Cyber Seminar Transcript  
Date: 05/06/15  
Series: HEC  
Session: Specifying the regression model  
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 Phibbs: My name is Ciaran Phibbs and I am one of the economists at HERC and this talk is about specifying the regression model. One comment, Heidi may have mentioned this before if you have questions enter them in the online system, Elizabeth will be monitoring them and will interrupt me as appropriate. This lecture will focus on the independent variables, on the right hand side of the regression model. Regression models make several assumptions about the independent variables and the purpose of this talk is to look at some of the common problems and methods of fixing them. Some of these things may or may not be covered in a standard econometric class. As a matter of fact, I know that some of them are not and it is part of what a previous advisor of mine referred to as the art of econometrics as opposed to the science in terms of how you actually do some of these things.

The topics that we are going to cover are Heteroskedasticity; Clustering of Observations; Data Aggregation; Functional Form and Testing for Multicollinearity.

Heteroskedasticity is actually very easily dealt with and in the standard regression model where you have independent variable, intercept and Beta-X (βx) is a matrix of your independent variables and your error terms it assumes that the air terms are independent of the xi. A common pattern that occurs often is that as an independent variable increases that the error terms get bigger and does not always happen in that relationship, they can get smaller, there can be more complex relationships, but the problem is that when you violate this assumption the standard errors are biased. Parameter estimates are unbiased, heteroskedasticity has not effect on the parameter estimates but they are what is referred to as inefficient but the most important thing is your standard errors are biased.

Fortunately, there is a very easy fix for this. For those of you using Stata there is a robust option in essentially every regression command in Stata, which uses the Huber White Method to correct the standard errors. So you get a robust standard error it basically corrects standard error for this problem so that you do not have, your standard errors are correct. The other thing that you can do in some circumstances is the transformations of the variables may break this relationship. For example using log (X) instead of X may correct the heteroskedasticity problem. My recommendation is estimate the model as is appropriate and just use Stata or the equivalent to fix the standard errors.

The second problem that I want to talk about is clustering. This is something that we encounter a fair bit in healthcare. The assumption is that the error terms are uncorrelated. But clustering for example patients are clustered within hospitals, they can be clustered within clinics or within providers as some common examples of clustering. And when this happens, the error terms are uncorrelated if you have a bunch of patients that are seen with the same physician, those error terms are technically correlated. As a simple example here, I am running a model where x1 is a patient level variable, and x2 is a hospital level variable but it could be any of those common clinical aggregates. And when you estimate this model, the model is going to assume that you have as many hospitals as patients, but that is clearly not correct. When you have patients clustered within a hospital, the standard errors for beta-2 (β2) are too small because as you may remember that standard errors will go down as the number of observations goes up. But for beta-2 (β2), the number of observations is the number of hospitals not the number of patients. And so the standard estimation method is going to have assume any more observations then there are and as a result your standard errors are going to be too small. Again, there is no effect on parameter estimate. The parameter estimate is correct it is just that your SQL significance could be off.

General estimating equations within SAS can correct the standard errors. Again, as one may know from economist like Stata in State for virtually every regression command there is a cluster option and it uses the same Huber-White correction method, it corrects for heteroskedasticity to correct the standard errors for this clustering. And the nice thing about Stata is that, I do not have the command here, but basically in the previous example, I would just put cluster equals and I would have the hospital ID as a variable in my data set and it will just do it and this is available for almost all of the regression commands in Stata. It is not available for all of them but it is available for the vast majority of them.

Edward Norton who is now at Michigan actually did a formal comparison of the various different methods of the statistical packages have for correcting for the cluster problem and they are all virtually identical. There are differences where one uses N and the other will use N (-1) in the math so you get essentially the same answers.

I sort of mentioned Hierarchical Models in passing and that was a separate lecture in this series about Formal Hierarchical Models. Whether that is using more information on the structure of the hierarchy whether you can get somewhat different answers there both in terms of parameter estimates and standard errors, with using a formal hierarchical model, compared to just correcting for the clustering. How the answers change between those two methods will depend on the structure of the data. A lot of times, you get a fairly similar answers especially with big samples.

Just continuing, I did not advance the slide, I am sorry, issue of whether you need this or not will depend on your data and somewhat thinking through the problem. I am going to give you an example of how these errors change and this is from a *New England Journal* paper I had a few years ago. It is on a newborn intensive care units, not a VA example. I apologize but I had all the data worked out for this and it is a nice example. I am going to use that. I want to iterate in terms of this that if one looks in the literature, it is getting rarer, but there are lots of examples especially in the older literature of people running regressions with a fail to correct for this problem. The extent of the correction on the standard errors will vary with both the sample size and with the number of clusters relative to the number of observations. When you have big samples, the effects tend to be fairly small.

In the example, I am going to show you I had almost fifty thousand observations and over two hundred hospitals and ten years of data with repeated observations. What I want to show you here is if you look at the first one and the descriptions here are relatively unimportant. What we really want to show is what happens with the standard errors and you can see this was a fairly big sell and the effect on the standard errors between the corrected standard errors and the unadjusted standard errors is relatively small. You see here that this standard error interval is a little bit tighter than this and it had no effect on the statistical significance. When I come down here to this third row where it did have an effect and what happened was that this particular cell there was a very few number of hospitals that were in this cell. It was three or four hospitals over several years. So the number of patients relative to the number of hospitals was fairly large and as a result in the unadjusted, we see the results as statistically significant but in reality it was not and the standard error, difference was fairly large. Again, this just highlights the point if you have a lot of observations especially if you have lot observations within the cluster whether it is the hospitals or physicians etcetera within the group you are looking for then the correction will be relatively small.

No questions yet Elizabeth?

Elizabeth: There is one clarification question about hierarchical.

Ciaran Phibbs: Yes.

Elizabeth: They just asked if you were referring to mixed effects models.

Ciaran Phibbs: Not necessarily. There are a lot of different ways of doing this and mixed effects is one potential way of looking at this, but there are a lot of different ones. I was not referring to any particular model, there are a lot of different ways to do it. There is more than one way to address a hierarchical structure on a model and that is a topic of another lecture.

Another thing I want to talk about is data aggregation. Many times, we may have a choice in how to organize our data. You have observations and do I want to use in the example I am going to talk about we are looking at nurse staffing and do I look at the number of nurses per patient for the whole hospital or for the units and do I use it by a month or by the year as some examples. In terms of how I am going to measure it. In this case, I am measuring your staffing, well how am I going to measure it. How careful am I when I measure it. This is a choice that is fairly common in healthcare when you are trying to measure something over some period of time and the thing to remember is that as you increase the aggregation using a year instead of a month or a week you are going to reduce the variants. The aggregation can change the relationship that is observed between the variable of interest and the dependent variable.

As you can smooth over patterns, in this case in the example I am going to be talking about nurse staffing and this from a data stat from a paper that we published the reference if someone actually wants to go look at the data. If I use nurse staffing at the unit level and there is in general for that one unit the staffing was basically good or maybe a little better than average for the entire year. But there was one month when it was really bad, well patient outcomes might be bad in that one month if use monthly data I might detect that signal but if I use annual data that average staffing is going to be fairly average so I will miss that signal. That is the underlying idea of data aggregation. The point that I want to make here is that this can really, really matter.

We just ran the model, we ran this model with a whole bunch of different levels, I am going to go back here, where we look at data at the unit versus the hospital and the month versus the year. Instead of putting up a whole slew of numbers I am just going to show the effects of data aggregation in terms of this is the effect on the aggregate nurse staffing on patient length of stay and you can see here the range is about a fourfold almost fivefold variation in the point estimates depending on the different models. I am not talking about which is which using the percent of this nurse staff that were LPNs we see a huge variance from an insignificant negative to a fairly large positive. For the use of AIDS we see about almost a full threefold variance and for the use of contract nurses we see not only big differences, but we actually got a sign reversal depending on the models we use. I will just note that this one and this one were both statistically significant. Estimating the model one way we had a statistical significant negative effect and another way we had a statistically significant positive effect. The point I want to make is think about when you are making aggregation choices think about what you are doing because they can have an effect and in general you want to use more disaggregated data. Again, it may depend on the question and what you are doing.

If someone is interested, it is still a work in progress but I have a whole paper examining this issue in detail that we are working on and that I will be presenting at the VA HSR&D meeting.

The next topic that I want to talk about is functional form. If one looks at a regression model where the standard regression model Beta-X (βX) is assuming a linear relationship between X and Y. This is not always the case and if it is not the case you have what is called a mis-specified model.

You should always check the functional form of every non-binary variable in your model. There are formal tests for model specification, some of which you may have been exposed to in class, the strength of these tests varies but the tests just tell you do you have a mis-specified model, they really do not provide much more guidance beyond that. I want to talk about a method that I have used and have found very useful in terms of determining if I have a problem with functional form and what I should use. The idea is that if you look at the distribution carefully for each variable and you can do this by using sets of dummy variables to examine the functional form. Think of a common continuous variable something like age and is a frequent control variable in healthcare and may have a non-linear effect. Middle-aged a few additional years may not have much of an effect but as you get older, the risk associated with age may go up in a non-linear manner.

Look at the distribution of your variables so you can create a set of dummy variables with reasonably small intervals. I like to do this with no excluded category and then run the model of no intercept, you can make an excluded category. In the intercept the problem is you have more than one of these variables that you are trying to test you could be forcing more than one in the intercept whereas if you do it separately for sets of independent variables for the model with no intercept then you do not have that problem.

I am actually going to show some data again from the NICU data set I referred to before and what we were looking at was the effect of patient volume on mortality. So what you do here when I was trying to look at the effective patient volume at a NICU is you graph out the estimates and this gives you an idea of what kind of function you are looking at. You can use this to determine which functional forms would be a good starting point or where to make cuts for categorical variables if you determine that is appropriate. You see here in this case for the effective of the number of variable birth weight infants treated at a hospital in a year you get what is a fairly steep functional decline and then what is actually a continued decline here but it is a very slow decline. These changes here this one could think of it as a spline but it is actually a non-linear function here if one looks at it in more detail. Because we were also concerned about levels of NICU care here we get this that relationships on this so the same for different levels of care.

It can be that you have instead of continuous thing there may be categories or you may have spline relationships and it may be difficult to use a continuous function to predict across the range and get it to line up. Categorical variables may be easier to present to medical audience and that is actually what we did in this *New England Journal* article was we combined volume groups with the NICU levels of care. But what is important is that we carefully tested it and determined an appropriate set of classifications.

Because I have sort of blasted through this, are there questions on functional form? Elizabeth.

Elizabeth: There is just one question about the difference between high and low aggregation, if you could describe that.

Ciaran Phibbs: Okay back there, high and low aggregation was the question.

Elizabeth: If you could define the difference between high versus low aggregation.

Ciaran Phibbs: High versus low aggregation. I am not sure that I understand the question but I think what they are asking is basically what I was saying is that you want to have the data as disaggregated as possible. If you are thinking about it in terms of time, very aggregated data would be looking at annual data and disaggregated data might be looking at daily or weekly data. To the extent that you can, you want to use the lower level of aggregation or a weekly aggregation or a daily aggregation or whatever it is that you can feasibly do. Because then you are not going to suppress the variance and using higher levels of aggregation like annual data may suppress variance that masks real effects.

Elizabeth: Thanks.

Ciaran Phibbs: Again, in terms of the functional form I cannot stress enough how important it is to deal appropriately with the functional from anytime you have a variable that is not a continuous variable. Do not assume linearity. When I am reviewing articles all the time I see this happening, you see it when reading journal articles that have slipped through the review process. When you have non-linear relationships you can get very misleading results, when you are dealing with a non-linear relationship and you model it as a linear relationship.

Elizabeth: One question about functional form. You advised two sets of dummy variable to examine functional form simultaneously and one test for functional form.

Ciaran Phibbs: If I have two variables that I want to look at the functional form for, I have done it both ways. The thing is especially if they are interrelated, you need to look at them carefully. The answer is going to depend on the data and this all comes back to something that many people have repeated Jack Needleman says it most eloquently about know your data. Before you even run a regression look at your data carefully in lots of different ways and understand it. Then when you start running, the regressions you still need to look at your data in different ways so you understand what is going on. There is no correct answer as to whether I should do them all independently or separately. One of the things looking at a lot of them at the same time may be complicated so you can run a regression. I am going to make up an example here where let us say I had a model where I had both age which we know tends to have a non-linear effect and let us say I had income in the model which also has a non-linear effect. I might get a rough idea of what the income effect is in terms of its non-linearity and control for that roughly when I do the set of age dummies so that I am not getting biased estimates because age and income can also be correlated. I might want to have a ballpark of information or if I start with one and then specify that as a function somehow, let us say I decide I am going to log-in \_\_\_\_\_ [00:27:06] or whatever. I specify that function and then when I do the dummy variable test for age I have already controlled for them on the other. You are going to have to play with your data in terms of how you approach that. We have not given you the answer you are looking for but you have to dig.

Okay, Multicollinearity most people are probably familiar with Multicollinearity as a concept in that if x1 and x2 were strongly correlated regression can have trouble trying to figure out what part of the variance is due to x1 versus x2. This can have two affects. First, it can affect the standard errors because of the uncertainty you will get bigger standard errors. The second thing is it can affect the parameter estimates and this can even go off the charts. If you have two variables that are highly correlated you can get examples where and is it one variable will be assigned a big negative parameter and the other variable will be assigned a big positive parameter as they try to offset each other if they are highly correlated. Now there will be the joint effect so it will not net to zero because the joint effect will be included in that positive and negative. You can get these offsetting effect phenomena affecting your parameter estimates. So you need to be really careful with collinearity.

If you have strong simple correlation, you know you have a problem. If you have two variables that a correlation of .9 the regression is not going to be able to sort that out and you are going to clearly have a problem but there can be more subtle problems with moderate levels of correlation . The Variance Inflation Factor or the tolerance, which is inverse of the Variance Inflation Factor, are options. They are in SAS, they are in the regression diagnostics in Stat. They are probably in most standard statistical packages and that measures the inflation of the variances of each parameter due to collinearity among the repressors. The general rule of thumb is that if you have a variance inflation factor greater than ten you have a significant collinearity problem.

Simple rule of thumb – if you have a correlation greater than .5 you are going to need to really dig, you probably have a collinearity problem. I want to note that you can have a collinearity problem with correlations of less than .5.

To understand collinearity think about it. Just conceptually, the relevant collinearity is not just a crude correlation because regression breaks the analysis into a bunch of regression planes depending on how many variables you have. Collinearity can arise if you have a correlation within one of those dimensions given how their regression is estimated and I will give more of that later.

I am going to work through some numbers here in terms of correlation. In the study of nurse staffing and patient outcomes as we were, exploring our variables, we had one variable that we called registered nurse tenure, which is how long the nurse had been working on the unit, and how long nurses had been working on the units was correlated with age of the nurses. The simple correlation was .46, depending on the models we ran if we included both tenure and age, we were getting variance inflation factors of eighteen to thirty depending on the model and the subset.

In terms of fixing Multicollinearity more observations is good because that helps the regressions sort out, it gives it more dimensions in variance and helps it sort out the relationships as long as there is not perfect correlations. You may have to revise the data in ways that reduce the correlation. In this example, we dropped age from the model.

Just to show you the effects here of things that can go on with the correlation of .46 if we ran the model only with tenure variable we got a significant number. If we ran it only with age we got a significant number. If we ran it with both, the magnitude of the parameter estimates for both was reduced and neither of them were significant. This is an example of where it is not only affecting the standard errors, which is what people, tend to think of in terms of collinearity but it was also affecting the parameter estimates such that neither of them, not only were the parameter estimates getting smaller in this case, but neither of them were statistically significant.

I referred to this earlier and I want to come back to it. Regression is an end space and essentially, what regression does for as many repressors as you have is they are creating what is called an Eigen vector which shows how much of the variance is explained by each variable in each of the different end dimensions of the regression based on the number of variables. Intuitively you can get a collinearity problem if you have a correlation just in one of these dimensions. There is a nice option in SAS called the Collin Option in Proc Reg so what you do in SAS you say model and you put your variables and then you do slash colon (/: ) and it will give you a breakdown of what is going on with in each of the regression planes. I went back to the newborn data because when you are dealing with low birth weight infants, birth weight and gestational age are very correlated and they are both important varies. I ran a very simple model just for this example of where I included birth weight, gestational age and race. When you are interpreting this Collin diagnostic and I will show it to you in a minute, if you have what is called a condition index is greater than ten similar with a variance inflation factor and a condition index greater than a hundred is an extreme problem. What we have here it actually gives you the Eigen vectors values, which is showing you how much is being explained here and the different variables. You can see down here in the bottom where we have a condition index greater than ten and we can see that this one where it is loading mostly on gestational age but it is loading a lot on birth weight, those correlated variables and is causing a problem.

Now, there is a way to get around this, collinearity you may have to change your model. What we did in this case was we used sets of dummy variables, we used small intervals of birth weight. We also used two week intervals of gestation again in dummy variables. In this particular model because of the nature we estimated again in terms of functional form the effective birth weight is different for males than it is for females than it is for multiple births. So we used separate, again this is a functional form thing coming back in, we estimated separate birth weight functions for each of them. When we did this where before we had a serious collinearity problem, the maximum condition index was less than eight. Again, if you use variance inflation factors in this model, we did not have the problems and our predictions actually improved. I just put up a graph of what they looked like just for the heck of it. Again, what we did and this is a common thing that you can do with this type of relationship is by resorting to sets of dummy variables instead of the continuous variables you may get enough independent variance. Again, with tightly assigned small groups such that you can actually reduce the collinearity problem so that you can retain both of the variables in the model.

It does not always work but sometimes it does.

Then I want to close in terms of that in terms of it is an old text now, but this is a classic text on regression diagnostics. There is new stuff available since that came out but I like to cite classic texts. The newer econometric books will have a lot of this stuff in it. Are there any questions on Multicollinearity Elizabeth?

Elizabeth: There are a few that came in late about functional form and a few on Multicollinearity.

Ciaran Phibbs: Okay.

Elizabeth: For functional form, once you see the plot of the estimates of the semi-variables, what are the common functional forms you would consider for different types of plots?

Ciaran Phibbs: Just look at the data and you know what a logarithmic curve looks like, what an exponential looks like. And there have been cases when you look at the data you say okay that is a log relationship and if you plot out log of age it sort of matches perfectly with that curve looks like you just use log of age. By looking at the date, when you plot out those dummy variables, let me go back here, here I have a few very tight intervals and I can look at the relationship and sometimes you can say oh that fits a standard, I can just use this function for that and it will work. Then that is your functional form. Other times you have functional forms that are really messy and you can say okay I cannot just use a simple transformation I am going to have to use, to come back to this particular example it is not well defined here in this graph. I have a better graph I should use in the future, but it is really what the shape looks like is sort of a progressive curve here and then above here it is actually a slowly declining linear relationship. From this case, one possibility would be to use the spline where you would use some sort of non-linear specification here and then a linear specification above fifty.

Elizabeth: Can you repeat what the X and Y axes are?

Ciaran Phibbs: Okay so this is the vertical axes are the odds ratios associated with mortality and the horizontal axes is the number of patients treated. You are plotting out what you are interested in in small groups so I was making groups of the number of patients treated and what the mortality risk was on average or hospitals that had those groups. Again, it is a matter of looking at your data and saying what can I do with this. Sometimes when you look at the data, you can say okay I can just use this simple transformation, E to the X log X whatever square root. There are some common transformations and you can actually test them in terms of plotting them to see if that will transform it to give you the linear assumption, that regression wants to see. If not you may have to make more complex adjustments.

Elizabeth: Okay.

Ciaran Phibbs: Another question?

Elizabeth: Not on functional form, the next ones are Multicollinearity.

Ciaran Phibbs: Okay.

Elizabeth: Given the risk of multicollinearity, do you recommend parsimonious approach or kitchen-sink approach when specifying the model?

Ciaran Phibbs: What I recommend is before you estimate the model that you think through what variables should conceptually be in your model and include those. Now, there are times when the conceptual model ends up with a huge number and then when you have that you are going to increase your likelihood of getting collinearity if you are just throwing lots and lots of variables in. But, what it boils down to is are you data mining where you want to throw everything in and see what pops out. Or do you have conceptual hypothesis that you are testing. There are two different approaches, those are two different conceptual approaches, which have valid applications. If you are data mining and you are looking for associations, you need to be really careful when you get an answer that says there is no association because the no association could be masked by some collinearity. I do not know if that helps feel free to ask a follow on question but it depends in general if you are testing hypothesis you want to carefully think about your model, figure out what should be in the model and what should not and then start testing.

Elizabeth: Alright. It says does Multicollinearity matter if the purpose of the regression is prediction not an explanation for the estimation of data?

Ciaran Phibbs: Yes because to come down here, when you affect the parameter estimates this could affect when you have the collinearity and it cannot separate out the variance and it does not know what it is this can affect the predicted value because you are affecting the parameter estimates. If you have things that are very highly correlated, it can end up affecting the predicted value. You do need to do it even when you are just trying to predict things and you are not concerned with the underlying relationships.

Elizabeth: Thanks. Question about Stata.

Ciaran Phibbs: Yes.

Elizabeth: Does Stata automatically drop one or more of the correlated variables?

Ciaran Phibbs: No.

Elizabeth: Alright.

Ciaran Phibbs: You have to tell it, computers do exactly what you tell them to. If you tell them a model and this is your model, it is going to estimate that model. You have to test for the collinearity and adjust the model accordingly.

Elizabeth: Alright. When looking for multicollinearity is it okay to ignore clustering?

Ciaran Phibbs: The clustering will not in terms of the mass clustering does not really affect the collinearity, the effect of the collinearity. Again, in your initial data builds your initial explorations when you are looking for things like collinearly you are really not concerned with what is significant and what is not significant, you are looking at the underlying data structures so you can get an estimate. Whether or not you correct the estimates for the standard errors for clustering really does not matter because that is not what you are examining.

Elizabeth: Alright. In the nurse staffing example, why did you decide to drop age and not tenure?

Ciaran Phibbs: In this case because we were conceptually interested in the fact of how long a nurse had been working on the unit affected outcomes. You could have a new grad nurse who was twenty-two, started working and then was there for ten years, she would be thirty-two. You could have another nurse who did not start working in that unit until she was forty, worked there ten years she would be fifty. The tenure affect was the same but the ages would be different and the bottom line is that conceptually we really were not interested in age we were interested in tenure. We thought we should control for age just because you might have different levels of productivity at age and then the collinearity said you cannot have them both in the model so we kicked out age.

Elizabeth: Alright. A question about aggregation.

Ciaran Phibbs: Yes.

Elizabeth: Would it be a problem to aggregate if my RHS variables are state level, the outcome variable is individual level and aggregate to state level.

Ciaran Phibbs: Okay so it sounds like you are running a regression where you have data on individuals whatever it may be, but you also have some state level variables in your right hand side. Now, if the question is – should I reframe it to just look at state level aggregates of the individual outcomes I would say it would depend on the data you have. But if you had a model where you had some individual level variables on the right hand side as well, then I definitely would not start aggregating because you want to get that variance and you were looking at some you might have that you will meet at the state level. One of the things that we have to face is that it can be the case that for some of the control variables we get we cannot get them at the individual level, we can only get them for aggregates whether it be zip code, county, state whatever it is. Even in that case, if we have in terms of we have some data on the individual let me just take an example. You pull data from the CDW or from the medical SAS files so you have data on individuals in terms of their age and conditions and so on and you are estimating mortality or whatever it is, you are estimating. There are some variables that you can only get these control variables at the state. I am going to invent an example, this may not be case but you can only get a variable on the rate of people that do not have insurance at the state level. You can actually get it at a smaller aggregates but I am just sort of going off the cuff here. So you have these variables that you can only get at the state level that might have an effect. You can still include them just realize that you have different levels of aggregation and it would be better to run that model at the individual level than to aggregate it on up because then you are going to mask a log difference by the reduced variant. In general keep things as disaggregate as you can is a good rule of thumb. Even if you end up mixing levels of aggregation and then you just have to adjust accordingly because individuals are clustered within states for example.

Elizabeth: Is there a way to assess multicollinearity between repeated measure or time dependent/independent variables?

Ciaran Phibbs: That gets more complicated because of the structures, I do not have specific tests to recommend off the top of my head, but one needs to look at that. Sorry I cannot give a more specific answer.

Elizabeth: Okay. What type of correlation analysis would you recommend for testing multicollinearity?

Ciaran Phibbs: Well the simple one is to use the variance inflation factor. That gives a very accurate test as to whether the variances are inflated for specific variables. How one proceeds beyond that if you look at the simple correlation and you say the correlation between these two variables is .7 I have to do something to fix that that is one issue. When you look at the correlations in the patterns and you do not see something but you are still getting the variance inflation factors that is when you might want to go to the Collin option so it can show you what is happening within each of the regression planes because there could be more subtle relationships.

Elizabeth: Okay.

Ciaran Phibbs: Is that it for the questions?

Elizabeth: There are a few more. There is one question about a graph but I am not sure which. It says I am assuming the odds ratio associated with mortality in the previous graph as shown was the response correct?

Ciaran Phibbs: Yes.

Elizabeth: Great.

Ciaran Phibbs: Alright.

Elizabeth: I think.

Ciaran Phibbs: That is it for the questions.

Elizabeth: That is it.

Ciaran Phibbs: Thank you for your attendance and next week there is another lecture where I am on again and we are talking about limited dependent variables. It is going to talk not just about logistic but also things where your variables are not a big continuous variable so something like count data where you data are small counts as one example. We are going to talk about some of the different models there some of which may or may not be covered in a standard econometric class. I think that finishes it we are done a few minutes early, you found some time.

Unidentified Female: Fantastic.

Ciaran Phibbs: Heidi do you have anything to jump in here.

Heidi: I was just going to jump in, I am going to close the meeting out in just a moment. When I do that, you all will be promoted with a feedback form. If you could take a few moments to fill that out we really do read through all of your feedback and you get to make changes for our current and upcoming sessions. I want to thank everyone for joining us at todays’ HSR&D cyber seminar and we look forward to seeing at a future session. Thank you.