Cyberseminar Transcript

Date: 04-05-2017

Series: HERC Econometrics with Observational Data

Session: Limited Dependent Variables

Presenter: Ciaran Phibbs, PhD

*This is an unedited transcript of this session. As such, it may contain omissions or errors due to sound quality or misinterpretation. For clarification or verification of any points in the transcript, please refer to the audio version posted at* [http://www.hsrd.research.va.gov/cyberseminars/catalog-archive.cfm](file:///\\vhaslcnasres.v19.med.va.gov\l%20)

Dr. Ciaran Phibbs: So my name is Ciaran Phibbs. I’m one of the economists at HERC, and today’s talk is continuing our econometric series on limited dependent variables. And I’ll note that, just reinforcement that Jo Jacobs will be monitoring the questions, and she has authority to stop me if appropriate.

So what are limited dependent variables? A classic limited dependent variable is a zero, one, or it’s something with a small number of options. So two or three choices, four choices, or small counts. And what we’re really dealing with is when the dependent variable is not continuous and not even close to being continuous. And what happens to a normal, a classic [inaudible 1:04] regression when you try to apply it to these problems and alternatives that are more appropriate. And in general I’m going to be talking about three different types of models. Binary choice models when the outcome is either a zero or a one; multinomial choice, which is looking at models where you have more than one option; and counts. So when you’re looking at counts of events but the numbers of counts are small. And in general you can think of this as that we’re looking at a framework where we’re looking at probability models where we’re looking at the probability that an event occurs.

I’m going to start with binary outcomes. These are very common in health care. One could come up with a very long list of things in health care where we look at, the outcome we’re interested in looking at is a binary outcome. The classic one is mortality, but there are a lot of other ones. Does the patient have an infection? Did some sort of a patient safety event occur? Or an adverse event? Was the patient readmitted to the hospital within 30 days? Did the patient decide to seek medical care or not for a particular event? And, you know, I could make a huge long list. That’s just the general idea.

What we’re concerned about is any kind of an outcome where it’s a simple yes/no. The event happened or it did not happen. And if you look at the classic OLS model, and let’s just, to keep things simple think about we’re trying to model patient mortality. So, and we’re going to model this, it has a zero if the patient lived and one if it died. Theoretically you could do it the other way around, or you know, you can classify it, but it’s basically, it’s a yes/no type of event. And the classic regression model, we have our intercept, we have our predictor variables and our error terms, and the thing is that probably we’re essentially modeling because of the zero-one dichotomy, the probably that Y is equal to one is a function of X beta and Y is equal to zero is one minus the function of X beta. If you use OLS to estimate this, which is also referred as the linear probability model, your error terms, your Ei’s are heteroscedastic because they depend directly on X beta. And the other problem with running OLS is that the predictions are not constrained to match the actual outcome. So you can get predicted negative values, or if you think of a probability, if you think of it in probability terms, you can have predicted probabilities of survival of being greater than one or negative, neither of which are really possible. And so when you try to make predictions from an OLS modeling this event, you get predictions that don’t make any sense.

Probably the most common approach, especially within health care, to this binary choice issue is logistic regression. Formally, in terms of modeling it, the probability that Y equals one is equal to the exponent of X beta divided by one plus the exponent of the X beta. This is a very common model in health care.

There, one thing to remember is that logistic regression is, the distributional assumptions are for rare events, and we will talk later about what happens when the events aren’t rare. One real advantage of using this in health care is this is the most commonly used distribution for this type of problem in health care. And most readers of healthcare journals are going to be familiar with the odds ratio, which is what you get when you exponentiate the regression coefficient, and that’s how the results are most commonly expressed. Technically it’s the law of odds of one event versus the other.

Another classic approach is a probit regression. This tends to be used more commonly in non-health applications. The classic example of this is the decision to make a large purchase. It is expressed in terms of, you know, does the event happen as a function of beta X plus E where Y equals one if Y is, star is greater than zero, and if it’s less than zero it’s equal to zero. That’s just the formal expression of the model.

There are other methods and distributions. In general, all of these approaches will give about the same answer. One of the things and reasons why people used to favor the probit in the past for some applications was it used to be easier to calculate the marginal effects with probit than it was in logistic but with modern packages such as Stata. That’s not an issue anymore.

But again, and when you use one of these models, it’s appropriately designed for this distribution. They’re designed to constrain your choice to, you know, they are specifically designed to model these binary choices so you don’t get predictions that are outside of the range, and your error term, you don’t have problems with your error terms.

One very important thing about odds ratios versus relative risk is that the standard method of interpreting logistic regression is the odds ratio. And what this is, is you get the regression coefficient from the formal model and you just exponentiate it and these are all relative to one, so any number greater than 1.0 is an increased odds of the event occurring. Less than one is a decreased odds, and they can go down. You know, it can’t go below zero, and it can go up to, it’s not constrained from above. And the thing that you’re really interested in is well, what is the risk associated with that? So let’s say we get a mortality coefficient of 1.5. The common interpretation of that would be a 50% increase in the mortality risk. That is not technically correct because you are converting the odds ratio to a relative risk directly to a relative risk. That said, that approximation works if your event rates are very low. If you’re dealing with a sample where your mortality is 2%, if you get an odds ratio of 1.5, you can safely say that you have a 50% increase in mortality risk. This approximation starts to break down at about 10%.

And here, this is a graph that shows what happens. Here is the odds ratio of one, and it shows what happens to the actual odds ratio as the incident rate goes up. And so you can see at 10% there isn’t much distortion. There’s a little bit distortion here for very high risk ratios. At 5% there’s, the distortion is minimal except for very large ratios. But at very low event rates these, there is no difference essentially between the odds ratio and the relative risk. But if you get up, if you’re looking at an event, say you’re looking at nursing home readmissions to the hospital, which can be 20-30% and you have big odds ratios, those odds ratios are really distorted. An odds ratio of 10 would only correspond to a relative risk of three in general at, up there. So you basically, as these event rates get high, these things just exponentially explode in terms of their distortion of the fact. And what’s really problematic then is in either direction you’re overestimating in some sense the risk of that event occurring when you do the standard conversion of an odds ratio to a risk ratio. Fortunately, there are ways around this, but I think that this a really important point because, in terms of not wanting to overstate your results.

There was a paper published almost 20 years ago now in JAMA that showed, gives a way to correct for this problem. And what you get is that the true relative risk is the odds ratio and you take one minus P zero where P zero is the sample probability of the outcome. So if we were looking at mortality and the mortality rate in our sample was 10%, the P zero would be 10%, would be point one, and point one times the odds ratio, and this gives you a correction. This is only an approximation, but it does work pretty well.

To give you an idea of the magnitude of these effects, this is from a paper of mine from a few years ago where we had about a 20% mortality rate in our sample, and when you apply the risk ratio compared to the calculated RR odds ratio, you can see that some of these effects are fairly big. The effect for small odds ratios are very small, which is consistent with this picture here that, for the, the distortions are much smaller for small odds ratios than they are for big odds ratios, and with intervening things. And I would also note that the extent of this will also depend on underlying distributions. But the point is that this can be fairly significant changes in your reported event.

As I noted before, this is only an approximation and many journals, especially epidemiology journals, now want you to direct, want direct estimation of the relative risk. And I know of cases where someone had done the odds ratio even with relatively low events in an epidemiology journal. I had one recently where we were dealing with event rates of about 5%, and they said no, you can’t use the logistic, you need to use the more appropriate option. And one option is to estimate a Poisson with a robust error variant. And basically when you’re doing, the Poisson has this very restrictive assumption that the mean and the standard deviation are the same, so you would underestimate the variance, but there are ways around this. And because you’re doing this with a binary outcome, the incident rate ratio that you get out of a Poisson is really a relative risk, and you can run this in most statistical packages. Stata has a Poisson command, and to get the correct variance you just need to use the robust option. And you can get the same thing. You can do this in SAS with Proc GENMOD. And if you do this, you will then get the correct risk ratios directly estimated, and you will make the, you will avoid that overestimation problem. And if you’re going to an epidemiology journal or you have an epidemiologist that’s really concerned about this or other statistician that is very concerned about this, you may avoid having the step of going back and re-estimating your models.

I want to note, remind people from that graph that I showed before except for very large odds ratios or when the event rate is large, the effects of the correct estimation of relative risk are relatively modest. But it’s better to do it correctly and then because, as I noted before, you are then not overestimating your reported risks.

And in terms of this binary data, there are lots of variations of this model that can be applied. You can do these models for panel data, for grouped data. In panel data you can now estimate these with both fixed and random effects. This is, I didn’t update this, but it’s like a couple of years ago I actually went through this data manual and looked at the number of commands, the related commands that were for binary data, for binary outcomes, and found over 30 of them for these different versions of grouped data, panel data, you know, logistic, etc. And you know, given what’s happened with Stata, I’m sure that the number of commands is actually greater than that now. And there are continual developments in terms of these, oh, I’m going the wrong way here, in terms of these estimation commands so that, in general at this point, you can estimate a binary dependent variable model for just about any situation.

One concern with, that you should think about is goodness of fit tests. There are several tests for goodness of fit. Probably the most commonly reported statistics for logistics regression or similar regressions are the area under the receiver operator curve, or ROC curve. And just to note in SAS, when you do Proc LOGISTIC, that’s the C-statistic that sits down at the bottom of your regression output. And you need to remember that the range for the area for the C-statistic is point five to one, and intuitively what you’re doing with this statistic is you’re comparing how well does the model predict compare to random assignment. So with random assignment you’re going to be right half the time. And so that’s point five, and a value of one is you’re perfectly predicting in between, so you see, you know, an R-squared of point six is a pretty good R-squared, but it’s not that good of a C-statistic. And you need to, you know, just remember that in terms of interpreting this. Good C-statistics are point nine and above, are really good C-stats.

The other commonly reported test for goodness of fit in logistic is the Hosmer-Lemeshow test. And what this does is it breaks your sample up into N groups, usually 10. Some programs, I know Stata will let you do this, will let you choose the number of groups that you want to break the sample in, or it’s an option. So you’re breaking the sample into equal numbers and it’s comparing the observed to expected events in each event.

And if you’re, one issue with this is that if your model predicts really well, all of your predicted events are going to be in the highest risk groups if you have a situation where you do have a lot of discrimination and the events and the risks are highly correlated. And so you can get a situation when you’re running this Hosmer-Lemeshow test when, if you’re looking at 10 deciles in terms of sample size and mortality where almost all the mortalities in the 10th decile and some of it’s in the ninth decile. And so yes, you’ll pass the, you may pass the test, but you’re, if you’re concerned about more discrimination, you aren’t really getting much information about how well your model is predicting. And so you’re passing the statistical test and making the reviewers happy, but it’s not providing as much information as it could about how well your model is really predicting among those patients that are at risk.

And several years ago, just to help us improve our understanding of how our model was predicted, we played around with this, and this is how Luft and a statistician at Stanford, Bill Brown, and myself, where what we did was we, instead of dividing the sample into equal groups, we split the sample by events. So in this case we were trying to predict mortalities. We split the mortality so the number of patients in the groups was different. The 10th decile had lots of mortality but not as many patients as the first decile. And then we re-ran the whole Hosmer, you know, by hand. We re-ran the Hosmer-Lemeshow test. And I want to note this is not a valid statistical test because we’re trying to see how well it predicts mortality and we’re using mortality to define the decile. So on a statistical perspective, it’s not valid. But in terms of understanding how well you’re actually predicting within the patients that are truly at risk, it is a useful diagnostic in terms of trying to validate your model. You can run a Hosmer-Lemeshow to, you know, calculate the statistic on it. It’s not totally valid, but it is a useful way of examining how well your model truly is predicting among the patients that are at risk for mortality if you really are concerned about that. That doesn’t apply to all applications, and it’s not something I would report. It’s something that you can use, it’s a trick you can use in your model building to help you build a better model and to compare alternate specifications of that model.

I, reminder and I mentioned this last week when I was talking about the right-hand side, but I’ll re-mention it now, if you have really big samples and you’re running a maximum likelihood routine, it takes much longer to run than running OLS. But the X matrix is the same. And so if I’m running, estimating a logistic model and I have a really big sample, it’s just, you know, 10 million just to be really, go out there. The P values for the OLS model are approximately the same as the logistic model, and the interrelationship of that X matrix in terms of things like collinearity, functional specification, and all that are the same. You’re dealing with that same beta X in a logistic regression as you are with, in an OLS. And so if you are dealing with these really big samples, you can do a lot of your model development just running OLS. The parameter estimates are not, cannot be interpreted, but they’re reflecting a relationship, and they can be very useful in terms of saving you time for a fairly large subset of the model building when you have really big samples and you’re dealing, to do, use OLS compared to maximum likelihood. And the other thing is that for many statistical packages the regression diagnostics in OLS that are easily available are more robust than they are for some of the other procedures. And so that’s just another trick of the trade when you have really big samples is that you can use OLS to help you with your model development just to get the models to run faster.

Alright, I’m going to change directions now if there are no other questions on this. And we’re going to talk about when we have more than one choice. With logistic and probit, etc., we’re talking about a zero, one. What if there are more than one outcome? And there are, you know, there are models that are ordered probit models that, where, you can, you know, there is a much smaller class of models. But there are classes of models where you can address this. And so what we’re dealing here is, and we’ll be very [inaudible 24:12] multiple decisions each with two alternatives, and then there are two different logit models for a single decision with multiple alternatives.

And these are also something that we encounter fairly common in health care. A classic example of this is the choice of what hospital do you choose, does a patient choose to go to among those that they could possibly choose in the market. You can also, from a VA perspective, you can think of well, you know, the Veteran can go to the VA or they can go to other hospital, several other options. It can also, you can also think of this in terms of the treatment among several options. For many conditions, there is, when a patient is, presents there’s more than one treatment option. And if you’re trying to study those, you have several different discreet options in terms of the treatment options. And if you’re trying to study and how things affect those or how outcomes are affected by that, you’re dealing a multiple choice. And these are choices are sometimes ordered and sometimes they are unordered. You know, it could be that there, by ordered I mean that there is a ranked preference to them. And other times, like when you’re going to a hospital it’s just which hospital do I go to. It’s not really an ordered choice, whereas the treatment options could be ranked by burden of the intervention or how it affects probability of outcome. There’s different ways that one might rank those.

In terms of the logit model for multiple choices, there are two different models. There’s the conditional logit model which was developed by Dan McFadden. He was actually sited as the primary reason for his Nobel Prize, and this is for unordered choices. And then there’s a multinomial logit model which is a different specification, and for that the choices can be ordered but they do not have to be. And the point that I’m making here is that when you have these, which model you choose is going to be very much determined by what your question is.

In the conditional logit model, just formally, although it’s of less importance here, you’re estimating the probability that you choose choice I and you have your X mention of the X beta and then this is a summation of all of those exponated X betas where the Js are the different choices.

If you’re looking at something like choice of hospital, and especially in terms of these choices, one of the advantages of the conditional logit model, which is also known as the random utility model, is that it is directly derived from consumers theory in economics. And in terms of grant proposals, one of the issues is what’s your conceptual model. And I got pushed back on this and had to re-submit, and I could say look, this is a model that won a Nobel Prize. It is well developed. And you have lots of references that you can point to in terms of that. And it was derived in terms of how do consumers choose from a set of options. The actual inspiration for developing this model was examining a problem of, how does a person get from home to work. Do they walk, do they drive, do they take public transportation? And those different choices had characteristics, different costs, different times, etc. And this model was driven by the characteristics of the choices and in that thinking of how an individual decides to get to work. The individual characteristics essentially cancel out because it’s the same for all individuals.

But you can, within this model, get characteristics back in it. For example, if you’re looking at, you know, from where a person lives, which hospital do they go to, so you have different travel times, and the different hospitals may have different characteristics. You know, university medical center has higher tech stuff, but you’ve got to deal with the teaching. You know, and whatever those different characteristics are, but you can back individual characteristics back into this model by interacting. In this example, I can interact with distance. So I can say does the type of insurance a patient has affect their willingness to travel, for example. Or does age affect your willingness to travel. So you’re bringing that individual characteristic partially back into the model, and the results are directly expressed as odds ratios.

In terms of estimating McFadden’s model, there are some limitations unless it’s changed since I last tried to estimate one of these models. SAS will require you to have the number of choices equal across all observations. And there’s a package out there called LIMDEP that has an NCHOICES option in this model that lets you set the number of choices for each observation. And in terms of looking at, to continue my example about choice of hospital, the number of hospitals that you have to choose from varies by where you live. If you live like I do in the Bay Area, there are lots of hospitals. If you live in a small, in an isolated rural area, you’re not going to have much choice. So this is a very useful feature. I have been told but I have not confirmed that you can do this in Stata with the clogit with the group option.

The multinomial logit model is specified somewhat differently. And the key thing here is that the choices are all expressed with respect to a reference choice. So you have to make one of the options the reference choice, and everything is estimated relative to that. And you get a set of parameters for each of the other choices. So in terms of this hospital choice example, this is not really a good choice because you’d have lots, you know, in markets with lots of hospitals you would have lots of different things. But if you were able to group your choices to continue along the same logic, say you were to compare VA hospitals versus private not-for-profit hospitals versus for-profit hospitals versus public hospitals and compare, you can estimate the relative attractiveness of those different types of choices. And you can still have, you know, many different hospitals in there. You’re just lumping them in groups, and you can still observe the distance. And this model allows, because it is not choice the way it is structured, you can include individual characteristics in that model. So you know, being VA centric and I’m looking as I just laid it out in terms of those different groups of hospitals, I can include the patient’s age. I can include the patient’s service connected status, etc. So I can put those other individual characteristics that may well affect the choice directly in the model so there’s trade-offs. You can’t have all the different discrimination in terms of comparing each choice by itself, but by, if you’re able to group the choices, and this depends on how your problem is formulated, then you can directly include individual characteristics.

And just to put a specific example of choice where the choice was not only could I group the hospitals but I could order them is that in newborn intensive care there are formal designations for the level of care in higher, you know, higher versus lower. And I actually did a paper looking at this where, so we took the highest level of care as the reference category and then we compared the other levels of category, we get coefficients on them, but it’s clearly an ordered choice as well. It could have been formulated as an unordered basis, but this was one where there was, where the choice was ordered in some sense and allowed us to compare the outcomes of very pre-term infants born in lower versus higher levels of care. And I’m just putting it up here. There’s a typo there I noticed, of an example of how to do this in a clinical journal. I put that reference there in terms of how to report this type of a model in a clinical journal.

For all of these choice models, there is something known, an assumption that’s embedded in them called the independence of irrelevant alternatives. And that is that the results should be robust to varying the number of alternative choices. And so in the McFadden model its actual choices. In the multinomial model where you may be grouping choices, you know, what happens if I exclude some of the options but not all of them? And the way to get at this is to re-estimate the model after leaving some of the choices. And there are, you know, you can do it in somewhat random way or a systematic way. And McFadden developed a regression-based test that allows you to specifically test to see if the model fails the independence of irrelevant alternatives assumption. I’ve put the reference here. Basically it’s a way of doing the test that’s outlined in the model where you run a regression after you’ve deleted some of the choices and compare the statistics of one model versus the other.

And I’ll note that one of the common reasons that you will fail the independence of irrelevant alternatives is that you really have a nested set of choices, and so you may need to estimate a nested set of models. So estimate one model and then the results of that to estimate a second model.

One thing and just as a general thing, this McFadden test is also structured so you can use it to test for omitted variables. And I will note that if you use these models, you’re going to want to do these tests for many health applications. It really doesn’t matter in terms of choosing, for hospital choice example or various types of provider choice. These things are very robust because distance has such a dominant effect that throwing out some of the choices really doesn’t affect your estimated results. And I’ll also note that this test is a fairly weak test, so it’s fairly easy to pass. So that this gives you an option of satisfying the reviewer that you’ve met the IIA, but you know, it is a weak test. So it still theoretically could be an issue, but you have something to cite that you’ve passed it.

Moving on to count data. Count data are integers. The same problem in terms of the dependent variable is that it can only assume specific values, can or cannot be zero. Counts can be zero. The problems diminishes as counts increase. As a general rule of thumb for count data models, you want to be using count data models if your counts are under 30, but that varies depending on the distributions.

Again, count data is something that we have encountered with some degree of regularity in health care. If your dependent variables, you know, examples would be if your dependent variables, the number of outpatient visits that each patient makes within some time period, like number of times you come to primary care in a year or the number of times the prescription for a chronic disease medication is refilled in a year given the, you know, that’s probably going to be a maximum of 12 there. The number of adverse units, events like the rescue events that happen in a unit or a hospital over some period of time. There’s a lot of things that we’re looking at where our dependent variables, you know, they’re not zero-one, but they’re small counts. And we have the same problem that we had in OLS in that the dependent variable can only assume specific values. It can’t be negative. And so we have heteroscedastic error terms, we get biased estimates. And there’s a special problem with count data in terms of really breaking down if you have lots of zeros. I’ll come back to that.

The classic model for count data is the Poisson distribution. There is a problem that the, it has this very restrictive assumption that the mean and variance are equal. For that reason in using counts, a better choice is usually the negative binomial. And Stata actually embeds these two packages together so that it’ll formally say, okay, we’re going test one model, test for the assumption, and bounce you to the other model if appropriate. The results are the same, they come out, you know, the interpretation of the results is the same. Estimating models with Poisson and the negative binomial get very, very similar answers. In terms interpreting the data, the way these models are formulated is the dependent variable, the log of the event rate is equal to beta X, and so you get what’s called an incident rate ratio by exponentiating the betas, and the interpretation of that is similar to an odds ratio.

In terms of using count data models, it is more common, most people know that if they have binary that they can’t use OLS. You tend to see, you do more commonly see OLS misused for counts. And one of the problems that, is that OLS will really misbehave if you have lots of zeros, and this can be the case even if your, even if you have a distribution where you’ve got lots of people that don’t have anything but the other people have something that happens, even if it gets up above that magic number of 30, you know, that rule of thumb number of 30. Even if some of your counts drift up into 40, 50, 60 but you have lots of zeros, OLS does not handle that well. And if you have something like that and you estimate OLS, you actually will reduce the statistical significance of your results and you can actually have messed up parameter estimates as well. I mean there’s even cases where the parameter estimates have different signs with count data models versus a properly specified versus a misspecified OLS model.

And so it really behooves you to use one of these models if the, if your data is counts. And the other thing to remember in terms of looking at the counts, you know, I’ve talked about this rule of 30. Look at your data carefully. You always want to look at your data and understand your data before you start estimating models. And also think about your data and consider the distribution of the data generating processes because frequently when you’re doing something like this you may actually have two different distributions that are generating the data that, and you may need to split your sample in some set and estimate separate models for them.

I’m going to give you a newborn example simply because it is, you know, really pronounced in terms of, if I look at the, consider the length of stay of all babies that are born, there’s really two different data generating processes involved here. You have the normal well babies which are about 90% of all births, and all of them stay less than five days, or occasionally someone will drift into five days. So that’s clearly a count, and if you try, and then you have the sick newborns who can have stays of up to three months. So if you just look at the distribution, you could say well, I’m over, you know, I’m over 30, I should be okay, but you really have two different data generating processes, and you have this very small tail that of, with very long lengths of stay that are really generated from a different data generating process. And so if you think about it, you shouldn’t be modeling these two sets of, modeling the length of stay for these two groups together. So again, it’s just an example of think about the data generating process. You may have two data generating processes, and you shouldn’t be combining them. You should be estimating separate models for those two different groups. And in this case you would certainly want to use a count data model to model the length of stay for the well babies, and for the sick babies you would, you might be able to use the OLS. You’d want to look at your distribution. You might be wanting to use a count data model still. And you know, you may want to run some model specifications for that. But again, it’s something where the point, this is non-VA example, but it’s a really clear example of where there’s two different data generating processes and mixing them can cause problems for your modeling.

I want to note that new models are being introduced all the time. There are more and better ways to address problems of limited dependent variables, and this includes, there’s been a lot of growth in terms of semi-parametric and non-parametric methods that I’m not going to go into within the scope of this lecture. But you know, talk to your biostatistician in terms of what the most appropriate model is for your data. And I can’t emphasize enough that the model, you know, the model that’s the appropriate choice is a function of what does your data look like, what are the data generating processes, and how does it match the question you’re trying to ask. And there can be lots of different questions.

Okay, at the end here I’ve cited a few texts that are relevant, two good textbooks, and I’ve also repeated again the McFadden reference and the Zhang reference because those are useful references.

We’ll now open to questions, and while people are thinking about questions, I’m just going to note that the next lecture which is next week is the Part 1 of Cost as the Dependent Variable that Paul Barnett will be giving.

And, Jo, are there any questions?

Dr. Josephine Jacobs: Nothing yet. We can give it a few minutes.

Dr. Ciaran Phibbs: Alright, I got stumped. I mean last week, you know, I think about the same amount of talk time and we actually went over the top of the hour with people asking questions, so in terms of trying to target things, that’s... If not, people can be released a little bit early if there are no questions to come. Again, if there, if you think of a question later, feel free to shoot…

Dr. Josephine Jacobs: Oh, wait! We do have a question.

Dr. Ciaran Phibbs: …shoot me an email or shoot an email to the HERC Help Desk, and we can try to, okay, go ahead.

Dr. Josephine Jacobs: Okay, we got some questions coming in.

Dr. Ciaran Phibbs: Okay.

Dr. Josephine Jacobs: Why do OLS and probit estimates get similar sample sizes get larger?

Dr. Ciaran Phibbs: OLS and…

Dr. Josephine Jacobs: Probit.

Dr. Ciaran Phibbs: I mean it’s not OLS and probit or Logit getting similar, it was that, I’m trying to go back here. I’m not sure if you can ask, if whoever asked that question would add some more clarity. I’m not sure what you’re asking there.

Dr. Josephine Jacobs: Right, maybe if they can clarify in a second. I have another one that came in. If you have a lot of zeros, what’s your view on two-part models?

Dr. Ciaran Phibbs: That’s, if you have a lot of zeros it would really depend on the question you’re asking and the data generating processes. And so the, you know, what the question is referring to, the classic two-part model is, say, predicting hospital, you know, hospital expenditures. Most people don’t go to the hospital, don’t get sick, so you model, that classic two-part model is you model the probability that you’re going to have a hospitalization, and then you only model length of stay or costs or whatever for those people who actually go to the hospital because if you try to model it with everybody together you’re going to have a problem. And so that’s a case where because it’s a rare event it may be appropriate, but again, it really depends on what your question is as to whether that’s appropriate and what your data generating processes are. But it can, you know, for many applications, it certainly is appropriate.

Dr. Josephine Jacobs: Great. Thanks.

Dr. Ciaran Phibbs: Any other questions?

Dr. Josephine Jacobs: No more coming in. Maybe if there’s contact information or something you want to leave in case some people want to clarify.

Dr. Ciaran Phibbs: Yeah, you can always send a question to the HERC Help Desk and those get triaged to one of the HERC economists. You can send an email if you have a question about some of this directly; my first name dot last name at VA dot gov is my email.

Dr. Josephine Jacobs: Great.

Dr. Ciaran Phibbs: Are there other questions? If not, there is the, don’t jump off this just yet because there is the survey that CIDER likes you to take. And I will turn it back to CIDER to put that survey up on the screen.

Moderator: Thank you, Ciaran. Thank you, Dr. Phibbs and Jo, for another very interesting session. And I’m going to go ahead and close out, so audience members please, once again, stick around and fill out that survey form. Thank you, everybody.

[END OF AUDIO]