Cyberseminar Transcript

Date: January 23, 2019

Series: HERC Econometrics with Observational Data

Session: Econometrics Course: Introduction and Identification

Presenter: Todd Wagner, Phd

*This is an unedited transcript of this session. As such, it may contain omissions or errors due to sound quality or misinterpretation. For clarification or verification of any points in the transcript, please refer to the audio version posted at* [http://www.hsrd.research.va.gov/cyberseminars/catalog-archive.cfm](file:///C:\Users\VHASLCMyersK\AppData\Local\Microsoft\Windows\Temporary%20Internet%20Files\Content.IE5\73RXHSNF\l)

Dr. Todd Wagner: At least good morning to those of you on the West coast, like myself. Just wanted to say thanks for joining us today for the first class in our Econometrics with Observational Data. So I have the honor of just introducing briefly the course as a whole but then also to jump right into the first lecture. Let me make sure I can advance the slides and you can see things. Hopefully that works for you.

So the goals of the course are really to enable researchers to conduct careful, quantitative analyses as they think about how to use existing data in the VA, and those might be VA datasets and they might be combined with non-VA datasets. This is sort of the big data era, and VA has many datasets that are over a billion records, and we can do some really cool things with them, but we want people to think carefully about that first.

So in this course we are going to describe different econometric tools and highlight their strengths and limitations so that you can think about how you might use them in your research. Then we’re going to enforce the learning through some examples as we go.

What you are going to see is 11 different classes that are taught on different days. So today is just the introduction and we will talk a little bit about identification, but then you get to see sort of the classes. Wei is going to be talking about research design next, and then we have one, a fun one on propensity score. That one always seems to elicit lovers and haters alike come out in droves. Then we get into two or three ones that I would really recommend people come for, which is the Difference-in-Difference, Regression Discontinuity, and Instrumental Variables. These are courses and techniques that people frequently use to analyze observational data, and then you can see the remainder of the course, so it’s just a great lineup.

I will say, as we move on to today’s class, that we have a very wide audience and we solicit your feedback at the end. You’ll see that as we go. For each of the classes, we will do that. And I apologize ahead of time; about 5% of the class wants us to go super fast and go deeper on each of the things, and then about another 5% in the historical context says we’re going way too fast, slow down. So it’s really hard to teach a class like this nationally that hits perfect levels for everyone. So if you’re finding this totally boring, you can always contact us and we can throw you some great additional readings. And if you are totally lost we can try to work with you as well to get you up to speed.

So today’s class, I really want you to think about causation and thinking about how to model causation with observational data. Then I am going to talk a little bit about what is an equation? Economists uses equations. Statisticians use equations. If you’re not familiar with using equations, they can actually be very helpful if you are familiar with them, and so I just want to run those through those because we will use them throughout the course. I am going to give you an example. And then I am going to get into some of the assumptions that we use in a classic linear model. There are five sort of classic assumptions, and whether you hold on these assumptions really matters in what you can say with the data.

Sort of true with this class, but with almost all the things we do in Health Services Research, is that there is going to be these barriers in technology and terminology. So when we think about terms, it is easy to get into flame wars, whether you use multivariable or multivariate, endogeneity or confounding, interaction or moderation. I definitely have my personal preferences. I will try not to push those on you, but I would encourage you to think carefully about these. But also they are just words, and so part of what we are trying to come up with is a better understanding of each other, and so maybe we don’t always have to attack each other for different words here.

What I will point out is this great paper by Matt Maciejewski and colleagues, in Medical Care Research and Review, where they are actually trying to walk people through the vernacular here so that it is sort of cross-walking and helping people in different fields.

So I have some polls to get us off the ground here, just to make sure I understand where you’re coming from. The first poll is I am interested in understanding your background with analyzing observational data. And Robert, who is helpfully here, is going to push open the poll. What I am trying to get a sense on is if you are totally a beginner, you are going to click number one. You might say, well, I’m modest experience. You might say I’m number three. You can go in at two if you’re in between that. And then you might say hey, I’m reasonably advanced, I’ve used statistical methods to control for unobserved heterogeneity/endogeneity, and maybe you want to present, you can click five.

Rob: That poll is up. And Todd, I apologize, I didn’t quite understand what you were getting at with 1-2-3-4-5, and I just made three options.

Dr. Todd Wagner: That is fine, too. So then there is just three polls and you have to force yourself into one of those; as a beginner, a moderate, or an advanced.

Rob: Right. And those votes are coming in. We’re over 80%. That’s usually where it levels off and it seems to have leveled off around 86%, so I am going to go ahead and close the poll and share out the results and tell you that 22% say that they are a beginner who understands averages, medians, and variance but don’t run regression; about 58% say that they have modest experience, they are familiar with linear or logistic regression; and 20% say that they are reasonably advanced, have used statistical methods to control unobserved heterogeneity and endogeneity. Now would you like to move on to the next poll?

Dr. Todd Wagner: That would be great. Thank you, Robert.

Rob: Sure.

Dr. Todd Wagner: So that is, right away I am amazed at the diversity in backgrounds, so thank you. The next poll is about your training in economics, if you have any advanced training in economics. You can say yes/no. I also gave you an alternative, which is it was so long ago I can’t remember. I feel like every day I am getting closer and closer to answering that.

Rob: And those numbers are streaming in as well. We are way up over 75%. It’s going to level off any second now.

Dr. Todd Wagner: Awesome.

Rob: Yeah, it leveled off around 84, 85%, so I am going to go ahead and close the poll…

Dr. Todd Wagner: Okay.

Rob: …and share out the results. Yes--14% say yes, 76% say no, and 11% say it was so long ago they cannot remember.

Dr. Todd Wagner: Thank you. And then those people will like this next slide, which is the last poll right now, which is the years since your last degree. And I am trying to get a sense on if you are entering this field or if you are sort of at the twilight of your career. Maybe 8+ is not the twilight, but it gives me a sense of where you are.

Rob: It is taking a little bit longer for people to make their choices this time, but we are up over 75% again. And it looks like things have slowed down quite a bit around 85, so I am going to share the results. Seventeen percent answer one, 16% answer two to three, 12% three to four, 14% five to seven, so pretty even across the board there, but then 40% say 8+.

Dr. Todd Wagner: Awesome. Wow, so we have a large number of people who are entering the system and are just going to be amazed, especially when they see the VA data, that are just going to be blown away with it. But it sounds like we have a lot of people who have been in the system for a long time. I recognize that there might be some people here who aren’t using VA data or are interested in using data but don’t have access yet, too, so thank you. Appreciate your help, Robert.

So one of the things that we are going to be trying to do and talk to you through this class is this idea of causation and how do you understand causation. Historically, the way that we’ve modeled causation is through randomized clinical trials. So I am going to take a few slides just to walk you through what we think of as sort of a well-done clinical trial and then sort of pivot and say what is it about observational data that doesn’t follow a clinical trial. So historically, randomized control trials, or RCTs, are the gold standard for assessing causality because the experimenter is manipulating something that wouldn’t otherwise have happened. So that really is what is unique about the clinical trial is that you have got this experimentation control. This treatment/exposure is randomly assigned, so in the end you say the only thing that would have happened that is different between these two groups is this random assignment to this treatment. Otherwise they are the same.

And I put in parentheses (good) here, because not all randomized trials work out the way that they expect them to work out, whether it is poor randomization or something else that happened in this study. So one is, if you do it well, you can make causal inferences, and this is one of the reasons that FDA strongly requires randomized control trials for testing the efficacy of new medications because they want to understand the causal inference.

Random assignment really does distinguish experimental versus non-experimental designs. What the course is going to be talking mostly about is how to use observational data, so obviously there is a pivot there. And before we pivot there I just want to make sure people understand that when we talk about randomization that sometimes there is confusion about, do we mean random assignment versus random selection, and those are very different. In polling, like political polls, you often hear things like we randomly selected 100 people. That just means you are trying to say something that is generalizable to the broader context or community, but it is not experimental design.

Now you could do both; you could say we’re going to randomly select people and then experiment on give half XY, half an X intervention and half not an X intervention. But they are not the same, I just want to make sure people understand that. So it is the random assignment that is required for causation, not the random selection.

There are a lot of limitations with randomized trials. There are papers describing these. First off, because often it is not a random selection of people that go into these, there tend to be very select people who enter randomized trials; it’s that they are not generalizable. And if you look at, for example, many trials, you would see, wow, how does this very select group of people inform how the broad population should take this medication? Unknown.

There is also this thing called a Hawthorne effect, and that is that people behave differently when they are being observed. When you put somebody in a randomized trial they know that they are being observed, and so they behave differently, and this affects both arms. That is a bias that we have to be careful about. Then randomized trials are expensive and slow, so many of the trials, and I have been involved in over 10 or 12 in my career, most of them are well over a million. I have been involved in some that are over 10-20 million. So they are very expensive in sort of the scheme of research. We are now wrapping up a trial, hopefully this year, that was started in the 1990s, so there is also slowness that happens in some of these things.

And then finally, maybe it’s unethical to randomize people to certain conditions or treatments. I guess there’s, I would say there is a large role for observational data to sort of fill a void here, and we can learn from these data. I should have noted that Robert has pushed out the slides, but there was a commentary in Lancet yesterday, and I want to bring it into the slides, so he is going to re-push this out. He didn’t have the chance to get this before I added it today. That talks about the role of research control trials and randomized control trials. Now, if you pay attention to Twitter, this was like a firestorm yesterday. So here is the last sentence in this commentary: “In the absence of randomization, analyses of most observational data from the real world, regardless of their sophistication, can only be viewed as hypothesis generating.” So economists went berserk. They were like people have won Nobel prizes on this stuff. Epidemiologists went berserk. So I don’t want to get into the argument of is this a correct statement? What I would say is, there is a lot of garbage observational data studies out there. So the key word here is, let’s just say most. Most observational data studies are not well done. We are not trying to understand causation.

So what I would encourage people to think about and why I hope you are coming to this course to learn about these techniques is to say how do we get beyond just correlations, which can be wildly biased if you are trying to understand causation. So if you are part of the Twitter world, you probably saw all of this yesterday.

So can secondary data help us understand causation? I put up some headlines. I love coffee. I roast my own coffee. People tend to know that about me. So I put up some ones that have been in the news about coffee. It is everything from, you believe it is like the panacea to your life to that it is going to kill you. So obviously there are some challenges there.

So observational data; why might we use them? Well, first off, especially in VA, we have great historical data going back to the 1990s that is widely available. Like I said before, these are big datasets, so if you are into the lab data and you can look at lab results and lab tests, we are talking about billions of records. We can also do very intense quick analyses at a relatively low cost. So when I said, for example, that these clinical trials are 10-20 million, you might say for many observational data studies, it is considerably smaller than that. It is hard to say without knowing the scope, but it might be in the $100,000 to $1,000,000 range, depending on how detailed it is. It might be much more realistic and generalizable because you are using data on entire samples of populations. But the key that we have to pay attention to is we often talk about that correlation doesn’t equal causation, and it is because your key independent variable may not be exogeneous. It may be endogenous.

Now this might be a new term for you if you had training. One of the reasons I asked about sort of your backgrounds in economics--and most people said they didn’t have that--this is a term that people in economics use all the time; exogenous/endogenous. And let me explain it to you and hopefully it is clear.

A variable is said to be endogenous when it is correlated with the error term, and I will get into that a little bit later. But let’s think of it this way. If there exists a plausible loop of causality between the independent variable that we are interested in and the dependent variable, then it is endogenous and it creates a problem and we are not going to be able to understand causality in this sense.

That probably didn’t make sense. I have a couple more slides here that hopefully will make more sense. So one is just to talk about where does endogeneity come from? It can come from a number of sources, so it is not always from one source or another. It can come from measurement error. Maybe we have biased measurement error. If you are using panel data over time, you can get things like auto-regression, things that are just the way that the data move in sequence and time. You can get simultaneity, so things are sort of happening at the same time, so supply and demand. The most typical one that we think about, which we will talk a lot about, is omitted variables. Sometimes we are just not controlling for everything. I’ll give you an example in the next slide about smoking. But there is oftentimes where we just don’t know what is happening on all the variables. You can get these weird sample selections, these different confounding variables. So when we talk about endogeneity, one of the terms that you might also hear is unobserved confounders.

So let me give you a final example of what I mean by endogeneity. This is a common one in smoking. So we all know that people smoke. People’s decision, sort of that choice to smoke or not smoke is affected by many factors, and almost any study that you can think of is not going to measure all of those factors. You can think of the family influences, the friend influences. They might be different influences at different points in time. If you are a parent like me, you might think of your kids as they are growing up, and my son is now in high school, very different influences in high school than there was in junior high and elementary. There are genetic factors. There are neighborhood factors. We live in Northern California, where for a large part, cigarettes are viewed as a terrible thing. There is a lot of stigma attached with cigarettes. So imagine running a regression where you are interested in understanding heart attacks. So heart attacks is your dependent variable, and you are interested, does smoking cause heart attacks? Well, right away you see that there are a lot of things in sort of what is affecting the choice of smoking that could also be affecting sort of what is going on with the heart attacks, and we are just not going to observe that, so that falls into that error term, and that creates that endogeneity that we are going to talk about.

So a lot of what we are going to spend time on in this course is trying to separate out and think about how do we get good identification of a causal effect on something like smoking. I will say that there are times where it is just not possible, so there is no always silver bullet. It goes without saying that if you observed everything in the world, you were omniscient, then endogeneity is not a problem. But that is not the case. We don’t observe everything. There is no dataset that observes everything. There are different approaches that we will talk about in this course. One is to control for the observables as best we can. And this is why that class that I give on Propensity Scores creates such fur is that you have some people say maybe with propensity scores we can do a much better job controlling for observables. We can get into some really big, wide dataset analyses and do a great job on that hand.

Then there is another group of people who say maybe we should focus on the variation that is not endogenous, that is exogeneous. We’ll talk about how to use things like instrumental variables and regression discontinuity to get at that because if you go back to that smoking sample, you might say, for example, that there’s many reasons that people smoke or stop smoking. Many of those are endogenous, but there might be a couple of them that weren’t the person’s choice. Maybe they lived in a state that just changed its tax rate. And you might say, well, that wasn’t their choice. They have been living in that state for a long time. So maybe we can explore using that change in tax rate as an exogenous change, and so you will get things like that in this Instrumental Variables and Regression Discontinuity classes.

So there’s different terms when you are in econometrics versus statistics, and I mentioned a couple of them. We say endogeneity but people also say unobserved confounding. I don’t want to split hairs there, but I just want you to understand that those similar problems are going to come up when we try to figure out is this a causal effect.

In economics, there is a general philosophy, if it smells endogenous treat it as such. If it seems like it is an endogenous or a problematic variable, just don’t assume that it is going to be okay. But what we really have to pay a lot of attention to is what is generating the data, this underlying data generating process, and trying to figure out what is causing it. In economics, we also place some additional assumptions on that. We tend to think things like organizations or people are profit maximizing or maybe they are quantity maximizing or time minimizing. So we try to do some things in econometrics that might be a little bit different than in statistics. But many of the times, we are using very similar terms here. But almost all of us, whether you’re an econometrician or statistician or just a health services researcher with a biostat background, you are going to rely on equations. We are going to spend a fair amount of time in this class talking about equations, so I want to run through equations so that you feel comfortable using them. I think that if you don’t use them already you will like them, but it does make it sort of necessary to talk about them.

So terms, first off, I just want to be careful here. When I say a univariate, it just means a statistical expression of one variable. I might take a sample of this group, and we said that there was a beginner, intermediate, and advanced training in economics. We could calculate the mode of that variable, and that is just the univariate statistic.

Maybe we are interested in a relationship between--and I am going to have some slides on it later--income and height. So that is really an expression of two variables, so that is a bivariate relationship. Then you might say, well, I am interested in income versus height, but I want to control for a bunch of other stuff, and so that is going to get us into the multivariate world.

So here is, when you look at an equation if you are not familiar with it, on your left-hand side is your dependent variable, often known as your outcome variable. You are going to have on your right-hand side an intercept. You’re going to have a covariate. And then you’re going to have an error term. If you have kids and you have been helping them in geometry, you are going to be like, wow, that is the function of a line. Yes, you are right, that is the function of a line, where you have an intercept and a slope, so that is perfect. We are trying to fit lines to things when we are doing a lot of the statistics we are going to be doing.

So keep in mind that “i” in this case is just an index. It could be any type of index, but if we are analyzing peoples data, then that index typically refers to the person, and that is our unit of analysis in this case. If the “i” was referring to an organization, then we would be saying that unit of analysis is the organization, and there might be many other index, but typically “i” is used when we are referring to individuals.

This is just an expansion of that. Here we have the same dependent variable. We have an intercept. Hey, now we have two covariates; we’ve got beta-1, which is related to the X, and we have beta-2, related to the Z. So this is a multivariate model because we have two covariates, and again we still have an error term.

But we can even expand it a little bit more. And here what I’ve thrown at you is a curveball. Now we have an unknown number of covariates. We have two that we are particularly interested in, X and Z, but then we also have this other sort of sum or vector of covariates that I have just sort of, with this sum sign there, so you can get a sense on what that looks like. So people often use this when they are wanting to say something like, we are really interested in height and gender and how they affect income, and we are going to control for a whole bunch of other nuisance parameters that are the beta-ijs.

All right. So error terms. So in all of those equations that I had before, and if you go back to geometry with your kids, there was no error terminal in the line model, it was just a line. But hey, in statistics, there is an error terminal. An error exists because we are not perfect. We are not tracking perfect lines to everything. We are not measuring things perfectly. There might be omitted variables. There might be measurement error. There might be just human indeterminacy in what is creating this. One of your goals in life, if you sort of have New Year’s resolutions with statistics, it would be understand the error structure. And we are often going to try to do things that minimize error or understand what is going on in that error.

So let me give you an example here. This will become more concrete if we talk about something stupid. So this is not meant to be high sophisticated math or statistics here. Let’s just talk about height and income. So here is our equation. You might say, for example, that you are interested in understanding the relationship between income and height, and you have a hypothesis that height is not related to income. So here is a question to you all, a rhetorical question: If beta-1 equals zero, then what is beta-0. And hopefully you are thinking like, well, if beta-1 is zero, then what we have here is your income with an intercept. So that is just your average income. So hopefully that makes sense to you. So if beta-1 is zero, then what you are going to come up with is just your average income for Y. And that’s that intercept for your line.

So here is some data. I made up some data for you that looks at, on the y-axis is income. This is annualized income in thousands of dollars. So you’ve got people at $50,000, $100,000, $150,000, $200,000. Then you’ve got height in inches. So you’ve got somebody who is at 60 inches, that’s five feet tall, all the way up to somebody at 75 inches. I am 5’10”, so I am somewhere there in the middle, like a lot of people.

So how do we want to describe these data? You can start to see that there does appear to be, at least for part of this, an upward slope between income and height. So let’s think about, maybe we want to fit an estimator.

So what is an estimator? It is a statistic that is going to provide us information that is going to summarize the information of the parameter of interest. In this case, it is height. And we are going to apply a function to the data that is going to tell us a little bit about this. So here is some common estimators. In the simplest world, just think of mean as being an estimator. You are trying to understand, in sort of a univariate sense, what would be the average income. That is a univariate estimator. You might be just saying, well, let’s just take two variables, income and height, and let’s just for different heights look at the average income, sort of the bivariate relationship, and then you might say, well, I want to be a little bit more sophisticated here and I want to have the mean of income by height controlling for other variables in sort of a multivariate sense.

So that gets us until the Ordinary Least Squares world. So let’s just talk about Ordinary Least Squares. This is the multivariate. We have got this line and this immediately is the line that is getting plotted here on your height and income. And the hard line is your fitted values from this regression, and your dots are the original data that I showed you. So we are using, in some sense what we have drawn is a line and so that is just what we were talking about before is that an equation for a line looks like this. But of course there is error around it, so this is the relationship that we are getting out of this height and income regression, and of course there is error.

You might say, well, what about other relationships here? Maybe this isn’t fitting so well. You’re right. You might say, well, maybe we could do some fancy regressions; least absolute deviations or maximum likelihood. And one of your jobs out there is to choose an estimator. And you say, well, there’s no super estimator, so maybe let’s think about different ways of choosing an estimator. And you are going to start running through all of these common criteria. Least squares, you are trying to sort of minimize the distance between the line and the data. Maybe you are particularly curious about unbiasedness or efficiency. There is a whole bunch here. There is a very large literature, and I will get to some citations later if you are interested in choosing an estimator.

What we often see people do in health services research is they start with OLS, and it works pretty well. And what is OLS trying to do here? Here is the data with the OLS. It is that line and you have got the green arrows, which is the measurement error between the line and the fitted value, the fitted value and the dot, I should say. It is trying to minimize that line. So that OLS, by definition, is the line that minimizes those values. People like it. It works pretty well. It is easy to compute. It is relatively fast in computation. So if you are getting to the world of millions of records, or hundreds of millions of records, or billions of records, running some of these fancy models can take minutes, days, depending on how you do it. OLS tends to be pretty darn quick, relatively, so people tend to use it as a staring point. It also tends to work pretty darn well.

Okay, so let’s extend this model then. What about gender? What we looked at so far is just income, and we said if we go back a slide, we’d say hey, there’s a pretty good relationship between income and height. But what about gender? We haven’t talked at all about gender in this relationship. Could gender affect this relationship between income and height? Maybe we want to include just a different intercept for women versus men. And I apologize, my son would yell at me and say, dad, you are being so gender binary. In this regression world, I apologize, I am just taking on sort of a gender binary approach. Maybe you want to create an interaction effect and say what we really are interested in is the relationship and different slopes of that line between men and women.

So here is what it is going to look like if we just take a gender indicator variable. So here is your regression of the line. B is your intercept. X is your height. What we have now added is a separate gender intercept. So what does this look like on that equation? What you are going to end up with is a different intercept for one of the groups than the other group. So now the blue line is the men and the red line is the women. But the way we have created this regression, it forces the slopes to be the same, so that is why they are looking parallel. By definition, this regression forced them to be parallel.

If you want them to have different slopes, that is an interaction effect, so then you are into this equation. It looks a little bit similar to what I have shown before. Now all of a sudden you will notice that we have this interaction, this timesing of the height times the gender. That allows that slope to be different. And here you come up with again, here is this sort of depiction that it matters more for men than it does for women. Now I should say this was made-up data. This is not real data. I was just trying to drive home a situation. I also chose height specifically because we are sort of, genetics and nutrition and so forth, predisposed to certain heights. But it could really raise a question of, is height endogenous? There might be many reasons why height would be endogenous, especially if people are getting poor nutrition that is causing them to be sort of stunted growth. So that may not be as important in some parts of the U.S., but in certain neighborhoods it matters, and especially outside the U.S. that matters.

So hopefully that sort of walking you through the equation of a line is helpful. We are going to talk a lot, you’re going to hear economists talk about the identification. Is the association meaningful? Should we change behavior or make policy based on associations? Generally speaking, most people say no. And if you go back to that Lancet article, what they were saying yesterday was they were saying let’s not create policy based on associations unless they are super, super carefully done. Remember that term “most” I have highlighted. I want you folks, as you are heading down the road in the future, to think about how do we really push this, so we are thinking not just about associations but causal.

So to be able to say something that’s causal, we have to meet five assumptions in the classic linear model. So right now we are about 35 minutes into the course. I have about 25 minutes to talk to you about these five assumptions. That is about 20 more slides, so hopefully you will hang along with me, but I want you to be familiar with why this is so important.

I am going to give you an example here, and I am going to pick on myself. So I am an avid cyclist. I have been biking for 27 years. I used to race a lot. I ride to work almost every day, in part, because I live in Northern California, but of course that is endogenous because I chose to live here. But I am going to pick on something and that is bicycle helmet laws. I bet there are people out there in this audience who think that bicycle helmet laws should just be mandated. We should just create laws that everybody has to use them because we know, in laboratory experiments when you drop a helmet with a bowling ball in it, it protects the head. But it doesn’t translate to real road. What you will notice is that there have been studies done on this. Do bikers behave differently when they are wearing helmets? They actually do. What is more fascinating to me is that drivers behave differently around bikers who are not wearing helmets. If you do studies when they are paying attention to drivers, they actually give people who are not wearing helmets, they give them more space, so we are building in biases to these analyses. And then of course you have these other problems that if you create laws, for example, on these things, you have unintended consequences. So it is a low uptake of bike share.

You can tell this is a hot topic. I am not saying I am advocating one way or the other. I am not going to tell you if I’m wearing a helmet on my way to work. But I do want you to think carefully about the analyses and the assumptions you are making in the analyses. I am sure I am going to get all sorts of hate mail from the 20 or 30 of you out there, so I’ll just apologize for that in advance.

So the classic linear regression. Let me walk you through the five assumptions here. First off is just to say there is no single model that is going to be the best model in every circumstances. I will say that the classic linear model that we typically use is the starting point, and we talked about OLS earlier. So there’s five assumptions and the variations in these assumptions will guide your choice of where you want to go next and what you are concerned about and I am going to walk you through them.

The first assumption is not that strenuous. The dependent variable can be calculated as a linear function of a specific set of independent variables plus an error term. So you are sort of saying implicitly, if you are using a linear model, that the line is the best way to draw this data. And it can be lines or line segments, so be careful there. It doesn’t have to be a single line. You could have little lines. But those little lines are a fair representation of the data. If you know that what you are looking at is nonlinear, then it might be a violation of assumption one and you might want to move on to something else.

There are common assumptions or violations to assumption one. One of these violations is you know you’ve got omitted variables that are nonlinear and you can have nonlinearities, and some of these nonlinearities can be gotten away with, with transformation. So in cost data, a typical situation, we have got these very nonlinear sort of dependent variables and some people transform them to get them more linear in function.

In general speaking, this assumption doesn’t tend to be the assumption that everybody gets obsessed with. There are empirical tests for it. So the theory-based transformations, so if you are coming at this from an economics point of view and you say, hey, I am really interested in understanding the production function, sort of the most common theory-based model, there’s the Cobb-Douglas, which is a log-based model. So might say right off the bat I know that what I really need to be is in the log world. Then you can say maybe I am going to transfer and use the linear model once I have a log transformation.

There are some common sense approaches, but then there are these the Pregibon Link test, Ramsey REST test. What they are typically doing is looking your residuals in your model to try to figure out other problems in your residuals that tell you that you didn’t do a good job with the way that your model is fitting.

I just wanted to be careful, in this sort of--what we’ll typically hear, because one of the assumptions if I go back up is to say that we have omitted variables. There is always this balancing act between how do we develop parsimonious models that are easy to publish in a journal, that are representative without omitting everything. Some people are going to want to have more variables. Other people are going to want to have fewer. And it is going to lead to this question of parsimony in your model.

I still see people using stepwise regression models, so this might be backwards or forwards. I would generally say don’t do that is my recommendation to you. There is little penalty for including a nuisance variable. There is a big penalty for excluding something that matters. There is a lot of literature that shows these sort of stepwise approaches are biased and wrong. And if you are really interested in this, you can get into sort of the newer techniques like the lasso techniques, if you will, that try to give you a penalty for new variables versus old variables and sort of where you cut them off, but that is outside of this class today. Bit I would say avoid stepwise.

So here is the bias, just be careful. Here is the bias if we ignored gender. So if we submitted the line, that red line is if we didn’t take into account gender. Now if we did just a gender intercept, you can see the bias that is created. It does matter. One has to be careful in estimating these things. So I will move on to assumption number two.

We assume in sort of a classic linear model that the expected value of the error term is zero. Remember, we have this error term because we don’t measure things perfectly. We have omitted variables. There is indeterminacy and sort of human behavior that happens. We often make assumptions, but that expected value or that mean value is zero. If we break down that assumption, if that fails, then we are going to have a biased intercept. This is very common when you have cost data where the data are very non-normally distributed. So you end up with these worlds where a lot of the people, 80 of the people might be between zero and $5,000, but you have got a small percentage of people whose annual spending and healthcare are 80, 90, 200, $300,000. That is going to end up leaving problems with your expected value and your error term. So people often work with trying to transform the data to get back to this problem and those are known as sort of log costs and there are smearing estimators. Again, if you are interested in this we can get you to it, and there are going to be two classes later in the course on analyzing cost data.

So assumption three is that your individual cases are independently and identically distributed, as are their error terms. So I am going to give you a slide in a second that is going to show problems with that, but there are two common reasons why you are going to violate assumption number three, sort of this IID assumption.

One is that you’ve got autocorrelation. You’ve got data that represents people over time, and people over time tend to be autocorrelated where their current behavior is correlated with their past behavior and we’ve got to take that into account and that creates bias if we don’t. Another one is that we assume that data are homoskedastic and that the errors are identically distributed, and here is an example. When we look at something like length of stay in a bed, and your cost, your national cost, where you can clearly see the longer you stay in the bed, the more variants you have. There is clearly a positive relationship here, but also notice that the spread gets bigger as you get further out in your sort of your length of stay. I should be careful here. Length of stay in bed section, the bed section is a VA vernacular. So if you are not familiar with that term, we have different bed sections and you can say I’m interested in the length of stay in that bed section. But just to note that that sort of differential spread is going to create problems for your statistical model.

What are the costs of violating assumption number three? The first is to note that your estimates of the coefficients are not biased themselves, but your standard errors are biased. And so you are going to get significant effects that are not always significant or null effects that are not always null.

So in this case, I would say plotting is often very helpful, especially if you are working on sort of a bivariate case where you can just sort of draw this out, if you will, and you can say, hey, let’s look at these things. We’ve clearly got a problem here. There are also different statistical tests for heteroskedasticity, group-wise heteroskedasticity. There are many of these that are out there that tend to have limited power, and that is a general truth in almost all of the empirical tests for our assumptions is that we have tests for many of these things, but they typically have limited power. So you have to have very large samples to be able to test the assumptions themselves.

So if you go back to this slide here, you would say, okay, so what are we going to do to fix this? There might be one easy thing to say. Maybe if we transform the dependent variable we can fix this. So maybe we are not interested in cost by bed section. Maybe we are interested in log cost. Maybe the log cost problem will fix it. Another way is to say maybe we need to work on our standard errors to get more robust standard errors and those are typically known as Huber White or sandwich estimators. You can do that, too.

There is a bunch of papers by Gary King, a statistician and political scientist from Harvard, who has been talking a lot about our over-reliance on sort of just Huber-White standard errors, so he is calling them whitewashing. So just be careful about--there is no easy fix to many of these things, and I encourage people just to be careful when they are working through their regression analyses. There’s reasons why people end up with PhDs in these things; it’s because you have to be rather careful in doing it.

Assumption four, so observations on independent variables are considered fixed in repeated samples. So this is largely going to be, although it sounds sophisticated, it is going to get back to our problems of endogeneity. So a lot of people spend a lot of time trying to figure out this endogeneity. And the second bullet there is one of the important ones, is that you’ve got a correlation that is between your independent variable and your error term, and you can’t test this assumption. Your regression model assumes that this doesn’t happen, but it still happens. Why? Because you might have things like you are not measuring everything you need to measure, so think back to the smoking example. There are many reasons we can’t measure all the reasons why the person chose to smoke, and that is going to be in the error term. We assume it is not, but if we make that assumption and it is not a good assumption, all we are going to end up with is a correlation and not a causation. So we spend a lot of time trying to sort of suss this out. So there are many typical violations. You are going to end up in this errors and variables world. You can sort of measurement error type world with bias. You can get into auto-regression that creates this problem. But a typical one is things like simultaneity or unobserved confounding.

I would say that this is one of the few errors of the assumptions that people spend a lot of time on, and one of the reasons that people are very interested in things like instrumental variables and regression discontinuity. So if you are particularly interested in this set of assumptions, I would encourage you to come to those classes.

So just to sort of harken and push this a little bit further here, so when we talk about errors in variables, the measurement error, so the dependent variable is maintained in that error term. So when we got back to those lines and think back to the height questions and income is that the error term was holding that measurement error. And that OLS assumes that the covariates are measured without error. But if they are measured in error that is correlated with their covariates, that can be highly problematic, and we don’t know how much error and how much bias that is implicitly generating. There is no metric for saying is it a little bit of bias? Is it a huge amount of bias? People have shown that when you do it the right way, things like instrumentals, you can get entirely flipped signs. So it is the question of how much do we want to press on sort of our measurement error here?

So there are many ways that we end up with common violations. Lagging the dependent variable is almost a no-brainer that creates problems because now you’ve got the dependent variable on the right-hand side, too, that is going to be creating all sorts of problems. You are going to get into this world of contemporaneous correlation. You can do things like this Hausman test trying to get around it, but that is also weak in small samples. Again, I had mentioned earlier that you end up with empirical tests that are not perfect themselves.

And then you end up in this world of saying, well, maybe we need an instrumental variable. Maybe we need to focus on the variation that is truly exogenous. So if you believe that that is a potential solution to this, I would really encourage you to come back to that IV class that we teach in a couple of weeks. There has been a lot of interest in using instrumental variables in health services research. Common instruments are things like driving time, taxation rates, and there’s a bunch of other ones that have been floated. They don’t always work, so you can’t just say just because the instrument worked in one situation it will work in other situations too.

Assumption number five is relatively simple. It just says we can’t create a regression model if the number of observations is not bigger than the number of covariates. So if you are trying to model more covariates than you have observations, you are going to have a huge problem and you can’t invert the matrix.

Another problem is that you have to assume that there is no perfect multicollinearity, and if you end up in a world, so I know SAS does this; if you have two variables that mean the same thing, just named differently, and you try to put them in, SAS freaks out. Stata says it can’t do it and drops one of those variables and tells you which variable it drops. But it assumes there is no multicollinearity. If you have multicollinearity but it is not perfect, the regression can still happen. You can still invert the matrix. You are just going to be with widely imprecise or wildly large standard errors.

We always joked in graduate school is you paid attention to perfect collinearity and then of course the professor said if you feel like you have got multicollinearity but it is not perfect, just increase your sample size, often because they could laugh and say that is always easier said than done. So there is often--it is not easy to increase your sample size, but be careful of collinearity.

Assumption five is coming up more and more now in the world of prediction. And I just wanted to say a couple of words about that. So we have created a Risk Score in VA called Nosos. This is trying to, in some sense, predict what the person’s estimated or mean spending is this year or next year, depending on the model that you’re using. And we use a huge number of right-hand side variables. And there are like 83 different medical variables. There are 50 or so mental health variables. And people say aren’t you worried about the collinearity between some of these variables. My general answer to this is I am not trying to say anything about the specific beta coefficient for any one of those variables. I am just interested in creating the best prediction that I can. So in this world, where it is about prediction, I am not worried about specific interpretations of specific variables. If I were, it would be highly problematic because they are correct, there are many sort of tightly colinear variables, specially on the mental health side, and I would be very worried about the sort of imprecision there. But as a single model, it generates an estimate and that is what I am interested in. So in the world of prediction the multicollinearity might not be a problem that you are worried about.

I know we are running out of time, so I am going to give you some references and then move to questions and answers and ways to contact us, because it is really hard to do this and answer all of the questions in the time given.

So for those who like econometrics, Kennedy has this book that is very easy to read. I have an older version on my bookshelf, “A Guide to Econometrics.” I loved it. Bill Greene, NYU econometrician, has sort of what people think of as sort of the standard econometric text analysis. A lot of these are going to be heavily equation-dependent. And so one of the reasons, talking about equations, is that’s a common--math becomes sort of this common vernacular that we can all get around. So that is a great book, but it is dense.

And then Wooldridge at Michigan State has done a great book on cross-sectional panel data models; people love that book. And then there’s a number of other ones, depending on your background. So years ago when I gave this class, people were like you should really approach this from a different background, not just economics, and I’m like, well, that’s my training. But here’s one that is sort of bringing at this from a sociological point of view. I think it’s a great idea to sort of understand that whether you are an economist, or you’re a political scientist like Gary King, that many people are struggling to understand causality with observational data. If you are an epidemiologist, Miguel Hernan.

So if you’re not on Twitter, you will find many of these folks on Twitter and you can follow them and then have fun. If you want, any questions, I would say there’s a couple of ways to reach out to me specifically. My email is there. You can see the general HERC email. You can see our HERC VA will be announcing all of our future courses on our Twitter account, my Twitter account, and I would encourage people to spend time in this because it is complicated, but it’s worth it.

So not a whole lot of time, so I apologize for that. Jean [phonetic], do we have any questions? We probably have a gazillion that I skipped over.

Jean: Yeah, we do have a few questions. If we don’t get through all of them, I encourage you to contact Todd or HERC directly. So the first question asks: What is your comment on an ever-burgeoning opinion that data for a linear model is overly manipulated and therefore makes the results questionable at best? Therefore, nonlinear models are far more valid and actually reflect more truthful depictions of data.

Dr. Todd Wagner: Wow, that is a great comment. So you could also go back and say every model is wrong. But some models are useful and so what you’re going to get into this world of is trying to come up with useful models. There may be times where nonlinear models are very different in their interpretation and use than a linear model. And that’s going to be on your shoulders to create. If you are using a lot of logistic regression, Ciaran Phibbs is going to be talking about that in a different class. Logistic regression is when you have a 0-1 dependent variable, and maybe it is for something like are you smoking and it is a question of yes or no. There’s a lot of ways to interpret the coefficients on that nonlinear model. One is to say, hey, we are interested in odds ratio. There’s relative risk. There’s marginal effects. They don’t all mean the same thing. There’s a great paper by Edward Norton that came out in a commentary in JAMA recently that is talking exactly about this comment about how do we best interpret these things from nonlinear models, where they might mean different things depending on where the person is on the curve. So that is the problem with nonlinear models is they are not an overall mean, is they might say, well, for certain people here’s the effect, for other groups it’s this effect. So I want us to be a little bit careful about saying one, that nonlinear models are always better. Sometimes they are much harder to interpret. Maybe it doesn’t matter all that much. Maybe it matters a huge amount. So hopefully I answered the question. If not, I will get vilified, I’m sure.

Jean: Okay. The next question asks: Will including robust standard errors fit multicollinearity?

Dr. Todd Wagner: No. So robust standard errors is just trying to get a way of this problem of your standard errors being potentially biased. But it is not going to be a solution to your multicollinearity. Your multicollinearities, you either have to drop one of the variables or increase your sample size, or if you are in the prediction model you just don’t care about the individual coefficients, so it doesn’t matter to you. But for most of the time, no, it is not going to solve your problem for you.

Jean: Okay. The next question asks: Can you explain exogeneity more precisely and the same for autocorrelation?

Dr. Todd Wagner: Okay, great questions. So everybody has--and there’s different ways that people think about endogeneity and exogeneity. So let me pick another one that people seem to like. Every one of us has different hormones in our body, whether it is testosterone or progesterone or different hormones. There is a big distinction between somebody who is naturally high on these things, let’s just say naturally high on testosterone, versus giving someone a shot of testosterone. So we can do a whole bunch of analyses that look at, for example, natural variation in testosterone. That is endogenous. We don’t know why the person has a naturally high or naturally low variation in the testosterone. That is just what we observe. It might be due to things we just don’t even know in the science yet. So because we are not observing those other variables, we can attribute something to the testosterone that is totally different than what happens when you inject someone with testosterone. So the goal really in trying to translate endogeneity or exogeneity is trying to figure out is there a way to translate sort of the endogenous information to understand, maybe some of it is exogenous. And so when we talked about the taxes and smoking, you might say, well, there’s many reasons people smoke. But there might be a small amount of that variation that is related to something about taxation. So maybe if we can isolate that small amount of variation, we can talk about the exogenous effect of stopping smoking. Hopefully that worked with that example. If not, you can follow up with me.

Oh, autocorrelation, right? You want me on the autocorrelation?

Jean: Yes.

Dr. Todd Wagner: So the autocorrelation is really just a matter of understanding sort of points in time. And if you think of it, whether it is people and sort of what they do today as a function of what they did yesterday, and so if you are tracking and you are modeling data on people over time or organizations over time, inherently there is a shadow of what they did in the past period. And so we have to be very careful if we are analyzing things over time that we are controlling for that shadow. Otherwise what we could be doing is just measuring the shadow itself.

Jean: Okay.

Dr. Todd Wagner: And Jean, you are always welcome to jump in with other answers if you feel that I’ve missed it.

Jean: Okay. The next question asks: Can you explain heteroskedasticity again?

Dr. Todd Wagner: Yes. So heteroskedasticity, you end up with this variable that has got an error distribution. I am going to go back to this slide here, where your variants, and Robert, I don’t know if you are still on. Is there a way that I can do, oh pen, that is what I’m looking for. There it goes.

Rob: Got it?

Dr. Todd Wagner: Yeah, I got it. So if I take this slice here--I apologize, my pen--versus here, as we start down on the low side here, it is very little variation. But up at the top end here, let’s see if I can do this with my mouse, it is a huge amount of variation. And that is heteroskedasticity. So the error is a function of your x-term, if you will, and that creates problems with the x-term. And remember, it’s not that it’s biasing your value, your beta coefficient, it is affecting your standard errors because our assumption is that it would be a cloud that looks something like this, which would be homoskedastic, that there would be no variation that is across your x-term. Sometimes it is easier to plot and think of it visually than it is to describe it. Hopefully that helped. Great question.

Jean: Okay, great. Another question asks: You talked at the end about multicollinearity and not caring so much if you care only for predictions. Can you explain that a bit more?

Dr. Todd Wagner: Yeah, and this will probably be the last question and then we will have to take them offline. I know people probably have appointments that they have to go to. Multicollinearity is really a problem when you are trying to interpret two, or different coefficients on your right-hand side variables. So let’s just say we went back and we are interested in understanding cost being our dependent variable and then you’ve got all these regression coefficients, and maybe you’ve got a bunch of them that are like specialized anxiety and depression, depression alone, and different clinical conditions that are very tightly connected. There might be very few people who are in one of those groups but not in other of those groups. And so it might be hard to tease apart the individual coefficients on those. But maybe those three work together in such a way that we can get a pretty good estimate of overall prediction on how does mental health and these conditions affect total spending. So on the total sort of prediction of total cost, they work okay. I just can’t suss the individuals out because they are too tightly related.

Again, I apologize if I am confusing people on that, but I can always do more later on.

Jean: Great.

Dr. Todd Wagner: Just thank you so much, Jean, for your help on all of this. Just to come back and say please come to our future classes. This was really just the intro to a bunch of much more fun, especially if you are going to get interested in the world on things like instrumental variables and ways that you can use data. And I know Jean is doing a fun one on difference-in-differences, so sort of understanding the data. And thanks to Jean and Robert for all their help running this. It takes an army to keep these things going, so thank you all.

[ END OF AUDIO ]