Robert: And we are recording. Well, Dr. Phibbs, can I turn things over to you?

Ciaran Phibbs: Yeah I'm only, and just a question for you, Robert. Because I only see live attendees, which seems not right. Okay now it's starting to go up. Maybe, should hold off for a second, like, people are having trouble getting in?

Robert: It's possible people having trouble getting in. We have an an e-mail address that we can receive from them at, and we'll take a look at that, and make sure that people are able to get in.

Ciaran Phibbs: Yeah. Because I mean, we're up to seven now, but that seems low. So maybe we'll hold off –

Robert: It does.

Ciaran Phibbs: – For a second here,

Robert: Okay. We are recording. And a lot of times people join. Like, they register so that they can get the links that get sent to them later, so that they can come watch the recording. So it's up to you.

Ciaran Phibbs: Maybe we'll give them another minute before I start.

Robert: Okay.

Ciaran Phibbs: Yes.

Robert: Do you, do you have any jokes?

Ciaran Phibbs: The numbers, okay, the numbers are starting to go up. Okay. So I'll – they're still small but they're starting to go up. Alright so I'll start. So this is part of the, part applied econometrics seminar series. And remind people that this is not a comprehensive course in econometrics but more of a selected topics for things that are relevant to health economics. This one is on the right-hand side, variables and is model.

And just some background as to why we're giving this lecture as this, when we were starting to put the course together years ago, we were talking. And one of the things that came up was the fact that all too often as journal reviewers or reading journal articles, we were seeing significant mistakes in how people were specifying the right-hand side of their regression models.

And because, there were lots of focus, okay, get the correct regression model depend, given the nature of your dependent variable like Poisson or logistic but or an OLS\_\_\_\_\_ [00:02:28], et cetera. But much, many people were paying much less focus to adequately address the issues about how to specify the variables in the model. And if those variables are properly specified to meet the model assumptions?

And the issue is is that regression makes a number of assumptions, including some reasonably strong assumptions about the variables in the model. And the purpose of this talk is to address some of these issues for the more common problems, and talk about some methods of fixing them. It's not intended to be all encompassing. And several of these things are things that should have, may have been not covered early given the cursory coverage in the standard, sort of, MPH-level regression classes.

Those with more advanced training statistics have probably been exposed to these. And so the issues that I'm going to talk about are heteroskedasticity, clustering of observations, data aggregation, functional form, and testing for multicollinearity. If one thinks of – and we're gonna just, in terms of talking heteroskedasticity, but just think of basic regression model where you have a dependent variable. You have an intercept term. You have a matrix of axis \_\_\_\_\_ [00:04:21] to your regressors, and a, a vector of coefficients, and your error terms.

And regressions, the regression model assumes that the error terms, the EEI [PH] are independent of the XI. And we have a very common, the, a common pattern that occurs a lot of times is, is that as x gets bigger, the error term gets bigger. It can go in the other direction, too, theoretical. But this is actually quite common when you have things with very skewed distributions.

And in healthcare with the hospital costs are very skewed, like, so stay is very skewed. Severity of illnesses is very skewed. Just to quote some examples, so and if you are looking, if you were looking at hospital costs and severity of illness was one of your regressors, you almost certainly would have, if the error term gets bigger as severity goes up in terms of costs. And so you would have heteroskedasticity there.

The, so why does this matter? It biases your standard errors. The parameter estimates are unbiased and varied. Technically inefficient, but the point estimates are unbiased, but your standard errors are off. Fortunately for heteroskedasticity, there's a simple solution, and the, in Stata it's the robust option, which uses if you a Huber-White method to correct standard errors. There are equivalent options in most regression packages, so that you can just specify if you're in Stata or whatever. You just specify robust and you don't have to worry about it.

You may want to worry about the efficiency of your estimates. In terms of understanding your variables, you may want to look at them if you think it's a problem. It may be more appropriate to specify log of an XI instead of X as the right-hand side variable. But we'll talk more about that when we talk about functional form in terms of specifying the variables. But it's certainly something to consider. Is that you may want to include some specifications that reduce that.

And if you, the log specification because the outliers, the deviation in them is less of a unit so that the, that in and of itself will reduce the heteroskedasticity. The, before I move on to the – because that's because there's a simple solution, you should be using standard errors as, sort of, a default. It, before it moves on, if there's any questions, enter them in the chat. Joe will let you know.

So next topic it, that occurs a lot in healthcare that isn't as well-addressed in the basic econometric, of course, is the issue of clustering. Regression, the same regression model, it assumes that the error terms are uncorrelated. But the, that assumption is violated if you have some option, observations that are clustered within others.

So we have, in healthcare there are many instances when we have data where we will have variables or patients and providers. And the same provider is treating many patients, be it hospital or outpatient provider, or stuff. So there's a lot of instances in healthcare where we may have clustered observations. And to put, continue this in terms of this, if we had where the the beta 1 is our patient-level variables; and beta 2 represent hospital-level variables or another, it could be any sort of provider. Where you have patients clustered within providers.

In this case you have an issue. The issue is, is that regression assumes that there, that each observation is independent. And the provider, there's a problem with the provider-level variables. Essentially, it's assuming in this case you're a provider or a hospitals, that you have as many hospitals for providers as you do patients, which is not the case. And remember that as your sample size gets bigger or a given, you, you were, your standard errors will shrink because you have a more precise estimate.

And the, if you don't adjust this for the provider-level variable or variables, the standard errors are going to be too small because the regression is gonna think that you have as many providers as you do patients. And so this is a more precise estimate. And so there's absolutely no effect on the parameter estimate, but the standard errors are artificially small. There are fortunately ways to address this, generalized estimating equations are one method.

There are also formal hierarchical methods that can be used where you're actually modeling the structure. That's a different issue that I'm just going to mention, but not take on here. One alternative that works very well in Stata is that there is a cluster option which uses that same Huber-White sandwich estimator to correct the standard errors. The same estimate, the same, it, it's slightly different but it's the, the, in terms of the underlying math is the same as for correcting for heteroscedasticity.

What you're doing is telling the, the, telling the regression, where, what, what the cluster is. So that what you do is you have to include a hospital ID or a provider ID, and tell them that the observations are clustered at that variable. And you include that in your function specification. And then the, your standard errors or automatically fixed. Both the general, GEE, and this cluster option yield essentially the same result.

That may not be the case if you formally model a hierarchical model. Sometimes you get very similar results. Sometimes you get different results, that will depend on the structure of the data as to whether you actually – and that informs whether you actually need to do a hierarchical model. That's beyond the scope of what I'm going to talk about today. So I'm just mentioning it.

It's worth something to consider in terms of what your approach is and whether you need to have a formal hierarchical structure in your model. Give you an idea of the magnitude. I'm sorry, I just jumped ahead of myself. So and we'd want to note in terms of these hierarchical \_\_\_\_\_ [00:12:18] linear models, that you often get the same, very similar answers. And it really depends on what you're doing.

And I'm going to give an example of just to show you what happens with clustering. This is because it's a nice example, but it's not directly VA relevant. This is about NICU patient volume, and NICU level on mortality from a paper I had several years ago in \_\_\_\_\_ [00:12:49] journal. It's a good example where I had – it's the best example I had. So even though it's not VA relevant, I'm using it.

And failure to make this, and I want to emphasize that, even today, you see a lot of, you will see a lot of stuff where people haven't addressed this because they haven't thought through the problem and the assumptions of the regression model. And the extent to which the correction will, has an effect will vary with the combination of the sample size, the number of clusters, and the relative, and how that relates to the number of observations.

So if there is more of a discordance between the, the the clustering. So if you have only, if you have lots of patients clustered within a few providers, that's going to make the provider variables have more of a shrinkage in their standard errors. And then the other thing is, is as samples get bigger, because estimates get more precise as your observations increase, the effects of clustering will get smaller.

So in the example I'm going to use, I'm going to use a fairly big sample ant almost 50,000. And there were only about 200,000 hospitals. So with ten of years of data with repeat observations, so there weren't that many clusters relative to the number of observations. So what that meant is is that the cluster, that the each observation had lots of patients within it, and so that they, you have the opportunity for biased standard errors.

But this was offset by the relatively large sample. So you can see here, the the middle column that says 95 percent confidence interval is the adjusted standard error. And then off to the right we have the unadjusted standard errors. And you can see here that they're fairly close. The one exception where they got a little bit different and I got a switch in the statistical significance was in this group, which was sort of an oddball group in that there were only three hospitals in this cluster.

And so they're relatively small number of patients. So even if this hadn't been an issue, that relative to the point estimate, the confidence intervals were large anyway. And you can see that there is some movement, but almost all of them remain the same, but that was for a large sample. But there was some change in the confidence intervals, and in this very large sample.

And if you were, if I'd had a sample instead of 50,000 it had been 4,000 or 2,000. The effects on the, what was logistically significant, and the effects in the standard errors would have been larger just because the estimates weren't as precise. Because you had fewer observations.

But the, the, so again think about your structure, if you have clustering, you absolutely must control for it. And there are methods to do this that are relatively straightforward, and can be addressed within any of the statistical packages. And I noted the one for Stata, but there are equivalent ones for other programs. Any questions, Joe?

Josephine Jacobs: Not yet.

Ciaran Phibbs: Alright. Next I want to move on to data aggregation. And this is something that all too often people don't think about and can have a real effect. Because many times we have a choice in terms of how we organize our data. In the big secondary datasets we tend to use we have lots of data, and we want to, we want to measure something. And do we measure the annual number of patients treated, monthly number of patients treated, the weekly number of patients treated, the daily number of patients treated?

And the thing is, is that as we aggregate more or do we aggregate data across the entire hospital? Or do we look at specific units? And the more you aggregate in terms of those averages, you're going to be masking variation. You're going to reduce the variance. So just if I look at the daily census of a unit compared to the annual census of the the unit, unit, the the daily census is going to vary a lot more.

And depending when you're looking, what you were looking at, that variance can matter. And it can even change the relationship between the variable of interest and the dependent variable. And I'll show you an example below where science are changing and also changes in the magnitude. So this data aggregation is important. And so one needs to think, when you're specifying your model, you need to think about is this going to matter, or test to see if it's going to matter?

This is, I'm going to give you an example of data aggregation where it comes from a couple of papers I did on how nurse staffing affects patient outcomes. The Winter paper that's the second paper here is actually directly looking at the effects of data aggregation on reporting this on how this affected the outcomes. And these are patient level regressions looking at the effects of nurse staffing on patient outcomes.

And we aggregated at the end, compared both looking at unit level data versus across the whole hospital. And we also compared month and the year. For what I'm going to report here for simplicity, I'm just gonna report how the results changed with aggregation across units. And so just to look here, HPPD is the hours per patient day of nurse staffing, and how that's associated with patient outcomes.

And you can see that the hospital effect is, there is a difference between looking at different types of units. So if I estimate this at the unit level, and not only do that but do it separately for ICUs versus acute care, I get a very different answer for acute care versus ICUs. The effect in, in acute care units is about twice as big as it is in the ICUs. And ironically, when I aggregate it for the whole hospital, the estimated effect is even smaller than what you get in the ICU.

So not only are you masking differences between your different units, but by aggregating across the whole hospital, you get it, you get an answer that is different than the subcomponents, and it's not in between. And this continues here, so that by looking at the percent of the nursing staff that were LPNs we see we get, overall we get a negative. Here we get a small negative, which is nowhere near as big as that, but we get a positive in the ICUs.

So in other words this aggregation in the patterns of the aggregation can really distort your estimates, if you're using an inappropriate level of aggregation. And if we get these similar types of things in terms of these out of range things, the effect of contract nurses, which in this case is worse outcomes, more like they have the outcome, is significant and big here.

But the combined estimate, although significant, shows that it's almost, it's like almost half the size of the actual estimates when we just aggregate the data. And so this, this example really estimate, it really emphasizes that aggregating the data can really distort your findings. And so to the extent possible, you want to disaggregate as much as you can so that you're not masking differences.

And so we think about going on here, this, why this is happening? Is that acute care units and ICUs are very different units on many different dimensions. And especially in the severity, and their needs for nursing, and the levels of nurse staffing. The ICUs have much higher levels of staffing. And so with the higher staffing levels, you get some smaller effects in the ICUs. But then combining them masks bigger effects in acute care units.

In the Winter paper, we also showed that we got pretty significant differences by aggregating for months versus years. And that was that, sort of, what I was referring to earlier about masking the variance. And so again, when you're looking at at aggregating data you need to think about this. And in a more general thing, aggregated data are gonna yield less precise estimates because they're asking variance. But they also, also note that, as I showed in that example above, it can actually, sort of, distort your findings in terms of – I got sign reversals in some cases in the disaggregated data compared to what you get in the aggregated data. And I would get, I was also getting cases where the effect in the aggregated data.

Just to go back up here, in the aggregated data the, the combined effect for the hospital was outside the range of the two disaggregated effects. So there is really this distorting thing. And so the extent that you can, and if you can't disaggregate the data because you, you need to acknowledge this and the limitations when you report your findings Gee, we're reporting this but we were aggregating data. And that could be biasing our results.

The next thing I want to talk to, and this is something that, the problems that I see reviewing papers and grants is the one that is probably ignored the most is functional form. The standard regression model in general, some of them, some of the specifications that are inherent nonlinear specifications, that it may be something else. But the general regression model assumes that X has a linear relationship with Y. And this is not always the case, and can result in mis-specified models that can actually have meaningful effects in your, on your findings.

When you think about it, just think about the effect of age and increased risk. There is many conditions where the risk increases with age, but between 20 and 50, there is essentially no effect. And then it starts to creep up. And then as you get older, it will go up even more. And so and if you're just put in a, a linear variable, you're overestimating the effects of age until you're elderly, and then underestimating the effects among the elderly.

And so, all too often people do not consider this. As a rule of thumb, you need to check for functional form for every nonvariable \_\_\_\_\_ [00:25:54], nonbinary variable in your model. And one of the other things that you see, especially in the econometrics, and economics type literatures, "We did this functional form test and it passed." Because there are formal model specification tests. You may have been exposed to some of these in your class.

There's two problems with these tests. One, they don't really show you what you're looking at,just is the model mis-specified? And more importantly, these tend to be what are referred to as fairly weak tests. So it's very easy for the, a mis-specified model, it's relatively easy for a mis-specified model to pass that test. So they're only going to flag models that are very mis-specified. And so what can you do about this?

Well, one, first, you should look at your data carefully and try to understand your data. But I haven't seen this in any class, but there's a simple way to actually look, and see what the relationship is. You carefully look at the distribution of each variable. And then create a set of dummy variables with small intervals, and then run the model with no intercept. And you have no excluded category in the intercept.

So for age, you create dummy variables in one year, or two years, or three year, maybe up five-year intervals to use that example, and then you run the model. And you don't just do this for age, you do this, or for whatever you're doing, you do it for all of your nonbinary variables all at once. So you're jointly estimating it. And then you look at the relationship and the pattern of those dummy variables which are essentially non – you're making it a nonparametric specification of age because each one can change, will show you what the functional form looks like.

And then you can then go on and make the needed adjustments to properly specify the model. Okay, I can do a linear specification. Or I can, gee, this looks like a logarithmic relationship or an exponential relationship where I need to use a spline function. Or there's really, sort of, this dichotomy and it has one. And maybe I just specify a couple of dummy variables.

But by looking at this, you can understand what the relationship is for all of the nonbinary variables in your model, and figure out how to specify the model so that you're actually fitting the data. So I'm going to give you an example of doing this. It's the same, same there, data set, that same NICU data set I had before. Because I had to address this for that. And so I'm going to show you in terms of how you graph out the parameter estimates, and then how to move forward.

So here is, I'm looking at the effect of volume on patient outcomes. And you can see that there is this very steep decline. Then a pretty flat relationship but still going down, and then it gets flatter. And one could model with trying to do a spline. So I could have one here, one here, one here, or one here, or this might just be an exponential decay. That's almost linear so you could have another linear there.

So there's several options. In this case because I was also, I'm just showing this slide and how to look at it. So I had the groups in relatively narrow intervals and I looked and saw what the pattern was. I actually did a more sophisticated look because I also had NICU level in addition to that. So I mapped this out for each NICU level.

And for some applications, you may want to just use dummy variables instead of continuous variables. Because especially when you're dealing with complex relationships it can be very difficult to get a continuous function that actually predicts, sort of, the entire range.

And as an aside, categorical variables are frequently easier to present. So in this case, and this is from that New England Journal paper, what I did was I made different volume groups for each level of care. You see that the, I have level, and here it's less than ten. Here to let you know, here it's less than 25, 25 to 50, 50 to 100. And these groupings were informed by the looking at that type of plot for each of these different levels of care.

And I haven't even put up all the parameters because there's more parameters than this that were in the model. But the whole point is, is by looking at it I was able to say, okay, well, for this group, for all those that are less than ten, they're at about the same level. And I can group from 11 to 25 within this group. And but it was by carefully looking at the data and you say, okay, I can do this grouping. And that gets a whole lot easier than trying to estimate the effect of increasing by one day be, the \_\_\_\_\_ [00:31:59], your volume by one in different functions.

It just was easier to present. It just yield just simply, essentially the same result as doing a bunch of the continuous function or having several continuous functions. In this case I would have had to have splines. It would have been very complex and hard to present. And so this made it easier, but it was informed by this deep dig into the data.

You don't just arbitrarily go and pick volume categories to say, okay, I'm gonna look at less than 50 patients versus more than 50 patients. That cut point needs to be informed by careful examination of the data in terms of specifying your functional parameters.

Josephine Jacobs: We do have a question about the dummy variable approach. Do you want it now or?

Ciaran Phibbs: Yeah.

Josephine Jacobs: Okay. Somebody asked if, including all of the dummy \_\_\_\_\_ [00:32:57] stuff, are you including all of the dummies at the same time? And can this inclusion of dummy variables introduce convergence issues within the clusters?

Ciaran Phibbs: It could introduce convergence issues. That is true. And yes, you'd need to consider everything because if you have two of these variables you're trying to specify, when you estimate the model with these set of dummies, if you only include one, and just include -- like I just included age as mis-specified. And you include something else. I have a, I still have a mis-specified model, so that that set of dummies could be biased by the mis-specification of that other variable.

So you need to do it all at once. And hopefully that answers the question. And yes, you can encounter convergent intervals. If you have convergent intervals, the first thing I would do is make the groups a little bit bigger. And that'll give you an idea of what to look at so that you can have an approximately thing, and then maybe you iteratively -- okay you figure out how to specify one variable.

Then you specify that variable. Then you go and look at the detail for another variable; or just have that, that you could have bigger sets of dummies for all the other variables where you get the detailed on one. Once you specify that you can iteratively move through if you have several of these mis-specified variables that you need to do this specification test for.

I'm presenting this. I haven't seen this in a textbook, but it's the best way that I have come up with to really address the functional form. Because yeah you could try all these different; okay, I'll try age, age name squared. I'll try log of age. I'll try and you can specify X, whatever it is. You could try lots of different functions and see what works best. But you're just sort of guessing at it. And by doing this, all these dummies you get to look and see what the specification looks like.

And one of the things that you may happen is when you look at that set of dummies, and then you say "Okay, that looks like a quadratic." So you just go specify a quadratic. You don't have to go to these dummy variables, but the sets of dummy variables will tell you how to specify it.

If there's no other questions, I will move on to the next topic which is multicollinearity. And this is something that we encounter a lot in healthcare. What effects, if you have two variables in the model that are strongly correlated? Essentially what happens is the regression is gonna have trouble attributing or identifying what part of the variance is due to one variable versus the other? The standard textbook say, okay, this is going to increase your standard errors. But this can also affect parameter estimates.

And I've even seen cases, if you have two highly correlated variables; and let's say that true effect was a small positive effect. And if you have two highly correlated variables in the regression bouncing back and forth, you can actually get cases where you get one happening, a significantly larger than the true effect, positive effect. And the other having a negative effect because the regression just can't partition it out.

So that's a statement of the problem. And it's obviously because they can't understand. The regression can't understand how much of the variance is due to one variable versus the other. You get increases in the standard error. And that's the, that's the textbook, what the textbook say is the problem with multicollinearity. They don't address that they should – you need to realize that this can also affect your parameter estimates. And I'll show you an example of this in a minute.

So the first step, the first thing you want to do, look at the correlations. General rule of thumb is if, if you have two variables that are correlated with a correlation of 0.5, you're going to have a problem. But you can still have collinearity problems when they're, when the correlation coefficients between the two variables is less than 0.5. And I'll explain that in more detail in a minute.

Strong simple correlation, you have a problem, but hidden problems cannot be detected by simple correlations. Fortunately, there is a good regression diagnostic called the variance inflation factor or VIF. And it's, you can, it's an option that you can ask for in essentially every statistical package that's worth a damn. And what the variance inflation factor does, intuitively is it measures the inflation of the variances in each parameter due to collinearities among the regressors.

And some packages, I know SAS does this, will give you what's called the tolerance is another measure. And that's just the the inverse of the PIF, but. And as a general rule of thumb, if you have a variance inflation factor of more than 10, you have a collinearity problem. But this is in terms of your standard errors, and it doesn't, it will tell you that you have a collinearity problem. But as I mentioned, you may also have issues with parameter estimates as well.

And so I'm going to work through this little comparison. This is from that nurse staffing study that I referred to before. And some of the, in addition to looking at the amount of, the amount of nurses that were available, we were also looking at what we call the RN tenure, which is how long the nurses have been working on the unit, and age. And those two variables as a raw, had a raw correlation of 0.46, depending on what we're looking at.

We were in the exact model, we were getting variance inflation factors of 18 to 30 in this, in the models when we were testing that, which is serious collinearity problems. And and that was, I've alluded before that you can have collinearity problems when you have a correlation less than point 0.5. And then and it, what it does is it makes, this is less statistical significance. But I also want to address that it can affect the parameters as well.

So there's some fixes for ways to address more collinearity. More observations helps as long as there isn't perfect correlation, additional observations will help the regression sort out what part of the variance is due to one variable versus the other. And you can use that to get more precise estimates and to shrink your standard errors.

You can also revise the data in ways to reduce correlation. One is, just you would say, "Okay, we're going to drop something from the model because it's just not tractable." But there are other transformations that one could make. So just to, as I said, so, if we ran the variable with only age, or only tenure in it, we get these parameter estimates. And that's excluded with the drop to exclude.

But just to show you what happened when we ran both in terms of affecting the parameter estimates. Not only is it affecting the, is the corollary affect the standard errors, but both of the parameter estimates were actually, when we put both in the model, were smaller than if we included just one of them. So that it can have an effect on your parameter estimates.

And there are ways to address this. Just, there was another issue in this same paper where we wanted to control for how experienced the nurse was, and how, one, the nurse who had been working on the unit. But obviously, there was a correlation issue there as well. And so what we did was we broke it apart because we knew when the nurse, who had tenure, so we know when, we knew how many years' experience the nurse had. And we knew how many, how long that she had been working on a specific unit.

So we just included a variable for her, of her experience before she started working on the unit and tenure. And that broke – those or that broke it apart because we took out the overlap part of how many years of her experience were due to the time she was working on that specific unit. And so that made them independent and we were able to improve both the model with no problem.

So again, you need to look at your data carefully, think about it when you have problems. And there can be ways to transform the data to address the collinearity. So that was a transformation of the data that reduced the collinearity.

One other thing is that in terms of collinearity that you need to think about is that there can be hidden problems. And there, I'm gonna try to simply explain this. When you run a regression with many variables, you're essentially running, think of it as Ns, if you have N regressors where N is a reasonable number, 10,20, whatever, or how many variables in your model.

There are essentially as many planes in multispace as there are, as there are variables in the model. And the, and the eigen vectors that go into the, into the regression are looking at how much of the variance can you explain by each variable within each of those planes? And so you can get collinearity in one of those sub-planes, and these standard tests don't detect it. But there is a test in SAS that does. And it's the collin option in the regression package.

And in SAS, this is only in proc reg. It may be in something else, but it's not in most \_\_\_\_\_ [00:44:59], most of the packages in SAS. But it doesn't matter, even if you're supposed to be running a logistic model, the X matrix is identical. So this test works. The parameter estimates don't, these, but this test to see if you have collinearity in one of the regressors within your X matrix will work regardless of the model you're specifying.

And so intuitively, what you're looking at is the correlation in the Nth dimension of the regression. And to keep this tractable, I estimated a very simple model with the data I had. I was just looking at newborn mortality. And I had birth weight, gestational age, and I included if the patient was Black, just because to add another variables. And birth weight and gestational age of an infant are very strongly correlated. There is noise around it, so the correlation of 0.56.

In the terms of interpreting this, it will generate what's called a condition index. If it's 10 you have a collinearity problem. Greater than 100, you have a really significant problem. And the covariance, if two or more variables is greater than 0.5, in the regression point, so jus to look at this. What this is saying because it's very simple model. And we know that there's correlation between birth weight and gestational age, which is why I ran this simple example.

But what it says here, if you look at it, that in this, this is showing you the condition index for -- because there's four variables in the model that the intercept in those three variables. So you have four eigen values for, eigen vectors for the regression. And it's showing the condition index. And it's saying in this, in this fourth regression plan that you've got really strong correlation between birth weight and gestational age. And it's causing things to blow up.

And this is how you can test for subtle collinearities in the data that may be causing problems that won't show up in a standard test. Now, I mean, if I were to run this regression \_\_\_\_\_ [00:47:21] the variance inflation factor would really big because we know this correlation. But I'm just trying to show an example of how you work through this. And there are, so I have collinearity. What can I do?

I mentioned before about dropping variables at least specifying variables. And there is another thing that you can do in the same NICU thing, model that I did. What I did is instead of his birth weight and gestational age \_\_\_\_\_ [00:47:56], and highly correlated, is that I took, what I did was I specified the model with birth weight 100 gram intervals, and and gestational age in one week intervals.

I actually did separate birth weight dummies for different group, types of groups, singleton males and females, and multiple births because they were different. And when I did this, even though I had this serious collinearity problem, the maximum condition index dropped to less than eight. So there's no serious collinearity problems.

And so what I was, what was happening is that with each, in each of those groups, within each of those 100 gram birth weight intervals those infants were distributed across two or three gestational age groups. And so that created enough variants that with these binary variables, it could sort it out that it couldn't sort it out with a continuous variable across the whole range.

So that is another way of getting at this problem. I'm just, this is just showing you using these binary variables, sort of what those 100 gram birth weight intervals look like. But the point is, is that there is more than one way to transform your data. And if you have to think, you may have to look at your data carefully and experiment. But by doing this, you can identify ways to transform the data so that you don't have these collinearity problems, and you can get good parameter estimates, and reasonable standard errors.

The bottom line, and this is a quote that my colleague, Jack Needleman, likes to recite all the time, "Know your data." Before you even start running regression, run lots of diagnostics to understand the data, look at the data, look at the interrelationships with the data. And the more you understand your data, it will help you to properly specify your model and alert you to potential problems.

And so that you can address them, and then you can go through these steps. But too, all too often we get our data, "I've got my data set up. I want to run my regression and get my output." And that, and you get excited about it. But if you don't have a properly specified model, the model, the, as I've shown in some of these examples, you can get very different answers. And some of those answers are probably wrong if you don't properly specify your model. So I can't emphasize how important and neglected this is as a part of the modeling process.

I'll just note, it's a really good book on regression diagnostics. It's really old, but it's, there, or I'm sure there are newer texts. And I'll take questions and just note that the next lecture is me again talking about limited dependent variables next week. And are there any other questions, Joan?

Josephine Jacobs: Yeah yeah. Thanks, Ciaran. That was great. So we had one question earlier when you were talking about the aggregation problem. They wanted to know if this is also what epidemiologists called Simpson's paradox?

Ciaran Phibbs: I am not sure because I don't know what it they – it, it could be.

Josephine Jacobs: Okay.

Ciaran Phibbs: But yeah, you can get, with aggregation you can get results that are very different. And it's because you're masking differences.

Josephine Jacobs: And we have had somebody who wanted you to clarify something about the dummy variable approach again. They wanted to know, so for instance, if you were putting all of the dummies in at once for age, are you then putting the original variables for all the other continuous variables in the model? Or are those in dummy form as well?

Ciaran Phibbs: I I, those should be in dummy. You need to do it in dummy, you need to, all of the non – you need to do all of the non, non-binary variables as sets of dummies at once. And I addressed this with the previous question too, because if you don't do that, if you mis-specify the other variables, you could get biased estimates for the dummies, for the one you're looking at.

Josephine Jacobs: Thank you, Ciaran. Somebody asked, so the option you referred to from proc reg, does proc reg allow for clustering? If not – ?

Ciaran Phibbs: No.

Josephine Jacobs: – Do you run collinearity tests for only one cluster at a time?

Ciaran Phibbs: The clusters affects the standard errors or not the parameter estimates. So that isn't a concern for looking at the the collinearity.

Josephine Jacobs: Thanks. That's it. I think that's it for now. I guess if people have follow-up questions, maybe they can reach out to you directly. But I'm not seeing any more come in.

Robert: I'm not seeing any in the chat, either at this moment. But attendees if you have questions, feel free to put them into the questions panel. Perhaps….

Ciaran Phibbs: You can also contact me directly. I didn't put my e-mail down but it's Cphibbs at Stanford dot edu is the best way to reach me.

Robert: I'll put that in the chat. Do you want to go ahead and make closing comments?

Ciaran Phibbs: Just to reiterate what I said at the very end about know your data, and carefully test all of these assumptions . And it takes work but it can result in – more than once in the work I have done, it's caused me these careful tests, have caused me to change how I specified the model, and meaningfully change the results as as a consequence. And so it does matter or it can matter depend, it can matter depending on the structure of your data.

Josephine Jacobs: Thank you for your time, and for preparing, and presenting today, Dr. Phillips and Dr. Jacobs. Attendees, when I close the webinar, a short survey will pop up in a separate web browser. Please take a few moments and provide answers to those questions. Thanks, everyone.

Ciaran Phibbs: Alright, bye.

[END OF TAPE]