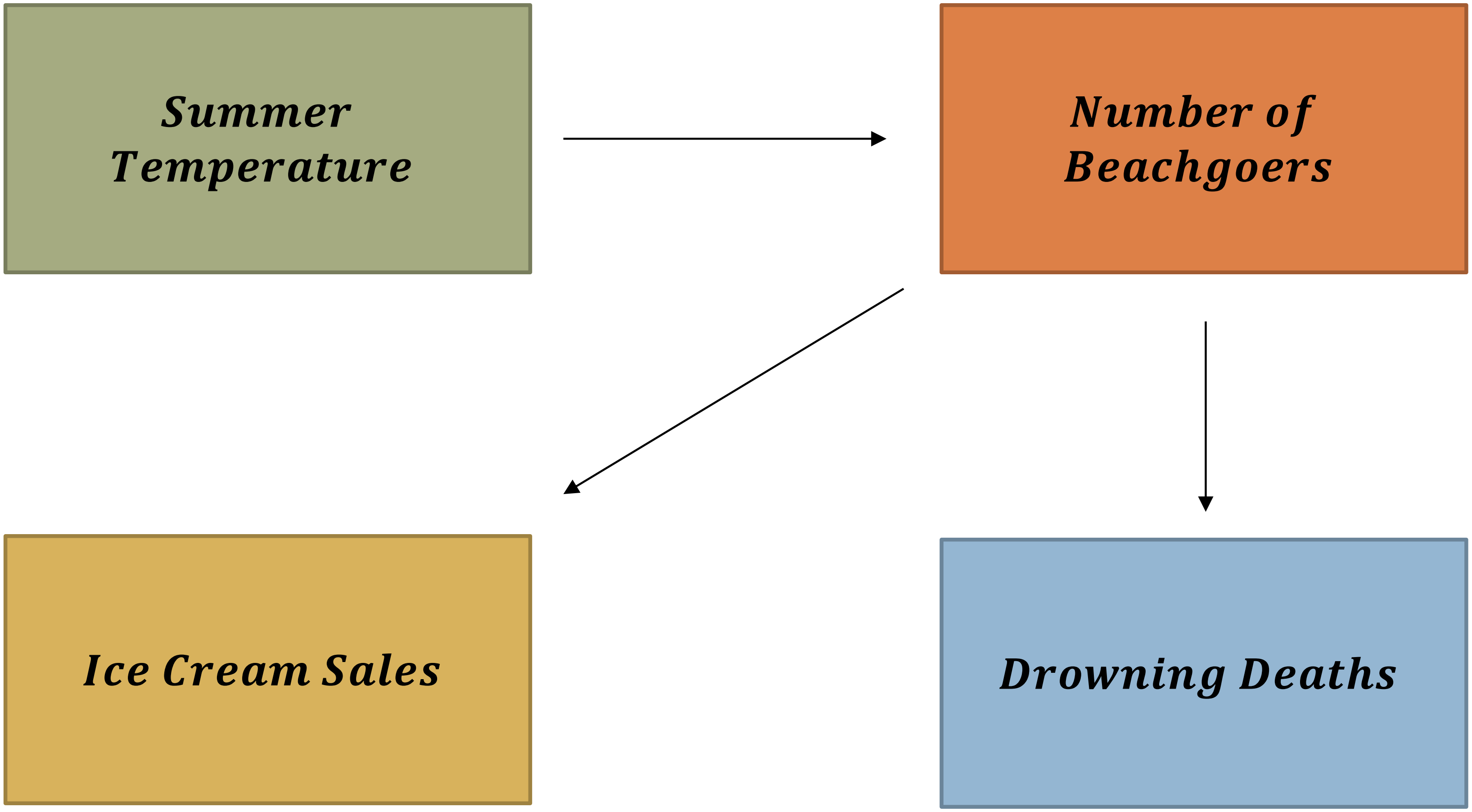
The Scatter Plot



- **The scatter plot suggests a strong positive relationship**
- **Is this causal? In other words, does ice cream cause more drowning deaths?**
  - Why not?

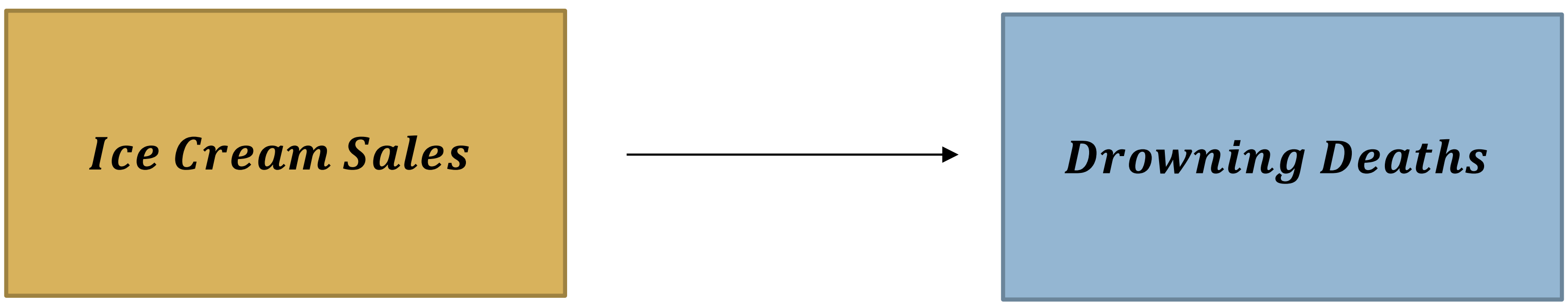
Endogeneity Continued

- Here's the Correct DAG:



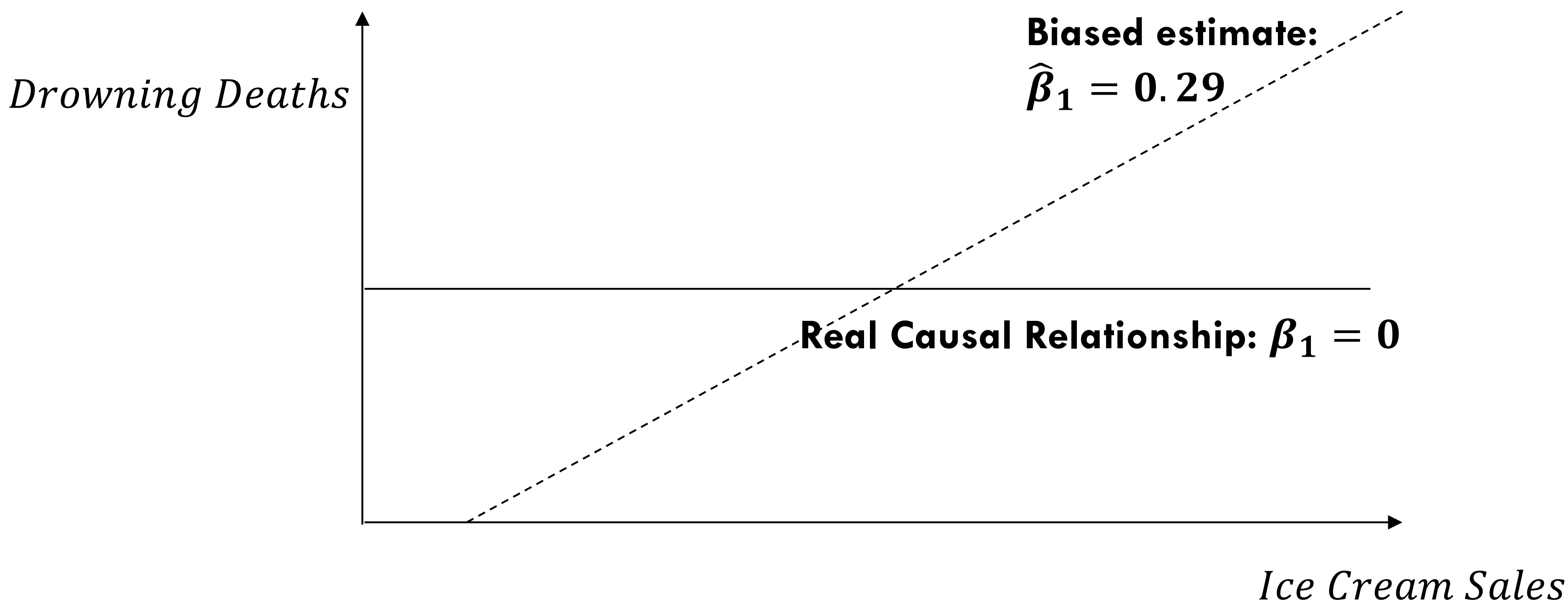
Endogeneity Continued

- Here's the Incorrect DAG:



The True Causal Relationship

- **The <u>True Causal Relationship</u> between ice cream sales and drowning deaths should be 0!**
- **But omitting this third variable (temperature) biases our estimates. Leaving it out gives us an estimate of $\hat{\beta}_1 = 0.29$**



Omitted Variable Bias

- **Having (a) variable(s) omitted like this results in a biased estimate. In other words, <u>the estimate of the slope parameter will not be an accurate estimate of the true relationship that we're trying to measure</u>**
  - This is called "Omitted Variable Bias" or OVB
- **Whether or not there is an omitted variable or not is again up to the reader to decide**
- **There is no formula for figuring out if OVB is a problem**
  - Use your economics theory, intuition, and social science logic

Solutions to Omitted Variable Bias

- **There are solutions to this problem**
  - Each solution requires certain assumptions in order to actually work
- **We'll now go through some of the most common ones**
- **Whether you as a reader of econometrics research accepts the results or not depends on whether you think these assumptions hold**
  - So you will want to think carefully about the research application and whether you "buy" that they hold or not

Solution 1: Add the Omitted Variables to the Model

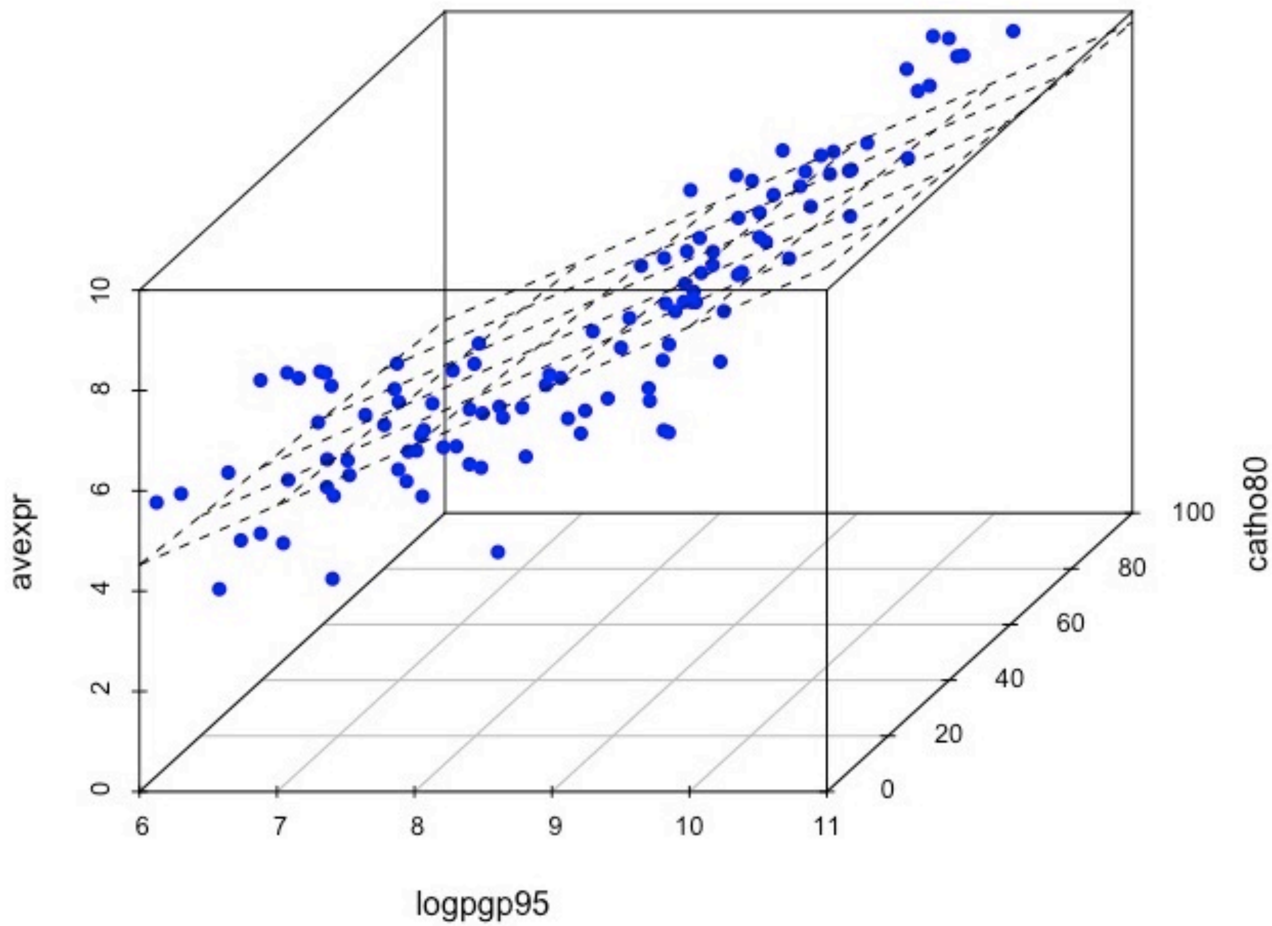
- **We can extend the model from a two-dimensional analysis to a multidimensional analysis**
- **So instead of estimating $\boldsymbol{Y_c = \beta_o + \beta_1 X_c + \varepsilon_c}$, we estimate $\boldsymbol{Y_c = \beta_o + \beta_1 X_{1,c} + \beta_2 X_{2,c} + \cdots + \beta_k X_{k,c} + \varepsilon_c}$**
- **In the context of AJR (2001) we could add a second variable to the estimation. Let's add a measure of the religious composition of each country. The simple model $\boldsymbol{GDPperCap_c = \beta_0 + \beta_1 PropProtect_c + \varepsilon_c}$ now becomes,**

$$GDPperCap_c = \beta_0 + \beta_1 PropProtect_c + \beta_2 PercCath_c + \varepsilon_c \qquad (6)$$

Solution 1: Add the Omitted Variables to the Model

- **Adding this variable brings it out of error term and OLS is able to estimate the effect of the variable on Y separate from the effect of the variable of interest**
  - If you like math, the Frisch-Waugh Theorem proves this
- **Before OLS estimated the slope and constant of a line, what is it estimating now?**
  - The dimensions of a plane! (next slide)

Solution 1: Add the Omitted Variables to the Model



- Here is the scatter plot of these three variables in 3-dimensional space
- **The plane of best fit** is now what OLS is estimating

Solution 1: Add the Omitted Variables to the Model

- The researcher can add more variables to the model, it usually doesn't stop at 2
- After three dimensions the surface that OLS estimates the dimensions of becomes impossible for us to visualize, but it is easy to handle mathematically
- Each variable that one pulls out of the error term and adds to the model should make us more confident that the results don't suffer from OVB

Solution 1: Add the Omitted Variables to the Model

- **Assumption needed for this to solve OVB:**
  - That all the necessary variables are accounted for in the model
- **Reasons this assumption might not hold**
  - The researcher doesn't have data measuring important variables. This could be because
    - The researcher's dataset didn't collect this information or
    - The measure is fundamentally not measurable
- **Note: This doesn't solve reverse causality**

Case when you don't have all the variables to include

- **Unfortunately, this is almost always the case**
  - Why do you think this is?
  - Because many many variables are not measurable
  - Example: Innate ability or talent
- **Think about a Mincerian regression:**
  - Outcome Variable: Wages
  - Ind. Var. of Interest: Level of Education

Solution 2: Add "Fixed Effects"

- **Fixed effects are a series of binary variables (variables equal to 0 or 1) corresponding to "membership" of observations to various groups**
  - For example, countries come from different continents. You can think of the continents as different membership
- **The AJR data will then have something like the following table (next slide)**

Solution 2: Add “Fixed Effects”

Country c)	$GDPperCap1995_c$	$PropProtection_c$	$Africa_f$	$SAmerica_f$	$Europe_f$	$Asia_f$	$Carribean_f$
Angola	7.771	5.364	1	0	0	0	0
Argentina	9.133	6.386	0	1	0	0	0
Austria	9.898	9.318	0	0	1	0	0
Burkina Faso	6.846	4.455	1	0	0	0	0
Bangladesh	6.877	5.136	0	0	0	1	0
The Bahamas	9.285	7.5	0	0	0	0	1
And so on…	…	…	…	…	…	…	…

Solution 2: Add "Fixed Effects"

- **Adding Fixed Effects changes the model to the following**

$$GDPperCap_{cf} = \beta_0 + \beta_1 PropProtect_{cf} + \beta_2 Africa_f + \beta_3 SAmerica_f + \beta_4 Europe_f + \beta_5 Asia_f + \beta_6 Carribean_f + \varepsilon_{cf} \quad (7)$$

- **This is almost always abbreviated to something like**

$$GDPperCap_{cf} = \beta_0 + \beta_1 PropProtect_{cf} + K_f'\Gamma + \varepsilon_c \quad (8)$$

Solution 2: Add "Fixed Effects"

- **Why do this?**
  - Adding fixed effects controls for all variables that are time-invariant (that don't change over time) for the group
- **Example of what the continent FE controls for are:**
  - Agro-climatic conditions of the continent
  - Distance of the continent from trade routes
  - Many others…

Solution 2: Add "Fixed Effects"

- **Assumption needed for FE to solve OVB:**
  - All endogenous variables are time-invariant and at the level of the group of the FE
- **Reasons this assumption might not hold:**
  - Endogenous variables are at a level below the group level (country level for example)
  - Endogenous variables are time-variant (change over time)
- **Note: This also doesn't solve reverse causality**

Special Example of Fixed Effects

- **If you have "panel data"**
  - Panel data is data from the same units over multiple periods of time

Panel Data Example

Panel Data Example			
Individual ID	**Ind. 1 Binary**	**Ind. 2 Binary**	**Ind. 3 Binary**
001	1	0	0
001	1	0	0
001	1	0	0
002	0	1	0
002	0	1	0
003	0	0	1
003	0	0	1
003	0	0	1

Special Example of Fixed Effects

- **If you have "panel data"**
  - Panel data is data from the same units over multiple periods of time
- **In this case the group membership is one's self!**
  - So you can control for any time-invariant attribute, observed and unobserved, at the individual level
  - This would solve the problem before of innate ability being unobserved because it is time-invariant
- **This can be very powerful**

Solution 3: Run an Experiment

- **This is the easiest of the solutions to OVB conceptually**
- **What if one could simply randomly assign different levels of property rights protection to countries?**
  - This would mean that there is no reverse causality, because the researcher is determining which country gets which level of property rights regardless of GDP per capita
  - It would also mean that there are no other variables determining property rights protections, again because the level of the variable of interest is solely the result of a random process determined by the researcher

Solution 3: Run an Experiment

- **If AJR could do this, then necessarily $PropProtect$ is independent of ε, and OVB is solved!**
  - Now, OLS will give you an unbiased estimate
- **This is the same idea behind lab experiments and clinical trials**
- **Abhijit Banerjee, Esther Duflo and Michael Kremer were awarded the Nobel Prize in Economics for their pioneering work in bringing this method to economics**

Solution 3: Run an Experiment

- **Problems:** Some things can't be randomly assigned, like the level of property rights protection in countries
  - Often random assignment is either unethical or impossible
- **Best Applications:** Measuring the effects of some or program on people
  - Examples:
    - Job-training programs
    - Incentives for teacher performance
    - Conditional and Unconditional cash transfer programs

Solution 3: Run an Experiment

- **Assumptions needed for experiment to solve OVB:**
  - The researcher successfully randomized who got the "treatment"
    - Technically this is untestable, but there are ways to support this empirically
- **Reasons this might not hold**
  - Researcher mistakes
  - While randomization was successful, follow up data collection loses people in a systemic way correlated to unobservables
- **Note: Experiments <u>do</u> solve reverse causality**

Reading Econometric Results

- **Econometric results (coefficient estimates) are always reported in tables in a paper. You should be looking here for the main learnings of the papers**
- **The text of a paper generally does the following:**
  - Describes why the relationship they are measuring is important
  - Summarizes the approach of the researchers
  - Describes how they dealt with OVB
  - Provides their thoughts on what the estimates in the tables mean

Example of a Table you Will See

TABLE III
RELATIONSHIP BETWEEN SLAVE EXPORTS AND INCOME

	Dependent variable is log real per capita GDP in 2000, $\ln y$					
	(1)	(2)	(3)	(4)	(5)	(6)
ln(exports/area)	−0.112***	−0.076***	−0.108***	−0.085**	−0.103***	−0.128***
	(0.024)	(0.029)	(0.037)	(0.035)	(0.034)	(0.034)
Distance from equator		0.016	−0.005	0.019	0.023	0.006
		(0.017)	(0.020)	(0.018)	(0.017)	(0.017)
Longitude		0.001	−0.007	−0.004	−0.004	−0.009
		(0.005)	(0.006)	(0.006)	(0.005)	(0.006)
Percent Islamic				−0.008***	−0.006*	−0.003
				(0.003)	(0.003)	(0.003)
French legal origin				0.755	0.643	−0.141
				(0.503)	(0.470)	(0.734)
North Africa indicator				0.382	−0.304	
				(0.484)	(0.517)	
ln(gold prod/pop)					0.011	0.014
					(0.017)	(0.015)
ln(oil prod/pop)					0.078***	0.088***
					(0.027)	(0.025)
ln(diamond prod/pop)					−0.039	−0.048
					(0.043)	(0.041)
Colonizer fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Number obs.	52	52	42	52	52	42
R^2	.51	.60	.63	.71	.77	.80

Independent variables

These two models have different variables, I just omitted them for clarity of the slide

Dependent variable listed above the columns

Numbered columns contain different models (configurations of variables)

Numbers contain values of coefficients

Numbers in parentheses are the standard errors

Stars indicate "statistical significance" of the estimated slope

Often an indication of variables in the model whose coefficients are not reported. Here Nunn indicates he has FEs for a country's colonizer

Number of observations used in the estimation

Percentage of variation is Y variable explained by the model. Generally ignored but always reported

Statistical Significance (without the Statistics)

- **Science wants to find evidence <u>to prove that something is always true</u>, BUT…**
  - "Always" is a long time! It involves many states of the world that we aren't able to observe in a sample from a certain subset of individuals at a given point in time, or points in time
  - If I take a random sample from a population at a given moment in time, or moments in time, can I really be confident that what I find from this sample will be true in the infinite number of samples I could draw from that population at other moments in time?
    - No!
- **So… Science satisfies itself by making a statement that research has found evidence against the claim that something isn't true**
  - What does this mean?

Statistical Significance (without the Statistics)

- **Let's say Nunn finds that the relationship between logged slave exports per million square miles and logged GDP per capita in 2000 is -0.128**
  - Assume this is a causal estimate
- **Is this evidence that the effect of slavery on GDP per capita is <u>always</u> -0.128?**
  - No, because it is based on one observation of the world out of the infinite observations one could make

Statistical Significance (without the Statistics)

- **So instead of trying to make an impossible claim about "always" we try to say something more modest:**
  - **Try to answer:** Is this finding by Nunn large enough and/or based on enough evidence to confidently reject the idea that there is no relationship (a coefficient of 0) between slavery and GDP per capita?
    - "based on enough evidence" means: "are the data powerful enough?"
    - The hypothesis that there is no relationship between slavery and GDP per capita is called the "null hypothesis"
  - We can answer this using statistical theory! But we won't here. You'll have to take Statistics and/or Econometrics for the machinery behind this

Statistical Significance (without the Statistics)

- **Intuition:**
  - If we find a really big (in absolution value) effect, then we are more confident that the effect isn't zero
    - So larger findings, all else equal, make us more confident we can reject the idea that the relationship is zero
  - If our relationship is derived from a larger dataset, it is based on <u>more information</u>, making us more confident that we are getting an accurate measure of the larger population
    - So larger (more powerful) data, all else equal, means we are more confident that we can reject the idea that the relationship is zero

Statistical Significance (without the Statistics)

- **Additionally:**
  - Less noise in the data makes us more confident our measure isn't zero, all else equal
    - More noise makes it harder to separate the signal (true relationship) from the random variation
  - More variation in our X variable around its mean (independent of the other RHS variables) makes us more confident, all else equal
    - More movement in X again means more information about how X moves with Y and again increases confidence

Measuring Confidence

- **The stars indicate how confident the researcher is (based on statistical theory) that the relationship they've measured provides evidence against the null hypothesis that there is no relationship**
  - *** usually indicates above 99% confidence
  - ** usually indicates between 95% and 98.999% confidence
  - * usually indicates between 90% and 94.999% confidence
- **Really, anything below 95% confidence is considered pretty shaky evidence for rejecting the null**
  - This is purely arbitrary however

Caution about Statistical Significance

- **It is important. Without it, an estimate could just be a misestimation of a zero relationship… Uh-oh**
- **But remember what can lead to statistical insignificance**
  - Noisy data
  - Small dataset
  - Small effect size
  - Little (independent) variation in X around its mean
- **If an estimated effect is big perhaps it is important to consider even if there is little confidence in the estimates?**

Measurement Error

- **Economic data are messy**
- **While it is valid to bring up sources of measurement error as worries in one's research design you should be cautious**
  - Often people discard good research just because the data are messy
  - If you have a reason to believe there is a worry about measurement you should always say explicitly why. Saying that the data are messy isn't sufficient by itself to invalidate one's work

“More and Better” Data

- **Last, more and better data won’t always result in better analysis**
  - Saying that the researcher’s conclusions will be strengthened by better data is sort of wrote and isn’t always true
    - Often researchers’ data are perfectly sufficient for their project. I’ve seen many great studies with small datasets (50-100 observations or so)
    - Better to speak of “limitations” of a researcher’s dataset if you think there are issues

Variable Transformations

- **Transforming variables allows for more flexible models**
  - OLS is not restricted to estimating simple linear relationships
- **They also change the interpretation of the resulting coefficients**
  - See table on the next slide for an easy list of how to interpret coefficients

Example: Log-Linear Model

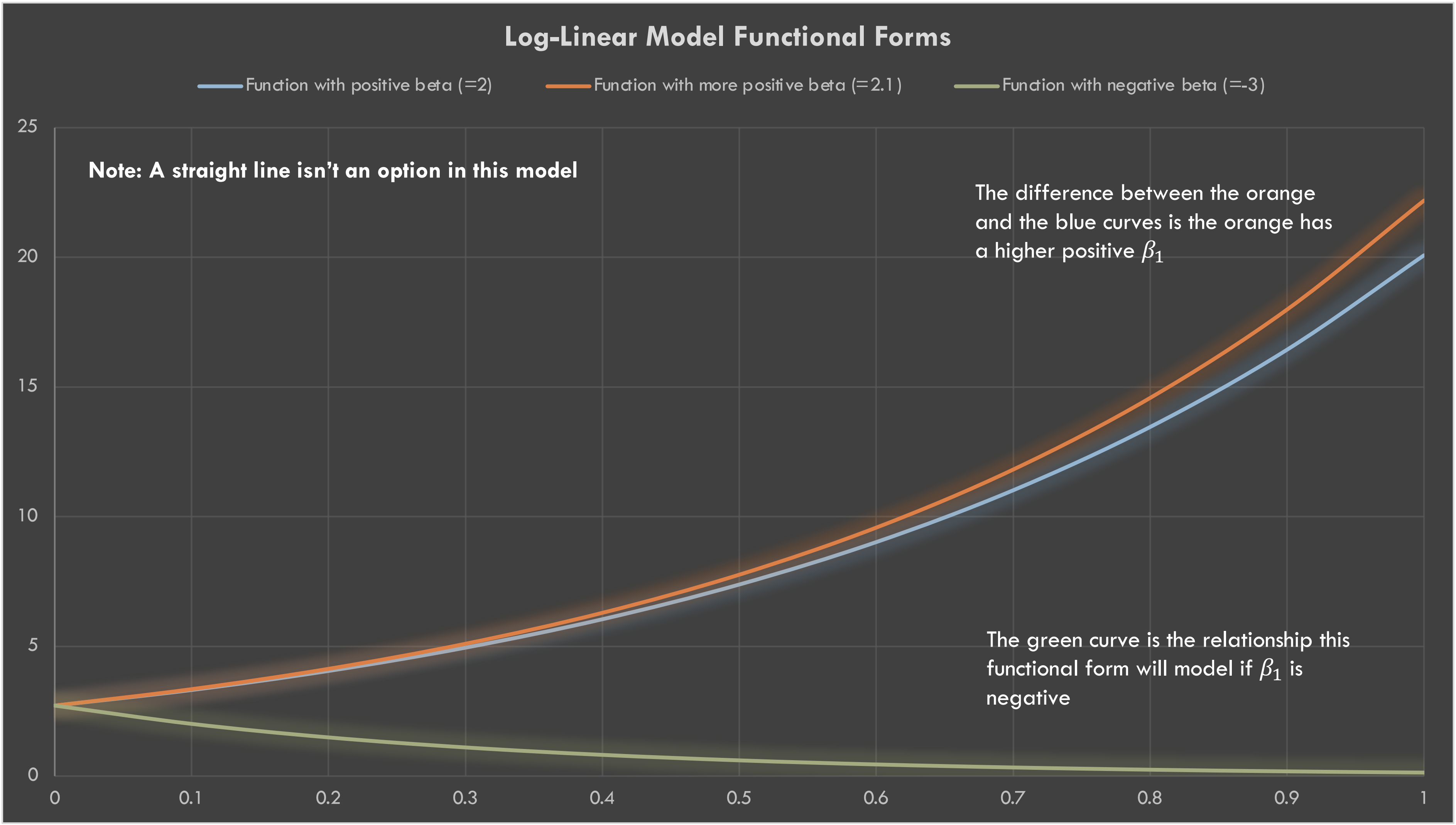
- These are models of the form,

$$ln(Y_i) = \beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i \qquad (17)$$

- You can rewrite this to have Y on the LHS as follows,

$$Y_i = e^{\beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i} = e^{\beta_0} e^{\beta_1} e^{\beta_2} e^{u_i} \qquad (18)$$

Log-Linear Model: Options Allowed



Variable Transformations

- **This really requires econometrics to get into in any real way**
- **Major points:**
  - Transforming variables with logs creates more flexibility for different relationships between X and Y
  - <u>The transformations result in different interpretations of the betas</u>

Different Functional Forms

Function	Form
$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i$	Linear-Linear
$ln(Y_i) = \beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i$	Log-Linear
$ln(Y_i) = \beta_0 + \beta_1 ln(X_i) + \beta_2 ln(Z_i) + u_i$	Log-Log
$Y_i = \beta_0 + \beta_1 ln(X_i) + \beta_2 ln(Z_i) + u_i$	Linear-Log
$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 Z_i + u_i$	Quadratic
$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \beta_3 (X_i \times Z_i) + u_i$	Interaction
$Y_i = \beta_0 + \beta_1 \frac{X_i}{Z_i} + u_i$	Ratio

Different Functional Forms

Function	Form	Approximate β_1 Interpretation
$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i$	Linear-Linear	The unit change in Y associated with (or caused by) a one unit change in X, holding Z constant
$ln(Y_i) = \beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i$	Log-Linear	The percent change in Y associated with (or caused by) a one unit change in X, holding Z constant
$ln(Y_i) = \beta_0 + \beta_1 ln(X_i) + \beta_2 ln(Z_i) + u_i$	Log-Log	The percent change in Y associated with (or caused by) a 100 percent change in X, holding Z constant
$Y_i = \beta_0 + \beta_1 ln(X_i) + \beta_2 ln(Z_i) + u_i$	Linear-Log	The unit change in Y associated with (or caused by) a 100 percent change in X, holding Z constant