Cyber Seminar Transcript  
Date: 05/13/15  
Series: HEC  
Session: Limited Dependent Variables

Presenter: Ciaran Phibbs  
 *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.*

Ciaran S. Phibbs: The title of this talk is Limited Dependent Variables. That’s a catchall category when your dependent variable is either a 0-1 or has a small number of options or is small accounts. And the real thing about this is that dependent variables is not continuous or even close to continuous, which is the underline assumption of ordinarily squares. And so the thing we’re going to talk about explicitly or some of them, is binary choices, which is 0-1, multinomial choice where you have a small number of choices and count where your data is a count and the counts are reasonably small. And most of these models have the general framework of a probability models where you are looking at the probability that an event occurs. And the – this is something that is relatively common in healthcare, if you think about the number of models that we have, that are estimated using things like, let’s just take models.

The basic problem is that we have heteroscedastic air terms and that the predictions are not constrained to match the actual outcomes. If you think about just the simple example of a yes/no, and either something happens or doesn’t, you think about in terms of a 0-1 interval, but you can only have a 0 or a 1. And with ordinary lease squares, you will have things, you can even have a negative predicted value where a negative number isn’t possible, that’s also true with counts.

The general framework is, and – I want to get the spotlight up here – is that if we have a general regression framework, where our dependent variable Y is a function of an intercept and a Bx matrix, and an air term, and let’s just take the zero – what is going. Y – excuse me, I’m not sure how to get rid of that.

Unidentified Female: That’s always a fun message to get, you want to turn the spotlight off, and close the program, and hopefully that will kick you out of the session.

Ciaran S. Phibbs: Yeah, okay, there okay, yeah, I tried to exit, okay, let me turn the spotlight back on. Okay, so I’m going to talk about this in terms of a 0 if you lived and 1 if you died classic mortality, which is a common application here. And what the model is going to estimate is the probability that Y is equal to 1, as a function of Bx, and the probability that Y is equal is 0, is then 1- the function Bx. If you were to run this in ordinary regular regression, ordinary leased squares, which is sometimes called the linear probability model, you have two problems. I have menti1d these before. Your EI’s are heteroscedastic because they depend directly on Bx, because your predictive probability is either Bx or one minus Bx, and so there is a direct dependency between the predictive value and the air term, in terms of how much you’re off.

And the second thing is that the predictions are not constrained to be 0 or 1. The predictive probability, if you were to run this OLS, you might – is you can get a predictive probability anywhere in the range, but you can also get negative values, and values that are greater than 1, neither of which are possible. And so running OLS really doesn’t work, is not appropriate when you’re trying to predict a binary outcome. And binary outcomes are very common in healthcare. I menti1d mortality but there’s a huge host of different types of things where we estimate this, where the model is 0-1 type of outcome. Did the patient get an infection, did a patient safety event occur, was the patient re-hospitalized in 30 days where that remote that that is actually treating a binary variable out of a more continuous variable in days to hospitalization. And did the patient decide to seek medical care for a condition. 1 could go on and on in terms of the list of different things that we commonly use in healthcare where it is a binary choice.

And the standard approach in the biomedical literature is a logistic regression. And most of you are probably familiar with that. What you’re modeling is the probability as a function at the exp1ntial of Bx divided by 1 plus the exp1ntial of Bx. That’s just the technical things you’ve all been exposed – or I’m assuming you’ve been exposed to the logistic regression. I’m going to talk about a few of the things that 1 needs to consider that are reflective of some common errors.

The advantage of logistic regression, two of the big advantages is it is designed for relatively rare events, which is frequently what we are dealing with. For most conditions, mortality is a relatively rare condition, readmission – hospital readmission rates are rare, patient safety events are rare, and so on. And the other – another major advantage is that it is commonly used in healthcare. And most readers of biomedical literature know how to interpret an odds ratio.

But there are other methods. Economist may commonly use a probit regression. This was actually developed, the classic example for which it was developed was looking at large purchase decisions where you had data and lots of individuals, but you only, say for purchasing a car, you only had the observation did they purchase a car, not if they didn’t purchase a car. And so, you observe 1 if you’re doing here, and you’re assuming that they didn’t purchase the car, if you don’t observe it. But you have data on all the individuals you are concerned with in the Bx. So you can go ahead and estimate this type of model. I think it is useful to understand it, how it was conceived, and that is somewhat different than a logistic.

But another method that is commonly used, in terms of binary choice methods, there are actually other methods that use different distributions. In general, logistic and probit will give you about the same answer. Many years ago, it used to be a lot easier to calculate the marginal effects with probit, but now, most modern programs will do that, so that’s not really an issue. That used to be 1 justification why economists tended to use probit over logistics, but Stata will give you those answers right away.

In terms of interpreting a logistic regression, the standard method, when you get the odds ratio, which is the experimentation of the regression parameter that logistic produces, and the standard method of returning that people do is they say, is use a percentage of that. So if you get a parameter estimate or an odds ratio of 1.5, you say that that is a 50% increase in the risk of the patient dying, or whatever the event is that you are looking at. But this is really an approximation of the relative risk. And remember that the logistic regression is developed for or appropriate for relatively rare events. And when the incidence of your outcome starts to go above about 10%, this starts to break down. And this graph is pulled from a JAMA article, I will provide the reference in a minute, it’s by Zangidahl, which develops a method of converting it. So you can see here, that, as an odds ratio, this is showing the odds ratio, and here we’re showing the incidence of the event, and you can see that these curves are relatively flat up to around 10%, although they’re starting for very big odds ratio to drift off at even at 10%. And that as they get bigger, you get more and more of a deviation in terms of the risk ratio. So here, odds ratio of two, and as you come up here, the risk ratio, for an odds – if there is a 30% mortality rate, an odds ratio of 3, really only corresponds to a risk ratio or an increase in risk of about 2. And you can see that these curves increase almost exp1ntially, especially for the bigger odds ratios.

And so that if your incidents of your event is higher, then you may need to make an adjustment. And we do apply more logistic regressions to situations where the probability of the outcome in the sample is more than 10%. And this – Zhang and Yu developed this relatively simple adjustment to give you the rich ratio from the odds ratio, adjusting for P-0. Where P-0 is the probability of the outcome in the sample. So in the example I referred to before, if you had – let us say that in your sample, you had a 20% mortality rate that you were trying – in your sample. That is a relatively high-risk group, but if you do, then you need to make this adjustment. And I have actually made a table here, this is from a paper I did that was published a few years ago, where the sample population mortality was about 20%. The sample was very low birthweight. Babies were admitted to newborn intensive care units which is a very high-risk population. And I’m just showing here what happens with a 20% mortality rate, the difference between the odds ratio and, what is the true risk ratio, as calculated by the Zhang formula. And as you can see here, with an odds ratio of 1.08, the adjustment is not that big in absolute terms, it is a 20% change. And as you get – go up here , those adjustments get bigger, so this odds ratio of 1.72 or over 170% increase in mortality risk, is really only a little over 100% increase in the mortality risk, a pretty big difference in your odds ratio. And that corresponds to here is the 20%, we are starting at an odds ratio of 2-3. Those curves are really starting to bend up. So that is what you are seeing here.

I think it is really important, and people do not always do this when they are reporting results, if their event rate is high, and if you are predicting something where you are looking at mortality, and the mortality rate on the sample you are using is 1%, you don’t have to worry about this. But as you get higher percent’s, you need to make this adjustment in terms of how you present the results. And this is something that is commonly not d1 in the biomedical literature.

Before I – is Christine on, or Elizabeth, are there any questions yet?

Unidentified Female: It looks like you guys are having a large spread in network issues than you thought. Neither of them are able to get into the meeting. I do have 1 question here.

Ciaran S. Phibbs: Okay.

Unidentified Female: Is the bias away from the null for OR versus RR different for protective versus harmful exposure?

Ciaran S. Phibbs: No, I mean it is purely a statistical relationship. And it depends on how you model the protective versus the harmful, so it is what you model is 0 versus 1, and this is a mathematical relationship, depending on how you formulated the model.

Unidentified Female: Okay great, thank you, that’s all the questions we have now.

Ciaran S. Phibbs: Okay, and I am going to try and remember to stop for questions on an ongoing basis. Okay, so just continuing with some notes about a logistic or probit type model is that there are now, for these 0-1 choices, there are now, all kinds of different things you can do. There is panel data for both random and fixed effects, there is all kinds of variations for panel and group data. This number is probably dated, probably three or four years ago I actually looked through this data manual to see how many different estimation commands there were related to 0-1 choice models, and found over 30 of them. And given how this data continues to expand, I’m sure that that has continued to grow. So basically, for just about in the application, there are specific tailored programs. If – that’s for people that you stated, you SAS, within proc GOM you can get it, just about anything you want, but you have to be a little more savvy in terms of how you actually specify things.

I want to also talk about the goodness of fit tests. Just a couple of notes. The most common reported statistics are the area under the ROC curve, or Receiver Operator Curve in SAS. This is the C statistic that shows up down at the bottom of regression printout. And just because people don’t always understand what the ROC curve is, it is the number that you – it is a measure of how well the model predicted, it varies from .5 to 1 because if you just sit and flip a coin, you’re going to be right half the time on average. And so that’s what the .5 represents. And then, this shows how many you predict, the curve actually, if you graph it out it is showing correct predictions versus false prediction, and as you go up, it is measuring how much better than just flipping a coin are you doing where 1 is perfect prediction. So it’s a measure, and numbers, if you are getting up .85 - .9 you have a very strong model in terms of the fact that you’re actually predicting 85% or 90% of the events.

The other model is the Hosmer-Lemeshow test. And just for reference I drew some numbers up here. So the Hosmer that model that I was showing numbers from, the area into the ROC curve was 86% so that is pretty good. And the Hosmer-Lemeshow test had a P value of .34, in other it passed the Hosmer-Lemeshow test. With the Hosmer-Lemeshow test, if you get significant P values, that means that there are problems with your predictive model. What the Hosmer-Lemeshow tests does, is it breaks your sample into end groups. The standard default in many programs is ten different groups. Some programs will let you vary the number of groups you look at. And it puts, within each of those deciles, it compares the observes, it continues the mortality example, it compares the observed number of deaths to the predicted number of deaths, and looks at how well you do. One problem that does show up, and this is not a formal test, but oh God – I’m still – I apologize for this, to stop the popups I tried to exit from my mail program before I signed on, but it keeps popping up because of the system wide problems we’re having with outlook.

And because it is breaking the sample into equal groups, so let’s say I have 10,000 observations, it is going to break it into groups of 1,000, and they’re going to be ordered by that predictive probability of mortality. And then it compares the observed to the expected number of events in each group. But one thing you can experience, is if you have a model that is very good at predicting mortality, that almost all of the events are going to be in the top decile, or the top two deciles. And your model will then pass the Hosmer-Lemeshow test in terms of – if you have a very good model it will pass the test in terms of having a high p-value, or a non-significant p-value. And so that statistically your model is okay.

But if you are really concerned about how well your model is discriminating, this is sort of an additional test for how well your model is predicting, that a biostatistician at Stanford, named Bill Brown and Hall Luff and I, with Pat Romano, developed many years ago. We never published it because the statistical properties of it didn’t work out exactly. But it is a useful diagnostic if you are interested in how well your model is predicting. And so what we did is, the same concept as the Hosmer-Lemeshow test except instead of dividing the sample up into equal numbers of patients, we divided the sample up by the number of predicted deaths. And looked to see if we had systematic bias, you know, within that tenth decile, were we over predicting those lower risks versus under predicting some with higher risk. And so this alternative specification, you have to code it by yourself. It’s not that hard, but it can be very informative in terms of how your model is predicting. If anybody is actually interested in the details of this, you can send me an e-mail. We had a paper that never got published because of some biostatisticians quibbling with us. But it can be useful for tuning your regression, and we even have a residual method for fine-tuning your regression here. But I just thought I would mention it, that the standard tests, with the Hosmer-Lemeshow test, the issue is, if your model predicts really well, you’re pushing all the predicted events into one or two cells, you will always pass the Hosmer-Lemeshow test. But you still might not be predicting as well as you might want to. And there are ways to deal with this. I am just mentioning this, and we will move on.

One final note in terms of logistic regression, although this may apply to other things, is if you have a really big sample. Let us say I am estimating something with a 0-1 event for – does among all enrolled veterans, how many of them actually use any medical care. I mean that’s not a good example, because that number is very high, it wouldn’t be appropriate for logistic maybe, but something where you had millions of observations, because the logistic, or any of these other things are maximum likelihood models, they take much longer to run than an OLS model. But I just want to make the note that the X matrix, and so your Bx is identical in all of these models. And so – well they are not exactly equal to p-values from an OLS estimate, are approximately the same as the logistic model. The parameter estimates are not right, but they tend to be correlated in terms of direction and how they interact with each other. And so if you have this problem, and you have some computing problems, you can use OLS for model development in terms of doing things like looking for collinearity, testing your model specification in other ways, and then only estimate the final model with logit or other maximum likely routines. This is something that many years ago, when computing power was much more of an issue, people would use, as computing powers increased. That is less of an issue today, but it is a neat trick if you are trying to run a really big sample on your own PC for example.

So moving on to what is called multinomial – actually are there any questions on logistic before I move on?

Heidi: The question I have here –

Ciaran S. Phibbs: Christine is in now too, Christine is showing up as a presenter if you want to try and add her Heidi. I see her in the names.

Heidi: Yes, I see her in the names, but I don’t know if she has called in yet.

Ciaran S. Phibbs: Okay. So Christine, if you can hear, try to call in to the call in number and Heidi will add it so that you can talk. So anyway, what was the question Heidi?

Heidi: The question, which should we report, OR or marginal effects, does this depend on journal?

Ciaran S. Phibbs: It will depend on journal and how you’re trying to describe things. You know, economics journals. In general, you are going to want to report the OR’s and it is just in terms of just talking about the marginal effect. That is more for an economics type journal, it will depend on the journal and what you’re trying to explain. So I can’t give you any hard rule of thumb, because there isn’t one. Biomedical literature, you tend to be using the OR.

Okay, moving on in terms of multinomial choice, where there is more than one choice. So logistic is yes, I live or die, the patient lived or died. What if I’m trying to predict if there are four different treatment outcomes, which is chosen. So I have got four. And the options are more limited and multivariable probit is where you are making, is it a way of modeling multiple different decisions, where for each of them, there are two decisions. And then there are two different logit models, where you are making a single decisions between multiple alternatives. A common application of the multinomial choice model is for which hospital did a patient end up in among the possible hospitals in their area. And within the logit models, there are two different choices, there is what is called the conditional logit model, which Daniel McFadden developed and was associated with the Nobel Prize, and that is for unordered choices. He developed this model for looking at, for a patient who is, for a person who is trying to get to work, and they can drive their car, they can take a bus, they can take a subway, so they have several choices, and they are not ordered in any way.

The multinomial logit model is a little more flexible. The choices don’t have to be ordered but they can be ordered. And, as I said, examples of these, what hospital you use, choosing among treatment options, there is other applications of this as well. It has been used some in health care. And the conditional logit model is expressed, here is the formal expression of it, it is exponential Bx divided by the sum of, or the individual choice pair – you know, individual I choosing choice J, divided among the sum of all of those possibilities. The details of this – this is also known as the RAM Utility Model, in that it was derived from Consumer Utility Theory, which is nice in terms of if you are looking at this, it gives you your conceptual model, looking at how consumers choose from set of options. The model is driven by the characteristics of the choices. The individual characteristics will cancel out in the calculation, so you are only modeling how the characteristics of the possible choices effect the choice that is made.

But there is a way, within the McFadden Model, to backdoor individual characteristics. In that, and I have actually done this in an example looking at hospital choice. We, because distance to the hospital is a very important driver, and so what we did to bring an individual characteristic back in is, we interacted, did the patient have Medicaid Insurance with distance. And so when you have this continuous variable, you can bring an individual characteristic in by interacting with, in this case distance, with something. So that can back door individual characteristics into the model. But you have to be doing that with a continuous variable, not a binary variable. And one of the nice results is that the results are expressed as odds ratio, so in terms of the final medical literature, people are familiar with them and applications of this model have been published in even general medical journals like “*New England Journal*” or JAMA. So you can get this more advanced model into that.

In terms of estimating McFadden’s model, this is a little more difficult. When you estimate this within SAS, it requires that a number of choices be equal across all observations. That can be a problem, because as you go across different markets, the number of hospitals are available, are not the same to continue the example I’m talking about. There’s a statistical program out there called LIMDEP, which is specifically designed for limited dependent variables. It was developed by Green, the same one who has a very common econometrics textbook. And in LIMDEP, there is an option, when you estimate this model, called end choices, that allows you to set the number of options for each choices, which is a very useful feature. I haven’t actually confirmed this, but someone told me that within Stata and in the C Logit option, that the group option allows you to do this, but I have not confirmed that.

Just to put up some numbers, I mentioned looking at choice of hospitals, just to throw some numbers up, I did a study looking at elderly service connected veterans and their choice of using a VA or non-VA hospitalized when they were hospitalized. And you can see here that all else held constant, that they were much more likely to choose the VA but that distance also had a very strong effect, so how far they were from the VA reduced that probability. Again, I am just sort of –

So moving on to the Multinomial Logit Model. What this does is, you have the probability of choosing option J and without going into the details, what you have to do with this model, and I am just going to skip the math here, is you identify a reference choice. And then you get a parameter estimate for each of the other choices relative to that choice. So if it is a relative attraction of choice J compared to your reference choice. And so that is a restrictive element of using that model. The flip side is that this model allows for individual parameter s to be included. In most cases, the model should include them. So if I am again, looking at the hospital choice and I am looking at do they choose the VA or do they choose one of the other hospitals that they could go to. I can include, for the individual, things like the individual’s income, is the individual service connected. I could include diagnoses or previous medical history, all of these things that may well effect that choice, can be included in the model. And that is a big advantage, but the disadvantage of it is that everything is relative to this reference choice.

One thing to estimate in terms of both of these choice models is something called the independent irrelevant alternatives. And that is that the choices should be robust to varying the number of alternative choices. So one common method of addressing this is that you re-estimate the model after deleting some of the choices, and look to see, are your parameter estimates robust, do they change a lot, if they do change a lot. McFadden actually developed a regression-based test and it is fairly easy to run. And if you fail the independence of irrelevant alternatives, what it’s telling you is that you may have a nested model in terms of the choices are someway nested, so that if you take away some choices, it is affecting others. So there may be a nested set of the choices.

The same McFadden model can also be used to test for emitted variables in these types of applications. And I will note, for many health care applications, the models are very robust. So to continue the hospital choice distance is a really dominant factor, looking at where people go. And given that, when you are looking at distance, you pass the independence of irrelevant alternatives test, no problem.

Before I move on, are there any questions out there?

Christine: Ciaran, there was a question about logistic regression.

Ciaran S. Phibbs: Okay, that was going back to two to go, but we can try to take it, go ahead.

Christine: Okay, otherwise we can save it for the end.

Ciaran S. Phibbs: Yeah, is it relevant to go in now, or do you think we should save it till the end.

Christine: Actually, let’s save it for the end.

Ciaran S. Phibbs: Okay, so you’re on now, okay. Moving on to count data. For count data, we have the same problem that we had with 0-1 or a few choices in that the dependent variable could only assume specific values. And account, I am talking about integer counts, you know, one, two, three, four, five. And the results cannot be less than zero. The problem of using OLS versus using a formal count data model will diminish as the counts increase. The general rule of thumb is that if your counts are under 30, or predominantly under 30, you should be using account data model, and not using OLS. But that is just a general rule of thumb, and part of it will depend on what your distributions look like.

Some examples where count data models are needed in healthcare are common applications. You’re trying to look at the number of outpatient visits a patient had in a year. And for most patients will see their doctor one, two, three, or four times, but you will have, in terms of what is applicable. Within the VA, if you have a patient who has some of the behavioral health issues, where they are coming in weekly or twice a week, they are going to have a very large number of visits. But with most patients only coming in to see the doctor a few times, that is a case where even though some of variables are greater than 30, you want to do a count data, because the vast majority of them are very small counts. Other example would be the number of times a prescription is filled, or refilled within a year. Given 30-day prescriptions, most of those numbers are going to be 12 or less, and so you want to use a count. The number of adverse events in a unit or a hospital over some period of time when you are looking at patient safety events in a particular unit or even in a hospital in a week or a month, there are not going to be a whole lot of those, so you are again, dealing with small counts.

Those are just some examples of this, there can be lots of them. And when you do this, you need to use a count data model And I will come back to this, but I just want to mention, that when you use a count data model or when you use OLS for a count data model, you can get numbers that are not correct. And it is not just that, in terms of Heteroscedastic of the error terms, for many of the applications, say, I am looking at number of adverse events in a unit, by week, and then there’s going to be many weeks when there are zeros. And if you are estimating a situation where you have counts and you have lots of zeros, and you estimate this with OLS, it really confuses OLS. And it cannot only give you answers where the, in terms of affecting your statistical power, but you can even get parameter estimates per person.

So if you’re estimating something where you have counts, you really need to consider using a count data model. The classic count data model is a Poisson. The Poisson distribution has this very restrictive assumption of mean and variance are equal. And therefore a much more common distribution, than per use in count data model, it is the negative binomial is a better choice. In Stata, the basic negative binomial is nbreg. And in Stata is a very nice feature in that they sort of combine the Poisson and the negative binomial, and, as part of it is a test for over dispersion as to whether, which you should use. Again, I’m not going into there.

Interpreting count data models, essentially you are estimating the log of the event rate as a function of Bx. So you have an incident rate ratio, which is the exponentiated beta, which is like an odds ratio, and has a very similar interpretation. And so just to put this up, I was mentioned adverse events in a unit, and so in terms of the monthly infection rates, and if you look at the effected RN tenure, how long the nurses have been working on the unit, and we got a beta of minus .26, and that is an incident rate ratio of .77. And so, what that is saying is that the infection rates are going down, as they number of years the nurses worked on the unit is going up.

In terms of using count data models, it is more common to see this happen when you see OLSU’s when you have counts, than it is for binary or very limited choices. I mean most people at this point in time know that if they have a 0-1 choice, that they should use a logistic or equivalent. But you do, in terms of misapplications, very often see, when the data is really count and they aren’t that big, people run OLS. And as I mentioned before and I want to reiterate, when there are lots of zeros, you really reduce your statistical power. And you can even get a switch of parameters here. And this is from that same study I was referring to in terms of count data model, where we estimated. We are looking at effective nurse staffing on patient outcomes, and just to show what happens in terms of statistical significance and false inferences that you can make, here, in this first one, we had an almost – we’re looking at nurse staffing. We had an almost significant variable of better, more nurses, reduces the infection rate, where in reality, these are the t-value. So the t-value, parameter actually went in the other way and was very, very small, so essentially no effect.

The nurse staffing in this particular model, for LPN Staffing, we had a very strong significant negative with OLS in a model where OLS should not be applied. And the effect is much smaller and not as significant with account data model. And then again here, we get another signed reversal from borderline significant negative effect to an estimate of a positive effect that was not significant. So I just reiterating that because of the nature where, and this was a sample where there were lots of zeros, so it is really confusing OLS. And not only is it affecting statistical significance, but OLS is estimating perimeter estimates that are actually in the opposite direction. And you can get things in the other – this is just an example I pulled up to show dramatic effect. The other thing to remember is that you can actually have effects in the – you can actually, in some cases, getting proofs statistical significance when you use a count data model compared to OLS. In this case the significance was being reduced, but it is going to be a function of the nature of the data.

Again, the rule of thumb is 30, but you should be considering account model if most counts are small. And you also need to consider the data generating process and the distributions. And if there are mixed situations, you may want to predict it, because if you have two different data generating processes that you’re co-mingling, you may be better off splitting the model. And this is not that much of a VA relevant thing, but if I want to look at the – a good example of a mixed data generating process, if you look at the length of stay for all newborns, well roughly 90% are normal well babies. And they all have lengths of stay less than five days, or less than six days. Where if they are vaginal deliveries it is all one, two, three day stays. And C-section deliveries mostly have four and five day stays. And that’s clearly a count and then sick newborns have much longer lengths of stay. If you look at the distribution, you have got – kids born at 24, 25, 26 weeks if they survive, staying three months or more. So you have a very different distribution.

And what you really have is two different data generating processes. And the reason this is an issue when you are thinking about this is, if you try to model the length of stay of all newborns, the 90% that are well babies are going to sort of dominate the sample and effect the relationship. You might really be interested in terms of the long lengths of stay for very sick newborns. And what you really have is two different data generating processes. And you want to split the sample and model this where the big prediction is going to be cesarean section and maybe some other things. And then this sample down here, which is going to be based on how sick the newborns are. And the point I’m trying to make is that if I try to do it all together, I can get estimates for these very sick newborns that don’t make much sense. Because this much larger group that has a different data generating process, dominates the sample. So it comes back to you have to understand and know your data and think through what you’re trying to estimate before you go and estimate the model.

And I just want to note in closing that there are new models being introduced all the time, and better ways to address some of the problems with limited dependent variables, and including that there are a lot of semi-parametric and non-parametric methods that may be appropriate for your particular use. And you just need to consult, if you don’t have a lot of expertise, you might want to consult with somebody that does have expertise in terms of what is appropriate for that analysis. And are there any questions that came up in that last section Christine?

Christine: Yes, so concerning your comments about having different generating processes, someone asked, what are your thoughts about running models on two different samples versus using a two-part model?

Ciaran S. Phibbs: A two-part model is one way of getting at the fact that there is different, that is a specific application, of a different data generating processes. And it would depend for the classic two-part model is for predicting healthcare spending, did you have any use, and then if you had any use, what did you spend? And as a two-part data generating process, and how – and that is structured for that particular different data generating process. And how one will formulate this will depend on what the two different data generating processes you’re trying to model are. So again, it will depend on the question as to what – how you address the problem will depend on the question. Are there any other questions?

Christine: Not concerning that most recent section.

Ciaran S. Phibbs: Okay so in terms of coming back, I’ll put this up, that I put in the end of the thing, a couple of various textbooks. I mentioned Green’s Textbook in terms of “*Econometric Analysis*,” and given that he had this software program that was focused on limited dependent variables, he has a lot of details in his textbook on that. Jeff Wooldridge has a very good econometrics textbook on the analysis of panel and cross sectional – cross sectional panel data, and then it is old now, but the classic – it was certainly the original classic textbook on “*Limited-Dependent and Qualitative Variables in Econometrics*” by Maddala, which is an old textbook, I am assuming it is still in print. But my copy is very old, from graduate school.

And so going back we had a question from before on logistic regression?

Christine: Yes. Yes, this question is about logistic regression. And it says, if the incidents and the outcome is too high, and logistic regression is not appropriate, what are the choices for modeling, and is there a method to determine if the incident and the outcome is too high for logistic regression?

Ciaran S. Phibbs: Well in terms of that, remember it maxes out at 50%, because then you can model the one minus incident. And you can – you can use logistic regression when the incident rate is very high, you just need to be very conscious that your odds ratios are not the same. You cannot take the percentage interpretation of the odds ratio and you want to use that adjustment for the risk ratio. And so, there is not any exact rule of thumb. I know people have used it when mortality rates or survival rates are fairly high. And it appears to work okay. I think the main thing is to remember that logistic is really designed for rare events and you need, in terms of how to interpret the co-efficient, you need to make the appropriate adjustment. And just in terms of that, I will not that – I also mentioned, have a couple of journal references here. I mentioned that I included, you can get, as Heidi said, you can get the slides from the mini URL. If you want the journal article for McFadden’s specification test, and I have also included the Zhang reference for more details on the method of adjusting the odds ratio to the risk ratio.

And the next lecture will be May 27, and it is on Cost as the Dependent Variable. And before everybody signs off –

Christine: Well actually Ciaran, we had one other question.

Ciaran S. Phibbs: Okay, so just make a note here before, and that is, once we finish, Heidi will be coming on and prompting you for a survey of – okay, go ahead.

Christine: This question is concerning multinomial logit models. And the person who asked the question sets up a scenario. So let’s say you’re examining how a pharmacological agent influences changes and cravings for a substance, say cocaine. Would you split into increase, decrease, or no change. And here you are comparing people who receive the treatment and people who do not receive the treatment. There, would you use a multinomial logit model?

Ciaran S. Phibbs: So that’s an ordered choice, and in that case, because it’s an ordered choice, an ordered probit may be more appropriate, because it’s ordered. Whereas the multinomial models are for unordered choices.

Christine: Okay, thank you, if the person has any questions to follow-up on that, you can feel free to send it in. But that’s all we have for questions, unless there are any others?

Ciaran S. Phibbs: Okay, and we’re almost done with the hour. So this is the next lecture, it is the first of two lectures. I am looking at the Cost as the Dependent Variable. Someone was referring to the two-part model, that’s the classic application of this. And then Heidi will come on and give you instructions for how to give the follow-up service.

Heidi: Yes, I am going to be closing the meeting out in just a moment, and at that point, you will be prompted for a feedback form. I know most of you have filled them out in the past, if so, you may not need to. We really do report these metrics, so we really do appreciate if you would fill it out for each session that you attend. So with that, I’m going to be closing out today’s session. Thank you everyone for joining us for today’s session, and we hope to see you a future HSR and D cyber seminar, thank you.