MCH EPI Conference

Time Trend Analysis for MCH Outcomes in Large and Small Populations

December 8 – 11, 2008

JEFFREY SHAFFER: To using the predicted IMR as opposed to the observed IMR. What do you guys think? Why would you want to do all this?

What are the advantages of this over the other one? Over the observed? Why not just take the observed? 11.04. Because you can't say with any reliability what it really is.

The idea is to say with a certain level of confidence what the true value is. If that's an estimate, then you don't have any standard error associated with it. So that's why you would want to do this.

Okay. Exercise 5.5, write a SAS program to generate a plot of the fitted values with a 95 percent competence band for the LOWESS predicted value generated in Exercise 5.3. The LOWESS procedure is one of the procedures that has one of these real spiffy plotting techniques of its own.

I'm going to show you how to do this. Some of the advanced plots you see in my programs I had to do manually. Now this, not all procedures allow for what I'm going to show you up here for the LOWESS technique. But the LOWESS technique does allow for this.

Let's go ahead and do this. I'm going to modify this technique, this procedure step right here. So the procedure step we just worked with, come right above it. Exercise 5.5 is just an extension of what we just did.

Type ODS. ODS is an acronym for Output Delivery System. ODS HTML, style equals minimal. ODS graphics on. And what I call this plotting technique is an ODS graphics option.

Now, some procedures have this, and they generate all these spiffy plots for you by default. They have prewritten routines to do this.

Okay and right below it type ODS graphics off. And write underneath that type ODS HTML close. And run the program.

And let's see what we have. It might take a second to generate these plots, too. If I did this right it's going to be coming out.

Okay. I think it's chugging away. I think these plots take a second to generate.

Okay. Change my color scheme. Still working. Now, if the procedure has this option, okay, it's nice. It's going to save you a bit of work in generating nice plots. But, like I said, not all procedures have this option.

What I come here to the viewer, the results viewer, notice what I get by default. A nice plot with a nice confidence band. Decent. Get a lot of plots for assessing the diagnostics by default.

Take a look at this stuff. Really good stuff. All by default. And there are many procedures coming out with an ODS graphics option. Like I said, go to the help documentation if you're unaware of which ones have it and which don't. I know proc core does. I know proc Linux does. I know some procedures that do. But, like I said, a lot of them don't.

But LOWESS definitely does. It's here. This is the one we've been talking about all the time. This is what you would report. This is a 95 percent confidence band.

The lines, the fitted line is what you guys see there. Those are the predicted estimates. They're smoothed together. But this is the one you'd be interested in. Some of this other stuff is there, too. Now, let's go ahead and go back and change the smoothing parameter. I think I asked you guys to change it to .75 and 1. So I'm just going to come back here. Just modify your existing code. Change .5 to .75. Because you are defining this, the number of neighbors.

So I mean you have to make sure you're choosing the right number there. And that isn't always easy to do, because sometimes it's just not clear of which number should be specified there.

Sometimes it may make sense based on the study. But sometimes it won't. Okay. Let's see. This is for .75. Now, notice here, notice my confidence band.

It's smaller. I'm including more observations in the moving average. The confidence band has to get smaller. That's what we get with a smoothing value or including 75 percent of the observations for each of these.

Now, what about these points lying outside of the confidence band? Are those important? Those are the important ones.

That's saying this is kind of an aberrant observation here. That might need attention. What happened in this year? It's not what we would expect with this model. It's outside of the confidence band. I would look at this observation -- some of these observations at the outside of that confidence band should be worth investigating. Those are the important ones there.

The other ones, what this is saying is the other ones aren't surprising. It's saying that these observed values are surprising based on this model.

You guys have any questions about this? Are we at break time yet?

UNKNOWN SPEAKER: I was going to say to note how this thing is splitting into two lines, which brings you to join points.

JEFFREY SHAFFER: That's a good point. Next we'll be talking about join point regression. Join point regression involves fitting multiple lines to a set of data. If you think about the linear regression case, we were fitting one line to a set of data.

Notice here, if we fit two lines to these data, one going like this and another one going like this, that might be a pretty good fit. Not only that, but if we fit a model like that, we would be able to detect where the point is. And that point is referred to as a knot or join point. Hence join point regression.

UNKNOWN SPEAKER: You get to worry about that year and what happened at one point and see if you can find explanations for why it was in such and such a year and grew afterwards.

JEFFREY SHAFFER: What time were we supposed to do break?

UNKNOWN SPEAKER: I don't know. About 3:30.

JEFFREY SHAFFER: Okay.

UNKNOWN SPEAKER: After you finish this chapter.

JEFFREY SHAFFER: We have a break scheduled for 2:45. I know we're running a little behind, but we will get caught up. If not today, definitely tomorrow. Let's go ahead and take a 15-minute break. It's 3:30. But we were slated to have a break at 2:30. Do we have our -- I think we should probably take our afternoon break now.

So let's go ahead and take 15 minutes and you can guys can ask me any questions you want now, but we'll go ahead and meet back at, let's say, 3:45. I'm showing 3:30 now. We'll say 3:45. We'll start with join point regression when you get back.

(Break).

Okay, you guys, now we're going to start join point regression. And, like I said, join point regression involves fitting multiple lines to a set of data.

Now, let's see where we are. We start here. And we ask ourselves, are we going to consider the outcome as continuous or discrete?

In this case, we're going to say no. Okay. We're going to actually consider this -- consider the join point data as discrete. Now, we're going to say no. And we're going to ask ourselves are the samples independent? Okay. And I probably could have put a segmented Poisson regression model here, and that's essentially what we're doing with join point regression is we're going to model these data using a Poisson distribution and perform join point regression. And so join point regression can be used with a few different types of response functions.

The response function we'll use is Poisson. So I'm just going to go ahead and lump that into this category of Poisson regression. Think of the join point regression we're going to do as a special case of Poisson regression.

Okay. So that's where we are in the flow chart. Now, once again, we've assumed in the flow chart that the samples are independent. No time dependence. We're going to talk about the trend over time. We're going to say the times are independent from year to year.

Okay. Now, let's talk a little bit about join point regression. It's also called segmented regression or piece-wise regression. And what it involves doing is fitting multiple lines to a single set of data. It's useful for describing shifting time

trends and abrupt changes in trends. And it does permit a Poisson model specification, and that's the specification we're going to use since we're working with rare count data, rare events.

Okay. So the join point regression model is defined like this. Let YI equal ZI over NI. YI still is the -- I've tried to be as consistent as I could with the notation through these models. YI is ZI over NI. And I think I have everything defined down here.

NI is still going to be the number of live births. YI is still going to be the number of infant deaths.

And we want this value to be distributed as Poisson with unknown parameter or unknown mean lambda I and number of live births NI. So we're going to make this model specification here. And this is going to be the fitted join point value for some independent time X.

And I haven't told you guys what X is, because the model is -- these independent times are going to be defined when we talk about the actual model. Now, here's what the model looks like. I know it's a little confusing to look at at first, but I want to show you guys that, first of all, you see an X here. That X is an independent time.

So that's some time. That could be some year. Maybe year one. Maybe 1990. You have it defined as that. That's what that X is. The Taus, these ones here are going to be the unknown join points, or where the two lines meet, the points. The intersection points.

Now, those are unknown parameters. This model involves fitting this model here and estimating these parameters. These parameters are of interest. This is where your join point regression lines are going to reach a peak, or something like that. Or reach the low point.

Okay. So that's what you're going to be estimating here. Where are these. The alphas and the betas, we've seen these before. Alpha and beta, these are unknown parameters. This is going to be an unknown intercept, the alpha. The beta is going to be an unknown parameter as well. Next to the independent time X.

Okay. So what we see here and the gammas are also going to be a set of fixed unknown parameters. So these are all the parameters you have to estimate. Now, I know this is a little confusing right now, and we're going to revisit this when we actually see some output. We're going to come back and substitute in here. And I think it will come into a little more clearer focus.

Okay. The join point regression software. We're going to use this today. Commonly used for these types of studies. It's a small convenient software package. I've installed it on each of the computers here. So we'll look at it. Some of the limitations of it -- now, this join point model, what we're going to is we're going to use a prewritten routine, written to perform this type of analysis, to run this type of model.

Now, notice here the response function. Log. We're going to model the log. The log of the events. Why is that? Well, for Poisson regression, it performs best if you define a log link. That means that all of what you see here is not linear. It's linear, you're assuming that's linear in the log.

So typically when you model Poisson data, you're going to indicate a log -- that's what's called indicating a log link. Meaning you're modeling the logarithm of the outcome.

Now, the join point regression software, we're going to fit, like I said, that model there. It's a small convenient software package. It's freeware. I've given you guys the URL, where to download it, in the online resources appendix of your manual. It's real easy to use. I suggest you guys go download it and use it.

Don't get intimidated by this model. It's not difficult to fit this model. It looks hard here but it's very easy to fit and interpret. And then come back to this and match up what you have with this. This is the hard part here. The easy part is fitting it.

Some drawbacks of this software. You have to perform separate analyses for controls. Meaning, if you were interested in race-specific or join point regression models, you'd have to do them separately. You couldn't just say I want to control for race and remove its effect and look at a single line.

So that wouldn't be able to be done. It does generate APCs automatically. So that allows for a straightforward interpretation, along with the confidence intervals.

And it's useful for detecting abrupt changes. Now, when you guys are trying to figure out when to use what, look at this model and try to see what the motivation is for this. What was the author's motivation for such a model? That's going to tell you whether it's appropriate to use or not.

To me, when I look at this model, I see a bunch of knots, I see a bunch of knots here and they're going to be join points. Knots and join points are the same thing. Some people call them knots. Some call them join points. And some call them intersection points and so on but I see a lot to do with knots.

So this is telling me that the motivation for this model is probably very much dependent on these knots. So the main motivation for this is to look at the knots and talk about the knots.

So join point regression is useful for detecting changes, abrupt changes, finding knots, where do the time trends begin to change? Where do they go up and down? Where is that point? That's when you want to use this.

Okay. Useful for detecting abrupt changes but does not allow for statistical subgroup comparisons. At least using this software. I shouldn't generalize that statement to all software packages. I've used this software because it's pretty convenient, and I know you guys will catch on to it right away.

If you why you a more complex model using SAS, if you write this code using SAS, I think -- actually I'm sure you can perform subgroup comparisons. But for this software, for this little software package, the covariates are included via stratification.

Now, a couple of things I want to mention before we go ahead and get -- we're not going to take too long before we get started on our exercises on this. I just want to kind of guide you guys through it.

Some of the things I'll mention up here, though. Specifying the model. These are some of the things you're going to see in the interface. I've just taken some screen shots here of some of the pieces you guys are going to see.

Okay, when you're modeling these data, well, I choose -- I say choose the log linear option, because I'm talking about infant mortality. The reason I choose the log linear model is because you model -- you perform Poisson -- this is a type of Poisson regression. You perform this using a log linear model.

I have rare events. So I'm going to use the Poisson approach. So, therefore, I choose the log linear model. And that's this LNY equals XB. If you guys weren't doing -- if you guys didn't want to use a log linear model, suppose you didn't have a rare event and you wanted to model these data as you did in regular regression, you could easily do that.

Choose this one. Okay. That says don't model it as a log. Model it as an extension of regression, of linear regression. Okay. Choose the Poisson model using a rate option. Once again, I've chosen this but for your analysis you might want to consider some of the other options.

Constant variance, Poisson model using a rate. That's exactly what we're doing. So if you choose this option, you're going to have to define a population. If you

want to use -- we want to model these right. We don't want to model these as rates if we can help it.

So what I'm going to do, I'm going to choose an underlying population so I can talk about my estimates as estimates of risk. So I choose this one. Poisson model using a rate. If I didn't have the populations, suppose I wasn't given the populations, I was just given the counts, you can model those, too. There's an option for that. Poisson model using count.

That says you don't know the underlying populations. This may come up. I can think of some cases. A common Poisson distribution is like the number of traffic accidents in an intersection or something like that. You may not know the population. How many cars came through that intersection. You may just want to consider modeling the counts.

So that could be used as well. And you'll need with unless variable to do that. But certainly if you have the populations, use them.

Some other things that are given in this software package are -- and you guys will see this referenced quite a bit in the literature. This is used quite a bit for these types of studies.

A maximum number -- a maximum number of four knots can be modeled. Meaning this is capable -- this software package is capable of modeling four knots. You can have four different changes. That's all it's capable of, whether you have more than that, I don't know.

This is what it's going to permit you to do. The default is three. Now, keep in mind that even though you want to detect the knots, it's much easier to explain something over a long period of time. So there's a balancing act going on here. You want to detect the knots but you don't want 100 of them because you're not going to be able to generalize over any time period at all.

So this allows for a maximum of four, the default is three. I usually leave the default there. But you can jack it up to four if you want. Covariates like I said are included via stratification. This is what you see here. We'll look at this in a second.

Model selection. Okay. I fielded an e-mail about this, so I went ahead and put this into the presentation, too. Okay. There are two different techniques for selecting the best model. One is the permutation test, and one is the bayesian client -- bayesian information criterion.

Okay. I have an affinity for the permutation test, and this is the reason why. Now, there are other people who probably like the bayesian, BIC better. I like this one more. And this is why I do.

First of all, the PT has the ability to increase the power of the test by increasing the number of permuted data sets. We'll see that later.

We can essentially increase the number of data sets it considers. We can indicate as many as we want and the test is going to become more powerful. We don't have that opportunity for the BIC. Talk about that later. It's kind of hard to articulate right now.

Second one, as per the join point documentation, I found this, what it said in there, they say the applications have shown that the PT approach works well for incidence of mortality data. They don't really provide any justification for that, but the authors of the software seem to like the PT approach for incidence and mortality data.

And last, which is really why I decided I liked this one more, it says here: Permutation tests are preferred -- this by Brendon and Bents, 2008, and this, once again, is in your set of references. Permutation tests are preferred due to nondifferentiability of the model and the existence of local minima in the sum of squares.

What that's saying is at the points, these points aren't differentiable and that's really hard to work with points when you're differentiating or doing any kind of integration, which is essentially what you're doing in using the bayesian information criteria.

Now, I'm sure one could perform a simulation study or something comparing these two. These certainly aren't -- this isn't chisel and stone. This is why I have a preference for the permutation test. But somebody else could easily have an argument for the other one.

Okay. With that in mind, let's take a look at some results. The nicest thing about join point is it's very easy to interpret. Very straightforward. Here is what you guys see up here is a fitted join point regression model to the cross-sectional IMRs. I didn't stratify about this by race. This is both races. I included them both in here just to give you an idea what this was like. What was detected was a single knot or a single join point or a single intersection or just a single point. Referred to all those things in the literature, right here. That's in 1999.

That says there was one abrupt change or one change, and that ended upcoming in 1999. There was a single line that could adequately describe what happened from 1990 to 1999, and there was another single line that could adequately describe what happened from 1999 to 2005.

Now, up here it looks like, if I look at these APCs, the APCs are reported right here. That says the annual percent change for line one equals negative 2.20. See how easy that is to look at? That's really nice in join point regression.

Very easy to interpret. Okay. So we're going negative 2.2 annual percent decrease in IMRs from this time to this time. Notice the asterisk beside it. That's telling you that is significant.

So that decline is significant. The second one says positive 2.2 three but there's no asterisk beside it so it's insignificant. So this is how you use join point regression. This is what it looks like.

Now, like I said, the main purpose of this, I think, is to detect the abrupt changes. This is saying you have a change in 1999.

Now, since the join point regression only allows you to control for confounders via stratification, at least the join point software does, that means you'd probably want to use this in practice in conjunction with another technique like regular Poisson -- full blown Poisson regression, as we're going to look at tomorrow.

Okay. Some join point output. This is the first thing you're presented with when you run a join point regression model, something that looks like that. That's the graph.

Now, some join point output, what you see here, is the parameter estimates. And P values. Confidence intervals. Notice down here the one at the output component at the bottom of the screen is very appealing. It gives you a confidence interval for your APC, for this time period.

It's saying that it detected one knot at 1999. And these are the confidence intervals for the individual lines. Okay. And I show you guys here, even though they've given you the APCs, they've given them to you here. I show you how to generate -- how to calculate those APCs using this output. It's always nice to know how they're getting these results up here.

Here's where they indicate the join points they found. It's saying here that you have a join point estimate of 1999. 95 percent confidence interval. That's important, too. It's saying that you're 95 percent confident that you have an abrupt change between 1995 and 2001.

Now, for these data, that's not very informative. That's a pretty large range when we're talking about only something from 1990 to 2005. But it is saying something. These are the only years we had to work with.

Okay. So it's one thing -- you essentially have two samples here. Don't forget that. You have a sample of live births and also a sample of years. So even though we have a lot of live births, we still don't have that many years. And that's why -- that's one of the reasons why this confidence interval is so large.

Okay. Let me move up. Okay. Here I'm just showing you guys how this number was calculated, this 2.20 based on the output. Okay. It's going to be EXP slope one minus 1 times 100 and you guys can look at that later.

This is how that was done, though. Here I've given you the output correspondents to go from the output to the model.

So if you were interested in plugging, basically what you see in join points output, plugging these numbers back in the model to write a predictive model, which you might want to do, you might want to predict a value.

Okay. So how you do that is right here. So intercept one in the output label is equivalent to the alpha in the literature. In the literature, the model's indicated like this from the authors of the software.

But if you want to express this model in terms of the output labels, this is how you would do it. Intercept one means alpha. Join point, where it says that, J, join, PT, where is that? Right here. Okay. There's its estimate, 1999.

So that's what we would plug in for Tau 1. That's the estimate for Tau 1. Slope 2 minus slope 1, right here. Gamma 1.

Now, some of these might be appealing. I mean, you may be interested in is the difference -- is this difference significant? I mean, that's not given in the APC, but if that's of interest to you, you can come here and see. Okay, there's a P value associated with that. .03. So a lot of these things beyond just the APC might be of interest to you as well.

Okay. And okay, so I go ahead and I do this and I say here for these data, here are your output components. Based on this little table I gave you to go from output component to model, or vice versa, depending on your perspective.

Okay. You're given these parameter estimates. You're given this model, because you're going to -- from these, you're going to plug these in here to get that. Once you get that then you plug these numbers in, these associated numbers, you get this model and then you're off and running.

You can calculate a predicted estimate. For a particular join point. So if you're asked to calculate the predicted IMR for year 2000, your X is going to be 2000. That's what's going to go right here. 2000 here. 2000 here. And calculate this estimate.

You're going to get the predicted values. What does this plus sign mean here? That says if that value happens to be negative, set it to zero. Meaning if I plug a value of 1998 in for X, that says make that -- wipe that whole term out. That's how it's defined in the model.

Okay. So nothing negative could go in there. All right. Let's go ahead here. And the predicted value at that point, if I wanted to calculate the predicted IMR in 1999, I just substitute in 1999. That's it.

UNKNOWN SPEAKER: So even though when you were setting up your model specifications you had the parameter of the number of knots from 0 to 3, through the different permutations when the program's running the different, I guess the algorithm, it then just finds one join point and that's all --

JEFFREY SHAFFER: That's all it found. Not only that, but it actually performed the analysis based on zero knots, one knot, two knot, three knot, four knot.

UNKNOWN SPEAKER: So this is considered the best fitting.

JEFFREY SHAFFER: This is considered the best one. This is the one it found and the one it selected.

Now, in practice sometimes you may see one that it found and you may not agree with it. I've seen a couple of those. It just didn't look like it fit very well. And some of the other ones looked like they should have fit better.

UNKNOWN SPEAKER: So is there a way to force it, or what --

JEFFREY SHAFFER: It's usually based on not having enough data. That's usually why that happens. Now, one way to do it is to go with more knots, to force more knots. But you're not going to end up with those being significant anyhow. It shows you all the different models. But what it's going to say is here you go. Here are your four knots but three of them are insignificant.

So I mean you have a different model and it looks like it fits well and you may want to use that as a predictive model, but you almost never want to go with something that's insignificant. That's the problem with that.

I don't know if I have a really good answer for that based on this join point regression package. Do you, Dr. Mather, have an answer for that?

I don't know if I do.

UNKNOWN SPEAKER: Because I'm just thinking about times where you might have a model that's .05 and another model that's .07 but because of the sparsity of data or something ambiguous, you know, the .07, yeah, it might not be statistically significant, but if you look at the underlying context for that model, you might want to pick that one because it actually does model your data better, whereas .05, yeah, it fits what you have.

JEFFREY SHAFFER: Well --

UNKNOWN SPEAKER: So just thinking about --

UNKNOWN SPEAKER: You're left shifting, right hand to the left-hand to see which makes the most sense. And I remember running a few of these. And as the join point or as that knot moves over and gives you less data on the far side, then that form doesn't have enough information to settle down.

UNKNOWN SPEAKER: Because when you look at the points, you kind of bottom out sometime around '96, that fluctuation between '96 and '99, and the real inflection point is '99 where it never goes back again.

So I can see why it picked this with one knot.

JEFFREY SHAFFER: Well -- but --

UNKNOWN SPEAKER: It seems like there's this little ambiguity, I guess.

JEFFREY SHAFFER: It's saying it goes from 1995 to 2001 there was a knot. That's essentially what it's saying. Which isn't exactly what you see on the graph. This is the best we can do is an estimate. But that confidence interval says that could be here.

UNKNOWN SPEAKER: Just as an ordinary line, you just don't have a lot of confidence in the line. I think I would agree that probably bottoming out in more or less the right place, but I just don't have a lot of confidence on what the deepness -- I'm sorry, that thing is not going up forever. There's a knot over there somewhere that it's probably going to settle down and that sort of thing.

UNKNOWN SPEAKER: Then it tells you for each knot its level of significance.

JEFFREY SHAFFER: Yeah.

UNKNOWN SPEAKER: Okay.

UNKNOWN SPEAKER: Yeah.

UNKNOWN SPEAKER: Where this got its start or I think the original set of data that they used was prostate cancer incidents. Remember in the early days, I think in '85, they went in to do the testing. They had done testing before. But they hadn't used it as a screening mechanism. They started to do it as a screening mechanism, and of course all these cases of prostate cancer came out of the woodwork, because they've just been there but undiagnosed in the pool of people before.

So what you had was the early levels, you know, low level, the steep increase in things and then it began to level off and drop and finally drop off again.

So they were interested in seeing whether the times were consistent with this push for screening. And you're usually looking at some policy change to see whether it was there but here are our -- I don't think that we have the cleanliness of data here for deaths, infant deaths that I'm pretending that cancer data is clean and I don't know whether it is either, but as that stuff goes.

JEFFREY SHAFFER: Just to clarify, what you're really saying for this model is you're 95 percent confident that you have a knot somewhere in this model, which isn't very informative. You're also saying that you're 95 percent confident that you have a decline of negative 2.20 percent per annum in IMR from 1990 to 1999.

And also you're saying if you're going to use this model to predict, the best you can do is what you see here. The best you can do is this. Now, how good is that? Well, we already saw that X goes from 1995 to 2001.

So this is the best predictive model we can get using this model. But I guess I would look at this with a little bit of caution here, because of that large range on X. I guess I'm not so confident that for prediction purposes I'd even be more concerned about this.

Typically with prediction, if you're interested in prediction, if you're interested in prediction, standard errors are going to be a lot more influential than if you're just interested in significance.

That's always the case. So if I'm interested in actually calculating a predicted value, I'm going to be kind of concerned with that variability on X. If I'm just interested in saying is there a significant -- is there an abrupt change, I'm going to be pretty confident saying there's an abrupt change between 1995 and 2001.

But that predicted model I'm not that confident about. And I see inflated standard errors, that's the first thing I think of is that prediction probably isn't going to be that accurate.

Okay. Let's go ahead and move on up here. I just plugged these numbers in, and I end up, I say here calculate the predicted IMR in 1999. Well, you're going to be given these predicted values anyhow for the software package is going to give those to you. It's nice to know how to plug those in if you need to interpolate, you'd know how to do it then.

If you're given all the predicted values already, this would be redundant here. Like I said, you may not always want to be given the point. You may want to predict a point you wouldn't have. You use that predictive formula to do that, in this fashion.

Now, keep in mind that the kind of data we have here. Even though -- let me back up. We're looking at these. We can plug any year value we want into that predictive formula. I could plug 1995 and a half in there. In other words, instead of -- let's say I wanted to plug the mid point for between 1990 and 1991 in there.

Now, would it make any sense to do that? You could do it. You'd get a predicted estimate. But it isn't going to make any sense because your reporting took place annually. Okay. Even though we've connected the dots here, it's kind of meaningless, if you think about it to -- this allows us to look at the trend. But what's going on between the years? That's not what the data really -- that's not what the data is able to show. That's just a convenient way of viewing it.

Keep in mind, nothing's going on between years that we can measure. Okay. Let's go ahead. Predicted IMRs. Here are the predicted IMRs for the whites using a join point regression model. Now, I stratified these by race is what I did.

Very similar pattern for just the whites. Now, notice both of these regression lines are significant. Both of them. So I'm 95 percent confident that the average annual percent change is this, negative .3 percent per annum, between 1990 and 1999. Looks like it picked up the same knot, same join point for whites as it did for both races combined.

The second one here is also significant. That wasn't the case for both races. Take a look at blacks. This happened sometimes in practice. This is what the software picked. Now, I'm not confident at all in that line.

I think what this is telling you is that a join point regression with four or fewer join points isn't the appropriate model for these data. That's what that's telling you. It didn't seem to fit very well to me at all.

So not all the time can you accurately describe a trend using four join points. Your data may not be like that. And that's what this is telling you here, that you better have another model specification, because this isn't the right one.

At least that's how I'm reading it. Another way of reading it is to say, well, I don't have the APC from here to here is not significant. Negative .5 is not significant. I can't get over how poorly that looks, the fit looks. I can't get past that. So I guess I'm saying that fitted join point with four points isn't going to fit these data for the blacks in this case.

Join point conclusions before we start the exercises. A few knots facilitate interpretation. Certainly the fewer knots, the easier it's going to be to describe over the entire time period. But additional knots lead to a better fit.

Further, there's something to say in just detecting a knot. Maybe if you're not interested in just talking about the lines but finding out where the changes exist, maybe many knots is appealing. So it depends what you're trying to do.

If you're trying to generalize over a long period of time, many knots is probably not that great. But if you're trying to identify abrupt changes, many knots or however many knots you have is good.

Okay. Race is clearly a significant predictor. Even based on this. Like I said, the join point software doesn't allow you to go ahead and control for race, but based on these two plots, okay, I'm going to look at the scale. If I'm going to compare races like this, I'm going to look at the scale of these IMRs here.

For whites I'm going between 5.8 and 7.7. For blacks I come over here. Notice, I'm between 13.8 and 15.2. Okay. So even there I can say that there's a difference. I just can't quantify it.

1998 to 2000 is the lowest point for whites. Now, let's take a look at that. I didn't give you the output for this. But that's a much stricter confidence interval than we saw before. That says it bottoms out in 1998 to 2000 for whites. Apparently we had a better model this time.

And the fit of the line for the blacks is poor. I don't think we can say too much about that other than the model specification probably isn't right. Okay. Now what I want to do before we go is let me see how many -- it's 3:30. Let's go ahead and at least start a couple of these.

UNKNOWN SPEAKER: 4:30.

JEFFREY SHAFFER: 4:30. Sorry. We'll get as far as we can. I think Henry wanted us to cut out here today at about 4:45 to go to the.

UNKNOWN SPEAKER: Downstairs so we can fall in the pool.

JEFFREY SHAFFER: Downstairs at 4:45. But we're going to do a couple of exercises before we go to join point. This is an important topic for this course. We'll pick up -- we'll finish the exercises tomorrow.

Now, exercise 6.1 says to use the join point software to model the cross-sectional IMRs from 1990 to 2005. What I've done here, in my manual, I have a guided solution for, I think, all of the -- at least some of the exercises, guided solutions.

Meaning everything's written out step-by-step what you should do and explanations of these things and so on. Now, I'm going to go over this on the screen up here. But like I said for future reference you guys can come to the manual for a more detailed explanation of the things I'm going to cover up here.

Okay. I'm going to be looking. I'm going to be going from Exercise 6.1 on page 33 of your manuals. So I've given you everything there. Now, I've set up these data for you so that we can use these data in join point. And I want to mention, because half the battle in join point is setting up the data.

So let me show you how you have to set it up. Okay. And there are a couple little tricks here. Let's see. Exercise -- not that one. Let me get the right data set and just show it to you.

There it is. Okay. Now, notice how this data set's set up. First I indicate year, or time, something like that. Deaths. No big deal there. Births.

Now, notice here that births looks a lot different up here than it does for the Louisiana data that we looked at earlier. The reason that is is because the join point software assumes you're putting your population numbers in terms of per 100,000. Does everyone understand what I mean?

If you were putting your data in terms of per 100,000 persons, you wouldn't need to do anything to the population variable. Now, we're putting our data in here in terms of per thousand live births to convert from per 100,000 or per 1,000 to per 100,000, that's what this program's looking for, is the number of births per 100,000 persons, I'm going to multiply our results by 100. And that's why you see the number of births like this.

I multiplied this by 100. That gives me the number of live births per 100,000, which is what the join point package is looking for. So take whatever you have. Get it in units of per 100,000 persons, and that's what you'll have to do first. No matter what you have. Just make sure it's indicated per 100,000 people.

So if I wanted to calculate an IMR here, it would be 796 divided by this one times 100. Is that clear that this has to be births per 100,000 people? Okay. Now, let me go ahead.

Let's go into the software. You'll be able to check that later. I know that's a little bit confusing, but I'm going to show you guys how to check it so you can't possibly do that wrong.

Okay. Now, let's go ahead and go into the join point software. And I'm going to come under join point. These are icons on your desk top. Go ahead and double click join point. Join point regression program opens.

What I want you guys to do is click "file", "open" -- actually, let's say "new session." Let's go to "file new session." Now I have this all chronicled in your manual. And I want you guys to click "other data file." I found the most effective way for performing join point regression was to store my data file as a text file. And that's this JPT that I've given you guys.

So click other data file. Click okay. And then it's going to ask you to find it. Okay. Well, I'm going to come and find it. Okay. Where is it? Right here. You're right. Student data sets. JPT.

And there it is. Now, I've set all this up for you. So this is what you're going to see here. Make sure it's set up correctly. I made it tab delimited. Missing character, I've indicated a space. And I did say file contains headers. Just click okay to that.

Okay, and you're going to be prompted to this kind of interface. Okay. Now the first thing I'm going to do is I'm going to make sure that LNY plus beta is selected. My dependent variable, go ahead and drag that down to -- what do you guys think that's going to be? Deaths. That's going to be the Y.

Independent variable. That's going to be year. Leave the shift data points alone. You guys see the purpose for that? You want to shift over, everything over a little bit to center something. You could do that. We're not going to do that but you could.

Next under heteroschedastic errors option, select Poisson model using rate. Under population come down here and select births. And leave the number of join points alone. Click the advanced tab at the top of the screen. And you guys can leave all the defaults in here. But I want you guys to take a look at them. You guys can see that -- leave grid search as the method. I'm not too sure about what differences come, arise from between these two methods.

Number of observations. You can leave that as the default. The min number of observations from a join point to either the end of the data. So something you might want to think about. Minimum number of observations between two join points. I usually leave the defaults for these things.

Permutation test is selected by default. And we're going to leave all those alone. After that I want you guys to click the little lightning bolt at the top of the screen and it says execute current session. That's the analog to the little running man for SAS.

Input file tab, all, variable.

UNKNOWN SPEAKER: Independent was not there.

JEFFREY SHAFFER: Oh, thank you. Thank you. That's definitely it. Okay.

UNKNOWN SPEAKER: Still didn't get it.

JEFFREY SHAFFER: Oh. Year. Okay. There we go. Thanks. Okay. Let's execute the current session. It will take a second. And there it is.

Now, I want you guys, the first thing you should do, because of that population thing is confusing, how you have to get it per 100,000 persons, is go over to model estimates -- not model estimates. Go over to data. Click the data tab, and look under observed Y value.

If that isn't the IMR or close to the IMR, the observed IMR, you did it wrong with that population variable. I had that problem in the beginning. And I know that's confusing to convert to per 100,000 persons, but this is how I fixed it.

And I know that the first IMR, observed IMR, without modeling is 11.04. So I come here. It's 11.04. Join point read the thing in correctly. And the estimates you guys see to the right are going to be the predicted values. Those are going to be all the predicted join point values.

Now, let's go back to the graph. That's usually where we start. What you're prompted with is the fitted join point model that the package chose based on the permutation test technique.

If you guys want to see other models, come up to where it says model. You see 1 with the asterisk? That's saying that join point chose that model. If I click the little arrow right next to it, it's going to bring me to a model that the regression package did not choose. Although it tried to fit. You can see two join points here. That's model two.

I'll come and click that arrow again. Notice here it tried to fit one, two, three join points to the data. And that's it because we instructed the package not to consider more than three join points. We could have gone up to four.

That's one thing I could have tried was going up to four join points. Now, with this model it turns out that none of the APCs happen to be significant. But the one it liked the most after comparing all these models was model No. 1 and that's the one it chose and it did find a significant APC.

UNKNOWN SPEAKER: Did the knot stay reasonably constant over all as well?

JEFFREY SHAFFER: Well, we can see. We can check that by looking at the model S -- not the model estimates, let me check. Yeah, the model estimates. What you guys see in a lot of my output for this might not look exactly as you see it on the screen, because this package was recently updated, very recently.

So I gave you guys the most recent version on these computers, even though it doesn't look exactly the same, the results are exactly the same. They just appear a little differently.

Now, right here is going to be the number of join points detected. It looks like it detected one join point in 1999 with this confidence interval. That's what we saw earlier in the presentation.

Let me move over to the next model. Let's see what it chose. Looks like it picked up two join points. Notice the confident intervals, though, 1992 to 2000. That's

horrible. Eight years and we only have 16 years in the whole time period. So that's not very informative.

But it did pick up two but neither one is very useful. Okay. For the third one it looks like it picked up -- I'm saying it picked them up, but it didn't really pick them up. It forced them in. I guess that's a better way to say it, because it doesn't seem like they're very informative.

A better way to look at it is they were forced in. The best model of the bunch was the first one. And that's the one that has the fewest number of parameters that best explains the data.

UNKNOWN SPEAKER: That's the zero one.

JEFFREY SHAFFER: That's the zero -- it's going to fit one with no join points whatsoever. Let's take a look at that. I forgot to talk about that. There is going to be a zero join point model defined by default. Now, if you come here and you see that the zero, the one with zero join points is the best model, which technique do you think you ought to go back to?

I'd go back to ordinary linear regression. That's what it's telling you you should do. A little easier to work with. You have fewer parameters. You don't have such a complex model. Might as well do that if you can.

Okay. But as soon as you bump up to one the next model then you can't do that. This is saying that -- see, if you run two regression models here, you're not considering the entire sample. So that really wouldn't be a good alternative.

This one's considering the entire sample data in one model. The other technique, I wouldn't advise that at all. All right. So that's how you get started in join point.

Another exercise or not?

UNKNOWN SPEAKER: It's getting pretty close. Isn't it a quarter to already? I think we should break.

JEFFREY SHAFFER: I think we're going to cut it off here today, you guys, and we'll pick up on this beginning with Exercise 6.2 in the morning, and if you guys get a chance, you can read through the rest of the exercises in Chapter 6, because that's what we'll be doing as soon as we get here.

And then we'll pick up with Poisson regression Chapter 7 when we're finished with that.