Cyber Seminar Transcript  
Date: 05/27/15  
Series: HEC  
Session: Cost as the Dependent Variable Part 2

Presenter: Paul Barnett  
 *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.*

Paul Barnett: My name is Paul Barnett. I am a Health Economist at the Health Economics Resource Center. I am going to talk today about a Cost as the Dependent Variable. What happens when you are doing a statistical analysis and you are using cost?

What do we mean by healthcare cost? Well, it could be at many levels. At the simplest level, it is just that one of the products, the cost of one of the products that is being used in delivering services; like a specific x-ray, a day of stay admit in the Operating Room, or a dispensed prescription. Or, it could be a cost of a bundle of products. All of the things that are being used to produce a particular outpatient visit, a hospital stay, or a treatment episode. All of the visits and stays for a particular type of patient over a period of time or a patient with a certain disease over a period of time. Another way of thinking about costs might be the annual healthcare costs incurred in a specific year or maybe in the participants in a study over the period of the study.

Here is an example of a cost to data from VA, a little dated now, four years ago – or, I guess five, really. This is the per person cost of all of the people he used at the Veterans Health Administration in the Fiscal Year '10. As you can see, a lot of people are up on the left-hand side with costs under five thousand dollars. But there are some very significant people out on the extreme right end of the tail. In fact, we did not show the part of the costs that are greater than thirty thousand dollars, but it is significant that people with thirty thousand dollars and above a year constitute a significant fraction of the percent also of the costs.

If we do some descriptive statistics on these data, we can see that the mean is quite a bit higher than the median mean, about six thousand dollars and median less than two thousand dollars. We also see the standard deviation is quite big. Then there are these two statistics, skewness and kurtosis, which are measures of the shape of the distribution. Skewness has to do with a right tail and the – excuse me, the symmetry of the distribution in kurtosis. How high the peak is. The skewness and what statisticians call the third moment of measure of symmetry. In a normal distribution it is zero. Here in our case, we have a positive skew, more observations to the right. Skewness fit statistic was 14. Our kurtosis was 336.

In a normal distribution, it would be three. Our peak distribution and the thickness of the tails are very much different from normal. It is what that kurtosis statistic says. Why is this? Well, the skewness has to do with rare but extremely high cost events. Only some individuals might be hospitalized in a year. Or some individuals have a very expensive event or expensive chronic illness. They are the ones that are the ones that are responsible for those right tails. We talk about positive skewness being evidence of skewed to the right. Now again, Heidi's help here in talking about the – ask you what matters in terms of the central tendency? What do we care more about? Do we want to measure the mean or the median?

Unidentified Female: Just click on the radio button next to what the answer you prefer. I am watching them come in. The answers are still coming in. I will wait for things to slow down a little bit before going through with the responses here. It looks like things are slowing down. I am going to close it out. We are seeing 33 percent of the audience saying mean and 67 percent saying median. Thank you everyone for participating.

Paul Barnett: Well, that is very interesting. Median would be kind of a way to get around that skewness problem. But actually if I am a policymaker, someone who is trying to allocate resources, I really do care about the mean. Because I have to pay for those very expensive, high cost events. The mean is the thing that we are worried about. Even though it is subject to that effect of the skewness, just because we need to pay for those expensive patients. There is another attribute of cost data. It is that there is often zero values.

To illustrate this we have done a graph where we have taken all of the people who use Veterans Health Administration in the 2009 Fiscal Year. We said what was their cost at Fiscal '10? As you can see, there is something around 16 percent or so of the people have no cost at all. This is a feature of many cost analyses is that we have lots of zero observations.

Not only do we have this problem of the right tail, we also have a truncated distribution on the left in many studies. People who are enrolled but do not use. There we are, just what I said. Enrollees who do not use care zero values, it is a truncation on the left side of the distribution.

Now, I want to ask another poll question just to say well, so what sort of hypotheses are you folks interested in testing? This is another poll to consider. Heidi, are they allowed to answer more than one option?

Unidentified Female: Yes, you can select all that applies. The options that we have here type of treatments, quantity of treatment, characteristics of patients, characteristics of provider or other. Because there are more one things that people can answer, the responses are coming in a little bit slower than usual here. But we will give you all just a few more moments before we go through the results.

It looks like we are good there. We are seeing 82 percent saying type of treatment; 65 percent saying quantity of treatments; 85 percent, characteristics of patients; 65 percent, characteristics of provider; and 29 percent to other. Thank you everyone.

Paul Barnett: Great. Well that gives me an idea of what sort of examples we might use in the discussion. The first thing I want to do is just talk about the Ordinarily Least Squares regression model Ordinarily Least Squares regression model, the classic linear model. Because we are going to sort of build on this knowledge. We have a dependent variable. We assume that in this case costs, that it could be expressed as a linear function of chosen independent variables. That X might be the thing that you are interested in investigating, the type of healthcare, or the quantity of healthcare, the characteristics of patient and provider.

All of those things predict a cost. What the Ordinarily Least Squares does, it estimates parameters and sometimes coefficients, the alpha and beta in that equation. It does so by minimizing the sum of the squared errors, that is the minimum distance between data points and a regression line, a fitted regression line. How well can we predict the data from how close it is by a finding a line that best fits it. The linear model that has cost as a regression variable – excuse me – as the dependent variable, it looks like this. It has this epsilon on the error term.

The beta in this case is interpretable in raw dollars. It represents the change in Y for each unit change in X. For example, if beta is ten and X is – the patient's age, there is a ten dollar increase for each unit increase in X in age. On the other hand, if X were say a type of treatment, did the patient get a particular type of treatment or not? X is a zero or one variable, then X is the effect. The beta is – excuse me – is the effect of having X set to one, and the ten dollar X for cost of having received the treatment.

Now, I want to talk a little bit about the expectations operator just because that is going to be important for the rest of the talk. If you are not familiar with this, this is a function. The expectation of a random variable, W, is simply the sum of the values of W times each values' probability of happening. Of course, probabilities are between zero and one. This is basically a probability of weighted average of that variable, W.

Just to give you an example; this is a little bit terse or dry definition. It is imagine the expected value from throwing a die. The die has six possible values, one through six. Each has equal one-sixth probability. We just sum up those values times the probability of each value occurring and get 3.5 as being the expected value from the whole die. This is how the expectation operator works on a random variable. We have that epsilon in our Ordinarily Least Squares termed, our error term. That Ordinarily Least Squares is built on the assumption that an expected value of that error term is zero.

The errors from different observations, which we have noted as I and J here, but they are not correlated. The expected value is zero. But the errors have identical variance. They are normally distributed. They are not correlated with the Xs, with our independent variables. These are the assumptions that need to be in place for us to use Ordinarily Least Squares. We find that many of these assumptions can be violated when we use a cost data. But most importantly, the assumptions that errors have identical variance; and that the errors are normally distributed.

The rest of today's talk and the talks I am giving next Wednesday really will be on how to handle this question of handling the non-normal data and data that has errors that are not identically – that do not have the same variance. Now just back to our Ordinarily Least Squares model for a minute. Say we have a dichotomous explanatory variable like the type of treatments. X is people who are in the experimental group and got the experimental treatment. Zero is the control group. They did not get the treatment. Beta is the effect of X. The predicted cost is we can estimate if X is zero, they are in the control group. Then it is just simply alpha.

The notation here is Y hat is the circumflex over the Y. It means a predictive value of Y. The bar means conditional on X being zero – is simply alpha. If they are in the experimental group, then the Y hat is a predicted value of costs. It is alpha plus beta when X equals 1, conditional on X being 1. That is a very straightforward interpretation analysis of variance really is this same thing, a regression with a dichotomous independent variable. ANOVA relies on Ordinarily Least Squares assumptions. Now oftentimes we want to find out well, what is the effect of X while we control for case-mix?

In this case, I am introducing a case-mix variable. That could be some health status. It could be age. It could be gender. Actually, you could have all of those things, of course in a multivariate regression. If we include the case-mix variable, often what people will do is calculate those conditional values. What is the predicted value of Y? Where they evaluate the case-mix variable at its mean value. The bar over the Z just simply means the mean value of Z. If we have and say, Z is age.

The mean age in the study is some people are 62 years old, then we would say well, we will figure cost at alpha at 62 years old. We will multiply at two times that mean 62. We will come up with a predictive cost for evaluating folks at the mean. We can do the same thing for the experimental group. We have a way of talking about costs. Our beta 1 in this case, of course, is still the difference between the groups. This seems very simple. But it is just a necessary underpinning for what is to come.

Our assumptions in Ordinarily Least Squares are about the error term. The residuals often have a similar distribution to the dependent variable, and in this case cost. Why worry about the violation of these assumptions? In the words of the late Will Manning, in small or moderate sized samples, a single case can have a tremendous influence. This is because there are no values on the left side of the distribution to balance those very influential outliers. In fact, it has been observed that in the Rand Health Insurance Experiment, that classic study, there was in one particular health plan a particular patient who was accounted for 17 percent of the costs. That is clearly the case of a rare event or lightning strikes and unexpected. But it has a tremendous influence.

Let us just graphically think about this. Here is a data cluster down around on the left. But there is this one influential outlier on the right. You can see our value of estimating what is the effect on this case X as a quantity? Say it is the quantity of treatment that was received. We said point X as an affective point, 0.88. If that one outlier had a slightly different value, we would less reduce our estimate of the influence of X by half down to 0.42. You can just see graphically how that outlier really draws the line towards itself. Because it has got this influence or this leverage in terms of minimizing the sum of squared distances.

One approach to dealing with outliers is to take cost and take its natural log, the log of transformation. This is not the law to the base-10. This is the Naperian or base-e log. We, at ten dollars, the cost is a value of – log cost is 2.3, et cetera. You can see this is a different scale. If we look at a log cost regression. Now in this case, our Y is log of cost. Our or excuse me, our dependent variable is log cost, this is exactly the same data. But now you can see that this one outlier has much smaller effect on our beta, a much smaller effect on how the line is drawn.

This seemingly really solves the problem of an influential outlier. Another way of looking at this, and this is back. This graph is the same one you saw previously. Our costs in the VA for patient costs five years ago. If we now take the log of the cost, you could see that this looks much more like the bell-shaped curve that we associate with the normal distribution. There are two data here. One with and without pharmacy.

We can look at our descriptive statistics, our measures of normality and see now that the mean and median, once we take the log costs are really quite similar. Our standard deviation seems much more reasonable. The skewness, remember zero is the skewness in a normal distribution. That is so it is pretty symmetric. Our kurtosis is actually shown less peak than we would expect in a normal distribution. Three would be for a normal. This has seemingly gotten us to a distribution of values that is much more close to normal.

We can run the regression of the cost, log cost just the way we would do in Ordinarily Least Squares. It has some desirable properties. But we have to think about well what does beta mean? The beta now has a different meaning. They also, they are not interpretable as raw dollars. The parameters represent the relative change in cost for each unit change of X.

This is based on the calculus. There is slide at the end of the talk that derives this, if you are interested in following up on this. But basically what this proportional change in cost or relative change is it means that for each change in X, if the beta is 0.1, that is a ten percent increase associated with each unit change in X. Or, if X is a group membership variable like they are in an experimental group, and the experimental group has ten percent higher cost. Now policymakers are not always satisfied with just a percentage, a relative percentage. They want to know that absolute amount of what the Y is.

Often, we want to do this in the context of case-mix. We want to say in this case, we want to find out what is the predicted cost given that they are in the group and they have the mean value of the case-mix variable? How do we come up with Y hat? Well, the great temptation is to take the anti-log of this regression value. That is we exponentiate it. We take it to the E to the value of this score, the right-hand score. Is that the value of the raw costs, the cost in raw dollars? Since we took the log cost out, can we take the anti-log of prediction and come up with and back to the transform – retransform it back into cost?

The answer is no. That is not a good idea. Here is the reason why. If we – and this relies on an analysis using our expectation operator. Our expected value of Y or of the cost is the expected value of the anti-log of that fitted value. The key problem is that the U, over there on the right-hand side, the error term is still a random variable. We would have to make the assumption that expected value of the anti-log of the error terms is one. Is that a reasonable assumption? Since it seems maybe it could be true. Because the expected value of the error term is zero. Does that mean that expected value of the anti-log of the error term is equal to one? That seems like straightforward algebra. Does it not?

But if you work through it, the non-linearities show that is not the case. Here is just an example. Say we only had two error terms. We have – that are one and negative one. In that case, the expected value of the error terms is zero, right. The mean value of one and negative one is zero. But the mean value of their anti-logs, it turns out to be 1.5. If we worked through it here. You can see that the non-linearities show us that the expected value of the anti-log of the errors is not the same as the anti-log of the expected value – and of the residuals. These are just not equal. You cannot make that assumption. If you do, you will be making retransformation bias.

There is a way to convert the logs into log cost, the predicted log cost into a raw cost. Retransform it back into the units that are of interest to your policymaker or a reviewer. You can work through this. Remember before we said that this value here – whoops, I am sorry. The expected value of cost is the anti-log of this value. The thing that comes out is we still have to take the expectation of the residual or the anti-log of the residual, excuse me; E to the exponent \_\_\_\_\_ [00:22:26] there. On the average of these anti-log of the residuals, it becomes a factor that I have circled there in red. We multiply that times our predicted score, the anti-log of our predicted score. We end up with something. We interpret it as cost.

This estimator is called – it is a smearing estimator. It is just simply we take the anti-log of the residuals for every observation and find their mean. It is usually a number that is greater than one. Most statistical programs after you have run our regression; so, we would run on Ordinarily Least Squares regression using log costs as the dependent variable. We would have a residual. We save that residual for all of the observations. What part is not explained by the regression? We find its anti-log and then take the mean. That is our smearing estimator.

Last time I gave this talk, people said well how would I do this in SAS? You have your regression, your model of log cost; and I just said let us make it real simple here. We will have an indicator variable called group. You are either one, you are in the group; or zero, you are not in the group. Then I would output the results to a file that I my file. Save all of the residuals; name them my residuals. Then a data step I use is take the anti-log of those residuals. That seems pretty straightforward, just create this new variable that I call the smearing estimator. That is a value for each one of the observations in the data set. Then I find the mean.

Now I have a single point estimate, which is my smearing estimator, that mean value of the anti-log of the residuals. Then what do I do with that, and that smearing? I save it. I just said well, save it as a single value – as a macro variable. I predict the log cost. That is the Y hat from the variable, and then transform it back into natural units. Once I have got my predicted cost, I can multiply it – or the mean of the predicted costs – I can multiply it by the smearing estimator.

Now I made a little bit of a leap here about how did I come up with that predicted cost? Let us talk a little bit about that. The simple approach is just to use the coefficients that you have estimate, the alpha and the beta. I am calling alpha intercept here. We could predict the cost of the people who are not in the group. That is alpha. The intercept, the people who are in the group, alpha plus beta. Each of these is a predicted log cost. We can multiply each of these by the smearing correction. We figured out the cost of people in the group and the people not in the group.

Now, what if there are other covariates besides group? Can we just do the same thing and do like we did before, evaluate the parameters with the covariates set to their mean? This is kind of – can we do this? Set it up so that the find the mean of the log costs controlling for case-mix just like we did before? This is, the example here is for the people not in the group. We just take the alpha plus the beta evaluated to mean of the covariate here, in the case a mean value in Z.

This turns out to be not such a good idea because of the non-linearities involved. It is basically back to that same sort of statement that the anti-log of the mean is not the mean of the anti-log. We worked through the example of why that is not true for residuals. It matters whether you do the anti-log before or after applying the expectation \_\_\_\_\_ [00:26:23] the same way here. \_\_\_\_\_ [00:26:28 ] there before or after you are taking the mean. When you have other variables, the better idea is to predict the cost for everybody in the data set. Once as if they were in the group and again as if they were not in the group.

There are essentially using the parameters to predict the log cost for everybody. Then we need to apply the smearing correction. Now we have the retransformed – take your anti-log, apply the smearing correction. Now we have the predicted cost for each person. Then we take the mean. We basically want to find the mean – the cost from every one of the observations in the group in the data set as if they were in the group and as if they were not in the group; and then predict the cost. This is kind of an important lesson about working with nonlinear costs and simulations in this way when you have a nonlinear specification.

Now, I will refer you to the original paper by \_\_\_\_\_ [00:27:38 ] and \_\_\_\_\_ [00:27:40 ] some years ago. I think this was part of the RAND Health experiment when they developed this. It is quite an old site now. Now one problem with the smearing estimator is that it assumes identical variance in errors. The property called homoskedasticity, which is one of the Ordinarily Least Squares assumptions. We will talk about next time about what to do when the homoskedastic errors assumption is not a good one to make. Retransformation comes into play when we are using a log model. The log model is useful because data are skewed. But those fitted values must correct for retransformation bias. I am just summing up here to say that is the important issue about log models. They have this advantage. But they have this limitation.

Another problem is that we have zero values in cost data. The left edge of the distribution is truncated. Remember that you cannot take the log of zero. That is just undefined. What do we do in this situation? If we are in a situation – have data like that example where you had VA data. Where there were enrollees who had no utilization. There is a great temptation and many people have done this. They say well it is – we cannot take the log of zero. But it is – what if we just give them a value like a penny, or a dime, or a dollar? That is pretty close to zero.

We will have some number to estimate. Why do we not just substitute that for these zero observations? Then we can move from there. Well, we will have a log value for that. You can do this, but there is a problem. Here is an example where we took all of this data set. This is another hypothetical data set. We substitute for the zeros, one dollar. Then you can see that the values of the parameter estimate 0.12 for having each unit of X.

Here we do it again with ten cents. We can see that it matters a lot. The parameter has changed. The line has changed based on how we – whether we use the dime or whether we use the dollar. This turns out that this choice of an arbitrary small positive value to replace the zero cost records can be very influential and kind of for the same reason that the outliers were important on the right side. Here they are on the left side.

Of course, it requires, in this example – the work is that there are quite a few zero observations that need to be accommodated. Let us think through why this is not such a good idea to substitute the small positive value. The log model assumes that the parameters, our alpha and our betas are linear in the log scale. Thus it assumes that a change from a penny to a dime, or a dime to a dollar has the same importance as a change from a thousand dollars to ten thousand dollars, or ten thousand dollars to a hundred thousand dollars.

That choice is pretty important. It is possible to use a small number – positive number in place of zeros. If there are just a few records involved and you should check that the results are not sensitive to the small positive value. But in truth, there are really much better methods. Some people use transformations that allow zeros such as the square root. Others use a two part model. The two part model is where you have – and we will talk about this next time. But the two parts are – one is a participation part. Did the person incur any costs?

Then another is a regression for a conditional cost. Given that the person had incurred costs, how much were those costs? That way you just handle the zeros as a separate regression. Then there are other types of regressions that allow zero values, these link regressions. We will talk about those next time. We talked about Ordinarily Least Squares and the whole problem with using it. If the costs are not very skewed, maybe it is okay. If there are not too many zero observations or if there is a big data set. They have the advantage that the parameters –are much easier to explain.

It is possible to have rather than a hurdle model just a single regression that includes the zeros. But I think in the modern world, the reviewers are going to raise their eyebrows at costs when they are Ordinarily Least Squares. You had better be prepared to defend it and also have said you tried some of the alternatives. It is just not the best way to handle costs. Because these cost are just – it is such a notoriously badly behaved variable.

In general, cost data are not normal. They are skewed by high cost outliers and many data sets that are truncated and have many zero values. Ordinarily Least Squares makes assumptions about the error term. That it is normally distributed and that it homoskedastic errors that often do not turn out to be true with cost data. Error least squares does not work out so well. It can result in biased parameters and have outliers that are very influential. Then this is especially true at small to moderate size samples.

Log transformation is a good approach because it can make the costs much more normally distributed. We can still use Ordinarily Least Squares. But it is not the only – it is not always necessary. It is not the only method of dealing with skewed cost. If you do, the meaning of the parameters in either of these models depends on whether you are using the linear that is raw costs like in Ordinarily Least Squares or a log dependent variable. In Ordinarily Least Squares, we have the raw cost. Beta is the absolute unit and change in X for – is a change in Y for a unit change in X. When you have the log of cost as the dependent variable, it is the proportionate change in Y for unit change it X.

The parameters depend on which model you are using. If you are going to find the fitted value, you could find a linear combination of parameters and variables. But if it is a log of dependent variable, you cannot simply take the anti-log of that linear combination of parameters and variables to say that is the raw cost. You have to correct for retransformation bias. One of the ways of correcting for retransformation bias is by finding the anti-log of the \_\_\_\_\_ [00:35:16 ] of the value; and you then multiplying by the smearing estimator, which is the mean of the anti-log of the regression residuals.

Cost data also have this problem often of having zero values, a truncated distribution. The log of zero is not defined. It is sometimes possible to substitute small positive values for these zeros. This can result in bias. They are really much better methods. The next session next week, we will talk about how to handle these zero observations with two part models. We will talk about regressions that can be done with link functions. These link functions have the same property of transforming the scale like using log or square root. But the way they are set up, they do not have the retransformation bias problem. Then they can accept zeros.

There is also a possibility of doing nonparametric statistical tests, which we will talk about. Then we will cover which of these methods is best. That is a little teaser for the next time. I will just mention that if you are interested in a good overview of what we just presented; not so much next week's stuff, but the review of what we just covered now. Paula Diehr's paper in the Annual Reviews of Public Health is a good orientation. If you have a VA e-mail address, I could send you a copy of it. If you do not have access to it; and send us a message at HERC at VA dot gov. If you are not the VA, well you are on your own. Copyright precludes sharing that.

There is some supplemental reading. I mentioned this Naihua Duan article from 1983. Will Manning has written on alternatives to the smearing estimator. There are other ways besides the smearing estimator to work. But it has really moved on to these link functions now. This is that promised slide about the derivation of the parameter meaning in the log models. That is what I have. I wonder if folks have any questions?

Unidentified Female: Yes. There are a couple of questions here. Someone asked about the log model. What if you add a value say one to all data points? It feels that this could fix the log issue.

Paul Barnett: Right. I am hoping that was asked before I showed how it matters a lot whether you are adding one cent, or one dime, or one dollar. If there are just very few observations, maybe you could get away with it. But I would check, if I were doing that. That your results are not sensitive to your choice of the small positive value.

Unidentified Female: This is \_\_\_\_\_ [00:38:17 ] –

Paul Barnett: I do think that would be difficult to get through a review in the modern error. That you are going to need to use one of these more robust methods that we will be talking next week.

Unidentified Female: This person just wrote in. They meant add one to all Y values and not just the zero value.

Paul Barnett: That does not solve the problem. You still have…. If adding one to a thousand dollars is – or just say a million dollar cost observation, obviously it will have very little effect on it. You are really just back in that same graph where your choice of whether it is one dollar, or one cent, or ten dollars is your small positive values. That choice can be tremendously influential if you have a lot zero values. It does not solve the problem on the right side.

If we just think and go back to this picture we had here. If I add in this picture, if I add a dollar on the right side to the thing that is up there at say a value of ten; well, that dollar is not going to move that dot at all. But it is still that whether I choose a dollar or a dime, it has a tremendous influence on the left-hand side of the scale. On the right side of the scale, it is out there in the sixth decimal point or something like that. It is just not going to matter.

Unidentified Female: This is another question asking how would you calculate the standard errors of predicted means after applying the smearing estimator?

Paul Barnett: The standard errors of prediction, it is a good question. I am not sure I can pull that right out. How do you do that, Jane? I am glad you are here.

Unidentified Female: I have typically done some \_\_\_\_\_ [00:40:15 ] to get 95 percent \_\_\_\_\_ [00:40:17 ] around a predicted mean.

Paul Barnett: Yeah. There is a situation where you sample with replacement from your data set. If we did that approach where we talked about where we estimated a value as if they were in the group; and again, as if they were not in the group. Now we have a data set of costs. I guess the question is are we are interested in what is the standard error of our estimate about and around the beta? Or, are we just interested in the standard deviation of the predicted value? It depends on what you are going to do.

But, say we are just interested in the variation of predicted values. Then we could sample with replacement from the data set and do that many times, say 1,000 times, and characterize the distribution that way. Traditionally, you sample, make samples of N, right out of a data that has N observations. But they vary. Because each time you pluck one, you put it back in, and take another one out. It has some distribution. Because they are not all determinately the same. They can get back to us, if they are asking about the other question.

Unidentified Female: Okay. Hopefully they will write in, if that is not clear. There are not any other questions at this time. If anybody else has any questions that they want to ask, feel free to type in.

Paul Barnett: Let us go back.

Unidentified Female: It does not look like there are any more questions.

Paul Barnett: This was that last question about when we predict costs, we can create these data sets with an observation for each person that is in our data set. Say we had a clinical trial. We could take everybody in the trial and give them their covariates and predict what would be as if they were in the experimental group; and do it again as if they were in the control group. Then predict costs for each scenario.

Unidentified Female: There is another question asking does a retransformation bias apply only to costs?

Paul Barnett: No. I think whenever you take the log of your dependent variable, then that problem is going to occur if you want to use the regression parameters to predict what the value would be of that dependent variable. You cannot just take the anti-log of your fitted value of the dependent variables. We, in our example – let me see, if can bring up the slide here. This is an example. We can predict what are costs from our regression parameters and our data. That this great temptation is to say well, I will just take anti-log of log costs. Or, it could be another variable that you have used the Y for; and take its anti-log. That is E to that, the right-hand score there, to that power.

That is the anti-log. That would be the predicted value. But because you are not doing the same thing with the residual, you are going to end up with bias. You have to multiply by that correction factor. The correction factor is usually some value that is more than one, a usually a small integer. In practical experience, when I have done this, it is a value like 1.5, two, three, something like that, in that neighborhood depending on how skewed the data are. But even if it were not cost, you are still trying to do this same thing of using a – some people call these semi-log models where the log is on the left-hand side. You want to be sure that you adjust for retransformation bias. Or use a general linear model that is what we will talk about next time – a model with the link function. It does not have this problem.

Unidentified Female: Somebody has followed up with what percent of zeros from the total observation can we tolerate?

Paul Barnett: Well, we will talk a little bit about what represents a good fit from the data. The tolerate, it kind of depends on what it is – what is the parameter that you care about? If your beta is what is the effect of being in this group? I do not have any hard and fast rule. But I just tolerate to what ends, to adding the small positive number, or any of that. I think that the things that we have talked about today Ordinarily Least Squares, log transformation.

I would say that any study that I was going to do costs, I might start with these as sort of the simplest analyses to get me going and get started, and take a look at the data. But that ultimately in order to get something published in the scientific literature these days, I think I am going to – I would have to use a general linear model, or at least convince reviewers that I had tried it. A general linear models accomplish much the same thing, but accommodate the zeros, accommodate the log costs without the problem of retransformation bias. They still have the same interpretation; say if you are using log scale as log transformation, the dependent variable, this the betas still have that same meaning as proportionate change. But those general linear models are really what you have to do now.

Unidentified Female: Somebody asked the question going back to that predict slide that you were showing. Why do we need to predict costs from the log model when there is only one predictor, case or control? Can we use the raw costs?

Paul Barnett: You certainly could do that. You could report their mean and this parameter. The beta that you are going to estimate represents – let us go back here. I think – well, yeah. That is without the, right – we are asking about this. The beta is the test of the difference between the two groups. It is more robust than if we had done it with simply using costs on the scale of cost. It is a better estimate of the statistical significance.

You might find sometimes that there could be a case where if you did cost as a dependent variable, and then you did log cost as a dependent variable, that you would find a different result for the statistical significance. Is the beta significantly different from zero? Log costs would be the better test. It is true that you could simply report the mean cost in each group and use this significance of the beta in that case as the test of whether the groups are different. In that simple case, very simple case, it should actually come out almost exactly correct when you do the – or the same as taking the mean.

Unidentified Female: Right. To the degree we were using – and usually controlling for a lot of different things, then we want a regression model.

Paul Barnett: Yeah. This simple model is like usually a trial where you really only have – you control for everything else by randomization. If you do not have randomization, then oftentimes, you have got a whole series of Zs. It is not just one covariate, but a whole bunch of stuff that you try to control for.

Unidentified Female: This is a question asking can we use mixed models instead of OLS for cost analysis?

Paul Barnett: I am wondering what is meant by a mixed model? I know there is a SAS procedure.

Unidentified Female: Well, sometimes they refer to multilevel models of mixed models. I do not know if that is what she is referring to? That person could clarify.

Paul Barnett: A lot of times we have our, every discipline has its own jargon. I am not quite sure if I know. I heard of mixed model as being both quantitative and qualitative analyses in the same study. There is another good term.

Unidentified Female: Okay. She says yes as in mixed means multilevel.

Paul Barnett: The multilevel model is more about the question of what do you do when you have…? One example would be longitudinal data. Say we have multiple observations from people in the study. I have got ten years of data from the same people. Well, I would want to – one way of saying these are clustered by person. I need to have a person level error term.

Another example would be I have got people. They all have the same doctor. Another set of people have a different doctor. They would be clustered by provider. We may not want to make the assumption that these observations are independent. My errors are correlated over time. Or, that my errors are correlated with those of other patients who have this same doctor. That is the kind of when you get into the – some people call them mixed models, or hierarchical linear models. Or there are other terms for them.

I guess we call them random effects. We can call them random effects models in econometrics. That is sort of a different – that is a different issue. Cost data could occur in that sort of situation where observations are clustered either for the same person or many people that share the same medical center, or the same doctor, or something like that. We want to account for that. But we are not trying to explicitly say what is the effect of having Dr. X versus Dr. Y?

Unidentified Female: One of the questions asks if I wanted an estimate for the absolute difference in dollars between two groups, is it okay to simply exponentiate the beta?

Paul Barnett: I think that is right. I think if you work through in the simple case. But in the complicated case, no. Because the beta will have a different meaning depending on the level of the covariate. Say that I have a Z that is very high. Z, I should have…. Then my beta 1 has got a very different effect, than if my Z is very low. It all depends on where you are evaluated because of this nonlinearities of the system. If my Z is zero, beta 1 has very small effect. If Z is a very large number, then it is a nonlinear. The effect of a group membership it becomes much larger.

Remember it is a proportionate change. If I am a high cost person and it is a ten percent extra cost, then it is ten percent of my very high cost. If I am a low-cost person, and it is ten percent extra cost, it is not so important. It is much less. Somehow that is this whole issue of having to do a simulation that looks at observations. Then taking the mean effect over all of these observations. What observation in this data set is going to be a low cost person? That ten percent effect is going to be and have one value in dollars.

Another observation, a different person may be higher cost. A different amount – and then we take the mean of all of these observations. This is precisely, if you have got covariates involved, why you cannot just exponentiate the beta, if you will.

Unidentified Female: Okay. That looks like that is it for all of the questions.

Paul Barnett: Well, it looks we managed to use up most of their time after all. I had this time for an hour. But I think I talked too fast. I apologize for that.

Unidentified Female: That is okay. It seems that the audience was still very interested in what you presented even if you feel you went a little fast.

Paul Barnett: I think that we got some good questions. It was worth leaving that time.

Unidentified Female: Perfect, and for everyone, we are continuing this session next Wednesday, part two of this session. Do you have that slide there, Paul?

Paul Barnett: Yes.

Unidentified Female: Yes, there we go. I should be setting the announcement for this with the registration link out this afternoon. Just keep an eye on your e-mail for that. For those of you who are not registered, you can get a last minute registration in there.

Paul Barnett: I should say, Heidi. I just wanted to say that I did make a change in one or two slides from the paper copy or the electronic copy that you distributed before the talk. We will have the updated slides on the archived website.

Unidentified Female: Yes. But when I send out that archive notice, the link to the updated slides will be available there. People will not have to dig around looking for that there. When I close the session out today, you will be prompted with the feedback form. Please take a moment and fill that out. I will be sure to get Paul all of the feedback in plenty of time, if he needs to tweak anything before next week's session. If you do need to send something in, please take to opportunity to use it.

Paul, thank you for preparing and presenting today. We are looking forward to part two of your session next week. Thank you to the audience for joining us. We hope to see you at part two of this session on next Wednesday. Thank you.

[END OF TAPE]