Cyber Seminar Trasncript  
Date: 3/11/2015

Series: HERC   
Session: Econometics with Observational Data

Presenter: Todd Wagner  
 *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 www.hsrd.research.va.gov/cyberseminars/catalog-archive.cfm or contact:* [***todd.wagner@va.gov***](mailto:todd.wagner@va.gov)

Todd Wagner: Hello, everyone. My name is Todd Wagner. I am one of the health economists here in the Health Economics Research Center. This is the first of the lectures for the econometrics series with observational data. This is the introduction, in which case, I'm going to ask a couple of questions about you as an audience and provide background. We will also start getting the discussion talked about and started when we talk about identification.

Like Heidi said, we had question and answers. I apologize that we cannot do everything open with open mics. It is just too much. If you can ask questions. If there are clarification questions, I will try to answer them as we go. If there are big questions that are much more philosophical, I will try to address those at the end. Often, we run out of time. Sometimes there are questions that come afterwards. I will try to address those at that point as well. For the goals…. Hey, Paul?

Paul Barnett: I am sorry to interrupt. This is Paul Barnett, another health economist at the Center. I will be here to help Todd direct his attention to the questions.

Todd Wagner: Thank you so much, Paul. I appreciate it. We have two goals for the course. One is just to enable researchers to conduct careful analyses of existing VA data and possibly even non-VA datasets. What we are planning to do is describe different econometrics tools and their strengths, and limitations. We are, throughout each of the classes, we are going to try to use examples to reinforce your learning and to make sure that you understand exactly what is going on.

Then, it is like I said. Some people have contacted me specifically to address specific questions that they are working on. That is possible as well. Here is the course schedule. As you can see, today is the introduction that we are going to talk a little bit about, identification.

Christine Chee, who is a health economist here is going to have a class on the 25th; so in two weeks talking about research design. I am then going to come back a little bit about PET, screening scores on April 1st. Christine is then going to talk about natural experiments and difference in difference models followed by a class on the 22nd where we talk about instrumental variables. We are then going to have Haley Hedlin who is the Stanford professor talk about mixed-effect models. That is actually a new class that we have added this year.

There has been a lot of interest in trying to understand mixed-effects. We then have Ciaran Phibbs who is the health economist here talking about two, or is doing two classes. One is Specifying The Regression Model. He is talking a lot about functional form and sometimes the assumptions that go into the right-hand side variables. Then we are going to talk about in a second class, limited dependent variables where you might do something like a logit model. You might have a choice model with three outcomes.

Then, Paul is going to finalize the course with the course with two classes on analyzing cost data. There is a lot of people out there who are interested in analyzing cost data. We are going to give that specific attention with Paul's lectures. The goals for today's class are a little bit more modest. One is just to talk a little bit about causation and understanding causation with observational data. As you all know, that is a huge challenge.

We are also going to introduce the idea of an equation. We are going to describe the elements of an equation throughout much of the class. You are going to see equations pop up. I am particularly interested in not having people have this like shock or scared reaction when they see an equation. But that there is some understanding of what we mean by an outcome variable, a predictor variable, or a right-hand side variable. We will give you elements of that equation and how we mean it. We are then going to walk through an example of an equation.

Then I am going to run through the assumptions in a class linear model. When I say linear model, it is the just standard linear regression. The terminology, and it is never easy. I will point you to a paper that Matt Maciejewski, who is at the Durham VA at Duke wrote with colleagues. Confusing terminology is a challenge for many disciplines. This is particularly true in healthcare where we are often working in teams. Epidemiologists might have a set of terms that are very similar, but might be slightly different than what economists are using.

You could have terms and differences. They could be multivariable and multivariate, confounding versus endogeneity. Then you get into some challenging ones; interaction versus moderation. Then, of course, I just was joking with the right or wrong. But that is, sometimes these terminologies have sort of taken on this well, that is right. This is wrong. Sometimes that is not the case. It is just the difference of perspective. I will point you to Matt's paper. It is a great paper and sort of walk you through the different terminologies in health services research.

I would like to start by understanding the audience a little bit. We have – it is always amazing to me – 135 people out there that are listening. We are going to do a poll question. Heidi, if you can help me with the poll?

Heidi: Yes. The poll questions are up on the screen. The question is what is your background with analyzing observational data? I am only able to put a small amount of the possible responses up here. Todd, you could read through a little bit more of what they are supposed to be here.

Todd Wagner: Sure. Yeah. What I have an idea here is five responses. This is just a Likert scale. You can define yourself as a beginner, which is category one; which is, more or less you understand averages. What a median is and variance, but you do not typically run regression. You are interested in learning more about that. Three is you have modest experience. You are familiar with things like linear or perhaps logistic regression; and two being somewhere between those. Then we also have five, which is consider yourself reasonably advanced.

You have used statistical methods. You are aware of methods for controlling for unobserved heterogeneity, or things like endogeneity. Perhaps you have read about instrumental variables and even conducted one. I am seeing the results\_\_\_\_\_ [00:05:56]. This is, it is an amazingly diverse audience out there.

Heidi: It is. We will give everyone just a few more moments here. I will close the poll out and put the responses on the screen. It looks like we have stopped. We are seeing around 11 percent saying they are beginners; and seven percent gauging themselves at about a two; and 30 percent with modest experience; 22 percent seeing about a four; and 31 percent seeing reasonably advanced. Thank you everyone for participating.

Todd Wagner: That is a great segway. Throughout the class, we are going to try address both end so the spectrum. But as you can imagine, it is hard to in a cyber course address issues that are focused more on the beginner, but also really push the reasonably advanced. I apologize if I lose\_\_\_\_\_ [00:06:48] of the distribution. What I would say is if you are struggling at either end. You are not getting enough information as a beginner, you can reach back out to us. Or, vice versa if you want more citations. You feel like you are reasonably advance.

We can also push you with more papers in that area as well. The next poll, I have another poll item, which is what is your formal educational training? Here I have three or five categories in mind. One is your Bachelors training, and perhaps your RA, or program coordinator, and another one is your Masters trained. The third group is you have clinical training. I am really focused on people with – and let us say your physician, a nurse, or pharmacist, or even a clinical psychologist where your goal really is you are providing clinical care or your background in clinical training.

Then you have the fourth group is – you have clinical training. But you also have training in health services research, and perhaps through a fellowship. Then finally, the fifth one is you have a PhD or equivalent in a social science, but you have no clinical training. I am a health economist. I would say I am the last, the PhD with the social science.

Again, it is just an amazing group of diversity here that is very impressive. Right now, in the lead are Masters. Heidi, did you share that yet?

Heidi: I am sorry. I did not realize I was muted there. We are seeing around six percent saying a Bachelor's degree; 40 percent saying Masters; eight percent, clinical training, MD, RN, Pharm D, or clinical psychology; 13 percent, clinical training with health services fellowship; and 34 percent PhD in social science with no clinical training. Thank you, everyone.

Todd Wagner: Yes, thank you very much. Again, much like the last slide where we were getting your background on using observational data for analysis. This also provides more contact. We will use this to sort of gauge the next classes as well. Alright, so let us go on to the final poll question. Years since last degree? I am trying to get a sense on how many yeas is it? Are we talking mostly to people who just recently graduated? Are people – I am assuming most of you are graduate and maybe there are some students out there. If you are a student, just put one and then all the way up to eight plus.

Heidi: The responses are coming in. We will give everyone just a few more moments. I will put those up on the screen here. It looks like we have slowed down. We are seeing 12 percent saying one year; 18 percent saying two to three years; nine percent saying three to four years; 16 percent saying five to seven years; and 46 percent saying eight or more years since the last degree. Thank you, everyone.

Todd Wagner: That is perfect. It looks like we have about a third of the people who are relatively new since their degree. Then about half of the people who have been, like myself, who have been out for a while and getting stale. Alright, let us jump right in. the key that we are trying to often do with observational data is understanding causation. As you all know, correlation does not equal causation. The gold standard here really is, in the medical field is a randomized clinical trial. In our randomized clinical trial or RCT are really what we think about is the best way to assess causality. I

will sort of pose the question to you all. What is unique about a randomized trial? If you can think about what is really distinct about a randomized trial? Hopefully you get the idea that the treatment is randomly assigned. If it is done well, the only thing different between the control group and the intervention group is whether they were randomly assigned a treatment. If the treatment has no effect, you would see no difference between the two groups.

Very typical, if you go to most clinical trials, you will have a table one; which describes the characteristics of your sample. One column has the characteristics that your treatment group. Column two has two characteristics of your control group. They say wow, the randomization to look to work well. There is quite a bit of balance between the groups. Not always is there a perfect balance. But there is a lot of effort to go in to making sure that it balanced well. What is the benefits of randomization? One of the key issues is because it has been done exogenously. The researcher is the one who assigned the randomization.

You can say something about causal inference. It was not the patient's choice to be randomized. They just agreed to be in the controlled trial. But it was not their choice to be in the treatment versus observational arm. If it was just their choice, we would not know why they chose it. You would get all of these weird confounders. But because it is assigned, we can say something about causal inferences. That issue of exogeneity and sort of the flip side of that coin endogeneity is going to come back to observational data.

It is going to be an issue that we deal throughout this course. Random assignment distinguishes experimental and non-experimental design. I just wanted to make sure people do not confuse this idea of random assignment versus random selection. Selection can be really important for issues about generalizability especially if you are doing national surveys. You might want to say we randomly selected survey participants. But the random assignment to a treatment or a control is what is really critical for understanding causation just so you are not confused on the two issues.

Now there are a number of limitations that come with randomized clinical trials. There has been push on this front. I will sort of talk a little bit about that to make some of these limitations less onerous. But one is that you might have generalizability problems. You might have enrolled a group of patients who are not representative of the national population. One of the questions and even though you have great information on causality, you might have low generalizability.

There is this other issue which is Hawthorne effect. That is the idea of when people are observed or know that they are being observed is that they sometimes change their behavior. That is known as the Hawthorne effect. That can effect both arms. It is particularly challenging in behavioral interventions where you have got the behavior is what is changing. But it can happen otherwise, too, even in a pharmacy clinical trial with pharmaceuticals. Because the patient might change their appearance. A real challenge is that RCTs are expensive. They are often very slow. A typical clinical trial might take a year or two to plan. Then the data will take three years to enroll all of the participants. Then another year or so to do the analysis. You can easily talk about five to seven year before you start thinking about a clinical trial before the final publication is out.

Sometimes health systems want more rapid knowledge about what is happening with their patients. RCTs are not always the way to go. There has been a lot of work trying to do what is known as practical clinical trials. I will sort of point out there is a professor at Stanford who used to be at the VA. Phil Lavori, who has been pushing this with some folks at Boston. Mike Gaziano and Luciori where they are trying to do the ideas; if you come in as a patient. You are interested. Let us say you are interested in statin medication. The doctor realizes that they might have three or four medications that are equally possible to give you. That instead of just being the clinician's choice, or the patient's choice, that it is randomly assigned.

We can continually learn. There has been a lot of effort in the VA trying to understand sort of these practical clinical trials. There is another issue is that sometimes it can be unethical to randomize people to certain treatments or conditions. That comes up often when we are dealing with certain types of trials. You would say well, it is not ethical to randomize people to not get cancer screening, for example. They need to get cancer screening. Quasi-mixed experimental designs can often fill an important in sort of plugging the limitations of RCTs. But what we are going to come back and what we are always struggle with is this issue of it not having exogenous treatment climate. Can secondary data help us understand causation?

I have just sort of thrown up some titles that are pulled from the media. For those who have heard me talk about these Cyberseminars before, I talk a lot about coffee. I like coffee. You can see there is a bunch up here. Coffee may make you lazy. Coffee and exercise may decrease the risk of skin cancer. Coffee, depending on your perspective is either a savior or it is going to kill you, depending on what data you use the analyze those things. I like this one. Coffee may make high achievers slack off. I had a pot of coffee this morning. I am going to slack off today. I do not know if that is true or not.

Observational data, there are definitely some benefits of having observational data. One is that it is widely available especially in VA. If you are in the VA system, you will recognize that we have the corporate data warehouse. We are talking about terabytes of data, millions and millions of records. Relatively easily, we could do quick analyses at a low cost to understand patterns of care. We can look at the entire system. These might be realistic and very generalizable results. But the questions that we might be interested in may not be a randomized assignment. The question of is it callable?

I will get to this issue. If it is not exogenous, it may be endogenous. That is the opposite. A variable is said to be endogenous. This is statistically speaking when it is correlated with the error terms. You might have a right-hand side variable that you are interested in understanding. I will give you an example in a second. You also have this error term. When those two are correlated, you have a problem of endogeneity. This is assumption four in the classic linear model that I will get in later in the talk. If there is this – what we think of as this loop between the independent and the dependent variables, it can create problems with our analysis.

Then you often hear this assumption correlation does not equal causation. Endogeneity can come from many sources. One, it can be just there from measurement error. You can get autoregression or autocorrelated errors. This is specifically if you are looking at people's behavior over time. You can get these issues of simultaneity and sort of systems happening. The question about should we have more ICU beds or fewer ICU beds depending on the number of ICU or the number of ward beds? Those things are determined simultaneously. You can have omitted variables where it just did not measure something.

That is what is creating this correlation. Or, you can get some very complex sample selection problems. Let me be much more\_\_\_\_\_ [00:18:03 to 00:18:04] give you an example. The question is does greater use of positronic emission testing and screening decrease lung cancer mortality? Perhaps you dig into the data. You observe people in facilities that do a lot of PET screening while others do little. Then you say well, let us compare patients across these facilities. You find that negative correlation between PET screening intensity and mortality such that the more screening you are doing, the lower the mortality rate.

It is not correct to say that PET screening is causally linked to mortality. Because you can imagine all sorts of reasons why PET screening might be endogenous. You might have PET screening machine at places that are higher quality. Or the physicians just recognize that this was the right way to go. Those physicians are better on a whole host of reasons than other places. It is not the PET screening itself. It is that the places have differences in quality. There are a bunch of reasons why that set up makes PET screening endogenous. That is a very common type of example that we get with observational data.

This is my last slide here before I dive back into sort of a higher level question, which is this equation. I just want to highlight some differences between econometrics and statistics. Often people come to HERC. They say you are a health economist. Can you just do the statistics for us? There is obviously a tight correlation between the two. We often use slightly different terms. It can be confusing a little bit. We use cultural norms and economics – excuse me – where we often say if it seems or smells like it is endogenous, then it probably is. We do not take it on face value that something is exogenous. Then we often assume that there is an underlying data generating process that is economic.

I will point this out. Often you can relax some of these assumptions and still come away with the right answers. But sometimes you may not want to relax some of these assumptions. We often assume, for example, there are rational actors out there that are really interested in whether it is profit maximization. That might not be the ideal in VA. Maybe the VA is interested in quantity or quality maximization; or trying to – that the access minimization or time minimization. But they are trying to all adhere to some sort of objective function that is consistent.

The economics is very interested in understanding that underlying data generating process. When we deal with an economics, we often start with an equation. If you have ever talked to me about a study design; and what we will often talk about is well let us talk about what do we want them study? That will be the dependent variable. Let us just go through the terms. Univariate is what I think of as a statistical expression of one variable. I can say show me the mean or the average, or the variance of a single variable. That is a univariate – what I think of as a univariate expression. I can say well, like give me an x, y plot of two variables. To me that is a bivariate expression of two variables.

You might have something like I am very interested in height and weight, and the relationship between height and weight. Then we can say what is the correlation? That would be a bivariate expression. Then multivariate, I will just say it is an expression of more than one variable. It can be dependent. You can have multiple dependent variables. Or, typically what is common is you have one dependent variable and many independent variables. That is sort of a multivariate expression. What do I mean? Let me just be specific here.

Hopefully this is not freaking people out. Yi is the dependent variable. Sometimes people call it – is the outcome variable. In economics this might be something like total cost. You could also have maybe it is something like length of stay. You could also have some – another variable that is an outcome variable or even mortality. In the equation and this, you are going to have an interceptive where this crosses the Y axis.

You are then going to have a bunch of covariates that you are going to include, and then an error term. The term and what I think of as covariate. That is the beta one. Sometimes it is called a covariate. Sometimes it is called a right-hand side variable. Sometimes it is called a predictor or independent variable. I do not often use, and personally do not use those terms frequently just because they assume endogeneity. If you say it is independent, well it may not be truly independent.

I often use the term right-hand side variable. You will hear me throughout the class talking about right-hand side variables or covariates. Note that this is very – if you can go back all of the way to when you learned about a line in geometry where it is a line equation. It is Y equals MX, plus B. this looks very similar. What you are really doing is you are going to be sketching a line through your data. You will notice some other terms here.

One is this I. I, I think of as an index. Often people who are interested in understanding or analyzing data from patients. I in that case would be differences in patients. You would have I, a number or I, equals one to N number of patients. It is just one way to identify the index of patients. But it could also be something like a facility. It does not have to be I. We can introduce more than one index. We can talk about time. Then you might have like I N T. There might be like I said, other indexes there.

Let us just expand this equation a little bit. Now you have your dependent variable. You have your intercept. Those are the same as before. But now we have two covariates and again your error term. This is just a slight expansion of that other model that we talked about. This is a different notation. It just feels more comfortable with this notation. This might be an abbreviation of you are really interested in beta one. Maybe that is the key independent variable or key covariate that you are interested in.

Then you have a bunch of control variables. Maybe you are controlling for an age and income, and gender. Those might be summarized by your sum and the beta I J, X I J term. Often you cannot put the whole equation on the paper just because it is too wide. The error term exists. Let me just run through about this error term. We spend a lot of time in econometrics talking about this error term and trying to understand the error term. Error exists because there might be other important variables that we are not measuring. That is even in the data or the sort of big data error, we have variables that are not typically measured.

There is also just often there is measurement error. If you do the same measurement; and people that you have to use a ruler just to measure height on somebody. You might get that they are 5' 10" one time. You do it a second time. Well, it is 5' 9.5". It is just measurement error. There is often this issue of human indeterminacy and sort of this structure of what is causing what? That can lead to some error term. There is a lot of effort that we go into understanding this error structure and trying to minimize error. That is often what econometrics is really trying to be doing.

There is also, you can think about it. Error can be an additive. It can be multiplicative. When Ciaran talks later on in the course about understanding different dependent variables, that is going to have the very different look on, ultimately on the error term. The same is true when Paul talks about analyzing cost data. When you think about and maybe you are interested in using a log to regression. That can have a very different function on your error term.

Let me just run through an example with these equations just so it is even more concrete. Is height associated with income? Do you have, for example, in this case – income would be your dependent variable? We are interested in understanding what predicts or what is associated with income? Then X is your height. Just to remind you that your beta zero. I do – and Heidi, remind me. I have to find my notation tools.

Heidi: On that tab, it is hanging out the left-hand side of your…

Todd Wagner: I got it. Yes, thank you.

Heidi: Yeah.

Todd Wagner: Hopefully people can see that I am circling now the beta zero. That is your intercept. I will explain what that is. That is really just kind of your average height for your sample. Then this is the variable that you are interested in, which is your height, your average height. Here is your error term that we talked about. You might set up a hypothesis that height is not related to income. That would be your B 1 equals zero. That means if B 1 equals zero, then what is B zero? Here is a question for folks. That B zero is just going to be your average income. It is just the average of your Y I.

We are now up to 177 people. One of our former technologies, we were able to get more interaction with people. I apologize that I cannot open up the phone. It is also really hard for me. I would love to see people's eyes at this point to make sure that you – I can see people not glazing over or checking the e-mail. But I cannot. I assume people are getting this. Let us plot our height and income on an X Y 1. How do we want to describe the data? Each thought is a person. Well, okay, let us look at – maybe we have an estimator to describe these data. An estimator is just this statistic that provides information on the parameter of interest, which is height. We can generate it by applying a function to the data?

Here are many common estimators. Even though the sound estimator, it might sound fancy, a mean, and which is an average; a median, these are all just univariate estimators. Ordinarily squares is a multivariate estimator. They are just different ways for understanding your data. Here it is. What an ordinarily square comparison of height and weight would look like. This is just trying to draw a line that associates height and weight.

Paul Barnett: You made this data up, did you not, Todd?

Todd Wagner: I did. But they highlight my point.

Paul Barnett: Well…

Todd Wagner: If you are a…. Let us just say you could be an outlier here. You might love these data. My wife who is down on the, well just say on the lower half, may not love these data. But that is going to get to a point I am going to make later on.

Paul Barnett: Are you going to control for gender or well, or what? Let us keep going.

Todd Wagner: We will get there. You are reading my mind. Here is just your linear regression model. This is what you would get, if you were to run that through. There could be other estimators, too. You might say well\_\_\_\_\_ [00:28:53] interested in a least absolute negations or a nonlinear model is what I am showing you on the right-hand side. It is a nonlinear model. You might say well, let us. There may be a fact unlikelihood model that were interpreted and right.

There are many different types of estimators that one can choose from when trying to estimate this model. There are different criteria that one can set up in choosing an estimator. You could say well, we are really interested in sort of the least squares and trying to minimize the variance. You could say well, I am interested in unbiasedness as an estimator. That is always a good thing to have, an unbiased estimator. Maybe we are interested in efficiency and trying to think about the minimum of the variance.

There could be asymtotic properties of estimators. There is a maximum likelihood estimators. There is goodness of fit. We are not going to get on to all of these nuances to discourse especially not today. But just to know that there are many ways of choosing an estimator. Throughout the course, we will talk about specifically when we get to Paul's discussion on cost data.

We are often struggling with choosing among estimators. Because there has been a large literature talking about how do we think about what is the right way to model the data? What is OLS really trying to do? What OLS is trying to do is you will see these green arrows to the line. It is trying to minimize that absolute distance. That is what the OLS is trying to do. In this case having the income on the left-hand side. You cannot just swap and say well, let us put height on the – as your dependent variable as income on your right-hand variable. It does not work that way.

What OLS is trying to do is minimize the distance between the height and income. What about gender? Paul raised this issue. How could gender effect the relationship between height and income? You might think well gender is a confound here. Most women are shorter than men. You are right. What happened? Now there we go. There is what we would think about going back to this equation. You would say let us include a gender variable in here. This is really a gender intercept. This is, I mean, are you familiar with the term dummy variables or indicator variables?

We have now included beta 2, which say dummy variable for gender. This is going to allow men and women to have a different Y intercept. What does this look like? This is what this is going to look like with your data. You are going to force – this model is going to force these two lines to be parallel. What and where they differ is where they intercept the Y axis. Now we have a line, that red line, which is your information that is largely for your women. Then the blue line is for the men in the sample.

The beta one is the slope of the line. It is still forced to be the same. We have not changed\_\_\_\_\_ [00:31:44]. But it is the same for men and women. Maybe you are so smart to realize, maybe it is an interaction. Maybe the height matters more for men than it does for women. That is really the interaction term. If you are more of an\_\_\_\_\_ [00:31:57] or a psychologist, you might use the term effect modifier or modifier. Notice that here we have the height and we have the gender main effects. We are still getting that separate Y intercept.

But now we have got this interaction term. It is the X times Z. What is this model? What we are really looking at here is that the slope is now allowed to be different. Now you have the red line, which is for the women. It looks quite flat. You might say and come away saying, well there is not really a relationship between height and income for women. But that it is much more severe for men. You can work through these examples and figure out how is the best way to express the data with your models?

Paul Barnett: Todd, we have a question that I think is –

Todd Wagner: Sure.

Paul Barnett: – Germane to that, which is\_\_\_\_\_ [00:32:46] asked how do we determine is an interactive term is needed?

Todd Wagner: There are many ways to determine if you need to do interaction terms. Perhaps it is theory that tells you. Perhaps you are using a health services research theory that says what really matters here is this specific relationship between height and women. It specifies that interaction term. It has been well published in the literature.

You are just following that model. I would sort of say that is a theoretical argument for an interaction term. You can also use much more of an empirical argument. You can say well, we are trying to model the data and trying to best fit the data. You were to explain these things much more in a exploratory manner. Or, trying to understand does the interaction matter? It depends on your perspective and where you are coming from that determine whether the interaction matters.

What you are going to see in your output of your regression is a P value. Or, you are going to get confidence and\_\_\_\_\_ [00:33:50] standard errors. Eventually, a P value for this interaction term; and just to note that there is often much less power for understanding interaction term in data. If you have a small dataset, you might come away and say no. There is not a gender interaction here. It is because that we have much less power for it.

You have to be a little bit more careful when you are dealing with interaction. That is especially true, if you are dealing with interaction in nonlinear models; which we will get into much later on in the course. When you are dealing with a linear model, you have not reduced power. Then you have different ways of choosing among whether it is theory or empirical resolution for deciding whether you get interaction. Hopefully that answers the question.

Paul Barnett: Then someone asked is an interactive term the same thing as using a stratified analysis?

Todd Wagner: That is different. A stratified analysis would be – and I will show you how they relate. A stratified analysis would be saying let us run the regression on just women. Then, run a separate regression on just men. That is what I think of as a typical stratified analysis. You run separate models for each gender. That is actually a very common way to look for interaction. Because what you are often looking at – and perhaps you are doing a very large dataset.

You are saying well, it is by facility. I am very interested in understanding the facility effect. You say well, let us run a different regression for each facility just to make sure the effects are consistent across each facility. Then you would see, for example, that maybe Palo Alto where I am looks totally different in that stratified analysis. It might give you reason to examine the context of why Palo Alto might be different. You might not have a specific interaction in mind.

You might have to explore different interaction. But you would do a stratified analysis as one way to explore the data and understand it. If you founded your stratified analysis that gender really mattered for this height and income example, you could then go back and build this master datasets or master model that has the interaction. That sort of includes the information that you learned from the stratified analysis. A great question….

Great, we have about 20 minutes left. What I want to do now is a departure from the examples that we have been providing here. There are a number of assumptions built into statistical models; and especially – specifically this classic linear regression model, CLR. I just want to run through these assumptions. These assumptions are often what is going to drive the remaining classes. If you are particularly interested in a particular assumption, you might say that is who I am going to tune in for that class later on. That is how I will think of this. Just keep in mind that in classic linear regression, there is no super estimator. If you wanted to go back and say is ordinary lease squares always the best statistical model to run? There is no always the best statistical model to run. There is no super estimator.

I would say like these linear regression models are often used as a starting point for analyses. There are a number of reasons for that. One is that they are often very quick to run compared to some models like maximum likelihood models where it can take sometimes hours to run the models. It is very under – easy to understand the marginal effect. There is an easy understanding about that beta coefficient to what it means in a linear model that differs in more complex models. You can even run it.

Let us say your dependent variable is\_\_\_\_\_ [00:37:28] and mortality. You can even run linear regression on mortality. That is known as a linear probability model. It is often very easy to understand and run that model. It may not be the preferred model. But eventually, you might want to use logistic regression, which is maximum likelihood. But it is often an easy starting place especially if you want to understand interaction effects. Keep in mind that there is no super estimator. I often use linear regression to start many of my analyses. It makes five really assumptions that I want to walk through.

Then these assumptions in sort of variations and sort of ways that we challenge these assumptions will guide your choice of the estimator. I joked you are at the happiness of your reviewers; but, the likelihood of getting published and so forth. Often, you are trying to push the data here to understand the role of the estimator. Assumption number one is that the dependent variable – you remember if I gave you that example. It was income. It can be calculated as a linear function of a specific set of independent variables plus an error term.

This is often not a challenged assumption. But you are basically saying that there is a functional form on this right-hand side, which is the Xs, and the Zs. The error term that come together in a linear function that equal Y. that is a function one. That is a pretty standard assumption. Violations to assumption one could be things like omitted variables that were missing something critical. If you came away and said that height in your standard assumption is that height predicts income; but you failed to include gender, you would be omitting a variable. That would violate that assumption. Because it is creating problems with that functional fit.

You also could have nonlinearity. You might say well what is really driving this is not a linear model. Maybe it is a log linear. What is really much more important is the log fits. Many would make the assumption that you transform your data. You would say well, we are making the assumption here that it is the log, the assembly of logs and variables in a linear function. There are ways to test assumption number one.

You could have a theory-based transformations. In economics, we often had production theory. We talked about things like Cobb-Douglas. That is a very standard sort of theoretical assumption about how firms or individuals produce goods. You might have empirically based transformation. You might have a common sense. Then there is often these things, empirical tests and things like the Ramsey RESET test. There is also a Pregibon Link test. Then I give you the citations on these.

I will just note that common sense does pretty well in many of these issues. That sometimes these specification tests are – and need big data sets for them to be powerful to work. If you have got a small dataset, maybe a dozen or a couple – even a couple hundred people. You might end up with a negative RESET test, which does not necessarily mean that your model is the right specification. It just might mean low power for that.

Paul Barnett: This is my personal question is this is like –

Todd Wagner: Sure.

Paul Barnett: You are looking at the ends of the distribution to see whether the linearity really holds. In other words, the very short people, are they further away from the line than the very tall people, or the medium people?

Todd Wagner: That is exactly what the Pregibon Link test is doing, exactly. It is actually working on second order in some sense interaction terms. If your residuals to figure out is there anything less in your residual that you did not account for that you could have accounted for? That is exactly right. Thank you, Paul. I just want to take a slight detour here and talk a little bit about assumption one and stepwise regression.

I assume some people are familiar with stepwise regression. This is a statistical software technical that allows people to build a regression model in a stepwise fashion. You might say I am interested in the ten covariates. I am going to let this regression model go through and only include ones that perhaps are significant at a certain P level. People often choose, for example, 0.2. I say we will include those, if they are a significant less than 0.2 in that regression model. I would say that stepwise regression; and this might be a difference between epidemiologists and economists.

I am often very worried about omitted variables. If I have a, especially if I have a theory that says these variables should be in the model, I will force them in the model. I almost never use stepwise regression. There is typically very little penalty for adding what I think of as a nuisance variable. A variable that is not significant. But there is a big penalty for missing an important covariate. That results in bias and everything else in your model. Just be careful, if you are thinking we will just run this regression and let the statistical software sort of create the model for us.

There is a bias. I show that here, if we ignore gender. The red line is the estimate, if we did not include gender. You would say well. You would say here is your intercept. Here is your relationship. You will notice both there is an error in where it crosses the Y axis. There is an error in the slope. The omitted variables are really problematic.

Assumption 2, the expected value of the error term. Again, we are back to this issue of error term. It is an important issue. The expected value of that error term is zero. If we violate this error term, we have a bias intercept. Now this is typically a challenge when we are analyzing very skewed distribution as often is the case with cost data. People out there may have heard this term, a smearing estimator. There is a – when you were dealing with cost data and you've got this very long right-hand tail. Because you have got some people in your model or your data that have millions of dollars of expense where most people have a couple of hundred or a couple of thousand. It creates problems for the data expectation of our error term. It is not zero. Paul, in his last two classes of this set – of this course will be talking a lot about different ways to model cost data looking at assumptions.

Assumption 3 is that there is this assumption that each person or case in your dataset is independent and identically distributed with error terms. Now there can be ways that you could obviously violate that. There are some very simple ways to think about it. Let us say we do a survey. Someone answers the survey multiple times. Or a number of people answer the survey multiple times. That obviously means that those are not independent. There is a correlation there. Or, if you followed people over time, you might say well we have to account for the fact that there is a correlation within those people over time.

The other way is that you end up with this issue of homoskedasticity and that the error are identically distributed. This issue of homoskedasticity, its counter part, which is heteroskedasticity is going to come up a lot. I am going to give you an example of that. Here is an example of homoskedasticity. What we are interested in this – or what we have depicted in this relationship here is length of stay and cost. You will clearly see that there is a relationship, a correlation between the longer you stay in the hospital and the cost of that stay in the hospital.

What you are also seeing is homoskedasticity. If you come out here and take a snapshot at 250-bed length of stay. You take a one here. You will notice that this is a very tight distribution at 50 compared to a very wide distribution at 250. This is a very classic form of homoskedasticity. This can create bias in our estimate. We have to take this into account.

Paul Barnett: This is real data?

Todd Wagner: This is\_\_\_\_\_ [00:45:54]. Yes, this is actually. I do not want to say if it is real data. But yes, it is data.

Paul Barnett: Okay.

Todd Wagner: Yeah. It is real data.

Paul Barnett: It looks plausible.

Todd Wagner: It is. It is actually, it is valid VA data. No, PHI in there though; I have to be careful about that.

Paul Barnett: No. unlike the income and heights data, you did not make it up.

Todd Wagner: Fair enough, thank you, yes. Violating assumption 3, so the coefficients are not necessarily biased, but it is inefficient. There are bias standard errors. We have to be worried about the standard errors when it comes to the violation of assumption 3. Then like what we just showed, plotting is\_\_\_\_\_ [00:46:36] a very helpful way for diagnosing heteroskedasticity. There are different tests for heteroskedasticity. The more specific the test is – and this is sort of an axiom in life. The more precise the test, the greater the power. If you are interested in just sort of a general test of heteroskedasticity, there are tests like that. There is a White test, for example. But it is actually slow in power.

You could say I am very interested. Is it the heteroskedasticity in this particular variable? That is called a group wide heteroskedasticity. It has more power for that than the White test. But even still you might come up with a negative significance on that test and plot it out. Say well, it clearly seems to be violating it. But plotting is still one of the ways to go. Fixes for assumption 3; there are different fixes for it. You will see these come up time and time again; when we deal with cost data. But there might be other times to deal with it. Transforming the dependent variable may eliminate it. If we are worried about this heteroskedasticity, maybe in the log form it goes away. There might be other ways of dealing with it.

You can also use what is known as robust standard errors. They come under different names; Huber White or sandwich estimators is one way of also dealing and correcting your standard errors. There is a literature that has been coming out in the past few years with Gary King and some others that has been talking about are we –? I think the title of that paper, are we whitewashing the data? Are we too quickly falling back on this idea of just using robust standard errors instead of really trying to understand the dependent variable and why we might have heteroskedasticity? There has been a push back on just sort of saying well, let us just use robust standard errors. If you are running Stata for example, you can easily include robust standard errors with a comma R. but what they would like you to say, Gary King and his colleagues is think about it carefully before you just rely on that technique.

Assumption 4, the observations are on independent variables are considered fixed and repeated samples. Here is the mathematical expectation between the X, I and U, I; and given X equals zero. There are many ways that you can violate this assumption. It comes under this broad category of endogeneity. You can errors in variables and how you are measuring your variables. There can be consist bias in how you are measuring your variables. Unless you control for that sort of error in how you are measuring your variables, you can get into this problem. You can get into an autoregression, simultaneity; but think of it as if you are\_\_\_\_\_ [00:49:16] – you are thinking more healthcare selection bias.

Selection bias is almost going to lead to some sort of endogeneity. This sort of idea of let us think about smoking. I ought to come back to smoking in this example. If you are interested in the link between smoking and cancer. You were 40 years ago studying that link, it would not be so obvious to them then that smoking causes cancer. There has never been a randomized trial proving that smoking causes cancer. The tobacco companies will tell you that. But we all know now that smoking causes cancer. We have it through biological plausibility. But we also have it through many different types of observational data.

But there is a problem when you do a model that has smoking on the right-hand side as a covariate and mortality on the right-hand side. There are a lot of things that create people's choice of whether that to smoke. It could be that their parents smoked. It could be that they had friends in healthcare who smoked. We do not observe all of those things. It can create a lot of problems in how we model the data and violate assumption 4.

Paul Barnett: I guess, Todd, the other – when you were talking about selection bias. The classic thing is the sickest people get the medicine. You might think that the medicine causes sickness.

Todd Wagner: A very good idea, yeah, a great example. Just to be more specific about errors and variables. We often have measurement error in how we measure the dependent variable. That is\_\_\_\_\_ [00:50:48] measurement errors maintained in the error term. We assume that it is the mean expectation that is zero. Well, let us assume that covariates are measured without error. But often the covariate also had error. If that error is unobserved and in certain directions, it can be problematic in how we are creating this model.

That error term in our statistical model does not account for the measurement error of our covariates just be clear. Here is our common violations of this assumption. One is, for example, lagged dependent variables of the covariate. You have a model that has a time. You said we are interested in someone's behavior in this time period. We include their period in last time period and the lagged dependent variable.

That creates a problem in our model. There are different ways to assign and assess this. There is the Hausman test, but a very weak power and in small samples. We are going to come back to this issue of instrumental variables as a potential solution for this violation. There is a whole class that Christine Chee is giving on instrumental variables as a solution for this error.

Assumption 5, and many people see this as a trivial assumption. It is that you have to have more observations that you have covariates in your dataset. There is no perfect multicollinearity. This is often seen as trivial. Because when we are working with observational data, typically at the VA at least, we are working with thousands, if not millions of records. It is hard to imagine having millions of covariates or having more covariates than an observation. But just and keep in mind is that you cannot have more covariates in your model than you have observation.

You can also not include two variables that measure the exact same thing. That is called perfect collinearity. Now if I put in a covariate in my model where it is income and gender. I said wanted to include a variable for males. I also wanted to include a separate variables for females. It would kick you out. You cannot do that. That is multicollinearity. We are already comparing men to women with the male variable. The women's variable is perfectly the same thing and just measured in the opposite direction. You cannot have both of them measured at the same time. The easiest way to solve is to not include both men and women and as separate variables; but just have the male variable relative to the women's variable. That fixes that.

The other thing is that perhaps there are times when you might need to increase your sample size. I have seen instances where people pulled data across years. Let us say you are interested in a rare disease like ALS. This is a neurologic disease. It is rare. It does not – you do not see a lot patients with this disease in any one particular year. You might pull data across many years to get a bigger sample so you could look at your associations. At this point, that is all for the class. We have about four minutes. I apologize for running out of time. But I would like to handle the questions two ways. One is just to keep dropping in your questions. Paul will read them to me. I know we are going to run out of time there. If there are other questions that come up, probably the easiest way to get back to me is Todd dot Wagner at va dot gov. You can e-mail me as well.

Paul Barnett: There are not too many questions. One is please explain further why including a lag dependent variable or variables as a covariate, it violates assumptions?

Todd Wagner: That is a great question. I was going to…. Let me give you a citation. The statistical software, SAS and Stata. There is a much more ground with SQL, but references. Kennedy, "A Guide to Econometrics," it talks about the problems with lagged dependent variables. Let me just give you the intuition. Then I will let you look it up as why it is a problem. If I am interested in understanding people's behavior; and maybe I have understood it – and understanding why there are smoking today. I include a variable that is why they smoked last year. You are going to see a huge correlation in part because there is an addiction between smoking last year and smoking today.

That lagged dependent variable is going to buy us the other terms of your regression estimator. It is going to create a bunch of problems for having that variable in there. One is it is, that past behavior is so strongly correlated, it is sort of absorbing the correlation of a lot of other variables. But it also sort of building in this dependency into your regression model. Paul, do you have any other ways that you would think about lagged dependent variables and the challenge to\_\_\_\_\_ [00:55:32]?

Paul Barnett: Well, I think that is what you said is just right. We do have another question, which is so is model calibration about how well the model fits the data? Is this different from the model C statistic?

Todd Wagner: There are different ways that people can fit the data. I am not going to spend time talking about different statistics. But to note that there are different criterion. There is the Hosmer-Lemeshow tests. There are\_\_\_\_\_ [00:56:02] statistics. There are many ways that people could think about fitting the data and choosing the right model. It can be a little bit overwhelming at times.

If you have a specific question about the\_\_\_\_\_ [00:56:15] statistic, I am happy to talk to you more offline about that\_\_\_\_\_ [00:56:19] statistic. But just to say that there are many different criteria. You can say mean, standard; mean absolute error is one possible fit criteria. You could say R-squared would be a fit criterion for the OLS. You could say there is a Hosmer-Lemeshow test that one can create whether it is a dichotomous independent – or dependent variable, or even with cost saving you could it. There are many different fit criterion.

Paul Barnett: Yes. Someone asked, so it observe versus expected value? I do not think that is quite captured yet. But you had a slide right at the start, I thought Todd of, showing what criteria people\_\_\_\_\_ [00:57:01].

Todd Wagner: Yes.

Paul Barnett: I might remember it wrong.

Todd Wagner: Yes. There was a slide early in the start about different….

Paul Barnett: About some of the criteria that people use. But it is right that a good model will use the, but on the basis of the covariates and the function that the model has, be able to estimate what the dependent variable is. That is true.

Todd Wagner: There has also been a large literature that talks about what to do when theory and empirics diverge? Let me be specific about that. You might have a model that says that what we are really interested in here is the production of healthcare. It is really the theory says maybe it has got to be a Cobb–Douglas functional form. But your empirics suggest that does not fit the data very well. Then what to do in that case. There are people, some people who believe that the theory trumps the appearance. There are other people that say maybe the empirics trumps the theory. There is not often a right answer for that kind of challenge.

Paul Barnett: Then someone asked if you would review the difference between bias, confounding, and endogeneity? Are they all the same things? Are they each a little different; so bias, confounding, and endogeneity?

Todd Wagner: A great question – so, bias, let us just think about it at the highest level. Bias, it just means that you come up with this estimate of an effect. Or, maybe it is to go back to our income and height issue. You do not put the gender in there. You come up with an estimate of the relationship between height and income. You know that is biased. That is an incorrect and inaccurate estimate of the true relationship between height and income. There can be many reasons why that, we could have that bias.

One of the reasons that we could have bias in there is perhaps it is a sample selection. Maybe you have got a really crazy sample selection in your – and you went out and sampled people in your data. You did that. Another reason, it could be endogeneity is one reason for potential bias. Omitted variables in this case is a challenge. Because you did not fully control for gender. That is what is causing this bias. There are many ways that you could think about why you are getting the bias. Endogeneity can cause one of that. There was a third part of that question, Paul. Bias, endogeneity, and…?\_\_\_\_\_ [00:59:31] Paul?

Paul Barnett: Confounding.

Todd Wagner: Yes. Confounding is often just used as a different term for thinking about what is creating challenges in understanding the relationships between variables. In that case, you would say what is confounding this relationship between age and income? You might say well, we need to understand gender as a confounder. You are right. There are different fields. I will point you back to that Matt Maciejewski paper who does a great job walking through the different terms and how they might relate to different fields. But then you feel a little bit more confident when you are talking to different people about what do we mean when we say X or Y? What does endogeneity mean? How does it relate to self-selection? How does it relate to confounding and effect modification, and so forth? A great question… I know that we are getting close to the end of our time, if not over. Are there any more questions?

Paul Barnett: Yeah. I think we have answered all of the questions. Someone else, someone asked the question of including last year's smoking as a variable in a way to account for its effect in a linear model. I think that we are not quite clear about what the dependent and the independent variable is.

I was typing an answer, but maybe I will just say it; which is if we had our income model where height is predicting income, we would be trying to use height to predict income while also using last year's income to predict income. Obviously, the height is going to make very small contribution in that case. Because we have got a controlling for last year's income; which also\_\_\_\_\_ [01:01:13] potentially reflects height.

Todd Wagner: Yes. That is a great example of how then that can create bias for your estimates.

Paul Barnett: I hope that is helpful to the question in that example. I was not quite sure what they meant with their example from smoking. I do not have any others. I think we should announce perhaps in the next –

Todd Wagner: Class?

Paul Barnett: Class, yes.

Todd Wagner: Do you want me to\_\_\_\_\_ [01:01:37] and go back up the…? I do not know how fast I can get up to the top of the…. Because I have a slide that all of the class.

Paul Barnett: Yes.

Todd Wagner: Here we go. The next one is in two weeks. It is research design and Christine Chee who is currently on vacation in Europe. We will be talking about research and design. Then we will come back with a propensity score on April 1st, April Fools.

Heidi: We will be sending registration information out for Christine's session out one week for today. If you are not already registered, just keep an eye on your e-mail.

Todd Wagner: That was awesome, Heidi. Thank you so much for all your help this morning. Thank you, Paul again for being back up for me.

Paul Barnett: Well, lots of good questions. It looks like people were engaged. It is a very clear lecture, Todd.

Todd Wagner: Thank you.

Heidi: For the audience, if you all could just hold on for a just another minute. When I close this session out, you will be prompted for a feedback form. We really do read through all of your feedback. If you could take a few moments to fill that out, we would appreciate it. Thank you everyone for joining us for today's HSR&D Cyberseminar. We look forward to seeing you at Christine's Research Design session. Thank you.

[END OF TAPE]