MCH EPI Conference

Time Trend Analysis for MCH Outcomes in Large and Small Populations

December 8 – 11, 2008

JEFFREY SHAFFER: We left off with rate reliability. And this was a calculation for rate reliability based on infant mortality rates. I'm going to move to a simple comparison for a single infant mortality rate to a reference mortality rate.

Now, this is perhaps the simplest kind of test that you can, hypothesis test you can conduct using an infant mortality rate.

And many of you will be familiar with a Z score, which is defined as you see here. Z scores are simply assuming a variable is normally distributed, follows a normal distribution, some type of normal distribution with unknown mean, mu, and unknown variance sigma squared. You can normalize this or standardize this to a standard normal distribution, which is defined as a normal distribution with mean zero and standard deviation of 1 as you see here.

So if you subtract the mean and divide by the standard deviation, assuming Y is normally distributed, Z will be standard normally distributed. And that's appealing because Z scores have a number of statistical properties associated with them that allow for direct, nice straightforward easy interpretation.

Using my notation here, ZI is going to be a Z score for the Ith year. YI is going to be the number of infant deaths in the Ith year. Mu is going to be the mean of the reference population. And sigma is going to be the standard deviation of the reference population.

So assuming this is the definition of a Z score. We aren't quite finished yet. If the sample is large, okay, and what do I mean by large? We'll talk about some thresholds in a second. If a sample can assume to be large, to be large, we substitute the Poisson mean. Now Poisson distribution is a discrete distribution, and it's very appropriate for data that have rare events and data that are in the form of counts.

So a Poisson distribution is used for modeling count data. So if the sample is large, substitute a Poisson mean theta. Let's call this unknown mean. Now, the parameter theta, of course it's unknown. If you knew it you wouldn't need to conduct the test in the first place. That's the unknown Poisson mean theta for mu and sigma theta. Now a property of the Poisson distribution is that the mean is equivalent to the variance.

So that says if I use -- if I assume a Poisson distribution, that says this is going to be this. Okay. Mu is sigma squared. So I can go ahead and make this substitution if I assume that theta is the mean for a Poisson distributed variable. I can make this substitution.

Now, what does large mean? The NVSS uses a threshold of 100 for a large sample. So for infant mortality, this would mean at least 100 infant deaths would be considered a large sample, saying this technique is appropriate.

Now, keep in mind this is a normal approximation of the Poisson distribution. And we're actually using a property of a continuous distribution when actually the counts of infant deaths are never continuous. We're talking about counts there so they're not continuous, so this is an approximation and it's appropriate when the sample is large.

And the rationale behind that is the central theorem says that for large samples, the normal distribution can be used to approximate many distributions. Okay. So if the sample is large, here's what we have. A Z score that looks like this. Now, Z scores have a lot of nice statistical properties.

One of the nicest properties is that you can easily look at a Z score and see if it's extreme. It can be shown that 95 percent of the quantiles, in other words, the ZIs, the Zs, should lie within two standard deviations of theta.

So this is, in other words, saying if ZI is greater than 2 this can be considered as extreme. Okay. So if these values are larger than 2, the absolute value -- I guess it wouldn't matter so much if this was negative, because this would say it's

extreme the other way, but if the absolute value of this is larger than 2, this may be considered as extreme.

Let's take a look at an example to bring this into clear focus. An example. According to the NVSS, the National Vital, do you guys know of the NVSS, National Vital Survey Statistics. I think I might have it right.

According to the NVSS the IMR for the U.S. in 1990 was 9.2 per 1,000 live births. I found this number out there. What I was interested in doing was comparing this number or comparing the number in Louisiana to that number.

So of the 72, consider the Louisiana data. We had 72,112 live births for Louisiana in 1990, from which 796 resulted in an infant death. Okay. So what I'm going to do here is I'm going to say since 72,112 divided by 1,000 times 9.2 is approximately 663, what I'm doing here is I'm calculating the -- first of all, I'm getting rid of the scale. That's why I divide by 1,000. Because it was reported per 1,000 births.

If I want to get rid of the scale I just divide by 1,000. Then what I'm going to do is I'm going to multiply that by 9.2. 9.2 is what we got in the U.S. If I had the same number of births, how many deaths would I get for the U.S., is the question. How many deaths would you expect to get?

This is what you would expect to get for the U.S. That's the reference population. How many did I get for Louisiana? 796. That was a given. So I calculate how many I would expect to get in the U.S. given the same number of live births, and this is how many. 663.

So with this information, I have what I would expect to get in the U.S. I have what I was given in Louisiana, and if I come back here, I can look at this Z score.

This says ZI equals YI which is the number of infant deaths in the population that you're comparing to something else. So this is the given number of deaths in Louisiana minus theta.

This is going to be the Poisson mean, the Poisson mean for the reference population. Well, this mean is given as 663. This is what we would expect to get for the entire U.S. Now, since the Poisson distribution has the mean equaling the variance, we can use -- this one is going to be a function of theta here and theta there, in the same place.

So I can simply plug that result in. I end up with Z as -- keep in mind the mean is the variance. This calls for the standard deviation. This is sigma. Now, to get sigma, since the mean is equivalent to the variance for the Poisson distribution, the variance is sigma squared. So I take the square root of that to get sigma.

Now, I get this. I get Z to be about 5.17. The nice thing about Z scores you can look at them instantly and perform a hypothesis test in your head.

Is that number extreme? Sure it is. Z values range about between negative 4 and positive 4. Okay. Anything beyond 3 is pretty extreme. So that's very extreme. And the way you interpret this is -- well, you can look at the numbers and see that 796 is much larger than 663. But this is telling you -- this is statistically different than 663 and it's saying that 5.17 is a significant Z score, because it's much outside the range of data.

Now, under anyone's threshold, if you guys think about the percentiles or quantiles of a standard normal distribution, if you wanted to perform a formal hypothesis test, the comparison value here is going to be either 1.645 or 1.96, if you guys have learned the values for standard normal distribution, that's what you would compare it to. But typically when you use these the Z scores you don't really do that.

You look at them and you say that's pretty extreme, I'm going to move on to something else. This is an exploratory technique and we like to look at these. A good rule of thumb in the literature, you're going to find a lot of rules of thumb. A good rule of thumb is if that value is at least 2, if the absolute value of that is at least 2, it's probably going to be significantly different.

And I would conclude that Louisiana had a higher IMR than that for the U.S. in 1990.

Now, what this is a one sample Z test for IMRs is what this is.

UNKNOWN SPEAKER: Could you just pop back to the previous slide? Shouldn't that be over theta squared or excuse me over the square root of theta.

JEFFREY SHAFFER: No.

UNKNOWN SPEAKER: Because it was variance squared.

JEFFREY SHAFFER: Theta equals theta squared. But this calls for sigma, not sigma squared.

UNKNOWN SPEAKER: Exactly. Could you go back one more?

JEFFREY SHAFFER: Back one more?

UNKNOWN SPEAKER: No, the other one. Sorry.

JEFFREY SHAFFER: That's okay. This one here.

UNKNOWN SPEAKER: That's what I was asking. Shouldn't that be Y minus theta over the square root of theta?

JEFFREY SHAFFER: What you're given is theta equals theta squared. We agree on that; is that right? For a Poisson distribution, if I square, if I take theta, if I square that, those are going to be the same.

Theta equals theta squared. The mean equals the variance.

UNKNOWN SPEAKER: The mean equals the variance but the Z status Y minus the mean over the standard error of the square root?

JEFFREY SHAFFER: This is the square root. Think about this as the square root of theta squared. By definition of a Poisson distribution, theta equals theta squared. And by definition, theta squared is going to be the definition.

UNKNOWN SPEAKER: Right. I agree that theta squared is going to be --

JEFFREY SHAFFER: Theta squared is a variance. So if I take the square root of that I'm going to get the standard deviation for a Poisson distribution.

UNKNOWN SPEAKER: Right. Totally agree with you there.

JEFFREY SHAFFER: Say it one more time then on this one. You're saying you think -- maybe you don't. You're wondering why this isn't theta squared, is that what you're telling me?

UNKNOWN SPEAKER: I'm wondering why this is not the square root of theta.

JEFFREY SHAFFER: Because the square root of theta would not be the standard deviation of a Poisson distribution. The square root of theta squared would be the standard deviation for a Poisson son distribution. These variants for the standard deviation Poisson squared. If I take the square root of that, which I could have put this step in here, the square root.

UNKNOWN SPEAKER: I got it.

JEFFREY SHAFFER: You got it? Okay.

UNKNOWN SPEAKER: Thank you.

JEFFREY SHAFFER: Sure, sure. Let's move ahead. This is a basic test. Do this in the beginning. In the back.

UNKNOWN SPEAKER: Don't you in the formula -- you have the square root.

JEFFREY SHAFFER: Right, because it's given to me that the mean is equal to the variance. It isn't that the mean is equal to the standard deviation. So what this is is theta squared. See, there's nothing that says that the mean is equal to the standard deviation for the Poisson distribution. I'm given the variance or I'm given the mean. I can say that equals theta squared. I cannot say that equals theta.

So I'm given the mean. I can say that equals theta squared or the variance, and then I take the square root to get the standard deviation. You guys want me to say that again? I know that seems a little -- that's a little convoluted. I can see why --

UNKNOWN SPEAKER: So the definition that we know with the Poisson distribution, the mean equals theta

JEFFREY SHAFFER: In terms of means, we're thinking theta equals theta squared.

UNKNOWN SPEAKER: Perfect. That's conservative.

JEFFREY SHAFFER: We're given theta. That's what this is. If I put theta here, I put it here and theta down here, 663, I'm putting theta squared down there. I'm not supposed to put theta down there. 663 is theta squared or the mean.

If the mean equals -- really does equal the variance, then if I put that down there then I would be putting the variance. All right. Let me move on.

Let's go ahead on to large sample. Now, this was a simple hypothesis test. This wasn't a confidence interval. Now I wanted to next talk about a large sample confidence interval for a single IMR. And this does have appeal. We haven't gotten into all the modeling and complex things yet. But these are things that you guys are going to report. So it's worth a little bit of time here.

For a large sample, a confidence interval for a single IMR is given as you see here. Okay. For a large sample. And once again you guys may see different thresholds. I guess we'll go with at least 100 infant deaths on this for IMRs.

So you need to see what constitutes a large sample for the data that you're working with. But for mortality, 100 is a pretty -- that's a pretty good amount. That's a pretty high number.

Okay. So using that threshold, a confidence interval for a single IMR is what you have here. Now, this is in its crude form, it looks like a simple confidence interval for a standard normal.

But it isn't. What we're going to do is we're going to use this formula, you see here, for the standard error of the IMR. Why am I doing this? I'm not going to talk

so much about the derivation. I might have it in the manual if you look there. But I want to focus more on why we use this standard error.

Why not just use -- well, why not use, for instance, why not use something else. Why this standard error? Why not the standard error for a normally distributed, a normal mean? A mean from a normal distribution?
That's not this. The reason is because this standard error was derived from that of a Poisson distribution. So this is going to be a little more accurate than putting something else in here based on a normal distribution.

This I did adopt from one of the reports. One of the health surveys, health survey reports.

And once again I have that in my manual exactly where I got this. But it's one that's readily used. It's used quite a bit. And once we have this, it turns into a basic plug-in. Z 1 minus alpha over 2. That's always going to be usually 1.96 for an alpha value of .05. So this is pretty easy to calculate. YI once again the number of deaths in year I.

This is going to be a confidence interval for a single IMR. So this has a lot of appeal. Numbers by themselves don't mean much unless you can say how reliable they are.

So reporting an IMR is one thing. Reporting an IMR is one thing but seeing how confident you are in the results is another.

So we have everything we need here to calculate a confidence interval for a single IMR. Just remember that this SE, this standard error, is derived from a Poisson distribution assuming count data.

Like I said, I have a lot more details in my manual. I don't really have the time in this workshop to go over the derivations and how they came to these things. If you want, read the manual. I go into a lot more depth in there.

Here's an example. Of the 72,000 live births for Louisiana in 1990 there were 796 infant deaths. 95 percent CI for IMR 90 is what you see here. I could do this for each of the 16 years.

No problem at all. It's simply a plug-in. I'm going to go ahead and calculate the IMR. That's going to be 796 divided by 72,112. Now, I think I'm forgetting to multiply that by a thousand. I'm pretty sure I did. There's no way that number -- I should have probably multiplied this by a thousand to convert it to a true IMR per 1,000 deaths. This is how we begin to calculate the IMR, plus or minus 1.96.

And this is going to give us the standard error for the associated IMR. Now, one thing, the first thing I notice when I see this standard error, numerator and

denominator, which I like because it incorporates that information from the underlying populations. That's what you want.

So if you see single number here without differing between, where these numbers don't differ between the numerator and the denominator. In other words, a rate here or something, then it's not as good. It's not going to be as accurate.

So this is what we want. And we end up with 10.27, 11.81. And we interpret this as we are 95 percent confident that the true IMR in 1990 is between 10.27 and 11.81 under repeated sampling. Okay, that's how you would say that.

All right. Everyone see -- this is essentially a plug-in, you guys. But if you want, if you're interested in pursuing this a little more, just keep in mind that this standard error is based on the actual Poisson distribution. This isn't just a confidence interval for a continuous number, or a continuous variable.

Okay. Now let's talk about -- we did assume we were talking about a large sample here. Now, like I said, there are different policies for what constitutes a large sample. We'll go with the threshold of 100, 100 deaths, that may change based on the report.

Let's talk about small sample. Okay. I use the NVSS stuff on this. And that's included -- every time I have an acronym like this, if you guys aren't sure about

what it is or where I got it, just make a little note of that to the side and go back to my references in the manual and you'll see it, because I did include all these as references.

And you guys can download these PDFs free. They're online. So they're all there for you to look at. Okay. For a small sample, we use, with different notation, I changed the notation on these just to make it easier to present, we use the gamma method as described in the July 2008 version of the NVSS.

And this is a given. I'm not going to explain too much of the details on how this was derived. I will say once again that we're not assuming continuous data here. That's important. When you see these YIs and NIs broken up like that, that's telling you that you're considering the numerator and denominator separately, which is good.

So this is the one they recommend. The NVSS, the people that wrote this publication, recommended this confidence interval for a single IMR for small samples, to be more accurate, than just assuming continuous data.

And here it is. A couple notes about it. Number one, gamma in -- this is the inverse gamma function. I will say that gamma is just another distribution. It's another distribution out there. It's a discrete distribution and it's often used in

conjunction with count data and Poisson analysis. So think about it right now as another distribution. This is the one they recommend.

I'm going to show you guys how to calculate this in SAS. This gam in is a distribution. And it's saying these are its parameters. So in the first parameter, alpha over 2, this is going to be the typical alpha you guys are used to seeing. The type 1, the probability of a type 1 mistake.

YADJ is going to be another parameter or another thing you need to get this quantile. This whole number here is going to be a quantile for that distribution or the associated number just like negative 1.96 or .196 is a quantile for a standard contribution. This number here is going to be a quantile for the gamma distribution. A quantile or percentile for that distribution.

I'll show you how to calculate it later. These are its three parameters. The first parameter is going to be the associated probability value. The second parameter -- now the probability value is going to be the probability of -- the probability for this quantile to the left of it.

In other words, the area under the curve to the left. If it's a valid probability density function, the total area under the curve has to be 1. So this is going to be the probability to the left of a given point.

And YADJ is defined as you see here. And there's a lot underlying stuff here I'm not talking about, but that's what it is. This is known as the shape parameter. And the scale parameter is right here. And for this confidence interval, it's going to be defined as 1.

Okay. You guys don't need to know all the details of this right now. Know this is a very commonly used confidence interval, though, for a single IMR with small numbers when you're talking less than let's say 20 deaths, something like that, this is about as good as you're going to get.

We're going to talk about this in an in-class exercise. You can generate this. So where does this come from is the question. How do you guys do it in practice?

Well, in SAS, SAS actually has a gam in function. So you can get this number. And it looks like this. This is the general form. The first part of it is going to be the P.

So for the P, you guys would indicate alpha over 2. We were talking about alpha equals .05. This P is going to be .025. The A is going to be the shape parameter. So for A I'm going to indicate this. I'm going to calculate this number and I'm going to put it here.

So where's the last parameter? Well, SAS assumes that the scale parameter is 1. So you don't have to worry about it in SAS for this confidence interval. So we'll talk about that in an in-class exercise. I want to make sure everyone can calculate that.

Okay. As an example, for the previous example -- now, here we're going to take a look at in a little more detail of confidence intervals for IMRs. Recall for the previous example we calculated a confidence interval based on 72,000 live births and 796 infant deaths.

You have an IMR of 11.04. So 796 divided by 72,112 times 11,000 is 11.04. That's the IMR. That's the estimated IMR in 1990 for Louisiana.

Now, consider a sample of 4530, now I just made this up, consider a different sample, 4530 live births from which 50 resulted in infant death. If you divide -- if you perform the calculation, 50 divided by 4530 times 1,000, you're also going to get 11.04.

They're the same. This one's based on a much larger number of live births. So when we calculate a confidence interval for the smaller one, 50, less than 100, if we're assuming that's the threshold of a large sample, we're not going to use the same formula as we talked about previously.

However, when we calculate the confidence interval, we would expect that confidence interval to be wider than the previous one. Because we're dealing with fewer infant, fewer live births.

So plugging this in, like I said, I'm not going to concentrate too much on the plug-in right now. We have an exercise coming up in that.

But I end up with, after performing those calculations, I end up with a confidence interval from 8.17 to 14.6. Now, this is saying that based on the smaller sample we estimate with 95 percent confidence that the true IMR for 1990 is between 8.16 and 14.6 based on repeated sampling.

Now, let's back up, because we had the same IMR here for the large sample. The same IMR. So that's this one. That's this one. We had the same IMR. Notice the confidence interval for this. It's much stricter. Okay. From 10 to 11. The other one's much wider than that and that's because we have a smaller sample.

But we have used the right formula for the small sample. So there's only so much information, what I'm saying is there's only so much information you can get out of a small sample.

Does this meet the objectives of the study? This is where your sampling issues come into play. What constitutes a small sample. Well, are you happy with this confidence interval? If you're not, maybe it's too small of a sample.

Do you guys see what I mean by that? That may not be very informative. There's a lot of fluctuation there. That's a pretty big range when you're talking about IMRs.

But we've used the correct formula. We've done about as much as we could with it, we just don't really have enough infant deaths -- well, not really infant, just infant deaths we don't have enough population to make that to have any more confidence in that.

So that's for a small sample. But I'm just showing you guys that they were -- both IMR were basis point 11.04. Both estimated IMRs were 11.04. We have different confidence levels.

Go ahead.

UNKNOWN SPEAKER: I have a question, are you saying we can compare small sample to the large sample [Inaudible] as a matter of fact we have a small population in our state that is always the case.

JEFFREY SHAFFER: Right. You're saying can you compare this to --

UNKNOWN SPEAKER: The larger sample.

JEFFREY SHAFFER: Based on overlapping intervals, is that what you're talking about? Because that's a little tricky there. You need to be careful with those overlapping confidence intervals. That's not always appropriate. I'm not sure that's what you're asking me, though.

UNKNOWN SPEAKER: My question how can we compare two populations [Inaudible].

JEFFREY SHAFFER: I think one of the more effective ways of doing that would be to model it using Poisson regression or something like that. It's pretty easy to compare populations like that.

See we're not really -- this is a small sample confidence interval for a single IMR. So we aren't comparing anything here. Now, one thing you could do, the only thing we talked about so far in terms of comparison would be back here, right here.

But we're going to cover your issue in a little bit with more advanced techniques. So we're going to come back to that, I promise you.

But that is a good question, and these are things that you guys are going to have to do quite a bit is compare these. And we're going to cover a lot of that with the modeling.

Now, there are more -- I've covered a single IMR. Let me see -- okay. I didn't cover a confidence interval for comparison of two IMRs, and that's what you would want to do. But we're going to cover that using modeling techniques.

The formulas are a little more complex, but they exist. Analogues to this for comparing. When you think about comparison, always try to relate it to things you already know about.

If you guys covered two sample T tests, that's -- probably the analog to that is what you're interested in for this. Okay. But I didn't give that to you in the form of a direct formula. I'm going to give that to you in the form of modeling techniques.

All right. Any other questions?

I want to go ahead and do a couple of exercises here. Like I said, the code for these exercises is given and all the rest of the code is given in the manual.

Exercise 4.1, write a SAS program to generate a time trend plot. I want to make sure you all can do that in SAS. This is really a SAS question more than anything else.

Now, the code's lot shorter now than it should be any how than it was for Chapter 3. I'm basically giving you guys little programs. At least I think I have.

We're on page 23 of your manuals, you guys. That's where I'm getting the code to do these things. I'm going to start with Exercise 4.1. Okay. Exercise 4.1: Write a SAS program to generate a time trend plot of the cross-sectional Louisiana IMRs from 1990 to 2005. Now, the data set for this is CS.

Now, what I'm going to do, you guys should always do in practice, when you're asked to use a data set, always take a look at it. Even if you think you know what it means, just go ahead and open it up. So I'm going to go to the student data sets folder. I'm going to come in there. I'm going to double click CS. It's a SAS data set. Double click it, open it up and take a look at it.

Or another way to do this is the way we did it earlier, we ran a proc contents, whatever you're comfortable with. This is another way of checking it out.

It might take a second to open up. Okay. Here's what it looks like. And I wanted to show you guys what this is and what I did.

Now, I showed you guys how to link the data and everything in Chapter 3, but I didn't do that here. We're working with the cross-sectional IMRs here, cross-sectional data. What you see in this data set for each year is the told number of births for each year and total number of deaths.

These aren't linked in any fashion. It's just taking birth certificate, death certificate, looking at how many of each you had and that's it for each year.

We're going to be working with this data set quite a bit, and we'll get back to the cohort IMRs a little later in this workshop. So that's what it looks like. Notice three variables. Years, births and deaths, that's it.

Okay. Now, I'm going to close that. Question one, okay, let's go ahead and go back. And you guys can copy these global statements if you like. They're going to be the same. You have the same lib ref, same library reference as before. They're all in the same place in that student folder.

But I'm going to put them back here because I didn't copy those. And these can go in any order. I might have had them in a different order the last time.

Okay. Let's put a global title. Let's come down here. How about a lib name statement. Okay. Let me find these. This option here is a gem, does everyone understand what this does?

This clears your output and log files after each submission. So there's no more right click, clear all. Right click, edit, clear all. Is that the way it goes? If you have to come over here to your log file and you're seeing old log files pop up, if you didn't have that option you'd have to right click, go to edit and clear your log, clear your output file after each submission so you're looking at the most recent stuff.

What that does, that clears it for you automatically. So include that in every SAS program you write. Okay. Let's get into the actual exercises. Okay. Time trend plot. Okay. So we're going to go ahead, for those who haven't had an introduction to G plot, here's your chance to see how it works.

SAS is a very extremely powerful program for plotting, okay, and graphs and charts. It's a little bit time-consuming to figure out how to get the plot the right way, but you can change just about anything with the plots in here.

You have an option to do everything. Just takes a little bit of research to see how to get it the way you want. So I'm going to go ahead and read in the CS data set.

Okay. That's it. The data set is read in. Now we're going to go ahead and generate a time trend plot for these data. G options, I like to begin these most of the time, I hope I didn't miss any, but I like to say in the beginning G options -- LS -- G options reset equals all. And that says reset all of the G options that I may have out there.

If I define lines certain way for another proc G plot, maybe I connected the dots, maybe I didn't, this is going to reset all those options.

So I'm going to say reset equals all. For this particular example that's a redundant statement because we don't have any other plots. But when you guys are working with SAS programs with 10 different plots, you're going to want to put that before every G plot procedure step.

Okay. G options reset equals all. Let's begin by defining the size of the plot. Now, you guys may not like the size I have for this presentation, but this is how you change it. You can make it smaller, bigger, depending on what you guys need it for. If you want to copy it over to a word document.

Maybe you want to make it smaller, bigger, so on, but this is how you change the size of the plot. I'm going to say G options. H size. Meaning horizontal size, equals 7.5. That's 7.5 inches. And I can say I-N directly, but I can tell you that's the default in SAS.

So if I don't say I-N it's going to do it anyhow. I'm just showing you guys, that it doesn't have to be in inches. SAS has options from changing it from inches to meters. That's how you do it.

V size equals 5.5. And no border. It's up to you. If you guys like -- I'm just showing you guys how to turn it off if you don't want to see a border around it. There's how you do it.

All right. Next line. Now what we're going to do is we're going to define the plot setup. That's what a symbol statement does. I'm going to say Symbol 1 I equals join. And what that says is to connect the dots. I equals join. V equals diamond. Meaning I want the plotting symbols to be diamonds. And C for color equals -- let's see what I let it be. Red. And I want my line to have a width of 2. I can't remember what the units are for that. But I know width equals 2 is the one that looked the best. Sometimes with this G plot you have to run it a few times and go back and change those numbers a little bit to get the plot looking the way you want it.

I think width equals 2 came out pretty good. Okay. Now let's go ahead and run a proc G plot. Proc G plot in its crude form is actually pretty easy to run. Proc G plot. Data equals Exercise 4.1. We're going to say plot IMR against year.

And I forgot one step above. But I'll go back and fill it in once I'm done with this. If I had the IMR in there, this would be fine. That's what I want to do.

But before we do that, we have to calculate the IMR, which I haven't done yet. I forgot one statement when I read in the data set.

Now, we read in the data set containing the year variable, births variable and deaths variable. Let's calculate the IMR here. So I'm going to say IMR equals deaths divided by births times a thousand. That's it. Calculate an IMR like that. And you guys can go ahead and run that and take a look and see what you have.

This one ought to be fine. That's what it looks like. Like I said, you guys can play around with the width and size, depending on what you're using it for.

You can clean up certainly the axis needs a little bit of cleaning. I don't know if I necessarily need the century part of that year in there. So this is how you'd get started.

And that's a time trend plot in SAS. And this is how you get started in G plot. Now, G plot also has, if you guys get a chance later, I'm not going to talk too much about the details of G plot in this course. But if you guys get a chance later, I made some really spiffy plots in SAS and all you really need to do is look at one of those. Look at one of those and see how the statements go.

You can use -- you can define whole statements that describe how your axes are to look. You can use axis statements to define how this is going to look.

I used the statement here called symbol. And let's talk about the symbol statement for a second. This defines your plot setup, your line setup. And when I say equals 1, that says look for symbol 1. So this one references this one. That's what it's doing.

It says plot IMR against year and let the plot setup be 1. Meaning connect the dots. Give me diamonds for my plotting symbols and make the color red and make the width 2. Like I said, I the 2 certainly isn't in inches. I can't remember what it's in.

But the width of 2 looks pretty good. You can move it up or down. It might look a little thick. Now, there are other statements just like the symbol statement for axis, that allow you to dig into the details and control that.

Like I said, look at my manual. I've given you the SAS code for everything I've done in this presentation. Everything's in there in an appendix. So I certainly haven't put in all the details in the exercises.

That's Exercise 4.1. Let's talk about Exercise 4.2, write a SAS program to generate a race-specific time trend plot of the cross-sectional Louisiana IMRs from 1990 to 2005. This isn't much different. I want to show you guys how to do this, though. It gets a little more -- just another variable.

Now, what we're going to do here, I'm going to copy all that for now, but we aren't going to use this data set again. This CS only has the births, the deaths and the year.

We also need race. So we're going to use a different data set. For this data set is going to be called CS race. I call it CS because CS is for cross-sectional. We'll be looking at cohort data later, too.

Let me open that up. Like I said, step one find a data set and open it up. Run a proc print or proc contents. Whatever way you're comfortable with. I think I'm just going to open it up again. I'll go back to my data sets and double click CS race, show you guys what I did.

This is what I did for the code at each end of Chapter 3. I aggregated these. And what you see here are four variables. I have a year variable, race M for the race of the mother, the births and the deaths.

So this is -- we have subgroups here now. Now we can compare whites and blacks and do things like that. Now, we have this also categorized by year. So we can also talk about time trends. So this is the data set we'll use. Let's come back to the program.

Next problem. Now, this is a different data set, so I'm going to need another data step. I'm going to say data -- what are we up to now? Exercise 4.2. Like I said, you guys certainly don't have to use my data set names. But I'm going to try to be consistent with what I put in the manual.

We're going to say set. Path.CS race. The data set is read in. Now I'm going to come down here and modify this plot a little bit. We'll keep the first statement. That says reset all the G options. In other words, forget about the G options I used above. They're gone.

A lot of these are sticky, sticky commands. They'll stick the whole way through until you reset them. What I'm going to do here is say forget about whatever I did above.

Next I'll use the size the same. I'll use the same setting there. Now under this, if I want to generate this by race, I'm going to do need two lines. So I'm also going to need two symbol statements.

So I'm going to come down here and I'm going to say symbol one. And right underneath it I'm going to say symbol two. I'm going to join both. If I didn't want to connect the dots I'd say I equals none. Here I want to connect the dots.

If you wanted smooth lines instead of jagged edges, it's I equals smooth. V equals -- I want to at least make the plotting symbols, well, did I leave them in the same? No, I changed them. I changed the plotting symbol for one of them to squares.

And I certainly want to change the color so I can distinguish between the two. I changed the color to green. I guess I can leave, probably make sense to leave the width the same.

IMR against year, change the one to race M. That's how you're going to generate a line plot with a -- two separate line plots for blacks and whites. IMR against year.

I think I have everything in there. I'll go ahead and run it. Okay. Scroll down to the bottom of your -- all right. Let me take a look at my log file. Okay. Race M not found. Oh, because this should be Exercise 4.2 that I referenced. Let's run that again. All right. .

Let's see what I have here. Okay. IMR not found. We didn't generate IMR for the second one. We need to do that again. We're given births and deaths. So we're going to need to come down here and say IMR equals -- just like we did above, same statement, deaths divided by births, even though we did that for the first data set, this is another data set. Times 1,000. We need to do that again.

Then I think we should be okay. And there it is. Okay. A race-specific time trend plot for IMRs in Louisiana in SAS. Colored. Like I said, the axis could use a little bit of work, but this is where you start with trend analysis.

Okay. Any questions about that? Let's see what else I have. Now let's get into some statistics. Exercise 4.3 write a SAS program to generate the Cochran-Armitage trend results shown in output 4.1.

That one, let me see exactly which one that was. Output 4.1 -- let's see what page it's on. Output 4.1 is on page 17. And this involves considering the data individually. So we're not looking at the state level IMRs. So the way we do that in SAS is we're going to use a different data set. Let me open it up and show you why.

We're going to use the data set called CS Armitage. But I want to show you guys what it is. I know everyone in here could calculate this, but I think it would be a waste of resources to do this in class.

What you see here is just a different setup of what you guys saw earlier. This data set could easily be derived from what you guys see in this data set, CS. CS, I know everyone understands what this is. That's a very straightforward data set.

What I'm saying is all I did to set this up is I went from this form of this data set to this form. And do you guys recall what I said earlier about the survivorship variable? Not the births and deaths, but the number of those alive and those and the number of those dead. The variable was slightly different. And we had to get that via subtraction.

Now, I did all that for you, and I just went ahead and set it up differently. I just want to make sure you guys know how to go -- because in practice you're going to be able to get this. You have to be able to go from this to this.

Okay. And it's setting it up a little bit differently so that we can use, so that we can use a sub-routine, pre-written routine in SAS to analyze it.

This is how the data is called for in this technique. Does everyone see what I mean? But I won't waste the resources in the workshop doing that. I'm pretty sure I put that in the manual how to go from one to the other. Now what I'm going to do, assuming that data set, so if you guys went through and set up the data set in the proper way, like I did there, and if you don't know how to set that up

right now, one way of figuring that out later is to start with the data set called CS and go to the CS Armitage data set and see how they differ.

They come from the same source. So exercise 4.3 I'm going to say proc. And I'm going to show you guys a little shortcut here. Proc freak and some of you guys might already know this already. Path equals path.csarmitage. You guys can reference SAS data sets directly in your procedure steps like this.

One way to do this is to read CS Armitage in as a temporary SAS data set. I've referenced it in the permanent data form just to save a little bit of time. And we're going to say tables. Year against death/trend. Weight, count, run.

Now, I purposely misspelled count as a variable name because I like to avoid using reserve words or keywords. In any type of computer programming language it's good to avoid doing that.

Try to misspell, if you are going to use c-o-u-n-t to count up the frequencies, suffix it with a 1 or something like that. If you know a word is reserved in a program just change it a little bit so SAS doesn't get confused.

So this is how you generate the Cochran-Armitage test statistic in SAS. We use the total number of deaths against the year, and we weight the counting variable. Now, what is the count variable? What do you think that's going to be? Number

alive. I've decided to call -- I could have called this alive. Count means the number alive.

Maybe I should have done that. Okay. Let's go ahead and run this and see what we get. Okay. Let's see. Year against death. What did I do here? Year not found, because it's probably birth year. Let me take a look at this data set again. Let's go back and look at the data set any how to show you guys once again what this is. But I think I have a variable name.

Yeah, it isn't just year, it's birth year I've called it. Okay. K-O-U-N-T is going to be -- oh, no, it isn't the number alive. This is the number -- that's why I didn't call it alive. That was my motivation for this. The way this data set is set up is the count is the number in that category. This means for those that died I had 796 of them.

Now, when I get down to the ends, I have four of those who -- I have for those who lived I have 71,000 of them. And for those who died
I had 796.

Does everyone now understand why I called that count as opposed to alive or something like that, because it's not really that. It's status or somebody said status, that's probably a real good variable name for that.

Well, status, I guess, would go in place of death. And count, I guess you can leave that spelled like that. But I probably would change death to status.

Okay. Now, this is how this is set up. The reason the program didn't run is because I didn't spell year the right way. It's actually B-i-r-t-h-y-r for birth year. Against death. Now I think it's okay. You guys can go ahead and run that and see what you get.

Okay. There it is. So these are the results I showed you in the presentation previously. What are we saying here? What does this mean? .0014 for a two-sided point -- it's a small P value. This is saying we have some evidence of trend at the individual level. This doesn't tell you how much. This doesn't quantify the magnitude of the trend. It's telling you you have a reason to go further.

Now, let's go right underneath that, and what I want to do is look at this at the state level. Not considering the numerators and denominators separately. And this is done all the time. Like I said, you lose a lot of that sample size. We're going to go from all those observations to only 16. But it has some appeal because we can interpret this at the statewide level as opposed to an individual level. I'm going to come right underneath this and say proc core.

Now, you probably could have done this -- I'm not sure if you could have done this using proc freak for the state level. I found this under proc core to do it at the state level when you don't have the births and deaths separated.

I don't know, you guys may be able to figure that out in proc freak. I'm not ruling it out. But I definitely found it in proc core. Proc core data equals -- what I did here is I say Exercise 4.1 because we've input this data already. Exercise 4.1, we don't need to set this data set up in any special way for this type of analysis.

Let's say VAR year and IMR. Run. Now, this is the correlation procedure that you guys have probably seen. The only thing we do to generate Kendall's Tau beta is say Kendall here. That's the option for it. That says give me Kendall's Tau beta and see if there's any relationship between year and IMR.

Let's go ahead and run that. And voila, there it is. The Kendall's Tau beta that you guys saw earlier in the presentation. Once again, it's not significant. Could be a function of how many -- we only have 16 observations. Notice you're in here. 16 as opposed to the over one million for the individual level stuff. We only had 16 years. We looked at rates and we combined the numerator and denominator into a single rate. So we lost all of our observations.

So this is how you do it in SAS. Any questions about that?

UNKNOWN SPEAKER: On the Kendall's value, does that interpret it like a formation coefficient, would it really be that strong? Can you do that?

JEFFREY SHAFFER: Oh, yeah, but this is coincident with Mono tonicity. So you would say they're Mono tonic, like an increasing trend, increasing or decreasing trend over time.

UNKNOWN SPEAKER: .9 that would be a variable.

JEFFREY SHAFFER: Definitely a very strong trend over time.

UNKNOWN SPEAKER: If you were to prevent say two stakeholder [Inaudible] something like that and you're trying to give them an idea of the trend over time and it's significant, would you recommend this, because it's [Inaudible] even though you lose the power or does the other one make more sense?

JEFFREY SHAFFER: I think this one's a little easier to interpret. I think that you'd probably want to interpret this. See, the other one -- there's no doubt it's more powerful than this. But the problem with the other one is you have to talk about the risk for a single mother. An individual mother. Here I can talk about all the mothers in a certain state.

So I'm limited -- I'm very limited in what population I can talk about here. Maybe I'd want to bring this down to some kind of local level or something like that to stabilize this, but I've lost all my sample at this point. I'm down to 16 time periods.

I think this one has the more appealing interpretation. And I think in practice this is the one you'd want to interpret. Okay. Now, we're going to go ahead and move right ahead to regression models. So let me -- actually, let me just minimize that.

Okay. Regression models. Now we're going to talk about predictive models, talk a little bit about covariates and things like this. Those kind of techniques are useful in the beginning when you're simply interested, maybe you don't have all kinds of covariates to work with. Maybe you're just given the births and deaths and that's all the data you have. So those may be appropriate there. Or, like I said, for a very quick analysis, maybe you have to have something done and you just want some type of exploratory analysis and you want to detect whether there could be some trends, okay, that's a pretty effective way.

Always at least have a time trend plot somewhere in your presentation. Somewhere, something like that. Everyone can relate to the time trend plot. And that certainly needs to be there. Now, it's not -- there may not be any statistical reliability associated with it in that fashion. But, like I said, it's kind of appealing because everyone can relate to it at least in the beginning of your work.

Now, let's go ahead. This is going to be Chapter 5. We're going to be talking about regression techniques. Now, we're going to talk about quite a few regression techniques.

So why regression? Okay? Why not this? Well, most of these techniques don't allow us to control for anything. And regression techniques allow us to control for variables. They allow us to write predictive models and they allow us to do a lot of things we couldn't do with the simple descriptive techniques.

Now, what I've done here is I've tried to kind of summarize all the techniques that we're going to be using in this course. Some of the more involved techniques. I didn't put all the basic stuff down here, but some of this stuff we're going to be doing throughout the rest of the workshop are up here.

And I think it's a good idea after you're done learning all kinds of new analyses. The first thing you want to do is figure out when to use which one. That's what I like to do.

And it's a good idea to put this in the form, some form of a flow chart. So even for this course you guys might want to go over this flow chart. These are the words I put it into. But once again important to take a flow chart like this and put it in your own words. Make your own flow chart. You can look at mine but make sure you understand how to go from which one to the next one.

Now, if I look at the top of this flow chart -- now a lot of these techniques we didn't talk about yet. So I'm going to keep coming back to this flow chart as we go through the remainder of the workshop.

Okay. I start here. The first thing I'm going to ask myself is is the outcome continuous. Now, we've answered yes and no to that today. When we consider the outcome in the form of a rate, that's a continuous response. When we consider the outcome in terms of the number of infant deaths controlling for the population, that's a discrete count and that's not continuous. We need to need to ask ourselves because the analytical techniques are totally different based on the response type.

And that's with any regression or any type of analysis. If you guys haven't seen some of the stuff we do today, then relate it back to the simpler stuff. If you guys have worked, for instance, logistic regression. If you guys worked with linear regression, think about the differences between linear regression and logistic regression.

If the outcome is dichotomous or ordinal or something like that, then you're not going to use linear regression. You're probably going to opt for -- you're probably going to have to use something like logistic regression. It's the same idea here.

How is the outcome distributed? That pretty much dictates which way you go with these techniques. You won't always end up at the technique of interest, but it will get you going in the right direction.

Okay. So the first thing we need to do is ask ourselves, are we assuming the outcome's continuous or discrete or what are we assuming here? Is the outcome continuous? If it is, the next thing we're going to ask ourselves, let's say it is.

We're going to ask ourselves are the samples independent? Okay. We're going to talk a little more about dependence tomorrow. We're going to talk about two types of dependence, in particular spatial dependence and time dependence, time dependence, of course, is very important for time series data. Are the samples independent?

Now, we're going to be assuming quite a bit that the samples are independent. So that assumption is sometimes made and sometimes isn't exactly true. So that's definitely possible that you can make this assumption and still have it be a little bit dependent. What do I mean by dependence?

Well, think about -- we had a hurricane in New Orleans in 2005, and think about how the reporting would change after the hurricane. Okay. So from 2004 to 2005 -- let's say 2005 to 2006, we're going to have reporting issues that were affected by things that happened in 2005 that wouldn't show up prior to 2005.

So that's saying that my years, that I have time dependence is what that's saying. I don't know how much of that we'll be able to flesh out here, but I can tell you that's what dependence means, is that there's time dependence. You cannot assume that these things don't depend on another. Maybe there's some major event or something that happens and influences everything after a certain time, there's time dependence then.

Okay. If the samples are not independent, we're going to use AREMA techniques tomorrow to model the time series data assuming time dependence.

Now, the idea is find the simplest model that best explains the data. So you don't want to jump right into a time series model unless you have to.

If you can use linear regression, that's great. You might be able to use it sometimes. If the samples are independent or not independent, we're going to consider two types of dependence. AREMA techniques for time dependence, spatial mixed models for spatial dependence.

If we can't assume the samples are independent, okay, we can come down here and then we ask ourselves, are the samples normally distributed? If they're not, we can consider a response transformation. We'll look at this tomorrow. We can consider LOWESS smoothing, we're going to look at that in a few minutes.

If the samples are normally distributed, we ask ourselves is the model linear? And you guys can go through the rest of this. Okay. I'm going to revisit this and tell you where we're at after each regression technique we look at.

We're going to start with linear regression. Notice that's right here. Okay. So I had to answer all these questions to get down there. Were they justified? I can't say in all the cases they were. I'm going to show you guys more appropriate techniques for analyzing these data. But we're going to start here and we're going to assume that they are or we're going to say that we can't do it and we're going to figure out why.

Okay. So if we use linear regression, instantly you know we're assuming the outcome's continuous, which means we're modeling these data as rates. So that flow chart you'll see quite a bit.

Okay. Let's start with linear regression. Here's what the basic predictive model looks like. It's IMR-T. Now, I used the T subscript. You guys see this for the first time.

The T subscripts are typically used like this if you can represent your data in time order. Now, if you have all kinds of covariates in your data, it's going to be difficult to list your entire data set in time order.

Meaning, if you include blacks and whites in your data set, you aren't going to be able to list all those, if you go from 1990 to 2005, you're going to have that by race. It's going to be in time order within a particular race.

When you see just a T subscript, what that means is your data set as it's given is in time order. Meaning 1990 to 2005, that's it. That's what that means when you have a single T subscript.

If you want to put the I subscript there like I had before, that's fine, too, if you're more comfortable with that. I just changed it to let you know we're considering the data, a single data set that's already in time order as a whole.

But no difference in meaning with the subscript at all, really, other than it indicates it's in time order. Okay. Here's what a simple linear regression model looks like. Beta 0 plus B-1-X. Slope intercept form. It's essentially the same thing. B-1 is the slope of the line. B-0 is the intercept. This is a pretty straightforward interpretation. T equals one to K. K is the total number of time periods, or in this case from my study, years.

So some people, now I do this a couple times in this presentation. Sometimes I'll indicate the years as 1 and the K. When you guys are programming, you want to think about how the including years, for instance, as individual numbers affects the precision of the calculations. It may not affect them at all.

I actually studied that for SAS for these years and it really didn't make much difference at all, whether I put the years in themselves or substituted those numbers as one, two, three, four, and so on. So I think you can use either one and get away with it. Your intercept is going to change a little bit. Actually, a lot, if you use the 1990 to 2005 numbers.

Okay. We're going to assume that XT is continuous. Okay. And that may not be appropriate. Okay. If you're assuming that XT is continuous, that's saying that you have a continuous, a constant linear trend over time. Okay. Later, when we look at Poisson regression, I'm going to show you guys how to consider this as discrete and continuous. So this assumption may not be valid. That's a pretty strong assumption for time series data.

You're going to assume a linear relationship between IMR-T and XT, meaning these are linearly related. And you're also going to assume ET is independent. Another way of saying that is that example I mentioned earlier about the dependence, you're assuming that doesn't exist.

And you're also assuming that the errors are normally distributed. Now, assumptions may be invalid. Like I've said, I'm going to start from the beginning and say this is a questionable model. I'm going to show you how to do it because it's used quite a bit. It may be appropriate for some of your studies and surely if

you can use it, use it. Because it has a lot of appeal and it's pretty easy to interpret.

Assumptions may be invalid for these data. When you're dealing -- for small strata, the distribution for rates of rare events is not normal. That's really the purpose of a Poisson distribution, it's to model rare events, rare count data.

Now here you're assuming that -- here you're going to model these as rates and you're not making any -- you're not really modeling these using the right distribution.

The Poisson is certainly more appropriate for count data. What you're going to do is you're going to assume these are continuous. Now, rates are never completely continuous. Because rates, well, not continuous over the entire numeric scale. Rates cannot be negative.

Instantly you're making an assumption that can't be right. But, like I said, if we can relax that assumption a little bit we can use this. The error terms are probably not independent. As soon as you guys see time series you should think could be dependence. That's a classical case of when you would want to consider dependence.

But nonetheless linear regression is commonly applied in mortality trend studies. And I've given you a reference here. It's in the manual, if you want to take a look at it, see how he does it, you can. You can look up the reference and go see his publication, because he did use it for that.

Okay. Linear regression results. Now, we're going to go back and replicate this in the in-class exercises. It's not too difficult to fit the regression models. Here are some of the results I got after running the routines in SAS.

Okay. The first thing I noticed, one of the first things you'll want to look at when you run a linear regression model is R square or the coefficient of determination.

The R square can be interpreted as I don't want to say a good fit because R squared is oftentimes overinterpreted. R square is a good indicator if your model fits well. It's not a formal test. You interpret that as being a good R square is something close to one. If your observations all lie on the fitted regression line, then your R square is going to be exactly one.

In this case, if it's close to zero, now R square ranges from zero to one, with one indicating a good regression line, or a regression line that's useful for making predictions. Now, I have an R square here of .18. This is saying that my regression line probably fits poorly. Now, this isn't a formal hypothesis test, but this is giving me an indication that it probably fits poorly.

Let's overlook that for now. But keep that in the back of your mind. We have other techniques to try, because this probably isn't an appropriate model.

Assuming that was correct, let's go ahead and just go and look at the rest of the results. I'm looking at this estimate here. The PRF. That's the P value for testing the global hypothesis of significance of all the covariates. Now, we only have one co-variety, the time co-variety. So this is saying that the time co-variety is insignificant.

The P is greater than .1. There's a little debate about the threshold of P. There's not much debate when it gets over .1, though. Typically that's considered insignificant. Maybe a little bit of debate between .05 and .1. But this is clearly insignificant based on anyone's threshold. So this is telling us there's no significant linear time trend.

Now, it's important not to forget linear. There could be any kind of time trend underneath that. There could be a quadratic time trend. There can be an oscillating time trend. There's no linear time trend.

Here are some more results. Here are the parameter estimates and here's how you write a predictive model. IMR-T equals 10.44. That's where I'm getting that. It's indicated next to intercept. Minus .06 times XT. That's how you write your

predictive model. In practice, if somebody asked you if this model was good, if it did fit well, if somebody asked you to predict the IMR.

Let's say you guys had all 16, but let's say we were missing year 1993. If you were asked to calculate what someone would expect to get in 1993 based on the rest of the data, you would say IMR 93 equals 10.44 minus .06 times, if I coded this as 1990, 1991, 1992 and 1993 you would put that number here. The way I have it coded is you're going to do 1991, 1992, 1993, you would put a three here. Depends on how you code XT. Is it coded by year, or is it coded with just an ordinary number.

The model seems -- some general conclusions. The model seems to fit the data poorly. So we're not going to look too much more at this. The coefficient beta 1, now like I said there are plenty of times you'll be able to use a model like this. The coefficient beta 1 is interpreted as the actual change in IMR-T in the IMRs per unit increase in XT.

And this change is insignificant. But we can't forget about that model fit. If the model fit well, this is what we would say. Now, the confidence interval, typically you aren't interested in that confidence interval for intercept. The one definitely for year one you are.

Why did I call this year one? Because I changed how year was set up, from 1990, 1991 to one, two, three, four. Like I said, a little easier to perform calculations on.

Now, I want to show you guys how to calculate an average annual percent change for linear regression. That's one -- like I said, that's a commonly reported measure. So let's see how to take these results and report an AAPC.

Well, the definition of AAPC, we looked at this a bit earlier. Let's look in a little more detail and see what an AAPC actually means. I think you guys will be able to relate to this formula even a little more than the one we talked about before.

If you think about changing from one year to the next, the percent change going from one year to the next can be written like this. IMR I plus 1, meaning the IMR in a subsequent year, minus the IMR in the previous year, divide that by the IMR in the previous year and multiple that by 100.

That's an AAPC. Does everyone understand this formula? Because you guys can apply this not only to linear regression, but a lot of other techniques as well.

If you guys understand this, that's what an AAPC is. So when you go off to other modeling techniques and things like come back to this. If you guys want to

calculate an AAPC for it. It won't be like this for everything that we do here. But I can tell you this definition won't change. The first one.

Now, what does that mean for linear regression? Okay, is what we're going to ask ourselves now. Well, if I back up to -- where was I here?
For linear regression, the IMR I plus 1, what is that? The fitted IMR for regression, just IMR for a particular year, if I back up and look at the definition of this, the fitted IMR is this. Or if you guys want to look at it in terms of Bs, we'll go back and look at that, too. Right here. This is an IMR for a particular year.

Now, the expectation of these Es is going to be zero. So that's essentially going to drop out. Worry about these terms. Your fitted model is going to be written in terms of this intercept and that slope.

Now, the fitted IMR or what you would predict, the predicted IMR is going to be written like this. So a given IMR can be calculated as you see here. Now, that's going to be B-0 plus B-1 is what that's going to be. And why is that? Because if we come back here and we say this minus this times this, we're going to plug something in here.

What are we going to plug in for XT? We're going to plug in 1 for XT. The given IMR for a single year. We're going to plug a 1 in. Okay. Now I'm going to add

one. Why am I going to add one? Because I'm going to add one year. I'm going to add one more year to that.

And then I'm going to say minus, what is the IMR for the current year? B-0 plus B-1, that's what this is, if I plug one in for XT. Divided by the IMR for the given year. B-0 plus B-1.

And simplifying I get B-1 over B-0 plus B-1 times 100. That's going to be the AAPC for linear regression. So that's how you'd do that.

Now, with that in mind, this is nice, I mean this is a nice -- if you guys want to go back -- I know you probably need to bring this into clearer focus, when you guys get a chance, please do. But this is a nice little convenient formula for calculating AAPC for linear regression.

So using the parameter estimates above or previously, the AAPC for Louisiana IMR is using a simple linear regression model is calculated like this.

I simply plug these numbers in. Like I said, for now you can think about it as a plug-in, but go back later and make sure you can do this because that way you'll be able to do it for other models, too. If you can take this definition and apply it to a certain technique you'll be able to calculate AAPCs for many different models.

UNKNOWN SPEAKER: For the last equation you had there.

JEFFREY SHAFFER: This one?

UNKNOWN SPEAKER: B-1 should that be a 1 because the previous [Inaudible], should you have 1 over.

JEFFREY SHAFFER: B-0 over B-1, you're probably right. Did I have that here? B-1 -- let me think about this. We have B-0 plus B-1 -- B-0 plus B-1 plus 1. Minus.

If we had it as 1 --

UNKNOWN SPEAKER: Substituting it for X.

UNKNOWN SPEAKER: Should be B-1 right there.

UNKNOWN SPEAKER: 0 and 1.

JEFFREY SHAFFER: No, but she's saying that these cancel out because it's minus B-0 minus B-1 and then I would only have 1 left in the numerator. There's no doubt this is right. IMR plus 1. Let me think about this.

UNKNOWN SPEAKER: The first term should be B-0 plus B-1 plus B-1 because you take the value of one increment. So the slope --

JEFFREY SHAFFER: You're right. Okay. Yeah, she is right. Instead of this 1, this should be a B-1. This 1 here in your notes, you can replace that to B-1 and the reason for that is if we add 1 -- I got it. Thank you.

Think about this, if we were to calculate XT plus 1 times this, we would need the parentheses there. And that's what I -- that's the mistake I made. I didn't have the parenthesis prior to XT. It should be XT plus 1. And that's where you get B-0 or B-1 plus B-1 again. So this should be B-1 then you'll have a B-1 left over and it will still be this.

So the final answer is right. Change the 1 to a B-1 in your notes.

All right. Plugging these numbers in, notice the number I get here. Negative .60 percent per annum. So what this is saying is assuming linear regression, I'm decreasing about .6 percent annually is how you'd read that. Think about this, because we want to -- when we look at other techniques and calculate AAPCs, we still want to remember this one. And if your results are drastically different, ask yourselves why. Now, one main reason could be that this model wasn't very good in the first place. So we don't want to forget that.

I'm going to talk a little bit -- any questions about this by the way? Okay. Next, I want to talk about quadratic regression. And I'm going to talk about this because you can easily extend the model. Not necessarily to quadratic regression, but some type of polynomial regression to fit a little better.

Typically the more terms you have in the model the better the model's going to fit. But the idea is to come up with a model, a very simple model, that explains the data. Okay. When you get a real complex model, it's hard to communicate.

So as you add terms, it gets more complex and harder to explain. However, what led to this? Okay. Why quadratic regression? You want to ask yourselves that. And I think I have a plot coming up. You always go back to the plot. Okay. And you ask yourselves is there anything that -- is there any reason I might want to consider a pattern like this?
A good way to start is with the plot. And it did have a slight quadratic look to it. Okay. So I went ahead and fit a quadratic model. Let's talk about what a quadratic model is.

It's essentially adding a quadratic term or a squared term, in this case, of time to your model. So I add -- I leave XT in there. And typically you do that in practice, if you're going to include the squared term. We also have the linear term as well because you really don't lose any degrees of freedom for adding it. So you might as well have it in there, the linear term, too.

And although you could just have the squared term it's better to leave it in.

Okay. Now, this is what the model looks like. One extra term, nothing else has really changed. We're still assuming that XT is continuous.

And we have a quadratic relationship. We're assuming a quadratic relationship between IMR and XT, or year or time, or however you want to consider it.

We're still assuming that the Es are independent, or you can think about these as the dependent IMRs or the dependent variable is dependent among the years. And they're normally distributed. So that hasn't changed. What we've done here is we've defined a different relationship between the IMRs and the predictor set. We're saying that's not linear. It has a peak or a nadere [phonetic] or whatever, but it's not linear, it's some type of quadratic.

Okay. Here are some of the quadratic regression results. Taking a look at the R square, okay, I like to look at the first R square. I like to look at the adjusted R square where I'm comparing models and I like to look at the first R square when I'm looking at an individual model.

Just looking at R square, that's a lot better. Okay. The .62 is much better than what we saw with the .18 for ordinary linear regression. Now we shouldn't be

surprised that it's a little better. But it's significantly better. There's a significant overall regression now that we didn't see before. So when I test the hypothesis of an association between XT and, well, the joint test of XT and XT squared being related to the IMRs, it's significant now.

It's a different conclusion. But we have a better fitting model. Okay. Here are the parameter estimates. Once again, this is how you guys write this predictive model out just as you see here, XT squared.

B 2 is significant. Take a look at the significance levels. Both year one and year one squared are significant. The estimate is positive. Positive parameter estimates indicate that there is a decline minimum in increase over time. Does everyone see why that is? Think of a quadratic curve. You go down. You reach bottom and then you go back up. That's what you're seeing there.

And that's what the negative estimate means. If it was a positive estimate, it would be going the other way. Okay. This is how we write the predictive model. So if you guys wanted to calculate a predicted IMR for a given XT, you could easily do it. Plug in for XT.

This is a more -- this is one of the better plots I have using G plot for this manual. So if you guys are interested in the syntax for G plot, take a look at the code for Chapter 5 in your index or your appendix and see how I generated this plot. Now

it's an extension of what we did earlier. We generated a line plot, a time trend plot for two lines.

Okay. I just embellished it a little bit, fixed up the axes, drew these horizontal lines and did some things here. A legend, things like that. Now, what is this plot saying? This is a time trend plot with the observed values, the observed IMRs. Those IMRs resulting from the fitted IMRs resulting from the linear model and the fitted IMRs resulting from the quadratic model.

How did I get the fitted IMRs, plugged in here. Use SAS to do it but it's simply a plug-in. I plug XT. If it was the first year I would put a 1 here, a 1 here and get a number. So that's all these are. The predicted IMR plots.

Now, which one is best? Okay. This one, if you notice the line, what you're trying to do, the red line going through the observed values. Doesn't look like it matches very well. A line just doesn't adequately describe the observed pattern.

Now, if you look at the quadratic model, that's the green one, okay, it seems to do a little better. Not perfect, but it does better. So that's where you're at with those three models. Okay, we're not going to talk too much about the ones I have up here, but these are some other functions that you guys might want to consider. Some other response functions.

The quadratic, we just discussed that. An exponential model might be important for mortality studies. I think you guys can see that when you're talking about mortality you may be talking about some type of exponential distribution, particularly at the ends and the beginning. Okay. So that could happen in practice.

This is how it's indicated. This is how it's defined, what you see here. We did this one. Exponential. Logarithmic. It may have some appeal. Power. Sin. This model would be useful for seasonal effects, things like that, things that oscillate over time, maybe a sin model.

Model comparison. Okay. I think we can see that the quadratic model is better than the linear model. But let's go ahead and justify that. We can't just say that based on the R squares. Although that's a good indicator that one fit's better than the other.

Let's go ahead and conduct the hypothesis test and see which model -- now, there's no doubt the quadratic model is going to fit better than the linear model. The question is, is the quadratic model worth the effort? Does it fit significantly better than the linear model? Because if it doesn't, you want to stick with the linear model, it's easier to explain.

Okay. So in particular what we're going to do is we're going to test H0 B2 equals 0. We're going to test the hypothesis that the quadratic term is insignificant.

Okay. And this is the F test. This is a standard F test in regression. You guys will run into this quite a bit if you do a lot of regression. This is called a partial F test, would be the name for this. And it's defined as the sum of squares error for the reduced model, minus the sum of reduced error for the full model divided by the difference in the number of parameters between the two models dividing that whole statistic by the mean squared error for the full model. And this is the distributed F, as an F distribution. Alpha and the DFMC minus DFMC -- the difference in the parameters is one way you can look at this.

An easy way to look at the denominator here this DFMC reduced minus DFMC full, a way to look at that is to consider the number of difference in predictors or the difference in the number of parameters. They're all going to give you the same result.

Formally in the literature you're going to see it written like that. And the DFMC full. Let's take a look at an example of that so we can bring this into a little better focus here. If you guys aren't familiar with sum of squares, just bear with me on this. You guys can go back and look at that a little bit later. But this is a very common test for comparing regression models.

Okay. So for our data, if I back up, let me show you where some of this stuff is, if you guys haven't seen this. I think I gave you guys some of the output here. Okay. I did. All of these sum of squares in this formula I've given you are going to be given to you in the output for these models. For instance, I'm going to pick a few here. You guys see mean square error. This is MSE. That's my notation for MSE. And there's the number. Now, this is for the quadratic regression models. These are the things you guys are going to see when you run the code. Let's find one of the ones we needed. We needed MSE full. What do I mean by full? The more complex model. The model that has all of the terms in it. So we need MSE full. I come back here, look at my output, what is it?

Now, the DFs are given, too. They're all here. So this is just a matter of finding what I need. This is the quadratic model or the full model. MS mean square error .20278. So I take that number. Once again, this turns into just a simple plug-in.

And there it is. .21 in the denominator. So this is just a matter of finding these numbers in the output. Even these are given, where I told you guys the shortcut way to remember that number. These you'll see in the output, too. So if you like these numbers, you can find them in the output. That's how these will be shown in the output is DFMC full and DFMC reduced.

Okay. Now, after plugging all those numbers in, I get this F statistic or the critical value or the test statistic, sorry, the test statistic to be 15.57. The critical value for

this F distribution with one numerator and 13 denominator degrees of freedom, just based on this I know I'm talking through that kind of fast, but that's all based on what I'm giving you here, after plugging those numbers in.

I get this critical value for this to be 4.67. So this is going to be definitely a significant term. I reject H 0 and conclude that I cannot, that beta 2 is not zero. I cannot drop it from the model and go down to the linear one.

So this says that the quadratic model is not only better than the linear one, it's significantly better than the linear one. Statistically significantly better.

Any questions about that? One more regression technique for this chapter, and that's going to be LOWESS smoothing. Now, so far we've made a lot of assumptions using regression techniques. And what we're going to do now is we're going to use a nonparametric regression technique. And the reason, the motivation for doing this is that nonparametric regression techniques require almost no assumptions. So we're making no assumption about the response surface. Before we assumed we had a linear relationship between the outcome and the predictor.

In the second model we assume we had a quadratic relationship between the outcome and the predictor. Here we're not going to make any assumption about the response surface.

The technique we're going to use is called LOWESS smoothing. LOWESS -- and I think I have the acronym spelled out here somewhere -- LOWESS is locally averaged -- what is it, locally averaged --

UNKNOWN SPEAKER: Locally weighted regression scatter plot smoothing.

JEFFREY SHAFFER: Locally weighted regression scatter plot smoothing. That acronym is spelled out in your manual. What this is essentially nonparametric regression. And what we're going to do we're going to locally average the predicted estimates, where the average age is based upon a user defined number of neighboring observations.

Now what this is useful for is defining something about what's going on with your data. And no assumptions are required about the response surface. No assumption of linearity. No assumption of a quadratic relationship, nothing like that.

Now, let's look at this and allow this to come into clear focus.

This is the model for LOWESS smoothing. And notice here I have this written in terms of a single predictor. We're going to look at time. And you can have a single predictor. And you can have either a single predictor or two predictors. If

you have a single predictor it's referred to as scatter plot smoothing, if you have two predictors it's referred to as surface smoothing.

Now, the draw back of LOWESS smoothing is that it cannot be defined in terms of a predictive formula. So with this you aren't going to be able to write out a predictive model. That's not what this is for.

So what is this for? If this isn't a predictive model? Well, it gives you an idea -- now, you can also generate plots of these predicted values, of these values, these fitted values when you're finished. But what is the purpose of this? These are predicted values based on LOWESS smoothing, but it doesn't give you the model in a predictive form so you can predict new observations. You can't write the model like that.

So why would we want this? Well, when you guys are interested in using any regression technique, one use of this is to associate a standard error with a particular point. So if somebody asked you to estimate the IMRs in 1997, okay, and you had a sample estimate. Like I said, you don't get a predictive formula here, you had a sample estimate but you had no standard error to go with it, well, this allows you to get a very accurate standard error. So that's why you would want to use it in practice. But once again there's no predictive model. So we can't talk about things that go on outside of the range of these data, if we don't have any data for a certain year, we can't estimate that using this type of model.

However, if you know absolutely nothing, making no assumption about your response surface, know nothing about that, and that happens quite a bit in practice, then this is a very good technique to use. It assumes nothing else about the response surface.

So let's take a look at these. What I did here is I generated a plot of the predicted IMRs using LOWESS.

UNKNOWN SPEAKER: How is that a moving average?

JEFFREY SHAFFER: It is a moving average. This is a type of moving average. And what you're essentially doing is averaging these over neighbors to calculate, that's what this is, LOWESS smoothing. By smoothing, what smoothing means is you're essentially calculating a weighted average based on the neighbors. And in this case it's going to be a moving average over time.

So this is essentially a moving average is the answer to that. The way LOWESS smoothing works, is it looks at all of the observations within a user defined threshold of time. So if I wanted to define the number of neighbors, and that's called a smoothing parameter, let's say -- now for a smoothing parameter, I don't think I'd want to say I have 16 years. What I want to say, let's say I want my

smoothing parameter to consider 16 neighbors, probably not, because I don't think this time's going to have anything to do with that time.

But keep in mind we only have 16 observations. So there are a certain number of neighbors that this technique's going to require to even calculate the estimate. So the fewer neighbors you consider, the harder it's going to be for this to perform its calculations. But what it is it's an average.

What I've done here is I've also generated a confidence band for each of these. Like I said, that's why it's appealing, is because it gives you a measure of the standard error for each of these points. Now, this would be useful for calling your attention to some type of underlying epidemic, if you saw a spike here and you didn't know anything about the response surface. See, if you assume something about the response surface, let's say you assumed linear regression here, and let's say there was an epidemic here and there's a big peak, you're going to draw a straight line through that, you're not going to see it.

With this one, you're not going to mask anything. You're going to calculate a moving average and if you make the number of neighbors small enough, okay, you're probably going to see that.

So this is very useful for things like that. Okay. Here I used a smooth -- what's called a smoothing parameter equals .5 and this says to include 50 percent of the

observations as neighbors. Meaning when I calculated the estimate for this observation, it considered the observation up here and averaged them all up.

I can set this at whatever I want. If I say include -- if I say smoothing parameter equals 1, that says include all of the observations in my moving average. May not want to do that. Okay. Now what we're going to do is we're going to perform some of these analyses ourselves.

So I'm going to come here and we're going to go ahead and take a look at several exercises. I want to go through each of these regression techniques and make sure everyone can do them. And the first one says we're going to start off with the simple technique for regressing IMR on year using linear regression.

And use the parameter estimates to calculate the associated AAPC. So let's go ahead and do that. Now, I want you guys to use the cross-sectional data. Once again, the details of the exercises are given in the manual. So I'm going to jump up to page 30 and the exercises are there along with the SAS code. Okay. I don't think I calculated the AAPC here. But I'll show you guys how to do it.

Let's start by running a simple linear regression model on these data.

I'm going to start from the beginning. We're using CS data set again. So I'm just going to keep some of the stuff I have. We did this previously. I won't call it

Exercise 4.1 but it is the same data set. Let's call it 5.1. CS data set, the cross-sectional data set. Okay. One little trick here that you guys can see is that I do this. IMR. After this I say year plus 1.

I say year. I don't think proc reg let me put in -- I have to square it there. I'm not going to be able to square that within the regression procedure. I went ahead and created years, even though I want you guys to run linear regression first, go ahead and create the square term while you're there.

Now, notice what I do here. And it's even more important to do this when you're using the square of the variable. Considering year as continuous. What I say here, now the variable in the data set is called birth year. B-i-r-t-h-y-r. I have that variable there and that goes from 1990 to 2005. What I do here is say year plus one, all it does is creates a variable going from one to the total number of observations.

In other words, what this is doing, since my data are already sorted from 1990 to 2005, this would create another year of variable that says one, two, three, four, and so on. It's easier to use, particularly because you have a square variable here.

You've got to square that 1990, 1990 squared, 2000 squared, that's not going to be good. I did test this for simple linear regression, okay, not doing this, not

substituting that with 1 through the number of years, and I didn't get any differences. It just changed the intercept. But I'm sure that you would get differences here.

And I'm sure that it wouldn't be as appropriate for the square term.

UNKNOWN SPEAKER: [Inaudible].

JEFFREY SHAFFER: Yeah, it very well could. That's a pretty big number. It may work. But even if it doesn't it makes a little more sense like this.

Okay. That's it. You're creating those two terms. Now, the regression in SAS is really easy. Proc reg. Exercise 5.1. The GLM procedure, some of you guys might use that, too, to perform regression analysis as well. That's fine. I'm just going to use the proc reg here yields pretty simple output. The output's a little simpler to look at.

Okay. And that's it for the simple linear regression model, that's it. And, by the way, this is a test for trend. Don't forget that. This is a formal test for trend. But you're assuming, you're testing for a linear trend over time.

So we had other tests for trend in the beginning that didn't have a predictive model, but that's what this is. Now, we concluded that the model didn't fit very

well. Let me go ahead and close some of this out and run this again. All right. Year one not found. Year one plus one. The reason I call that year one plus one is if you suffix it with a number like that you always ask yourself what you did to it, why you didn't call it year. And this will remind you that you did change how year actually looked.

Now, let me rerun that. Okay. And there it is. If you guys want to, go ahead and run a proc print on this data set, or open it up, either one. Take a look at it. See how that -- what this does is this says create a new variable called year one that just goes from one to the number of observations.

So we can take a look at that if you want, Exercise 5.1. And I come here. Here's what it looks like. Notice what I did.

I didn't square year one. Notice how big those numbers are. Right, exactly. So I think what I'd want to do is come back here and say year square equals year one square, so I did it in both places, and I might have done that in the notes, too. I think I did.

So you guys can correct that in your notes. Where I have years SQ equals years square, you guys can put a year SQ or put a year one before the double asterisk, right here, where I have this highlighted. It should be year one squared. Now, you'll get a regression model, okay, based on what I did before. It isn't going to

hang that up, but you have to ask yourself looking at those numbers, is it going to be appropriate.

Let's rerun this. Anyhow, that's why it's a good idea to do a proc print, proc contents, proc univariate, proc freak, things like that, make sure it's in there right. So year one goes from one to 16. The only reason I could do that in that easy of a form is because these data were already in time sequence. 1990 to 2005. If they weren't, I would have had to sort them beforehand or that wouldn't have worked right.

Okay. Now, let's take a look at some of the regression results. Here are the results that you guys saw earlier for the simple linear regression model. If you guys want to run the quadratic regression model, okay, it's just a matter of coming back here and putting in year SQ.

Year SQ. That's it. Add that to your regression model and run it again. And there it is. Those are the results we went over there. Regression's pretty easy to work with in this form.

Okay. Now, let's go ahead to Exercise 5.3. I say write a SAS program -- by the way, before we move on to that, I didn't mention the AAPC. Let me run the simple linear regression model again. How would you guys get the AAPC from this? That would be negative .06 beta 1. Beta 1 divided by beta 0 plus beta 1

times 100, that's where that would come from, right here, these estimates. You'd need only these two estimates to calculate that. And use that formula I have in the notes.

UNKNOWN SPEAKER: Just for the linear one, not for the --

JEFFREY SHAFFER: Right. A little more involved calculating the -- if we wanted to do that for the quadratic one, we would have to go back to my definition of an IMR and solve that little definition just like I did earlier.

So we'd have to do a little bit more work for the quadratic one. Any time you're interested in an AAPC, though, that's for linear regression, the one I did. Any time you're interested in AAPC, though, you just come back up to that one slide right here, and you say this is what an AAPC is.

How can I convert this into an AAPC? So based on the first definition you're going to have to go through that for quadratic regression and do what I did here, do what I did here.

As long as this meaning is intact, you'll get an AAPC. But I haven't done that for the quadratic.

UNKNOWN SPEAKER: And it will probably be a function of X.

JEFFREY SHAFFER: Yes, very well could be.

UNKNOWN SPEAKER: Of the year.

JEFFREY SHAFFER: Sure.

UNKNOWN SPEAKER: So you won't get a single number.

JEFFREY SHAFFER: No, it gets -- AAPC definitely gets a little more complex to calculate with more complex models. No doubt. But they can be calculated.

Okay. Now let's go ahead and Exercise 5.3. Write a SAS program to regress IMR on year using a LOWESS model. This is assuming we know nothing about the response surface, which is actually a good idea at the beginning of a study to make that assumption that you know nothing, because it allows you to go a certain way.

By the way, LOWESS smoothing can often be used to define another predictive model. It might give you an idea about your response surface. So if you look at the LOWESS smoothed estimates, you can look at that pattern and you can say, well, maybe this is a quadratic model. Let's go ahead and try the parametric form of that.

What do we have next? The LOWESS smooth model. This is going to be proc LOWESS. Same data set. All right. Let's talk about some of this syntax. Well, the first part of it is pretty straightforward. IMR equals year one. I've chosen the year that goes from one to 16 rather than the 1990 to 2005.

Okay. Degree equals 2. I'm seeing here now there are two degrees for the calculations in LOWESS smoothing. You have degree 1 and you have degree 2. My suggestion to you is to try both and see how much they differ. There's not a complete agreement about what this should be.

But I can tell you that it only has two ways of fitting it. One is with degree one. The other one is with degree 2. The thing about degree 2 is like a quadratic or curving. The other one is going to be in the form of a line.

Take a look and see which one is -- if they differ much at all. They may not. Direct, okay, that's almost always indicated. I guess you guys can think of that as part of the syntax. Smooth we're going to change quite a bit. We're going to change that from .5 to other things, because this we actually do have a lot of control over. Not so much the other things.

But smooth is going to define the number of estimates to include in an average. This is a moving average. This says include 50 percent of the total number of observations for each locally weighted average.

Now, we have to compare this, because we may not know how many to include. We as statisticians have to define that, but it's not always clear what should be put there.

So what we're going to do we're going to compare some. Alpha equals .05 just showing you guys how you could change the alpha value. These two last options instruct SAS what to print to the screen, what to put in your output file. This says I want to see the details. I can't remember exactly what that does, but I know it shows more output.

So if we took it out we can see what it does. Okay. Let's go ahead and run this. Like I said, we're going to be interested in changing this smoothing parameter.