Economic Jargon Cheat Sheet
ECON B225: Economic Development
Sebastian Anti, Assistant Prof. of Economics
Bryn Mawr College
Spring 2021

INTRODUCTION

This document is meant to orient you to important and commonly used terms in research using econometrics. It is meant to supplement class sessions on econometrics. It is organized into two sections. The first section will work progressively through econometric concepts and build up your intuition for the method. The second part is simply a glossary of terms you'll see often in econometric papers.

SECTION 1: AN OUTLINE OF RESEARCH IN ECONOMETRICS

1.a A Research Question

Econometrics is a core methodology in economics which seeks to measure relationships between economic forces.

Usually the econometrician is trying to measure a causal effect of one thing on another. For example, econometric methods are meant to answer questions like, how does a job-training program affect wages of participants? How does one's gender affect one's career advancement, or how do the strength of a country's institutions affect a country's current rate of growth?

Here is a good question that we'll work with for a while in this as an example: What is the effect of a country's strength of property rights institutions on a country's current level of development?

1.b The Hypothesis

For the sake of an example let's take that last question and walk through some of the steps that econometrics uses to answer it. First note that both a country's current level of GDP per capita and a country's property rights institution vary across countries around the world. Thus, they are both "variables", measures that change over different units of observations (countries in this case).

It's helpful to think about the question first in a Directed Acyclic Graph (DAG). Don't be afraid of this insane terminology. It just means a flow chart, which you should all be familiar with. The DAG represents how you think property rights institutions affect growth. Figure 1 contains a basic DAG outlining what seems like a straightforward relationship. It should seem pretty plausible that a

country's property rights institutions affect the ability of a country to grow its economy because private property (theoretically) enables free enterprise.

The DAG indicates this causal hypothesis with a little arrow indicating which force affects the other and the direction of causality. It is conventional in the literature to call the variable causing change the *x*-variable of interest, also called the "independent variable of interest", and the variable that is being changed the *y*-variable, also called the "dependent variable".

Let's now go a bit further. What do you think the relationship is? Do you think that as a country's property rights institutions get stronger this increases GDP per capita? If so then this means you think that positive change in institutions results in positive change in growth. This means you guess that the relationship is positive. If you think it's more plausible that improvements in the strength of a country's institutions causes growth to be negative, or in other words for a country's economy to get smaller, then you are suggesting a negative relationship.

I would say that a positive relationship seems more plausible since protection of private property is generally assumed to incentivize innovation, business creation, and entrepreneurship (even if you disagree, just go with me for the sake of the example here), so I have noted that in Figure 1 with a little "(+)" above the arrow illustrating causality. This DAG in Figure 1 maps out a hypothesis of what one might see as a plausible econometric hypothesis: That a country's institutions (x) have a positive causal effect on a country's growth (y).

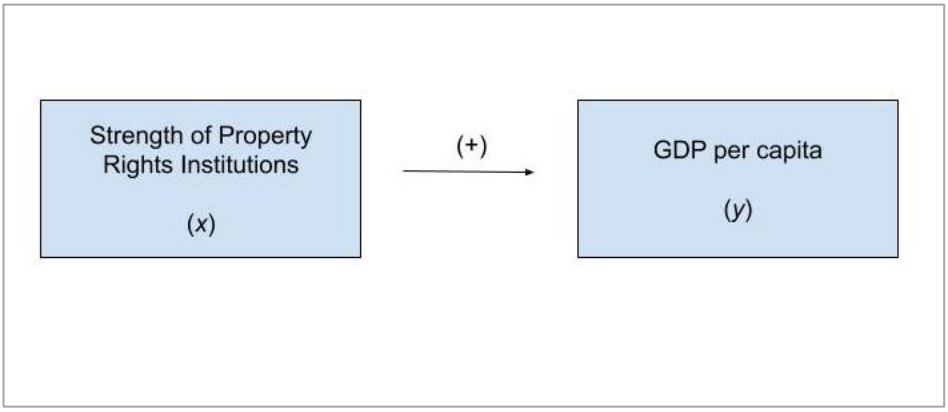

Figure 1: DAG of Property Rights Institutions on GDP per Capita

1.c The Data

The purpose of econometrics is to use data to measure this relationship we've just hypothesized. To do this we need data that contain the measures of a country's property rights institutions and their current GDP per capita. For this, I'll just use the dataset from Acemoglu, Johnson, and Robinson (2001), henceforth "AJR". This dataset contains a country's 2001 GDP per capita in 1995 and a measure of the country's protection of expropriation risk in 1995 as a measure of the strength of a country's property rights institutions. It also contains other useful measures, which I'll talk about in a later section.

1.d The Beginning of Econometrics: The Scatterplot

Let's start by just creating a scatter plot of the two variables in which we're interested. Figure 2 contains this. Here, the GDP per capita measure is actually the natural log of the GDP per capita measure. This is just a transformation of the measurement (log transformations are the opposite of an exponential transformation) in a way that makes it display a little better in charts and allows for a simple interpretation of our econometrics later, among other things that aren't important at this stage.

You can see that in Figure 2, it looks like there is indeed a positive relationship between a country's level of protection of expropriation risk and GDP, just as we hypothesized in the previous DAG. That is what we thought would be the case, so we're all done right?

Obviously not. Recall the point here is to measure this relationship, which we haven't done yet; we've just displayed the relationship in a visual. So how do we actually measure the relationship?

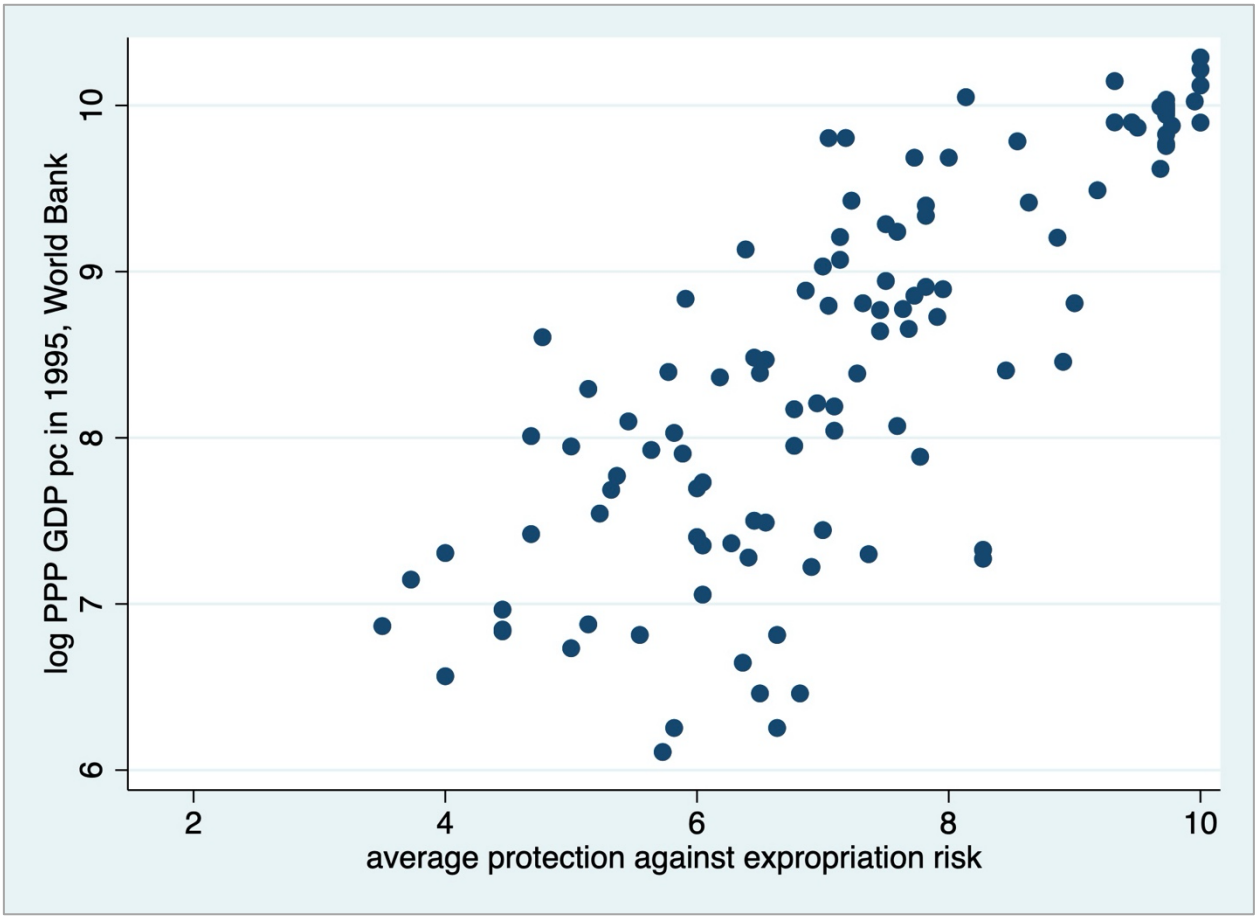

Figure 2: Scatter Plot of Protection Against Expropriation vs ln(GDP per capita in 1995)

First, notice that it looks like the data points are following an invisible line going through the middle of it. It looks almost like that is the true relationship between the two variables and the variation of the points around it are just noise in the data. If this line is the real relationship, then its slope will describe the change in our dependent variable as our independent variable of interest changes, in other words: the relationship we want. So how do we figure out what this line is?

1.e The Line of Best Fit

Let's describe this line of best fit for the data that we want to find. Recall the algebraic notation for a line from previous math courses you've had,

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \tag{1}$$

where β_0 is just a constant, and β_1 is the slope of the line. ε_i accounts for the variation around the line which we'll assume for now to be purely random and independent of x. A restatement of what we want to find is the β_0 and β_1 that "best fit" the data used in the Figure 2.

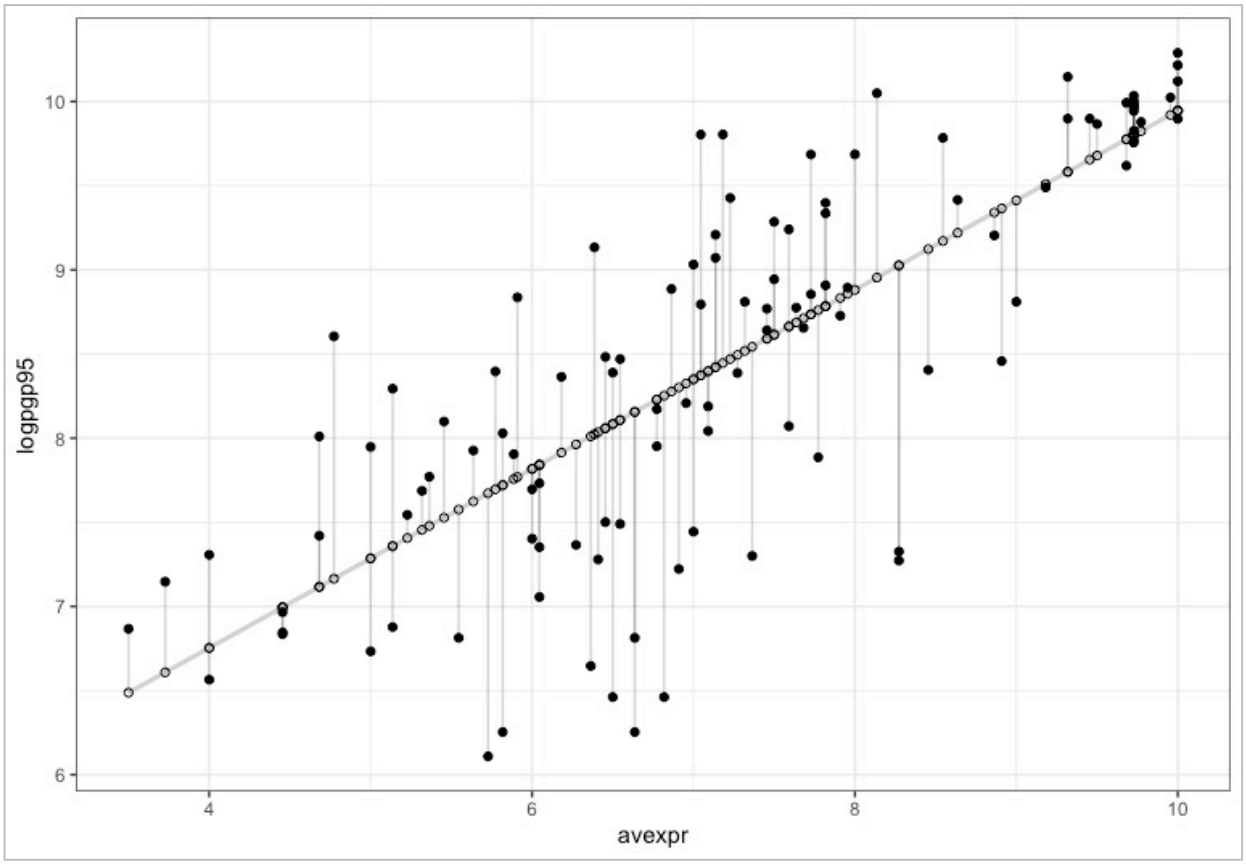

Figure 3: Scatterplot with the line of best fit, and depicted residuals.

The way we plot this line is we find the line that minimizes the distance between it and all the other data points. The distance between the line and a data point is called a "residual". The way this is implemented in practice is to find the line that minimizes the sum of the squared residuals.

We use squared residuals in order to ensure that the sum being minimized is a sum of positive values. This algorithm that does this is called Ordinary Least Squares (OLS), and you will see it in almost every paper we read. It is the core measurement algorithm in econometrics and is extremely powerful. It is also used commonly in predictive analysis, and all sorts of other data science applications.

Figure 3 shows the same scatter plot with the line of best fit, found using OLS, and the residuals (the vertical grey bars between the point and the line of best fit) also displayed. For this line, the measurement of β_1, denoted now as $\widehat{\beta_1}$, is 0.532. Since y is the logged value we interpret this slope as a percent change, so a one-unit increase in a country's protection against expropriation is associated with a 53.2 percent increase in GDP per capita in 1995 on average.

1.f Is $\widehat{\beta_1}$ the Measure of the True Causal Relationship between x and y?

Now this all seems good. We have a measure of this relationship we wanted to find and it is in line with our expectations for what the relationship would be. But let's think carefully about what we're doing before claiming we've "identified" the causal effect, or in other words, measured this causal relationship correctly.

In this regard, we have two major concerns. First, it is possible that protection against expropriation is perhaps simply correlated with the actual causal variable that is driving GDP. This would mean that we are incorrectly ascribing the effects of this important third variable to our independent variable of interest. You can think about many examples of these types of variables and we'll discuss them in class. When this happens we say that our measurement of $\widehat{\beta_1}$ "is biased" and that our model is "endogenous".

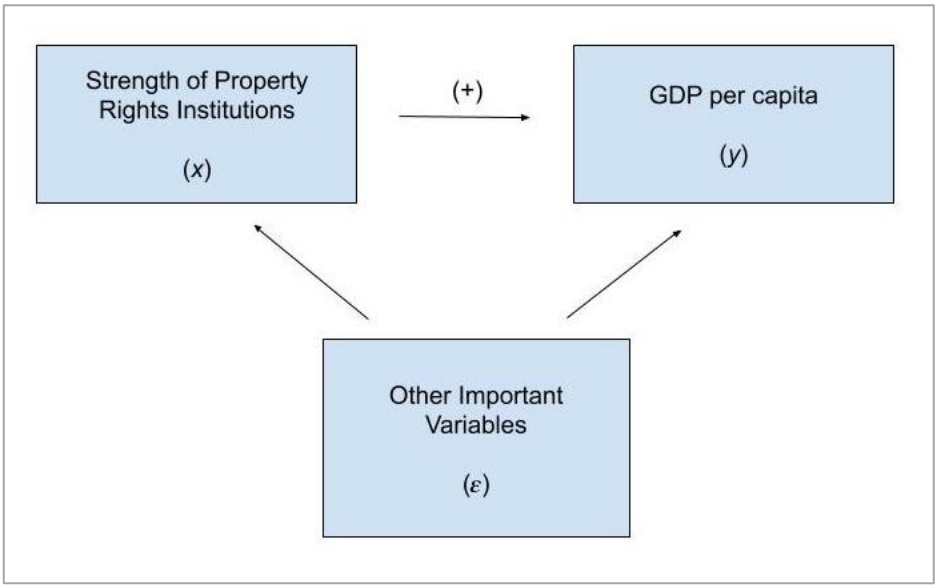

Figure 4: DAG with Omitted Variables

The reason for this is that we've omitted an important relevant variable in the system we are trying to measure. We call this type of bias "omitted variable bias". Another way to think about this is that our DAG in Figure 1 is wrong, and in fact what we should have drawn in something like the DAG in Figure 4. The way this changes the system we are estimating is that we add an ε_i to equation (1), which denotes an error term containing all the unmeasured stuff that we don't have in our dataset. This marks a departure from before where we assumed all the variation around the line of best fit to be benign noise. Now, we are saying that these other variables causing variation in our outcome variable are not random and in fact related to our dependent variable of interest and independent variable.

As you can probably imagine, this is a major problem for applied econometricians since so many things are simply not measurable in the world but might plausibly be included in the ε_i term. Many of the challenges to doing econometrics correctly and achieving an unbiased estimate, or "identifying" the true measure of how y changes as x changes are related to solving problems caused by omitted variable bias.

A simple and important way though to deal with this is to simply find a measure for your omitted variable and then include it in the estimation. Perhaps the omitted variable is that the religious composition of the country is both affecting protection against expropriation and GDP growth. As AJR note, the idea that the religious composition of a country's population affects economic growth dates back to Max Weber, and if it also affects the evolution of protections against expropriation risk, then we have an omitted bias problem. However, in our dataset from AJR, I have a measure for the percentage of a country's population that is Catholic. Therefore, I can control for this potential omitted variable and ask OLS to estimate the slope of the following plane (note that this is no longer a line since we have two independent variables and we are estimating the parameters of a shape in three-dimensional space),

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 w_i + \varepsilon_i. \quad (2)$$

Figure 6 shows the estimated plane of best fit. OLS is powerful and flexible and can accommodate as many control variables like this that you would want to add, however it is impossible to visualize the resulting plane beyond two independent variables in this intuitive way. Now $\widehat{\beta_1}$ is a partial derivative of y_i with respect to x_i and provides a measure of the relationship of interest controlling for w_i. This is pretty great, however as you can likely foresee there are limits to this. What happens when you need to include a variable for which there is no measure? We will get to this.

The second big problem is maybe GDP affects institutions and not the other way around, or maybe causality runs in both directions. This seems pretty plausible and is called "reverse causality". If it is present it changes our original DAG to look like the one depicted in Figure 5 with the arrow now going both ways. This also results in biased estimated of $\widehat{\beta_1}$. Unfortunately, adding variables to your estimation won't solve this problem and you'll need to use more advanced experimental and quasi-experimental techniques to claim this isn't a problem. We will get to this.

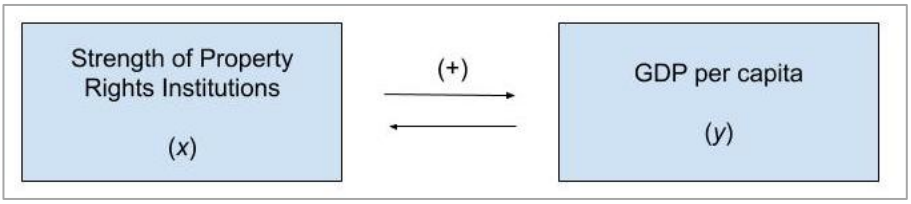

Figure 5: DAG with Reverse Causality

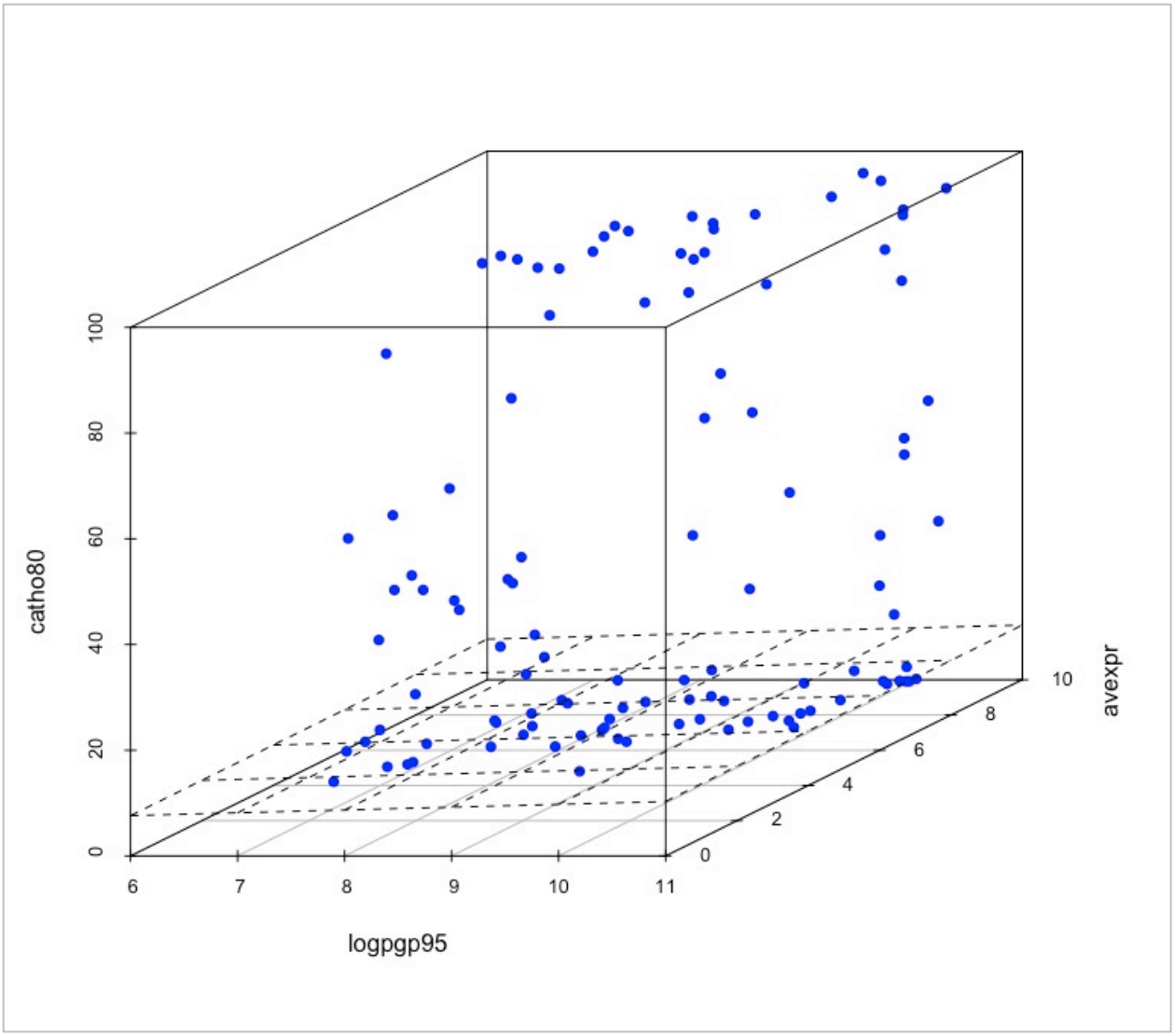

Figure 6: Estimated Plane of Best Fit

1.g How Confident Are We that $\widehat{\beta_1}$ is Close to the True Relationship between X and Y?

Another issue with estimating $\widehat{\beta_1}$ is a question over whether our estimates are reflecting a real relationship or whether they are just there because of some noise and randomness in the data. We gain insight into this by calculating the "statistical significance" of $\widehat{\beta_1}$.

Statistical significance as it is usually discussed in empirical work is essentially a statement of how confident the researcher is that the coefficient is not just zero (no relationship) given the levels of noise in the data, size of the dataset and variation available in their independent variable of interest. We find this by calculating the standard error for $\widehat{\beta_1}$ and then dividing $\widehat{\beta_1}$ by its standard error. This provides a t-value or t-statistic.

A t-statistic indicates how confident that the relationship exists and is not just a misestimation of a zero relationship. A t-statistic above 1.645 and below 1.96 means we are 90 percent sure. Above 1.96 and below 2.326 means we are 95 percent sure. Above 2.326 means we are 99 percent sure. Consensus in the discipline is that we can't really trust anything below 90 percent confidence.

Often these measures are restated as "p-values", which contain literally the exact same information as the t-statistic. A p-value below 0.1 and above 0.05 means we're 90 percent sure. Below 0.05 and above 0.01 means we're 95 percent sure and below 0.01 means we're 99 percent sure.

1.h Robustness and How to Read Regression Tables

I've hopefully made the point clear that OVB is a huge problem when researchers are trying to identify causal effects. Since no one ever knows the true set of variables that make up the right-hand side of a regression specification, you'll often see researchers showing results for a bunch of different combinations that seem to make sense. The researcher hopes to learn from this whether the results on their variable of interest are "robust" to the changing specification. Robustness is shown if the coefficient value and its statistical significance change very little across all the different specifications. If the coefficient is not robust, it is up to the researcher to provide a reason why the coefficient is changing depending on the coefficients being added.

Results are usually presented in a table format where, from left to right, the first column contains the independent variable names, and each column to the right of it contains coefficient estimates and standard errors from distinct specifications chosen by the researcher. At the bottom of the table are often important meta information about the regression such as the total number of observations in the data used in obtaining the estimates and a measure called the R^2. The R^2 is a measure of how much of the variation in the dependent variable is explained by the model that you estimated. While usually a higher R^2 is considered "better", it is a measure that really provides no information on causal identification, which you should recall is the point of this whole effort. So, most people concerned with bias above anything else tend to ignore the R^2. If you are using OLS as a predictive tool, then the R^2 matters more.

Table 1 shows an example of regression output for the two specifications laid out above in equations (2) and (3).

Table 1: Example for Regression Output Table

Variable	(1)	(2)
Strength of Institutions	0.532***	0.534***
	(0.040)	(0.039)
Share of Population Catholic		0.006***
		(0.002

Constant	4.626*** (0.301)	4.422*** (0.002)
R-squared	0.61	0.644
Observations	111	111

Notes: From AJR (2001),
*** p>0.01, ** p>0.05, * p>0.1
Standard errors in parentheses

1.i Ways to Solve OVB and Reverse Causality

Randomized Control Trial

Often a researcher is interested in the effect of a policy on some outcome for people. Let's think of a new example here: the effect of participating in a job-training problem on one's wage. An easy thing to do might be to simply find people who participated in the job training program and some that didn't and compare their wages.

What is wrong with this? Well, the obvious thing is that the people who decided on their own initiative to participate in the job training program are likely different in some systematic way than those that did not participate. They are likely (on average everything else constant) more ambitious, perhaps more confident and talented, perhaps they are simply people that have more spare time to devote to some extra training and this is related to them being better organized and better at managing their lives. As a result, if you make that comparison the wage difference you are measuring between the two groups will capture the effect of the job training program *plus the effect of all these other omitted variables* on their wages. This is just another example of omitted variable bias, and in cases like this where the systematic differences are introduced through a selection process, it is called "selection bias". Solving this is a massive challenge in econometrics research.

Randomized control trials (RCTs) achieve this. They require that the treatment studied, in this case the job training program, is applied in some random way to the population being studied. To do this the researchers literally randomly choose who gets to receive the job training program. This randomization means that there are no systematic differences between those treated and untreated and the difference in wages is just the effect of the treatment. This is the same methodology used to test drugs and medical interventions. In terms of the specifications I laid out before, RCTs ensure that our variable of interest, whether someone received the treatment (D_i) is independent of anything that could possibly be in the error term. In math this is,

$$D_i \perp \varepsilon_i \tag{3}$$

While this is an extremely effective way to achieve identification, it has the following limitations for use in economics:

1. Many treatments of interest cannot be randomized (an obvious example would be property rights institutions from above)
2. It is expensive and takes a long time to run an RCT
3. Often people drop out of the study over time. This is called sample "attrition" and if people are attritting based on systematic unobservable variables then you have OVB despite the randomization

Instrumental Variables (IV)

So, let's say we are interested in the effect of something that cannot be randomized, so doing an RCT is not a solution to OVB.

A classic example here is the research question: what is the effect of having a third child on a woman's wages? Of course, we can't just compare women who had third children to those who didn't because those who chose to have that third child likely differ from those that didn't in many systematic ways.

Therefore, we know that if we did that we'd just have selection bias messing up our identification. Furthermore, obviously, us researchers can't find a bunch of women with two children and tell a random selection that they must have a third child and the other random selection that they cannot. This is unethical, and completely impractical. So, let's think of a clever alternative.

An IV is a clever alternative when an RCT isn't possible like this. It is essentially a variable (z_i) that is correlated with our variable of interest but not correlated with any of the omitted variables causing omitted variable bias. Yikes! That probably sounds confusing to many of you, but all it means is that the relationships between the variables look like the one in Figure 7.

Essentially it means two things: (1) it means that we have a measure that is related to our independent variable of interest, and (2) it has no relationship with all those systematic differences between women who chose to have a third child and those that didn't. The first condition is often referred to in the papers we'll be reading as the "instrument condition." The second is often referred to the "exclusion restriction."

Importantly the instrument condition is testable in the data while the exclusion restriction is not and relies on a rhetorical argument from the researcher to convince the reader that it holds. Arguing about this makes for some really interesting papers. If you are citing IV papers in your term paper, you should make it clear in your writing whether you consider the IV approach valid.

So back to our example. A famous IV proposed for having a third child is a binary variable for whether the first two children are of the same sex or not. Let's think for a second about why this may be a plausible IV. First let's consider the instrument condition. Do you think that the sex mix of a woman's first two children would have any effect on her decision to have a third? Of course! If a woman's first two children are the same sex, she is more likely, all else equal, to have a third in order to gain the life experience of raising children of both sexes. Therefore, there is a

relationship between the sex mix of a woman's children and her decision to have a third child. This means the instrument condition holds.

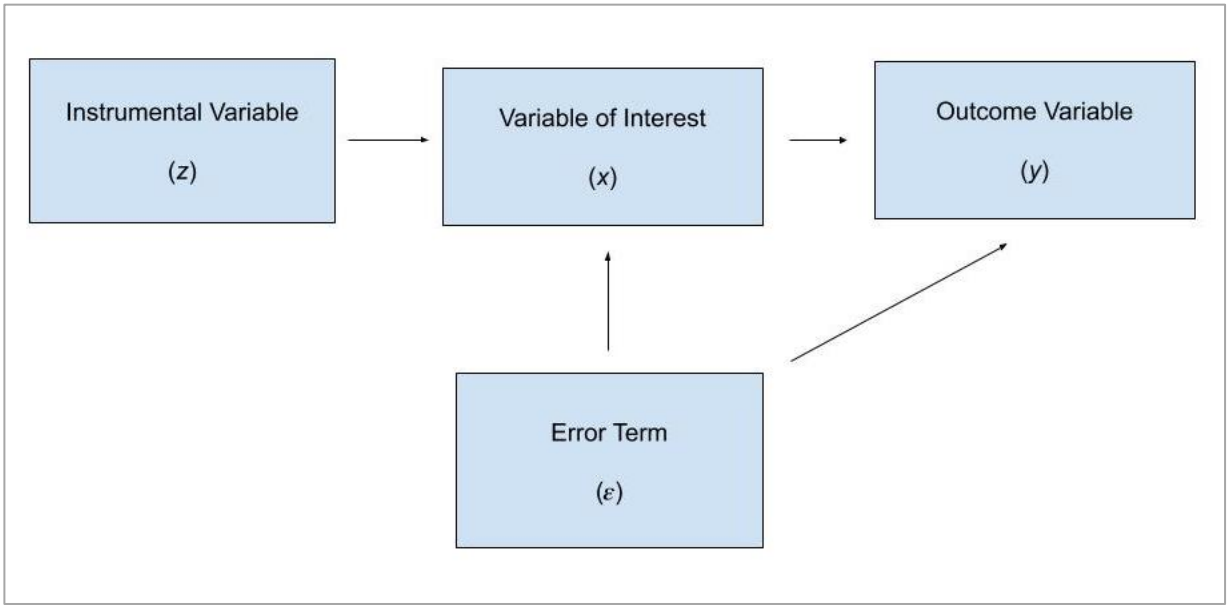

Figure 7: DAG for Instrumental Variable

Now let's consider the exclusion restriction. Does this hold? Well, a person's gender is a random biological process completely unrelated to a mother's wage. Therefore, a woman's children's sex mix is randomly determined and plausibly independent from all the other stuff that makes a woman select into having a third child.

The way an IV works is that you do two stages of OLS. For this reason it is often called "2-stage least squares" (2SLS). The first stage involves taking our variable of interest x_i and using it as an outcome variable and estimating the relationship between it and the IV, usually denoted z_i, as follows,

$$x_i = \alpha_0 + \alpha_1 z_i + \lambda_i \tag{4}$$

where the alphas are just the intercept and coefficient on the instrument and λ_i is just the error term in this regression. The estimate of $\widehat{\alpha_1}$ provides a measure of the relationship between the IV and the variable of interest. Generally, researchers consider this relationship strong enough to satisfy the instrument condition if the F-statistic for this regression is above 10 (Don't worry about why this is. It is just convention. Understanding just that this is a threshold for the instrument condition is satisfactory for now until you are able to take advanced econometrics).

Remember this is a measure independent of all the omitted variables messing up identification. Now here comes the tricky part. We now use the estimates of $\widehat{\alpha_1}$ and $\widehat{\alpha_0}$ to obtain the predicted values of x_i, usually denoted as $\widehat{x_i}$. These are measures of the variable of interest that are

predicted by the (clean) instrument and therefore cleansed of all the problematic correlation with those omitted variables messing up identification.

Note that we have $\widehat{x_i}$, we proceed to the second stage of 2SLS, and use this alternative measure as our variable of interest in the primary regression as follows,

$$y_i = \beta_0 + \beta_1 \widehat{x_i} + \varepsilon_i \quad (5)$$

Now the estimate of β_1 will be purged of the bias that self-selection is causing and we've achieved our goal!

Here are a couple drawbacks to IVs:

1. They are rare and you as the researcher need to be super clever to think of one.
2. There is never proof that the exclusion restriction holds so if you ever use one you'll always need to defend the assumption that it holds in your case from attacks from other researchers and sceptics.
3. It technically only identifies a Local Average Treatment Effect, a more limited measure than the true causal effect.

Fixed Effects

"Fixed effects" is a method that a researcher can employ when the data have a type of nested structure. For example, data of individuals where we know the village, district and province of their residence. Here, the individual is nested within a village, which is nested within a district, which is nested within a province. Table 2 shows an excerpt of some data with a nested structure.

Table 2: Fixed Effects Example I

Individual ID	Village ID	District ID	Province ID
001	001	001	001
002	001	001	001
003	001	001	001
004	002	002	002
005	002	002	002
006	003	002	002
007	003	002	002
008	003	002	002

For a simple example let's just consider a dataset with one level of this nesting structure: individuals and their village. When a researcher has data like these, they can create a series of binary variables. Each is for a different village in the dataset. Table two provides an example of this. The researcher then puts all but one of these binary variables in the regression they are running. These binary variables are called "fixed effects." Doing this controls for everything that in time-invariant within one's village. So, if the researcher is concerned about endogeneity at the village level in their analysis, they can control for all the time-invariant variables of concern by

including a village fixed effects. Think about why this is powerful. Time-invariant factors at the village level relate to anything that is not changing over time for villages. This could be local institutions, cultural beliefs and practices, or geographic characteristics, among others.

Table 3: Fixed Effects Example II

Individual ID	Village ID	Village_001_Binary	Village_002_Binary	Village_003_Binary
001	001	1	0	0
002	001	1	0	0
003	001	1	0	0
004	002	0	1	0
005	002	0	1	0
006	003	0	0	1
007	003	0	0	1
008	003	0	0	1

This method becomes extremely powerful when we have data of repeated observations of the same individuals over time. This is called "panel data" or "longitudinal data" and it has a structure where observations are nested within individuals. This type of data allows us to control for all time-invariant variables at the individual level.

A drawback of this method is that it does not resolve reverse causality.

Difference-in-Differences (DID)

One last incredibly powerful tool to deal with OVB and reverse causality is DID. Let's say you have a policy intervention like a job-training program that affects one group of people and not the other. However, the groups are not determined randomly. An obvious question would be: what is the effect of the job-training program on the wages of the people who received it? However, you can't just look at the before and after wages for the people who received the program since you are observing them in different time periods and various things changed for that group over that period of time, such as the general macroeconomic conditions in which they live, that might also affect their wages. If you did this, you'd be ascribing the effects of these omitted variables (the macroeconomic improvement) to the effect of the job-training program. This would be another form of OVB.

DID answers this by using the group that didn't get the program as a type of control group. If you have data from before and after the job training program came for both the group that received it and the group that didn't, you can look at the before and after comparisons for wages for the two groups. Now let's say you're looking at the average wages for the group that got the job-training program ($\bar{Y}_{D=1,t=0}$) and you then look at this measure in the second time period ($\bar{Y}_{D=1,t=1}$) and it looked like it went up! When you look at the same change over time for the group that didn't get the job-training program it looks like it stayed the same or went down a little bit. This is depicted in Figure 8.

An obvious idea is to say that the difference in the movement over time for these two groups is attributable to the job-training program. The difference in the movement over time for these two groups is just the difference between these two differences over time. This is the Δ^{DID} estimator and it is formally described as follows,

$$\Delta^{DID} = (\bar{Y}_{D=1,t=0} - \bar{Y}_{D=1,t=1}) - (\bar{Y}_{D=0,t=0} - \bar{Y}_{D=0,t=1}) \quad (6)$$

This method allows one to control for all the time-invariant factors associated with the outcome variable. The key assumption one needs to make is whether the outcome variable has the same trend over time for both the treated group and the non-treated group, which can be tested empirically if one has enough data from the period before the treatment (in this example the job-training program) arrived.

One can easily estimate this in a linear regression framework using OLS. This provides a standard error (and therefore statistical inference), and it allows you to control for time-variant factors. This is specified as follows,

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 T_t + \beta_3 D_i \times T_t + \varepsilon_{it}, \quad (7)$$

Where D_i is a binary variable equal to one if individual i is treated and zero otherwise, T_t is a binary variable if the observation is from after the treatment arrived and zero otherwise. In equation (8) $\Delta^{DID} = \beta_3$. They are both the effect of the treatment occurring over time.

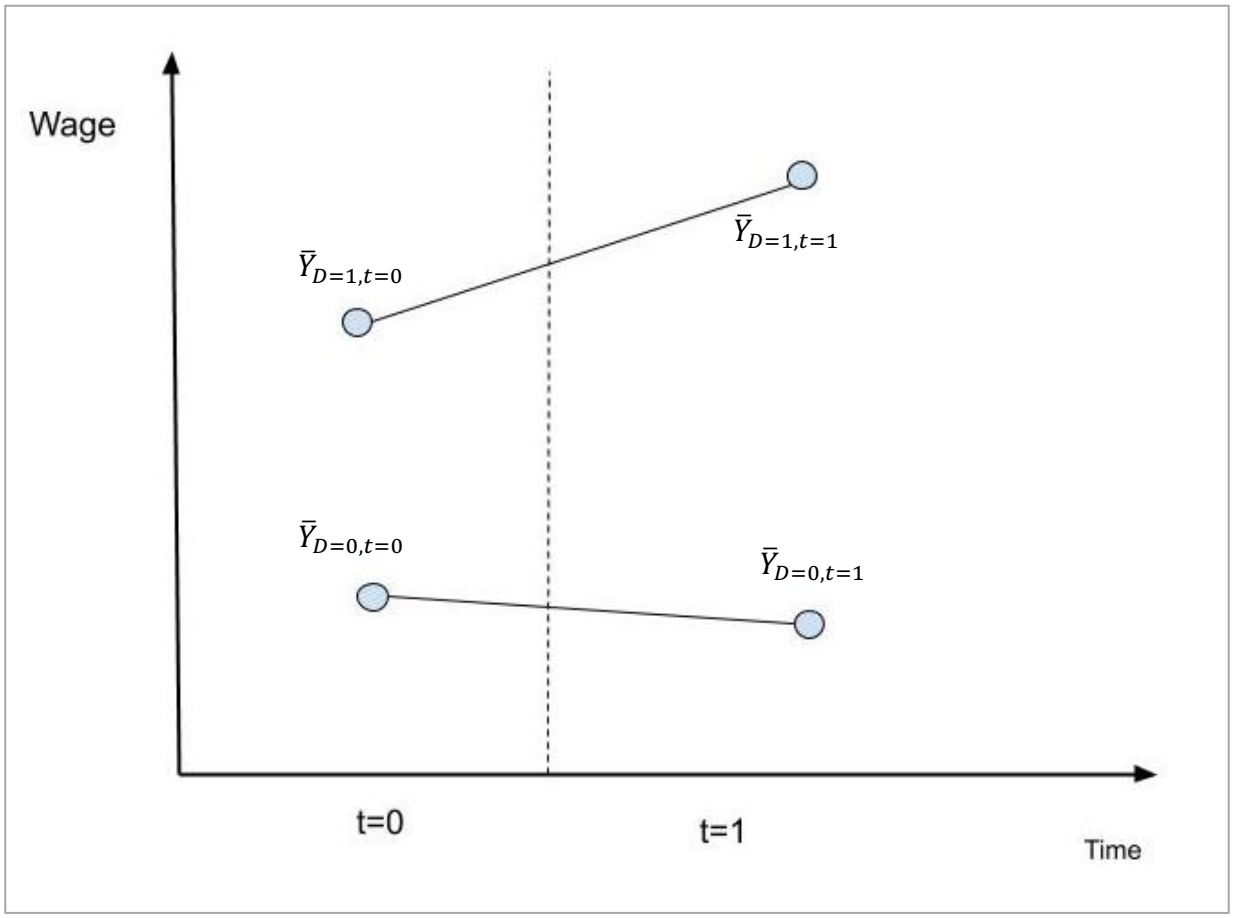

Figure 8: DID Picture

SECTION 2: GLOSSARY OF COMMON TERMS

Collinearity – exists when two variables are highly correlated. An example of "perfect" collinearity would be the relationship between a thermostat and the temperature. Collinearity presents a problem for distinguishing the effect of either variable on a third, dependent variable, because it is very difficult to tease out statistically whether the variation in the dependent variable is being driven by variation in one or the other, since they move together so closely.

Confidence level – the highest probability of Type I error (false discovery) that we are willing to accept to consider an estimated relationship between two variables as *statistically significant* (i.e. very probably not zero). In economics, we typically set a 5% or 10% confidence level, with a 1% level as an even higher bar.

Controls or control variables – are variables that are likely to affect the dependent variable but are not of central interest in the analysis. Because they may be partially correlated with the independent variables of interest they must be "controlled for" in regression analysis to avoid omitted variable bias.

Endogeneity – arises when some factor or relationship not accounted for in a model causes biased estimation of the hypothesized causal relationships between independent variables and the dependent variable. This is typically due to: i) reverse causality; ii)) omitted variable bias (with selection bias a subset of this category); or iii) measurement error. Mathematically, endogeneity arises when an independent variable is correlated with the error term in the statistical model. We call such a variable *endogenous*.

Fixed effects – variables that are 'fixed' over time or space, i.e. variables that vary across countries or individuals but not over time, or variables that differ over time but take on a common value for many individuals or countries. The term is also used to describe a regression that controls for all such effects, whether or not they are individually measured or even observable.

Identification – the statistical approach to identifying (quantifying) the *causal* relationship between two variables, rather than their partial correlation. Valid identification typically requires a clever way of overcoming problems of endogeneity.

Instrumental variables (IV) – Variables that are correlated with potentially endogenous independent variables in a model, but that do not otherwise have any direct relationship with the dependent variable. These variables can be used to predict values of the potentially endogenous independent variables. Because the predicted values will not be correlated with the error term in the regression (i.e. they will not be endogenous) they can be used in place of the measured values of potentially endogenous variables to identify the causal relationship between those variables and the dependent variable. This two-step process of identifying the parameters of interest is sometimes abbreviated ***2SLS*** (for two-stage least squares).

Maximum likelihood – a type of statistical analysis that provides estimates of the parameters of a model that relates independent variables to a dependent variable. For typical linear models the estimates will be equivalent to estimates obtained by using the OLS approach.

OLS – abbreviation for Ordinary Least Squares, a type of statistical analysis that provides estimates of the linear relationships between a dependent variable and various independent variables. These estimates minimize the sum of squared prediction errors from a model. This is the most common type of regression analysis.

Omitted variable bias – bias in the estimate of the causal effect of one variable on another that arises due to failing to control or account for one or more variables that are correlated with both.

Panel – a term used to describe data that covers multiple individuals (e.g. persons, households, firms, countries, etc.) and tracks the same set of individuals over time.

R^2 – a measure of the overall explanatory power of the statistical model, based on the proportion of the variance of the dependent variable that is explained by variance in the independent variables. The closer to 1 this measure is, the better.

Specification – refers to the specific form of a model, i.e. which variables are included and what mathematical relationships are posited between variables. Often researchers test and compare several alternative model specifications.

Standard error – a measure of uncertainty of the regression estimate of a given model parameter

Statistically significant – a parameter estimate is usually said to be statistically significant when it is very unlikely that the "true" value is zero (i.e. low probability of Type I error). More technically, an estimate is said to be statistically significant when the probability of estimating a value so different from zero just by chance if the true value were actually zero, is sufficiently low. We derive a measure of this probability from a *test statistic*, such as a t-statistic.

t-statistic – short for the "Student's *t*-statistic," which is commonly used to test the statistical significance of a model parameter estimate. When used to test the difference of a parameter estimate from zero the statistic is calculated as the value of the estimate divided by the standard error of the estimate. The probability of observing a value of the t-statistic greater than approximately 2 when the parameter estimate does not actually differ from zero, is less than 5%. In this case, we determine that the parameter estimate is statistically significant (or statistically significantly different than zero), assuming our required *confidence level* is 5%.