Cyberseminar Transcript

Date: June 20, 2018

Series: HERC Health Economics Seminar

Session: Modeling Longitudinal Semicontinuous Expenditures: A Practical Guide

Presenter: Maren Olsen, PhD ; Valerie Smith, DrPH

*This is an unedited transcript of this session. As such, it may contain omissions or errors due to sound quality or misinterpretation. For clarification or verification of any points in the transcript, please refer to the audio version posted at* [http://www.hsrd.research.va.gov/cyberseminars/catalog-archive.cfm](file:///C:\Users\VHASLCMyersK\AppData\Local\Microsoft\Windows\Temporary%20Internet%20Files\Content.IE5\73RXHSNF\l)

Risha Gidwani-Marszowski: Good morning, good afternoon everybody, depending on your location. I'm Risha Gidwani-Marszowski. I'm one of the economists here at HERC, and today I have the great pleasure of moderating what is sure to be a fascinating seminar on Modeling Longitudinal Semicontinuous Expenditures, a Practical Guide. We have the benefit of having two researchers here to give us the presentation. The first presenter is Dr. Valerie Smith who is a biostatistician and an investigator at the Durham VA Medical Center, the Center of Innovation there. She’s also an assistant professor at Duke University in the Department of Population Health Sciences. Her methodological research interests focus on methods for zero-inflated and semicontinuous expenditure data. She has many years of experience working with VA utilization and cost data in research studies. The next presenter will be Dr. Maren Olsen who is also a biostatistician and an investigator at the Durham VA Center of Innovation and an associate professor at Duke University, Department of Biostatistics and Bioinformatics. Her current method research interests focus on methods for complicated longitudinal data and heterogeneity of treatment effects and trials, and she has many years of experience working with VA patient outcome data. These two investigators are uniquely positioned to give us some valuable information about how to model costs, so I'm very happy to hand over the reins to them.

Dr. Valerie Smith: Thank you for that introduction and to HERC for inviting us to give this Cyberseminar. It’s a topic that we’re very interested in. And thanks to those on the line who are here today to listen. We’ll be talking about modeling semicontinuous longitudinal expenditures and hopefully taking a pretty practical approach to how you actually go about doing that.

So our agenda for today is to first review and compare strategies for analyzing what we’re calling semicontinuous healthcare expenditures that are collected longitudinally over multiple time points, so perhaps multiple years of data, multiple months, however you define your time point. We’ll go over model specification, what software is available to fit different models, and advantages and disadvantages of each approach. And I also want to point out that this talk is based on a manuscript that we published this year, a few months ago, and if you’re interested in anymore details in anything we cover that we aren’t able to cover in this talk, this manuscript is a good place to go for further reference. And it also includes SAS code to implement all the methods that we’re going to talk about that may be helpful.

Our goals for this Cyberseminar are to review some well-known strategies for analyzing expenditure data, also to introduce strategies that may be less well known, and then to provide a roadmap for when to use which approach, so that’s when it gets really practical, and review frequently encountered complications that may arise when you try to fit these models, because they’re not all that easy, and how to overcome them.

So before we get started, we wanted to get a feel for what the audience’s experience was with this type of data. So our first poll question is what is your experience with analyzing longitudinal healthcare expenditures? None or planned for upcoming work, some or less than one year, moderate or one to five years, or considerable, so more than five years.

Moderator: Responses are coming in. We’ll give everyone just a few more moments to respond before we close the poll out and go through the responses. And it looks like we’re slowing down here, so I'm going to close the poll. And what we’re seeing is 36% of the audience saying none or planned for upcoming work; 14% of the audience saying some, less than one year; 36% saying moderate, one to five years; and 14% saying considerable, greater than five years. Thank you, everyone.

Dr. Valerie Smith: Thank you. That sounds like a good mix of experience in the audience. So I'll start out with going over what exactly do we mean by semicontinuous data. So, and this is in the framework of healthcare expenditures, but it could be used for other types of data. So it’s pretty much any data that’s characterized by two defining features. The first component is often referred to as the binary component, and that’s pretty much a clump of zero values that represents a subgroup of nonusers. So these are the people who don’t have any healthcare utilization in whatever time period you’re looking at and therefore have no healthcare costs in that time period. And then this is coupled with the second component, often referred to as the continuous part that’s comprised of the positive values. So we have a continuous distribution of positive values among users. These are people who have some varying amounts of utilization. They could have a simple primary care visit or pharmacy fill if we’re thinking about any expenditure, or they could have a much large expenditure such as an inpatient stay. So we have varying levels of cost, and one thing we typically see when we’re looking at healthcare expenditure data in this distribution is that it’s very highly right-skewed with that long right tail. Most people don’t have very high expenditures, but you have some who do with those inpatient stays, things like that, the low probability events that are high cost.

And that leads to some modeling issues when we’re trying to deal with data that look like this complicated distribution. First, as I mentioned, the highly right-skewed distribution precludes just direct linear modeling. There’s two primary ways to kind of get around the fact that you have this very skewed distribution. The first is to use a link function, often a log link, but you can use square root link, things like that, in a generalized linear model, and that kind of helps pull in that tail. The second is to transform the data, similarly often a log transform, prior to using a linear model. So that suggests taking the log of your actual outcome values before you input it into a more standard modeling approach.

In addition, when we have this big point mass of zero, we have to decide how we want to accommodate the zero values. If we do the log transform option and want to use this more standard linear model, you have to add a small constant first if you’re just going to use a standard model. That’s just because you can’t take the log of zero. So you see people sometimes add one dollar or five dollars depending on how they want to apply that. If you’re using a GLM, you could just choose to not treat the zero values as any different than the others. The log link technically accommodates that. And a third option that we’re going to spend a fair amount of time talking about today is you can utilize two-part models to separately account for these two somewhat distinctive components.

So that’s generally covering semicontinuous data, and then you have the additional complication of when we’re talking about longitudinal data. Any time you have longitudinal data, you need to account for the correlation among repeated measurements on the same individual over time. The two primary ways you see of doing this is to include a random effect for each individual in what we refer to as mixed models, or to include a working correlation structure and fit the model with generalized estimating equations, or GEEs, as a secondary way of accommodating this correlation structure. However, there are some specific issues to semi‑continuous longitudinal data that are above and beyond standard longitudinal data or cross-sectional semi-continuous data. The primary one is that you need to account for correlation across the two components over time, and what we mean by that is that if you have a zero expenditure at one time point, that could be correlated with the level of expenditures incurred at another time point.

So it could be that a individual is very healthy, and they don’t have a lot of expenditures, and when they do, they’re very minor. It could also be that an individual is avoiding getting preventative care or going to the doctor for any reason, and then when they do, they end up in the emergency department, in which case you have a zero value to be correlated with a very high positive value. And that needs to be appropriately accounted for with this data structure. Additionally, the distribution and proportion of the positive values is going to depend upon the time frame under consideration, and that’s just because the longer your time frame, the more people have an opportunity to incur a positive cost and get out of that zero bucket in the first component. So more people will have visit expenditures in a year than they would say in a month. So these are the general frameworks we’re thinking about when we’re about longitudinal semicontinuous data.

As an example to kind of root this in real data, and we go through example in a lot more detail in the paper than we will in the talk today, but we use this study as an illustration looking at VA specialty care expenditures. In December of 2001, the VA increased specialty care visit copayments from 15 to 50 dollars, so a pretty big jump, and this allowed a natural experiment to examine how outpatient expenditures change due to a change in copayment. For the study, specialty care expenditures were calculated for each year from 2000 to 2003, which was two years prior to and two years following this copayment change, and there was a convenient control group available because Veterans who are exempt from copayments due to low income or service-connected disability didn’t experience this copayment increase while everyone else did. So for the original study, they did one-to-one propensity score matching resulting in a sample of 1,693 Veterans who were exempt from copayment and the same number who were required to pay who did experience that increase for comparison.

So to get into what this semicontinuous data really looks like, we can see over on the right-hand side of the screen histograms of these specialty care expenditures by year. We can see over here this big jump. Those are the zero values that we see, and then the tail goes out up to as much as 60,000, so that’s that highly right-skewed distribution. And looking in the table for specific values, we can see that the percentage of zeros ranged from 23% to 29% by year, so that’s a pretty substantial proportion of the sample having zero expenditures. We see that the median is typically between two and three hundred each year, so not huge. But then looking at the maximum, we see it goes up to as much as almost $60,000 but typically in the $20,000 range. So that’s again just reflecting that we have a small number of people who have very high costs leading to that long tail of the distribution that we need to account for.

So the four main strategies that we’re going to review today for analyzing data that looks something like this, the first is a one-part generalized linear model, and this is kind of described as treating the zero and positive values pretty much the same. It doesn’t distinguish them but treating them as coming from the same distribution. We’ll also talk about two-part uncorrelated generalized linear models. This separates the two components, looking at that clump of zero values and the positive distribution separately, but this requires strong, often unrealistic assumptions to be valid. And then we will move on to looking at correlated conditional two-part models that similarly separate the two components but don’t rely on such strong, unrealistic assumptions, with the second component being conditional on having a positive expenditure. And then lastly we’ll talk about correlated marginalized two-part models that sort of blends the one-part model and the correlated conditional two-part model by explicitly incorporating zero values but having the target of inference be the mean of both components as opposed to just the positive component.

So briefly to review the one-part generalized linear models, this may be more familiar to a lot of people. This just assumes the observed expenditures are from a single distribution and used as a link function to deal with the fact that we have that highly right-skewed distribution. We write it here with a log link because that’s what’s most commonly seen, but any link function that fits the data appropriately could work. And we just have the single mean function. It’s a pretty standard-looking GLM model, and it’s fit using generalized estimating equations coupled with empirical sandwich variance estimation. These are often referred to as robust standard errors. We’ll talk a little bit later how they’re not always robust, and so we just sort of stick with calling them by their formal empirical sandwich variance estimator’s name.

Advantages to one-part GLMs are that they’re very easily implemented in most statistical software, just sort of out of the box packages [unintelligible 13:31] use directly. SAS, you can do this with PROC GENMOD using a repeated statement or with PROC GEE. In Stata, you can use xtgee, or R you can use geepack. And I'm sure there are many others, but these are just to name a few. And the single component with the link function allows for very simple estimation of population-average effect on the original dollar scale, which is often what policymakers may be interested in when you’re trying to communicate these results outside of your research team. When a log link is used, exponentiating a parameter represents the multiplicative effect on the overall mean associated with a one-unit increase in the corresponding covariate. So this just provides a nice intuitive interpretation, easy to implement approach.

However, there are limitations to being able to use this simple approach. It’s not really suitable for data containing a significant proportion of zeros at any given time point, especially when your sample size is a little bit smaller. Simulation studies have shown significantly biased covariate effect estimates and high type one error rates. The type one error rate should be around 5% for most standard analyses, and simulation studies have shown that it can get as high as 50%, so very inflated if it’s fit to data containing 20% or more zeros. So the reason for this GEE methodology is asymptotically unbiased, but when we say asymptotically in the context of this very complex distribution, the sample size needs to be really, really big, so typically more than 50,000 when you’re talking about this many zeros. And the robust standard errors that are typically utilized with GEEs really don’t help to overcome this. They’re also asymptotic and also require this really, really large sample size. So in the context when we don’t have that, we need to look at perhaps the two-part model that’s more explicitly taking into account our data structure.

So I just want to pause here to go over kind of what we mean when we’re talking about separating our data into two components. So over here on the left with the purple graph we have, this is just the raw data as you would see. Some time points have positive values, and then other time points, this person with their fake data incurred no expenditures. And so to take this single stream of data and think about recoding this for how we want to think about it for a two-part model, we can look at this bottom part down here for the binary part. That just turns into zeros or ones depending whether there’s a positive response. So everywhere there’s a positive response in purple, we have a one in the red binary part. And for the continuous part, we pretty much just omit the zeros and take the positive values. And so what you can see is this induces a missing data type situation in the continuous part. Those zeros are now missing, and that’s okay, but it needs to be handled appropriately.

So the simplest two-model that we can review for longitudinal data is the uncorrelated two-part GLMs that typically via GEEs, and so the way this works is we just have two GLMs. The first is for the binary part and the second for the continuous part. The first one is pretty much just logistic regression, looking at whether or not there’s a positive value. The second component looks very similar to the one-part GLM using a log link, typically. The difference here is that the mean is conditional on having a positive outcome. We didn’t see that in the one-part GLM. And because it’s fit with GEEs, each part separately accounts for correlation among the repeated measures over time and fit via GEEs similarly to the one-part GLMs.

However, this approach has some pretty significant limitations, the first being interpretation. The binary component, that’s pretty much okay. That’s the population-average estimate of the probability of incurring positive expenditures for the entire sample at all time points. But assuming you’re also interested in that continuous component, that’s estimating the mean level of expenditures among a subset of individuals who incurred expenses at each time point. This means that the target population is changing depending on the time point. That’s rarely what you want when you conduct a model. You kind of want your table one, you want to see what your population is. You don’t want that population to change time point by time point depending on who has utilization at that time point.

Additionally, we kind of talked about that inducing the missing data when you take out the zeros, and the issue is that with this approach is that that really doesn’t account for the fact that you induce this missingness, and that’s because it’s pretty frequent that you’re going to have that correlation between whether or not there’s a zero and therefore whether or not it’s missing in that second component, and the level of expenditure at other time points. And if that’s not accounted for, then you get biased results, and that’s been shown in the paper reference there by Su, Tom, and Farewell from 2009. So this has a simple approach, but we can’t recommend it because it’s an awkward interpretation of who your target population is, and even if you’re okay with that interpretation, your results may be biased, and they may not even be correct.

So that moves on to how can we more appropriately handle the fact that we’ve sort of induced this missing data in that second component? And one popular approach for doing that is the correlated conditional two-part model, we’ll just call CTP for short here, and this uses two mixed-effects models, so we’re moving from a GEE framework to a random effect mixed-model framework. The first one is for the binary component and the second one for the continuous component. So this has a lot of similarities to the other two-part model. The first part is sort of a logistic regression with a random effect, just modeling the probability of a positive outcome. The second component is again conditional on observing a positive response, and in this case we’re taking the log transformation of those positive values and then using a linear mixed model for those log transform values. The reason that this works when the GEE approach with the uncorrelated part doesn’t is because this is incorporating those two random effects, and by doing that, we’re able to link the two components of the model. So these random effects are assumed to jointly follow a multivariate normal distribution. This both accounts for correlation among repeated measurements over time and the correlation among the two components of the model, so we no longer have this sort of informative missingness in the second component problem that we did in the uncorrelated model.

So the advantage to this approach, as mentioned, it’s a fully parametric model fit with maximum likelihood, so that allows us to have a bit more of a flexible missing at random assumption for the missing data. The GEE assumption for missing data is much stronger. Additionally, we get estimates of these variance components from the random effects to help quantify how the two parts are related, and that’s often of interest in and of itself. So for example, if you have a positive estimate of covariance between the binary component’s random intercept and the continuous component’s random slope, that would tell us that the probability of any expenditure is positively related to the amount of expenditures over time. So those who are more likely to have an expenditure are also occurring more. And that can be interesting from a policy perspective. This can be fit in standard software packages. It’s not really available in like out of the box like PROC GENMOD, that kind of thing, but it can be fit with PROC NLMIXED or Mplus, and the code isn’t that hard to adapt.

Some limitations of this model is that this more complicated random effects structure can be computationally challenging, which can be a little bit harder to converge and take longer to run. Additionally, that second component is conditional on positive outcomes, and because we did a log transformation and include the random effect, it’s not easy to convert predictions from the log dollar scale back to the dollar scale, which is often more interpretable and what people are really interested in. And also because we include random effects, that means that the estimates have a subject-specific interpretation, and that’s true with any random effects model, but we’re just pointing that out as a difference with this model.

And so then the last strategy we’re going to review is the correlated marginalized two-part, or we’ll call it MTP model, and this sort of blends the marginal interpretations from the one-part GLM with the structure of the correlated CTP model, the conditional model. The first component of this is the same as the correlated conditional two-part model. It’s a random effects model, just modeling the probability of having a positive outcome. The second component is where it gets a little bit different. This is going back to taking a more similar structure to that one-part GLM where we’re using a log link and modeling the overall mean, not conditional on having a positive value. The difference with the GLM is again that we have these random effects in the two components, and we’re linking them so that we’re accounting for the fact that these two components are correlated over time. Similarly to the CTP model, we assume that these random effects jointly follow a multivariate normal distribution, thus accounting both for correlation among repeated measures over time and a correlation among the two components of the model.

Advantages to this approach compared to the one-part GLM, it allows estimation of both the probability of use component and the covariate effects on overall mean expenditures, which may be of interest to policymakers or investigators, people who maybe would really like for you to do that one-part model except that it’s not going to be accurate if you have too many zeros. The parameter interpretation from that first component, if you exponentiate the parameter, it represents the subject-specific odds ratio for incurring positive expenditures associated with a one-unit increase in that corresponding covariate, and that’s the same between the MTP model and the CPT model.

The differences in the second component if you exponentiate a parameter there, it represents the multiplicative effect on the overall mean expenditures of the entire population, not just those who have expenditures associated with a one-unit increase in the kth covariate. And because of use of the log link, this has a dual population average and subject-specific interpretation, assuming that kth covariate is not a random effect itself, and that’s typically not the case. Usually when you’re doing this, you’re interested in the effect of an intervention, treatment, policy, something like that, and that’s typically not included as a random effect. You’re often including a random intercept, maybe a random float for time, and so that gives all the other covariates their parameters or corresponding kth covariates, that dual interpretation.

When it’s implemented in SAS PROC MCMC, which is how it was originally proposed, any value calculated from the parameters can be easily obtained with corresponding credible intervals or highest posterior density intervals, which are kind of a Bayesian analog to confidence intervals and can be interpreted similarly. And we originally proposed this with non-informative [unintelligible 25:53], so you should pretty much get the same estimates as you would with maximum likelihood, but it kind of adds a little bit of computational flexibility because one of the limitations is again this is computationally challenging. That cross-part correlation with the random effects makes these models a lot harder to fit and can lead to long run times and convergence issues, and it’s not as straightforward in standard software. We provide in that paper referenced at the beginning the SAS PROC MCMC code, and I also have NLMIXED code for fitting it with maximum likelihood that I would be happy, if anyone was interested to e-mail me, to send out. But it’s not incorporated in any out of the box standard procedure in any software.

So before we move on to sort of the practical roadmap, at which time my colleague, Maren Olsen, is going to take over presenting, we’re interested in getting what your experience is on any of the four methods presented, having used them before. So this is a check all that apply. Have you used a one-part GLM, the uncorrelated two-part GLM, the correlated conditional two-part model, and the correlated marginalized two-part model previously?

Moderator: And again, we’ll give everyone a few moments to respond before we close the poll out and go through the results. And it looks like we’ve slowed down here, so I'm going to close the poll, and what we’re seeing is 77% of the audience saying that they have used the one-part GLM, 55% of the audience two-part GLM, 5% correlated conditional two-part model, and 5% correlated marginalized two-part model. Thank you, everyone.

Dr. Valerie Smith: Thanks.

Dr. Maren Olsen: Thanks, Valerie, for that fantastic overview of all four of those methods. So as a collaborative statistician in the VA, when work on research projects and then presented with data that looks like this, a very first step of course is how to figure out which of these methods would be good to use. As a very first step, we recommend using descriptive statistics and plots to understand what your distribution looks like at each time point, similarly to what we showed you earlier in the presentation with the specialty care expenditures example. So it’s important to understand what are the percent of zeros at each time point and what is the degree of skewness and extreme values at each time point. We also want to understand what is our overall sample size and, importantly, what are the research questions of interest. You’ll see as we go through kind of the different scenarios of choosing which method is that they all answer actually slightly different questions, which I think are important to keep in mind.

So we have three fairly common scenarios that are probably most appropriate for choosing a one-part GLM as your primary analytic tool for this kind of data. The first scenario is when your sample size is in the smaller range, 200 to 1,000, and your percent of zeros is less than or equal to 10% at each time point. In the second scenario, we’re moving into the kind of medium to larger sample size at each time point, and that increase in sample size gives us some flexibility with increasing the percent of zeros at each time point to 10 to 15%. And then finally when your sample size is very large, say over 50,000 Veterans at each time point, then a one-part GLM could handle accommodating percent of zeros that are in the 15 to 30% range. Examples of research questions that could be answered with the one-part GLM include what is the effect of being required to pay a copayment on overall mean specialty care expenditures. Remember that the one-part GLM is not distinguishing between zero and positive costs. And another example could be what are the estimated or predicted overall mean expenditures for those with and without a copayment requirement in each year?

So if the data that you’re seeing falls into one of these three scenarios and the research questions that you want to answer fall into these categories of kind of looking at an overall mean, then a one-part GLM is the direction that I would go.

For scenarios that are more appropriate for pulling your data into two parts and using either a correlated MTP or CTP model, we then look at scenarios where you have a smaller number of patients but more than 10% zeros at each time point and a kind of medium range of subjects in your datasets and say over than, greater than 15% zeros at each time point. And finally the last scenario, even if you have a very large number of subjects that have over 30% zeros at each time point, then you really need to consider separating your data into two components and fitting a two-part model. The choice between using a correlated MTP or a CTP really depends upon your research questions and estimated quantities of interest, and I'll go into that in a little bit more detail, as well as the computational challenges of model estimation.

So, in comparing kind of the research questions that you could answer with those two different models, as Valerie mentioned, the binary part of those two models are parameterized equivalently. So the research question that you would answer with either the CTP or the MTP with the zero/one portion of the data would be the same. So an example is in each year, what is the probability of incurring specialty care expenditures for an individual required to pay a copayment as compared to if that individual was not required to pay a copayment? Then where those two models differ really is in the continuous part, right? For the conditional two-part model, not surprisingly, your research question is really conditional upon having a specialty care visit. Is there a difference in specialty log expenditures in each year for an individual required to pay a copayment as compared to an individual not required to pay? Notice that that interpretation included log expenditure, so the research question is being answered on the log scale, not the original dollar scale.

For the MTP, then, the continuous part of the model is giving us an estimate of the overall mean, and so our corresponding research question for that could be what are the estimated overall mean specialty care expenditures for those with and without a copayment requirement in each year? Additionally, as we noted earlier, you have some additional information that you can get from the variance components of the random effects for each of those models. In the correlated two-part model, you would be able to answer how the probability of any expenditure, that is the binary part, is related to log positive expenditures over time, whereas in the MTP covariance components of the random effects are going to provide information for you about how the probability of any expenditure would be related to overall mean expenditures over time.

Notice that we have used throughout the presentation so far a log on transform for our models, but there’s also additional kind of popular distributional choices. You could think about doing a log-skew-normal, and there are a few citations going into that in more detail. Additionally, Lei Liu has a fantastic JHE paper in 2010 where he uses a generalized gamma distribution and takes a Weibull and gamma and log-normal distributions as special cases. So we wanted to make sure to point out those additional citations for you.

So as we’ve alluded to several points throughout our talk, both the correlated CTP and MTP models are computationally challenging to fit, even from people like [laughs] Valerie and me who have expertise in how to fit them. You have to be really patient and try a variety of things. The correlated random effects does make it difficult for them to converge. In this day and age with very powerful computing needs, we’re not usually used to having procedures take many, many hours, sometimes over a day to converge, but these models can sort of fall into that situation. So a maximum likelihood solution as implemented in PROC NLMIXED can easily take that long. Both of these models can be fit with Bayesian methods, and our recent paper does have PROC MCMC code for the MTP model. Additionally, there’s a paper by Cooper from about 10 years ago that provides WinBUGS code for the CTP. Extreme skewness in the positive component also impacts model fit and convergence.

So we do have a number of tips that we have found to be really helpful when trying to use some of these newer strategies or I should say methodologies that those of you in the audience haven’t used as much. The first step is that even though we don’t recommend finalizing your model fit with an uncorrelated model, it’s often best to actually estimate the two parts of your model with this uncorrelated model to make sure that simple model converges. That way you’re kind of making sure that any issues that you have in the data would get worked out before you put that correlated random effect structure on there, because sometimes a model fit issue might actually have nothing to do with the convergence of the random effects. There might be something else going on, and you want to diagnose that before diving into one of these more complicated modeling strategies and waiting 24 hours to find out that [laughs] something isn’t working. Assuming, then, you get the uncorrelated model to fit, then what you’ll want to do is use the final estimates, the parameter estimates from that model then as starting values for your correlated model estimation. That will make things go much more efficiently.

Additionally, we have found it useful to try different software options. So because the underlying maximization and estimation algorithms differ slightly between those. There’s actually a really nice SAS white paper that includes some tips and strategies for mixed modeling that we have found to be really helpful when working with PROC NLMIXED. We have actually also used some freeware that, I think it’s called AML, that isn’t really maintained anymore, but again it uses some different maximization techniques, and we have found that will lead to convergence when NLMIXED won’t. We also will explore trying a different distribution choice, and I also cannot recommend enough too taking a more simple approach and actually just simply modifying your data or your research questions than to go into a scenario that’s more computationally stable. So for example, if you were originally looking at three- or six-month time periods of observations and had 30 to 40% zeros, that then you could increase your time period of observation up to a year. You might decrease the percent of zeros into the less than 10% range, which would allow you then to move into the one-part GLM territory. Additionally, one of the reasons why this is so complicated too is because of the longitudinal nature in that you’re maybe looking at kind of the repeated observations over time. One option could be to just simplify your study design to be a pre/post rather than repeated observations over time.

So I would say don’t forget to take a step back and think about if there are some kind of big picture changes that you can make if you’re running into really challenging computational situations.

So we’re at the end of our prepared slides. In conclusion, you can see that the model selection process is driven by both data structure as well as research goals. We have found that two-part models are often necessary when data have at least 10 to 15% zeros at any given time point and that two-part models also accommodate a variety of research questions. We have emphasized that correctly modeling longitudinal expenditures with many zeros can be computationally complex, but a lot of papers do provide code, and that has been incredibly helpful. We also want to point out that this is an active area of statistical methodology research and that that is also exciting in that the VA is kind of poised to be able to use these new innovative methods. We also at the end of the talk provide the citations that we have mentioned throughout the presentation for your reference and further reading.

And then finally, here are e-mail addresses for Valerie and me, and we are happy to take your questions via e-mail or now at the end of the presentation. Thank you so much.

Risha Gidwani-Marszowski: Thank you, Drs. Olsen and Smith. This is really very useful information. We do not have any questions right now from the audience. I will encourage all audience members to type in your questions if you have any. While we’re waiting for some of those to flow in, I have a couple of questions that I’d like to ask. So if we can go back to the correlated conditional two-part model, you mentioned in the beginning of the presentation that there is a missing at random assumption in that model, and I'm wondering how reasonable that assumption is. So when you’re modeling healthcare costs, aren’t zero values really due to a different health status, maybe a better health status, and that’s not really a random occurrence; that’s sort of a systematic occurrence.

Dr. Maren Olsen: Okay, can you actually just repeat the question one more time [laughs]? Sorry, I was\_

Risha Gidwani-Marszowski: [Unintelligible 45:48]

Dr. Maren Olsen: \_trying to figure out how to get back to that screen.

Risha Gidwani-Marszowski: Yep, this is right slide. So this missing at random assumption\_

Dr. Maren Olsen: Okay.

Risha Gidwani-Marszowski: \_for missing data, or I would say zero values in this situation, right? Is that really\_

Dr. Maren Olsen: Mm-hmm.

Risha Gidwani-Marszowski: Do you feel that that’s a reasonable assumption or is that assumption really, or is the missingness really due to something systematic, like a systematically better health status that means that the patient does not need to access healthcare?

Dr. Maren Olsen: Well, I should say that, so the missing at random means that your missing data can be dependent upon anything that is included in the model, either previous observations of the outcome variable or covariates that may be predictive of that. So I should say that I feel like the missing at random assumption is kind of with the caveat that you are including that information. So if you feel like there is something that is systematically different about your zeros, and lots of times the zeros actually represent kind of like true non-zeros, as well as people who are kind of potentially going to incur costs. And these models do not treat kind of the underlying cause of the zero any differently. However, including kind of that additional covariate information could help meet that missing at random assumption.

Risha Gidwani-Marszowski: I see. Okay.

Dr. Maren Olsen: Yeah.

Risha Gidwani-Marszowski: So that actually dovetails well with my next question, which is in the models where they are correlated, the two parts are correlated, some, let’s say zero values in time one may be correlated with high costs in time two because the patient has uncontrolled diabetes, versus for\_

Dr. Maren Olsen: Uh-huh.

Risha Gidwani-Marszowski: \_another patient a cost of zero in time one may be correlated with no cost in time two due to good health status. So in this situation, as long as you had included the covariates about their baseline health status, would that be sufficient or would you also want to be including time-varying covariates about health status?

Dr. Maren Olsen: I can’t say that I've ever been brave enough to include [laughs] time-varying covariates in one of these models. I think I would be, I guess I feel like that’s a, because it’s a random effect, it’s subject specific, and that that kind of distribution of the random effect will allow for both of those scenarios. So it’s really kind of looking at individual data, and from that individual coming up with a distribution, so I feel like the distribution of your random effect could incorporate both of those scenarios just by those realizations coming from a normal distribution. Valerie, did you\_

Dr. Valerie Smith: Yeah.

Dr. Maren Olsen: \_agree? Valerie agrees. She’s nodding [laughs].

Risha Gidwani-Marszowski: [Laughs] Okay, great. We have one question from the audience, and that is\_

Dr. Maren Olsen: Okay.

Risha Gidwani-Marszowski: \_if the research question is whether a policy decreased expenditures per person, and the data is repeated cross-sectional, which method would you use?

Dr. Maren Olsen: Okay, so I’d like a little bit more information about what you mean by repeated cross-sectional.

Risha Gidwani-Marszowski: So she saying that\_

Dr. Maren Olsen: [Unintelligible 49:45]

Risha Gidwani-Marszowski: Access to large claims data and the population who uses healthcare changes in each cross-section.

Dr. Maren Olsen: Mm-hmm.

Risha Gidwani-Marszowski: So the data is not longitudinal at the person level, so it looks like it might be\_

Dr. Maren Olsen: Right.

Risha Gidwani-Marszowski: \_some sort of panel where the panel membership changes over time.

Dr. Maren Olsen: So in that case you don’t have actually longitudinal data, as the person posing the question points out. So then I would be then thinking about analyzing each year separately using a general linear model. And Valerie, what are the, what do you feel like are the best citations kind of for that, because usually then the challenge is figuring out your best link function, if you’re going to keep it one part\_

Dr. Valerie Smith: Yeah.

Dr. Maren Olsen: \_or doing a two part.

Dr. Valerie Smith: Yeah, I mean I would be happy, this is Valerie. I would be happy to provide that audience member some specific references on that. The same idea would hold if you have a lot of zeros. You would need to look at a two-part model. The two-part models are a lot simpler to fit when you aren’t having to incorporate that random effect structure, so that really frees you up a lot in what you do. The guidance on the proportion of zeros I would say stays the same. If you have not huge amount of zeros and/or a very large sample size, then you could fit a one-part model. There’s a paper by Manning and Mullahy from 2001, I think in Journal of Health Economics, I'm not sure, that has a lot of good guidance on how to fit that kind of model. And then otherwise there’s, you can have the conditional two-part model or the marginalized two-part model both, just without the random effects. And they sort of answer similar research questions to what we covered here, and there’s code for both of those in various papers that I could point to you if you just want to e-mail me.

Risha Gidwani-Marszowski: Okay. All right, great. All right, so we’ll just have that audience member follow up with you individually. No other questions from the audience, so I will take the opportunity to ask another question. On one of the last slides you all mentioned that you could use final estimates from an uncorrelated model as a starting value for the correlated models to try to do a proof of concept to make sure the uncorrelated models were converging before you got into the more complicated correlated model structure. Can you expand a little bit more about what you mean by this? It feels a little Bayesian to me, but I feel like that’s not necessarily the case. We’re probably still on a frequency approach, and I'm not quite understanding how you’d use those final estimates to start a value.

Dr. Maren Olsen: Sure. So, yeah, thanks for that clarification. If you are using PROC NLMIXED, which does use a full likelihood approach, it actually requires that you specify starting values for each of your parameters that you’re going to be fitting, unlike PROC MIXED, which estimates them behind the scenes. And you can probably also with PROC MIXED specify starting values as an option; it’s just that people don’t often do that because it does it behind the scenes. But for PROC NLMIXED, that likelihood estimation does require that you specify starting values. So even with the full likelihood approach, I recommend that case. And then of course we also provided information about fittings, both the CTP and MTP, using Bayesian methods. And yes, you definitely need starting values to kick those off, to kick off that Markov chain Monte Carlo fit as well.

Risha Gidwani-Marszowski: Okay, great. Thank you.

Dr. Maren Olsen: Does that help?

Risha Gidwani-Marszowski: It does, yes. Thank you.

Dr. Maren Olsen: Okay, okay.

Risha Gidwani-Marszowski: So another question I have, so this log-skew distribution that you guys had mentioned versus using let’s say just a regular log-transform model, what level of skewness would you want to see in the data before you used a log-skew-model versus a regular log model?

Dr. Valerie Smith: That’s a good question. I don’t have a good feel for just looking at the data and knowing. I mean if you take the log of the data and do a histogram, you can kind of see if it looks normal or it still looks skewed. If it still looks skewed when you look a histogram of the log-transform values, the log-normal probably isn’t going to fit well. One of the nice things about both the log-skew-normal and the generalized gamma distribution is that they both take the simple log-normal as a special case. So if the log-normal really fits better, especially with the log-skew-normal, it adds an extra skewness parameter, and if that skewness parameter is, if you fit a log-skew-normal model and that skewness parameter is close to zero, then that just means that you really have a log-normal model, and you can simplify it.

Risha Gidwani-Marszowski: Okay, okay.

Dr. Valerie Smith: You can actually test that parameter, and if that skewness parameter isn’t significant, basically reduce your model to log-normal.

Risha Gidwani-Marszowski: Okay, great. And it looks like a lot of the programs to implement the CTP and the MTP models are in SAS, maybe in Mplus. Are there any Stata programs that you all know of you could recommend for implementing these types of models?

Dr. Maren Olsen: Neither Valerie or I know that, but I will be honest with you and the audience that coming from a statistical background that I am just not as familiar with Stata, but I have not seen that kind of coming in any of the literature.

Risha Gidwani-Marszowski: Okay.

Dr. Valerie Smith: I agree. I’ve seen [unintelligible 56:40] packages for cross-sectional conditional two-part models. I haven’t personally been aware of any for longitudinal models, and I feel pretty certain that there isn’t one for the marginalized model because that’s much newer.

Risha Gidwani-Marszowski: Okay. All right. So in the event that somebody wanted to operationalize these models, would SAS then be the recommended program to do that in?

Dr. Valerie Smith: Yeah.

Dr. Maren Olsen: Yeah, that’s where a lot of the manuscripts actually do have sample code in SAS, so that’s probably where I would start.

Risha Gidwani-Marszowski: Okay, great! If there’s any other questions from the audience members, please write them now. If not, I think we’ve got both e-mail addresses at the end of the slide deck from the presenters. This was really fascinating, very, very useful, especially for those of us that are kind of working in the trenches of cost modeling. It’s really nice to know some of the pitfalls associated with different approaches and also importantly, how to make the right inference based on the model that you’re selecting. So my great thanks to you both. This was really fantastic, and we’re very honored that you were able to present in our HERC Cyberseminar today.

Dr. Valerie Smith: Well, thank you.

Dr. Maren Olsen: Thank you. Great questions, and I'm looking forward to hearing if there’s any more.

Risha Gidwani-Marszowski: Wonderful. Thanks so much.

Dr. Maren Olsen: Okay.

Moderator: Just one final note for the audience. When I close the meeting, you will be prompted with a feedback form. Please take a few moments and fill that out. We really do appreciate your feedback. Thank you, everyone, for joining us for today’s HSR&D Cyberseminar, and we look forward to seeing you at a future session. Thank you.

[ END OF AUDIO ]