Jill: Great. Thanks, Maria. So, thanks, everybody for joining today's session on Limited Dependent Variables. Today, we're very excited to have Dr. Ciaran Phibbs presenting. Ciaran is a Senior Health Economist at the Health Economics Resource Center; he's Associate Director of the Women's Health Evaluation Initiative; and also Associate Director of the Geriatrics and Extended Care Data and Analysis Center. Ciaran is an Associate Professor of Research in the Department of Pediatrics Neonatology at Stanford University School of Medicine. And without further delay, I’m going to hand it over to Ciaran.

Ciaran Phibbs: Alright. Thank you, Jill. And so, the objective is this--this is a broad topic and the objective is to give a sort of a general overview, plus go into a few useful details for some of the topics related to limited dependent variables.

The standard MPH regression course focuses mostly on the ordinary least squares or linear regression model and limited dependent variables don't get much, except that most of them will expose you to a logit regression. This is a sort of an overview of a lot of the issues, there's a couple of things I’ll go into in a little more depth; but it is not an in-depth treatment, it's more to point you to the things that you would need to know--or should be aware of as you deal with dependent variables that are not big continuous variables.

So, by limited dependent variables, I mean one of two things: one is that you have a 0 or 1 choice, or a small number of options, or a small number of counts, but the dependent variable is not continuous or even close to being continuous and doesn't have the traditional normal distribution that we associate with standard regression analysis. Also, I’m going to make a note here--administrative note, forgot to make it previously--Jill will be monitoring the questions as they come in and she will interrupt me if they're clarifying questions that, in her judgment, I should address on the fly; and otherwise, we'll defer questions until the end.

So, in terms of types of dependent variables, the things that we can think of are binary choices: a yes, no, some sort of other things where there are just two options; a multinomial choice is a choice where there's more than one option, these can be either ordered or unordered; in other words, the order may matter or it may not matter; you also have counts where you have integer counts where they're small numbers, the number of primary care visits that a patient has in a year--most patients have zero, one, or two, some have more than that, but they are they're limited. Most models have a general framework of probability of models in terms of the probability that an event occurs, but that's not always the case, counts are not a probability.

We'll start with binary options in outcomes and this is very common in health care: Does the patient live or die; does the patient have an infection; did a patient safety event happen; was there re-hospitalization in 30 days; do the patients decide to seek medical care or not? There's a host of binary options and we encounter this all the time in healthcare. And in terms of a general framework, if you look at the classic regression model, Yi is β where your intercept and your βX is a matrix of variables in your error term, but instead of having the dependent variable be some sort of a continuous variable like income--that's a classic one for economists--it's a 0 or1; 0 if you lived and 1 if you died, and you could reverse that.

So, the regression equation becomes the probability that Y = 1 is equal to a function of the X's and the βs and the probability that Y = 0 is 1 minus that function. If you were to estimate an ordinary least squares model, which is also referred to as a linear probability model, your error terms are heteroscedastic and they can because they depend on βX; and your predictions are not constrained to match the actual outcome which is either 0 or 1, and an OLS model will give you a range of predicted values where some have a probability of 0.1, 0.7, whatever and you can also get predicted outcomes outside of the range; within an OLS model, you can get negative values and you can get values that are greater than 1.

So, the standard approach to this is, especially within healthcare, is logistic regression, and this just writes out the term here, it's basically just reflecting on a logistic distribution where the probability of these ones is the exponent of βX divided one by one plus the exponent of βX, that's just technical. And the advantages of a logit regression and things you need to remember, is that they are designed for relatively--the logistic distribution is based on an assumption of relatively rare events and it doesn't work as well for more common events; and one advantage is it's commonly used in healthcare and most readers of health care journals are familiar with the concept of odds ratios.

There are other approaches; probably the other most prevalent approach is a Poisson. This came out of the economics literature in the classic example was the decision to make a large purchase like to buy a car, yes or no? There are other methods using other distributions; in general, they tend to give about the same answer. One thing is it used to be a lot easier to calculate the marginal effects; with probit that's not so, there's a marginal effects command and Stata that gives you the marginal effects so you're fine.

And the reason that you want--and I’m assuming most of the readers are familiar with logistic regression, but there's some things that they probably may or may not be aware of. One is the odds ratio--and when we call the odds ratios is really a ratio of log odds and relative risks, which is what we're really interested in in terms of the standard method of interpreting logistic regression is that when you get an odds ratio--if you have an odd ratio of, say, 1.4, that represents a 40 percent increase in the risk where you're really talking about that's the relative risk. But the thing is that with a logistic regression, that is only an approximation and it starts to break down at about a 5 to 10 percent occurrence of the incident. So, whatever your dependent variable is, say it's mortality, if you have a population with a high mortality rate, those odds ratios are going to break down and that can apply to any of the outcomes.

And just to show, someone has graphed out the relationship between the relative risk and the odds ratio... So, basically, what this slide shows is what the odds ratio is and then the incidence ratio--they label this the incident ratio and the non-exposed percent--so, think of this as the percentage of the population that has your dependent variable, if you're dealing with mortality or whatever it is. And what happens is you can see, at very low rates--which is what the logistic regression is designed for--the relative risks are essentially the same as the odds ratio; and as the incident rate goes up, as you move to the right on the graph, those start to diverge away such that at a 20 percent incident rate, you're starting to get--and for the higher odds ratios, those numbers are quite different; at a 20 percent incidence rate, a relative risk of three is reflected by an odds ratio of five, if you just sort of project over to the horizontal axis. And so, the numbers start to break down and you see this, that people will run a logistic regression where the dependent variable is fairly common--20, 30 percent--and they're still interpreting those odds ratios as risk ratios, and that is not correct because they'll be biased.

And there was this paper that appeared in JAMA a long time ago, that gives a correction formula--I’ve listed the reference here--and the risk ratio is you apply this formula to adjust the odds ratio where P0 in this formula is the probability of the outcome, so your mortality rate or whatever.

And remember that this is going to be affected by the incident rate in the population; and just to give you an example of this--this is from a paper of mine from several years ago now--where I had a sample that was big, about 48,000, but the mortality rate in the sample was 20 percent. And so, you can see here how the odds ratios changed and that bottom odds ratio with 1.08 only changed to a 1.06, it was a pretty good approximation. But that top odds ratio where a was the odds ratio was 2.72 whereas the calculated relative risk was only 2.08, so that's a fairly significant reduction in the relative risk compared to the odds ratio, and it was because of the fact that there was a relatively high incidence rate of the dependent variable. And so, this is in terms of the magnitudes--a real example of how the magnitudes of those change for odds ratios for the relative risk.

This Zhang approximation is an approximation; it's not exact, but it's pretty close.

I will note that more recently, many journals, especially epidemiology journals, are now wanting to directly a direct estimate of the of the relative risk; and one option to do this is with a Poisson regression with a robust air variance; and if you do this for a binary outcome, the IIR or incident rate ratio of a Poisson is actually the relative risk. You can run this in most statistical packages, Stata has a Poisson command; you can do it in SAS within Proc GENMOD, so, it's relatively easy to do and depending on your journal, the editors may want you to do this if you were initially-reported logistic.

Jill: Ciaran, we have one clarifying question. "Can you correct confidence intervals with this conversion?"

Ciaran Phibbs: The Zhang conversion does not correct the confidence intervals--well, they wouldn't have to go back and look. You probably can, but what you can do in terms is talk about the relative risk; but as I said, it's now relatively easy to estimate these directly with Poisson. I mean the Zhang correction really, you can use that as a quick check to say, "Look, is this really going to matter in my sample?" To go back up here, if all your odds ratios are like the ones down at the bottom where they're not changing by very much or if you're estimating this on a sample with a 2 percent mortality rate instead of 20 percent mortality rate, the change between the odds ratio and the relative risks is going to be trivial, so you really don't need to worry--you can argue that you don't need to worry about it, but you can use that same formula as a quick check. I’d have to go back and look at the papers to see whether they have them for adjusting the confidence intervals, I just don't remember that off the top of my head.

Jill: Thanks, Ciaran.

Ciaran Phibbs: But the other thing is it's probably better to do the--since it's now easy, you may want to consider just directly estimating your risk ratios. And so, in terms of something with lower events, so just in terms of this--and this is for a paper where we initially estimated it with logistic and the editors--this is an epidemiology journal, they wanted us to directly estimate the risk ratio, so we had estimated it both ways--and I’m just sort of reporting the differences looking at the effect of active PTSD and when pregnant women veterans on gestational diabetes and pre-eclampsia. And the rates--we had about 20,000 observations and the rates were 4.5 to 5 percent; and you can see the odds ratios and the relative risks were almost the same, there was a slight change because we also--I didn't specifically re-estimate it for this because we had a slight change in the sample, but they're essentially the same, the confidence intervals are similar. And so, just to point out, as long as you're not in extreme cases, your results are going to be essentially the same.

In terms of binary data, I want to note that there are all kinds of variations; you can estimate this for panel data--where panel data is when you have cross-sections over time, so you're following a group of hospitals over many years or whatever. You can estimate them for group data; you can estimate them with both random and fixed effects models. This was a few years ago, so it may be a little bit dated because there could be more, but I went through this Stata manual and counted up that there were over 30 estimation commands for different types of models for binary outcomes. And so, just letting you know that, basically, in this day and age, there are [ten] programs that will address almost all of the types of models one would need to estimate with binary data.

One thing to note in terms of logistic is the goodness of fits tests for a logistic--and these apply to a lot of other types of models--the things that are commonly reported are the area under the Receiver Operator Curve or ROC curve, which is the C statistic in your STAS output, it's specifically reported in most regression analysis. And these range from--an ROC will range from 0.5 to 1, and intuitively, it is comparing how well your model predicts compared to random assignment. So, if you're just to flip a coin, you're going to be right half the time which is why that's the random assignment where you're right half the time, and this is how much better you do.

The other common statistic that most many of you may be familiar with is the Hosmer-Lemeshow Goodness of Fit, and what the Hosmer-Lemeshow does is it breaks the sample up into N--usually ten--some programs will let you vary these groups--equal groups in terms of number of observations and compares the number of observed to expected events in each group. And there's a formal test on this to see how well your model is predicting.

One thing that does happen in healthcare is that if your model predicts really well, the events are going to be concentrated in the highest risk groups. And so, if you're breaking your sample into ten groups and you're trying to take mortality risk, if you have a really good model, almost all the deaths are going to be in that 10th decile; and so, you may want to know a little bit more about how your model predicts. And with a statistician at Stanford, we played around with this and we never published it because there's underlying issues with our a priori selecting in terms of the actual theoretical statistics working out, but this can be a useful tool; and that is if you want to understand how well your model is predicting in the tail, how well you discriminate your mortality prediction where most of the mortality is in the top decile because you have a good model, is that you can--to see that, you can sort of flip this around and it's not a formal test but it's informative for you in terms of trying to understand how well your model is predicting. And that is to divide the sample so that the events are split into equal groups instead of the number of cases; and it's not a formal test as I noted, and I just made this hypothetical example of 100--if we had 100 observations and 100 events--so, deaths--and the Hosmer-Lemeshow test is going to--sort of what would happen in the Hosmer-Lemeshow test--and this is just an arbitrary breakdown, nothing specific, but it makes the point--there'd be a hundred observations in each cell; and as you can see, the deaths are all concentrated in the last two deciles, there's some sort of randomly distributed than the others.

And if you were to, instead, if you were to instead break it so that you had 10 deaths in each cell, you can get this kind of a breakdown like this, and this is an informal test, but one that--so, you can see in this case, how well and you can look and see, you observe or predicted, and how well your data is doing, and it can be informative. Now, this is one that performs very well, there's others that won't perform as well, but it gives you an idea of how your deaths are being distributed, or your events are being distributed, and can be informative for helping you understand how well your model is predicting.

I want to make a note about estimating with really large samples. We have 6 million, 7 million patients a year in the VA; you look at Medicare data, you have even larger samples. And when you have very large samples--all these models are maximum likelihood models, so because maximum likelihood is an iterative procedure to zero in on an estimate, they take a lot more computer resources and time to estimate than your regular ordinary least squares model. And a trick that you should remember is because the X matrix is the same, just comparing OLS to a logistic model for example, the P values for OLS are going to be approximately the same as what you would get in logistic or whatever is the correctly specified model; and because the right-hand side variables are identical--and that matrix is identical--in terms of doing things like testing for collinearity within your X matrix or just doing some general model development in terms of which variables are important, you can do some of your initial runs in OLS and get your results a whole lot faster--these are not going to be your final results you're doing it, but when you're just trying to understand your data and understand the relationships, running these things with OLS can save you a lot of time in terms of these initial exploratory things; and then when you get down to your final modeling, you can switch over to the maximum likelihood routines and this is a trick that can save you a lot of time because some of these maximum likelihood models can take a long time to run; and if you have very big samples, the VINCI data managers may get mad at you for hugging resources and so you can minimize when you're actually doing that.

Multinomial or discrete choice is the next thing I want to talk about. And what if there's more than one choice? When you have this outcome, your options are more limited; there's multivariate probit models where you can have multiple decisions each with two alternatives; and then there are two different logit models for a single decision with multiple alternatives; and you can also have an ordered logit model, so if your choices are sort of ranked, sort of 1, 2, 3 in terms of an ordered choice, that if you choose--in terms of the progression, there's models for that. I just want to talk a little bit about this.

There are a fair number of examples of this in healthcare. Which hospital do you choose among the many that you could go to or do you choose the VA where you also have several other options? And there's different travel distance to each of those options and characteristics of those options for the VA versus the non-VA options may vary in multiple dimensions including copays and perceived quality, et cetera. And, for many conditions, it's not simple; we can treat you with A or B; there may be several different treatment options and you want to and want to look at how those are choiced; and again, you have to deal with ordered versus unordered choices.

In terms of the logit models for multiple choices, there are two that are commonly out there: there's the Conditional logit model which was developed by Dan McFadden, who won a Nobel prize, and this is for unordered choices, the classic example there, and this was developed for--you have three or four different options for how you get from your home to your work, how do you choose and the characteristics of the choices where the common use in healthcare is that example I’ve referred to above, where you have a choice of many providers which do you choose, and there's different characteristics of choices.

And then there's the Multinomial logit model where you can also order the choices. I’ll give you an ordered choice in healthcare would be, in terms of emergency treatment, do you go to the nearest hospital--to the nearest hospital with a Level 1 or a Level 2 trauma center, and those are three different and there's ordered in terms of sort of just a lower-level trauma center, higher-level trauma center, so that's an ordered choice. And in the multinomial choice, you can reflect those ordered choices.

For the Conditional logit model one of the advantages in terms of application of this, is that we when we are writing our grant proposal, we have to have a conceptual model and this is--well, this is actually based on what is consumer theory and it's a random utility model, it's how individuals choose from a set of options. So, I’ve actually used this as the conceptual model in a successful grant proposal before as an IIR, so you can use this as your conceptual model if you're examining this type of thing. And the original model--the original McFadden model--only included characteristics of choices with no individual characteristics, and you can express the results as odds ratios. There are ways to include individual characteristics, you can't directly include them because the way the model is formulated, they cancel out. But you can interact things--I’ve included individual characteristics by interacting them with distance to the hospital in looking at a hospital choice model.

In Stata, there's a way you can directly--and with C logit, you can include individuals and choice level variables--by choice level I mean that would be the hospital characteristics.

Some notes on estimating this model. Unless they've changed it recently in SAS, you have to have the number of choices be equal across all observations; in the example I’m talking about with choosing among several hospitals depending on where you are, your number of choices may vary, both LIMDEP and Stata will let you vary the number of choices in a given observation which reflects reality, so unless SAS has added that feature recently, you would want to use another package.

With the multinomial choice, essentially what you do is you identify a reference choice and each model yields separate parameter estimates for each of the other choices relative to the reference choice. So, to continue the example of that trauma center, you might make the highest-level trauma center your reference choice and then you would get estimates for the lower-level trauma center versus just a community hospital, and you would get two sets of estimates for the relative difference of those. One of the advantages of the multinomial logit model is that the way it is defined, you can directly include individual characteristics, and so the model can include--the model would also include parameters for the choice characteristics to continue this hospital example, is the hospital a public hospital, or a private hospital, or is it a teaching hospital, or whatever characteristics, quality metrics of the hospital, there's various things that you probably should be including in the model.

The multinomial model can be estimated for either ordered or unordered choices, and in Stata, mlogit for up to 50 choices; if you have a big data set, that'll take a long time to estimate. And just to give you a concrete example with some numbers, something different, but something I’ve done is NICUs have--like trauma center NICUs, which are Newborn Intensive Care Units--have formal levels of care that define the patients they can treat; there's actually like a four or five graded level; I simplified it for this application into three levels, sort of low, medium, and high, and we were interested, because a lot of community hospitals are opening these mid-level NICUs, and we suspected that they were--that the patients that they were gaining were patients being shifted from the higher-level hospitals which had better outcomes, and that's what we were interested in showing. \_\_\_\_\_ [00:32:50] was a fellow of mine and we published this several years ago.

But just to give you an idea, we're looking at what happened when a mid-level hospital became a new choice and where are the births coming from? And you can see that they were shifting--for very preterm births--that they were shifting--oh, they were coming from both lower-level and higher-level hospitals, but because many of these, when we actually then did this next step of, "Okay, so what does this mean?" So, we then took the model and looked at predicting--with the model, predicting where the change is coming from?

And you can see that when a mid-level hospital opened in terms of the aggregate shares, there was a 17 percent in patients going to a mid-level NICU, but almost all of those were coming from a higher-level NICU and very few of them were coming from a lower level NICU, and you could argue that shifting patients from a lower-level to a mid-level NICU because they have better capacity, you'll get lower mortality; but the mortality is even lower in the higher-level NICU and that's where all these patients were coming, so this is sort of demonstration that this was actually causing patient harm as the conceptual. But the point is you can take the estimates from the top part of your parameter estimates, and you can then multiply those on through in terms of what is the impact of them.

One thing about these choice models is that both types of choice models have an assumption called the Independence of Irrelevant Alternatives, which is essentially that the results should be robust to varying the number of alternative choices, the idea that if you take one of the choices away, the results should be relatively robust. You can re-estimate the model after deleting some of these choices and see how much the results change.

McFadden came up with a formal regression-based test for this--I’ve listed the reference here--and essentially, this is a process of going through, estimating the model, deleting one choice, and doing that for several iterations and you get a formal test here; and if the model fails this independence of irrelevant alternatives, it may be that you need to think about estimating a nested model because it may be that there's some order in those choices. And one nice thing about the McFadden test is it can also be used to test for omitted variables; I will note I have estimated these choice models several times and the models appear to be quite robust, and I’ve always been estimating these in terms of choice of a hospital or choice of using VA versus non-VA providers. And in this case, the distance you have to travel is such an important predictor that excluding some of the options really doesn't change things very much; I will also note--because I’ve had reviewers complain about this, and you can say, "Oh, well, we did this test and we passed this test," and so that makes the statisticians happy--I will note that this is a relatively weak test, so it's fairly easy to pass, but what it's essentially testing is how you have gross assumptions of it, not finer gradation assumptions.

Next, I want to talk about count data or integers. This is a continuation of the same problem and the dependent variable can only assume specific variable values, and it can't be less than zero. So, when you have counts--number of visits you have, or number of medications you have, or whatever it is--where the dependent variable is a count and can only be integer values, the problem of using a regular or OLS progression on this diminishes as the counts increase. The general rule of thumb is that you need to use count data models for counts under 30, but this is not a hard-and-fast rule because--let me just give you an example of counts of the number of primary care visits that a patient has in a year--and for the vast majority of patients, those are going to be very small numbers--1, 2, 3--but there are going to be some patients that have more than 30 encounters, patients that have particularly complex and need a lot of care--and say "Oh, I’m over 30, I don't need to use this." Well, if the vast majority of your observations are still down under the 30, you need to use this model, and you especially need to use this model if you have lots of zeros.

Some examples where calculated models are needed are healthcare, mentioned the number of outpatient visits, the number of times a prescription is filled for a chronic--a prescription for a disease medication is refilled in a year; the number of adverse unit events that occur in a unit or a hospital over a period of time--those are fortunately pretty rare for most of these adverse events, so you're not going to have very many, so you need to use count data model.

The classic count data model is a Poisson distribution; it's a distribution for counts, it's specifically designed for counts, so it has that assumption. The Poisson distribution can deal with, "No, you're not going to have negative numbers and it's only going to be integers." The problem with the Poisson is a very restrictive assumption that the mean and the variance are equal, and that is many times not true. In general, in healthcare applications, a negative binomial is a better choice and there's an option for that in Stata--and actually, within Stata, part of the package for nb reg is to test if that overdispersion; for Poisson, there are other distributions that can also be used.

When you interpret a negative binomial or a Poisson model--the model is the form of the log of the event rate is equal to β x, so you get an incident rate ratio by explanation the β similar to what you do for an odds ratio for the logistic, and that's the rate at which event the event occurs. You can multiply the incident rate ratio by the exposure to get the expected number of events. And so, in general, you can think of the interpretation somewhat similar to an odds ratio?

Just in terms of count data models--I've alluded to this before--but you frequently see OLS models used or some regularity for counts, and it's for binary or very limited choices. And one of the real problems, when you're dealing with counts, is when you run OLS and there are lots of zeros, it really messes up the regression and it will reduce--it will actually reduce the statistical significance or increase the parameter the standard errors of your parameter estimates, and you can actually get results that are of the wrong side. It's theoretically possible. And so, the bottom line is that this can have real--when you're using a misspecified model or the wrong model for count data, you can get results that have real and large effects on both your standard errors and your point estimates.

As with logit, there are lots of extensions for this: Stata has a whole bunch of exceptions--you can do it for panel data; and not only for respect to distributions of the data; and again, all of these different--and this applies not only to negative--the nb reg is the negative binomial reg notation in Stata and there's a--GNB Reg is a generalized negative binomial, and which actually has allows you to do stuff with respect to the dispersion of the data. And the same thing for Poisson, there's different panel models and so on--fixed effects, random effects, all these types of things are available so that all of these tools that you may or may not be using in terms of different types of models can be extended to this limited two-count data.

One final note, when you're looking at this distribution, is to think about the data generating process. And that is it may be that you want to run--the data generating process may be mixed and you may need to split your sample. And I’m going to go back--because it's a particularly good example, I’m going to go back to a newborn example, that if you look at the lengths of stay that a newborns stay in the hospital after they're born--well, over 90 percent of newborns are normal, well babies and they're home--basically depending on type of delivery--in two to four days, and that's clearly a count there may be some variants. You also have some infants that have a minor problem, and instead of staying three or four days, they stay five, six, seven days. And then you have the extremely preterm infants--the infants that are born two and three months earlier are going to have a two to three-month hospitalization, so this very long length of stay.

And so if you try--and this is just sort of because it's essentially a bifurcation of the vast majority are have no problem, they have one process for generating the length of stay; and then those who are sick, who have a very different process for generating length of stay; and if you try to model these jointly, you're going to have real problems, whereas if you split the sample, so you split out the well babies and then you take the sick babies as a second sample--and it's just an example that sort of the data generating process for that length of stay is conditional on, "Is the baby sick or premature?" And this is just an example where you're better off trying to estimate two different models because it's really two different data-generating processes that have very different effects.

And that's just a point of when you're considering your dependent variable, think about it in terms of the data generating process, not just, "Do I have simple counts or a 0-1?", because if there's a different underlying data-generating process for different parts of your sample, you want to have separate models.

I want to note that this is an area where there continues to be a lot of innovation in terms of statistics with more and better ways to address the problems of limited dependent variables, including some semi-parametric and non-parametric methods, and so, it is, compared to other areas of regression analysis, an area where there's a lot more significant innovation and that these models are introduced all the time. Because I know you have the slides, I’ve listed some references that may or may not be useful.

And I want to close by noting, before we go to questions, that the next lecture will be on Fixed Effects and Random Effects and Joe Jacobs is going to give that lecture next week. And with that, we will turn to questions.

Jill: Great. Thanks so much, Ciaran. We had a few questions actually about sample sizes. So, one was in the context of binary outcomes, say logistic regressions. Somebody asked if it was difficult to get a large-enough data set for that; and the other was in the context of count data, similar lines, "Is it difficult or would you need huge amounts of data for count models?" So, do you have any thoughts on that?

Ciaran Phibbs: No, I mean like any regression model, you need a--you shouldn't be running regressions on samples of six, right? It's sort of meaningless. And, in general, bigger samples are good, but they don't need to be huge, and this applies to any--the running regressions for these types of models is no different than running regressions where an ordinary least squares model is appropriate. In general, the number of observations required is no different. But if you only have 100 observations or 200 observations, you can't put very much in your model; and one of the problems maybe--that actually happens in healthcare--is that there are more things that you really should be controlling for than you actually can feasibly include in the model. You have 180 observations and you really should be controlling for 15 different factors, because these are important covariates that should be in your model; you just don't have enough observations to include that many factors in the model, so you're going to have an omitted variable bias.

And that's a problem if you're dealing with small studies where you're actually collecting the data. When you're doing studies where we're using secondary data, for example, all the data from the entire VA from VINCI, you have large-enough samples that this becomes a non-issue. But you don't need huge amounts of data to estimate a count data model, it's essentially the same as any other regression.

Jill: Great. Thank you. And we have a "Great presentation" comment. Waiting for any other questions to come in, I wonder if there's any situations where you'd recommend a linear probability model or is it pretty much just, as you outlined on your slides, to test out significance before...?

Ciaran Phibbs: So, if you get really large samples, the problem gets less of an issue. There isn't any hard-and-fast rule that I’m aware of, but I would be--if I wanted to report that, I would also want to test the robustness of my findings with a model that was appropriate for the distribution of the data.

And that's the one thing, for some reason, you want it to be a linear probability model, well, turn around and just run a version of the model in logit and see if the results are similar. And there are cases where that is true, but you just need to be careful about using a misspecified model, which is what a linear probability model is, and you certainly want to test that to make sure that you're not going off the rails too far.

Jill: Great advice. And so, is there anywhere anyone can contact you if they have questions that come up after the fact?

Ciaran Phibbs: Yeah, you can send me email; it's best to use my university email during COVID because I’m on that a lot more than my VA email; both of them are in VA Outlook.

Maria: Great. Do you have any closing remarks?

Ciaran Phibbs: Just what we talked about here, and also, we talked about last week, with specifying the right-hand side variable--and this is really regression models have a lot of underlying assumptions in terms of the distributions of both the dependent and the independent variables, and make sure that you test them out and that what you're modeling matches the assumptions of the model, and test the robustness of your results.

Jill: Thanks so much, Ciaran.