Multiple OLS Regression
December 8, 2016

Erin Farley: Okay, welcome everybody. Good afternoon. We’re gonna get started. My name is Erin Farley, and I am one of JRSA's research associates. For those of you who may be less familiar with JRSA it stands for the Justice Research and Statistics Association.

Erin Farley: We are a national nonprofit organization dedicated to the use of research and analysis to inform criminal and juvenile justice decision making, and we are comprised of a network of researchers and practitioners which at the core include directors and staff from state statistical analysis centers.

Erin Farley: It is my pleasure today to welcome you to our webinar on Multiple OLS Regression and it will be presented by Dr. Ronet Bachman from the University of Delaware. Ronet is a professor in the Department of Sociology and Criminal Justice and she is a co-author of statistical methods for crime and criminal justice and a co-editor of explaining crime and criminology, essays and contemporary criminal theory.

Erin Farley: Her most recent federally funded research was a mixed methods study that investigated the longterm trajectories of offending behavior using official data of a prison cohort released in the early 1990s, and then re-interviewed and interviewed in 2009. So welcome, Ronet and before we go any further, I want to thank our partners at the Bureau of Justice Statistics for helping to make this webinar possible, and I would also like to take a moment to cover a few logistical items.

Erin Farley: So we will be recording today's session for future playback. The link to this recording will be posted on JRSA's websites usually posted the following day. If you have any questions for the presenter or would like to communicate with JRSA's staff, please submit all questions to the host using the chat feature on the right side of your screen.

Erin Farley: This session is scheduled for one and a half hours. If you have any technical difficulties or get disconnected during the session, you can reconnect using the same link that you used to join. You can also email Jason Trask at jtrask@jrsa.org. In the last five minutes of today's webinar, we will ask you to complete a short survey. The information that you provide will really help us to plan and improve future webinars and meet our reporting requirements. And we also are providing a document that Ronet would like everybody who attends today to have access to. And what we're doing is we are putting a link to it in the chat feature.
Erin Farley: So Jason Trask has just done that, and it was sent to all participants. So if you look into the chat feature, you should be able to see that cut and pasted and gain access to that document. This document will also be available when we do post this webinar online. And so with that, I want to welcome everybody and turn it over to Ronet.

Dr. Bachman: Excellent. Thank you, Erin.

Erin Farley: And there's a few slides you probably just want to breeze by those. Yeah, those are mine.

Dr. Bachman: I will and I want to echo, the last time I had a problem with my voice. So if some of you out there can't hear me or they're having problems, we tried to test it and make sure it was okay this time, but please email Jason and I'll try to do something with my mic. I also want to welcome you and I'm again, thrilled to be here. Some of you may have attended the last one I did on bivariate. And I just want to say how honored I am to be doing this, I worked for the Bureau of Justice Statistics, so all the state sacks out there, I've worked with you guys and it's just an honor to be here today.

Dr. Bachman: And also filling out those surveys after the end are really important because Erin told me not so many PowerPoint slides run at this time, focus on the SPSS. So I listened and this webinar will just go through a few PowerPoints. I think it's important to set up the difference between multiple OLS and the Bivariate case. So I'm just going to go through a few and then we're going to jump right into SPSS, going through multiple regression and for each sort of example, I'm going to record a couple of variables because I did get a request last time to go through a recode. So I've done that a couple times here too.

Dr. Bachman: So if you have a question, just provide that chat and I'll try to answer it as we go along. I know it's more helpful to do that. The lab link that I have is called the lab, just because that's the way I visualize this as a lab, I'm doing a SPSS lab taken you through an exercise to teach you multivariate regression, and it will facilitate going through the example I'm doing.

Dr. Bachman: Some people are visual, some people are auditory, so I provide both. So anyway, let's get started. So last time I went through bivariate, and any bivariate analysis that most people do, that's what typically the extent of our analyses when I was working for the Justice Department at BJS, that's typically all we produce in our products. But multiple regression at the state level, a lot of you will be interested in doing this because it all goes
back to determining causality. And there are generally three criteria that are necessary for causality.

Dr. Bachman: And the first is the simplest, right, establishing a relationship. It's easy to establish relationships between two variables. You can do that till the cows come home. The second is the correct time order and usually in research we assume a correct time order between the independent and dependent variables. That is that the independent actually preceded the dependent variable.

Dr. Bachman: The third is much more difficult to establish and that is that the relationship we see between our bivariate cases or anything we are assuming causes something is non-spurious. A spurious relationship is one that's really a false relationship caused by a third or fourth variable, and the example I typically give is the relationship between ice cream sales and violent crime. If I showed you a correlation between ice cream sales and violent crime, there would be a very strong positive relationship, as ice cream sales increase, so does violent crime.

Dr. Bachman: But we know that as people who work in this field, that is a spurious relationship. It's not that ice cream causes crime, it's that ice cream sales increase at a time when more daylight hours occur, when school's not in session, when there's fewer guardians of young people. So all these other factors are really the explanation of violent crime, not the ice cream sales. So that original ice cream sales, violent crime would be experience relationship.

Dr. Bachman: What we're constantly trying to do in research is determine what's a real relationship and what's a spurious relationship. Through methodology, most of you I'm sure have engaged in some sort of evaluation research where you're determining the effects of some program or policy. Typically the gold standard is a true experimental design where you can control and randomly assigned whatever you're examining to a control group and an experimental group. That gives you the ability to say pretty much unequivocally that is if there's an effect after the policy or program was implemented, the only thing different between those two groups was at one randomly received something and one didn't.

Dr. Bachman: So any change you see at the post test, you can be assured is a real non-spurious effect. In the real world, we often don't have the luxury or ethical considerations to be able to do a true experimental design. So what we use instead is statistical control and that is where multivariate,
hence Multiple OLS Regression that I’m going to concentrate on today comes in. It allows us to control for all the things that the literature and theory says is important in determining things that we’re interested in, like sentences received from an adjudication process, crime rates, et cetera, not of other important factors. So that's why we do multiple regression.

Dr. Bachman: The first criteria is easy, look at this. I can show you so many funny things that satisfy the first criteria of showing a relationship. This is per capita cheese consumption and number of people who died becoming tangled in their bedsheets. That's a clear positive correlation, but that's silly. It doesn't make any sense. We know that, you know, cheese does not lead to strangulation in your bedsheets, but these are the kinds of things that for years prior to computers allowing us to do these strong multivariate methods, we could find relationships for.

Dr. Bachman: We live in a multivariate world and whenever we look at something we're interested in the social world, particularly the things we're interested in crime and victimization and offending, there are many, many, many factors that need to be taken into account. It's hard to last time I went to scatter plots and I'm going to go through a couple of those today, but it's hard to visualize what's happening in regression when you're controlling for third, fourth, fifth variables because you can't really visualize, you're sort of looking at hyperspace, but it's easy to see through simple cross tabs what I mean by statistical control.

Dr. Bachman: So I just want to go over a couple of these bivariate crosstabs to show you what statistical control actually means. So here we have a typical crosstab between delinquency, which is the dependent variable and gender which is the independent variable. It's important to examine the percentage differences across the independent variable categories within the dependent variable category.

Dr. Bachman: So let's just look at those in the high delinquency category. We see that 33 percent of females engaged in high delinquency compared to 42 percent of males. This is a significant relationship. This is real data by the way. And so we would reject the null hypothesis here and conclude that females were less likely to engage in high delinquency compared to males, but we know that there are multiple factors that cause or create the conditions right for engaging in delinquency.
Dr. Bachman: We've typically thought gender is one of them, but more sophisticated research now is showing that once we control for other important things, that relationship typically goes away. One of those relationships is parental guardianship. This is a simple bivariate crosstab. I’m going to show you a partial crosstab table, that when I say partial it means it's partial out by a third variable, in this case, parental supervision.

Dr. Bachman: We asked kids on this survey, a ordinal question about how frequently their parents knew where they were in different scenarios and I’ve divided it up into those who said parents rarely or never knew where they were and I call that weak parental supervision compared to strong parental supervision, where kids said their parents generally always knew where they were.

Dr. Bachman: So what we see here, if we go back to look at the relationship between gender and delinquency, look at this week parental supervision crosstab. We see 46 percent of females engaged in high delinquency compared to 49 percent of males. Well, that percentage differential across gender has been reduced significantly, there's still three percent more males and females engaging in high. But the difference we really see here is between those kids who are strongly supervised by their parents and those who are weekly supervised.

Dr. Bachman: Look at this, 20 percent of females to 24 percent of males in the strong parental supervision. So this is illustrating very graphically what I mean by statistical control. You can easily see then that that original bivariate relationship here that we saw between males and females is really probably spurious.

Dr. Bachman: Once parental supervision is controlled, it's not gender that is creating this, it is parental supervision. And the bivariate case was probably the fact that we know females, particularly young adolescent females are much more likely to be supervised than their male counterparts. So that's what I mean by statistical control.

Dr. Bachman: I talked last time about bivariate OLS regression and you see here, and thank you Jason for redoing this, my first slide was messed up. The multiple OLS regression equation is a simple extension. The intercept, the first line here is the population parameters, the second line are the symbols for the sample statistics. It’s the same why the dependent variable is equal to A, a constant, but now we've just extended the
equation to include as many independent variables as our sample size will allow.

Dr. Bachman: How many is that? The general rule is you need at least 10 to 15 sample size, and for every independent variable. Now, I want to just say something quickly about model specifications.

Erin Farley: Ronet, can you hear me?

Dr. Bachman: Yeah.

Erin Farley: We have a comment somebody says the audio has been cut off. I can hear you, can just one or two other people let us know if you can hear Ronet or if you just heard her it might be [crosstalk 00:14:57] says yes. Okay, we got tons of yeses, you guys are awesome.

Erin Farley: Okay. So we recommend the person who says they have been cut off to just disconnect and reconnect and to let us know, just send us, from the audio conference and just let us know if you're still having problems. Okay, Ronet, back to you. Sorry.

Dr. Bachman: That's okay. Thanks. So generally the other thing I want to say about the number of independent variables is you want to put as many variables that are important from your literature review and from theory that are important in describing or explaining your dependent variable, and you don't want variables that are not important.

Dr. Bachman: So you need to really base the number of variables on what theory and the extent literature say are required to predict, because if you don't have an important independent variable, say for example, you're looking at the relationship between race and crime. If you don't have important variables like poverty, access to education, those sorts of measures, the relationship you're going to find is between race and crime is spurious because we know that that relationship is largely explained by economic deprivation indicators.

Dr. Bachman: So that's my little mini lecture on models. I don't have time to go into a lot of model fit stuff, I'm just going to take you through the basics. So I'm going to move on, but you can see here that this is a simple extension of the bivariate case, you can control for many, many, many important independent variables. Now, I also want to say we also went over these assumptions for OLS regression last time.
Dr. Bachman: There's one new one here, it's number six, and this is an important one. It's the assumption of no multicollinearity. A multicollinearity exists if you have two independent variables that are extremely related and generally that means a correlation of point seven or 75 or higher. If you have two variables that are so related, they can either be generally thought to be explaining the same thing in which you could create an index.

Dr. Bachman: Say for example, you were measuring economic deprivation and you have percent of families living below the poverty level, and you have infant mortality. Those two things are very highly related, so you would want to add them into an index or just use one of them to measure your construct of economic deprivation because it's really going to mess up the multiple regression analysis.

Dr. Bachman: And let me tell you why. I'm just going to show you these equations just to eliminate that the slope B, is being calculated not just by the relationship between any independent variable and the dependent variable. It's also being calculated by, and this is where controlling for a third variable comes in.

Dr. Bachman: It's also being calculated and taking into account any interrelationship that exists between all of your independent variables. So if you have two IV's that are extremely related, crib notes version is going to screw up your multiple regression equation.

Dr. Bachman: So let me take you through the output quickly and then I'm going to go to SPSS. This is what we saw last time, not this particular slide, but this is basic bivariate regression analysis, output from SPSS, and the dependent variable I'm predicting here, and this is back to that delinquency data that actually is real from a high school and middle school in South Carolina.

Dr. Bachman: The dependent variable assist delinquency scale and just like OLS, the dependent variable in OLS multiple regression is assumed to be linear and measured at the interval ratio level. So this is a scale that's comprised of several behaviors, whether students engaged in it and how frequently they did. The independent variable I'm using to predict delinquencies here is certainty of punishment. And that is a scale on how likely the kids thought they would be caught, and if caught punished for said delinquent behavior.

Dr. Bachman: Deterrence theory would predict that those kids who thought they were going to get caught for doing a deed would be less likely to do it.
Remember that in regression output, this first highlighted statistic R is the correlation coefficient. And I'm always going to use the adjusted R squared, which is the coefficient of determination, which tells me that when multiplied by a hundred, 6.2 percent of the variation in delinquency is being explained by certainty of punishment. This F test is significant telling me that this is a significant model predicting delinquency compared to just the general constant.

Dr. Bachman: This unstandardized coefficient B here, this is the bivariate regression equation. Y delinquency is equal to this constant, which you will recall is simply the intercept A which is that intercept where the regression line, the slope in this case crosses the Y axis when X equals to zero, we can interpret this. What we're really interested in is the unstandardized slope for certainty of punishment 2.85.

Dr. Bachman: And this tells us for every one unit increase in certainty of punishment, delinquency decreases by 2.85 units. That T value that measures the significance of the slope we call this null hypothesis says there's no relationship. I'm always going to use an alpha or significance of all five, which you guys are sol standard. So we can reject and say kids who thought they're more likely cut, or less likely to engage in delinquency.

Dr. Bachman: The next slide is stacks of respondent and this is a dichotomous variable coded one for males and zero for females. So we know that in this case, compared to females coded one when respondents were male, delinquency increases by 6.2 units and it's also significant at the 038 level. Now, what I'm showing you here is that these two variables, both at the bivariate level showed a significant relationships. So if we didn't go any farther, we would conclude that gender is significantly related to delinquency in the population.

Dr. Bachman: Males, you know those males out there engaging in more delinquency. But look what happens when we put it in the multivariate model predicting delinquency. Now, one of the things I want to stay, and I said this in the bivariate case. When you have multiple regression models and this R becomes the multiple R, that's why the R in this case never reveals a positive or negative sign because when you have a model, you have several independent variables. Some may be positively related to the dependent variable, some negatively. So R in this case simply describes the strength of the relationship between all of the independent variables and the dependent variable.
Dr. Bachman: So in this case, remember R goes from zero to positive one, negative one, positive indicating a positive relationship, negative indicating the negative relationship, but there is no sign here because sex of respondent is positively related and certainty of punishment is negative related. The multiple R squared tells us that 6.7 percent of the variation and delinquency can be explained by all, or in this case, both of the independent variables. This is R and the F test for R is a model fitness test, it simply telling us whether or not the multiple regression model explains a significant amount of variation in the dependent variable.

Dr. Bachman: So it does in this case, but notice we can't say anything about any particular relationship with the IV, any specific IV and delinquency, we have to go down to the final box that displays the regression coefficients. And just like for the bivariate case the constant is what SPSS labels A the intercept, and this is the intercept now for both slopes when both IVs or equal to zero. This is the intercept. Again, don't interpret this, but now we can say, let's start with sex of respondent.

Dr. Bachman: It's positive and notice that it's gone down from the bivariate case compared to females who were coded zero males engage in, have an increased delinquency score of 2.87 units, even after controlling for certainty of punishment. Now, that's what we can add to our interpretation net of the other independent variable. Another way of saying that is even after controlling for that other independent variable.

Dr. Bachman: And notice the T value for this particular coefficient. Now, the significance of this T value is now .33, meaning that if we reject the null hypothesis that states no relationship between gender and delinquency, we're going to be wrong 33.4 percent of the time. We're only willing to be wrong typically five percent of the time, so we have to fail to reject and conclude that after controlling for certainty of punishment, gender is no longer significant when predicting delinquency.

Dr. Bachman: Back to my big S word you can now use in scrabble, that was a spurious relationship like ice cream and violent crime. The second variable of certainty of punishment, however, retains its significance. Notice that it's significant at the OOO level, very, very significant still so we can conclude that the population that even after controlling for gender of respondents, kids in this sample and in the population, that's what we're doing this hypothesis for. Who believes they're going to get caught and punished are less likely to engage in delinquency.
Dr. Bachman: So basically, goodness gracious, I had an extra slide in here. This last slide, puts in the formal null hypothesis so you can have that as we move along. But I think that's the last slide and I'm going to move to SPSS now to go through a couple scenarios. And we might have to get on and share my laptop now. So I'm going to go to this. Hang on, I'm going to go to QuickStart. Forgive me, share my desktop, monitor one. Whoops.

Dr. Bachman: Okay, can you see my SPSS? No one can talk to me, I guess. Okay. I'm going to assume you can see this and if you had downloaded that lab multiple regression, that's exactly what I'm going to take you through. So I just want you to have it as a visual in case I go too fast. This first research problem that we're going to investigate is predicting state rates of robbery, and you can use aggregate data at the county, city, neighborhood, whatever level.

Dr. Bachman: This is just a great place to start because there are a few cases in the scatter plots actually allow you to detect bivariate outliers. So the research question we're going to examine is what factors are related to robbery variation across states? We know that there are several theories. Social disorganization theory, postulates several structural factors including economic deprivation, residential mobility is another one. If you have people moving in and out of a community that creates disorganization and people aren't effectively able to control their citizens.

Dr. Bachman: Another factor of particularly related to robbery is urban locations. We know they're more likely to current urban versus suburban and rural locations. So another important variable to control is percent of a population that resides in rural areas. And another general geographical control is just region of the country. We know that the south tends to have higher rates of violent crime in general, so controlling for regional location in that way is important.

Dr. Bachman: The first thing I'm going to do is go through the descriptive statistics, always, always, always. The most important thing you do before you jump into any bivariate or multivariate analysis is make sure that your distributions at the univariate level are not extremely skewed or contain outliers. Now, I know you've had webinars on this and you know this well, so I'm just gonna briefly go over this.

Dr. Bachman: I see that everything, and the other thing is that if you have dichotomies in a categorical level, you cannot include a categorical nominal level variable, that is a variable that is just categorical like region, religion,
gender, race, unless they're dichotomies. And those dichotomous variables have to be in order to understand the regression coefficient, have to be coded zero and one.

Dr. Bachman: So I'm looking at the descriptive statistics here. Whoops! I'm looking at my own web. I'm going to actually take you through how to get them. So you go into everything that you do in an analysis sentence under analyze, descriptive statistics, provide all various options. I'm just going to ask for descriptives and I've already got this in here. It gives me all of my independent variables along with the dependent variable of the robbery rate. And I click okay and it kicks me to the output window.

Dr. Bachman: Now, one of the things you may not ... one of the things I do if I'm doing an analysis is, first thing I do is take these notes. There's always some note, you can never see it in the output, but it's always there if you cut and paste it to another document. But I click on that and take it out and then since you can only open an output window in SPSS, I click on whatever output I want to use for future reference and simply right click and copy it over to a Word file so that I have access to it at all times instead of always needing to have SPSS available to look at my output. So that's just a word to the wise.

Dr. Bachman: So back to the output, I see that all of my variables are interval ratio. I see a mean robbery rate of 116, et cetera, but I see these four regional variable that goes from one to four. If I go back into variable view in SPSS and I look on region, I see that it is a four category variable here. You can't see it all the way, but I'm going to do a frequency so you can see this, how it's coded.

Dr. Bachman: I'm going to ask for a frequencies of region here. I just put regional which of the variables box, so I can see how this is coded and it's coded. The other thing I want that I always tell people to do is SPSS by default will generally give you the labels of variables and those are southwest, Midwest, northeast, but I always ask for under options to give me the numerical value because it just saves so much time.

Dr. Bachman: So I know that in these four regional variable one is coded south, and that's the variable I want to use for my dichotomy and code, recode this variable into a new variable called one for south and zero for all the other regions. Now, how do I do that? Every time you go into recode or compute a variable from scratch, you go under this transform. I don't know what this is called tab, I'm going to call it. And I use compute a lot, I
use recode a lot. Word to the wise and I think I said this last time. Never, never, never, never, never recode right here into the same variable, because you will lose instantly your original values.

Dr. Bachman: You never want to do that even if you're just changing it slightly. I always recode into a different variable and give it the same variable name with R behind it, so I know that that's been recoded. So I'm going to click on recode into a different variable and this is what I get. Now, I'm going to start from scratch here. The input variable is a variable you want to recode which in this case is region and then you have to give the variable a name. And this is just another case after years of doing analysis.

Dr. Bachman: When I'm doing a dichotomy and they all have to be coded zero and one, I called variable, I give it to name that one is coded so that I don't have to keep referring to code book like you do for gender. How is that coded again? Now, I know that this variable is a dichotomous variable coded one for south. I'm just going to remove this. This is the basic recode dictionary that you're telling SPSS how to use to recode this variable.

Dr. Bachman: This is another thing I always do, I know from my frequencies here that there are no missing values, I don't care, I do this, it's just an anal thing. All system are using missing, I add this the first thing just to make sure that anything that's missing is coded missing and not giving a legitimate value.

Dr. Bachman: Now, you see from the output that the value of one is equal to the south and we're going to keep that as one. So I just do value one equals one and I add that to my data dictionary. Think of this old to new as your dictionary telling SPSS what to do. Now, there are several I could keep going through this, two equals zero now, three equals zero, four equals zero, but that's three steps.

Dr. Bachman: The simplest thing is to say, "Okay, all other values and that's two, three, four in this case are going to be equal to zero." Now, add that and then I'm done. Click continue and SPSS has a little mechanism that wants to make sure that you want to change your variables. So notice that the okay icon I cannot click. It always makes you say, "Yep, I'm copacetic with that, change it," and then press okay. And what happens instantly notice is that this new variable called south is now under my variable view.

Dr. Bachman: So if you got out of this dataset before you saved it, you would lose that. Now, one final thing that I always do is I run a frequencies on the newly
recruited variable and don't just assume that my recodes are correct because at my age, about 30 percent of the time they're not. So this is an easy check. I look at south here my recoded variable, I have 17 states that follow their 33 percent of the cases, which is just like my original regional variable. So I know that [inaudible 00:35:25] my recodes of work and I'm ready to move along.

Dr. Bachman: Okie dokie. So, now with that, I'm going to take you through quickly looking at scatter plots. Now, remember that scatter plots are an important way to examine the bivariate distribution of an independent and dependent variable. Why do we look at this? It shows us the shape of the distribution and the strength, but perhaps more importantly, it shows us any outliers or any idiosyncratic variation that we need to take care of.

Dr. Bachman: I'm not going to go through all of them, but I'm going to go through one. There's this chart builder in SPSS that never works for me, I always go under legacy dialogue and just build my own scatter plots this way. All my grass for that matter. So I click on scatter dot, which is a scatter plot, it gives you several options. This matrix scatter is a nice thing, but it's so puny you can never really see what's happening.

Dr. Bachman: So I just do a simple scatter that's typically what we want, SPSS does a lot of stuff for businesses now that are just not useful. They look kind of cool, but they're just not useful. The old simple scatters, what I always use. And on your Y axis is always where you want to place the dependent variable. Your X axis is where you want to play the independent variable. I'm going to put percent rural here as the independent variable and my robbery rate, which is what we're attempting to explain here.

Dr. Bachman: I just want to show you here. Now this is the scatter plot that's created and we see here all of this flat scatter but look at this little dot here, this is a problem. This is a bivariate outlier and I highlighted that in the bivariate webinar, but now I want to show you what to do about it. I know, and I won't take you through how I know, but I'll tell you, this is the district of Columbia that's always included in state level data sets. It has zero percent of us population residing in rural areas, and it has a very high robbery rate.

Dr. Bachman: This is what is called a bivariate outlier and notice that remember that regression is trying to draw linear lines through these data and notice how flat all the other data are here, but this one thing up here because messing it up. So well you need to get rid of bivariate outliers like this.
you have many of them, it's probably easier to call it those cases missing, or do something else with them random in a little bit. But in this case I'm going to exclude the district of Columbia from all future analysis because I know that it's not just a bivariate outlier here, it's on a couple other cases.

Dr. Bachman: So if I go to data view and I have states as one of my variables, I see that the district of Columbia, which is an alpha numeric variable also is associated with an ID of nine, and that's unique identifying number. So what I'm going to do is under this data menu, you can also do a number of things with your data, sort your variables, merge files, et cetera. But this is an extremely handy function called select cases that you can select out cases or do subsets of the file.

Dr. Bachman: Oftentimes, I want to look at males and females separately, so I'll select out mails or select out females. In this case, I want to select out the district of Columbia. So notice when I click on that select cases, I get several options, I can do a number of things. But in this case, I want to do if condition is satisfied and then SPSS allows you to select what that condition is. If I wanted to look at just states in the south, I could say if south equals one, in this case I'm going to click on it. And what I put in is the variable ID not equal, that's the not equal sign in SPSS nine.

Dr. Bachman: So that tells SPSS, "Okay. From here on out, she does not want any variable included if the ID is equal to nine." So I'm going to say continue and okay. And now notice on my data that the district of Columbia is hashtag out, so I know when that hashtag out all the analysis that follow are not going to include it.

Dr. Bachman: So to make sure I want to go back and look at this scatter plot, which I'm going to do here. Simple scatter and I'm going to look at this relationship again without that pesky outlier and see what happens. And sure enough now I have a nice visual where I can see and discern the direction of the relationship. As theory would expect, states with higher percents of rural population you have one state here with 60 percent of its population in rural areas. Probably one of the square states in the middle of the US where I'm from.

Dr. Bachman: So states with higher percentages of rural population tend to have lower robbery rates. So that's just a little guide on how you should inspect your scatter plots before you move on to multiple regression. Now, the final thing you need to do before we move on to multiple regression is
examine the bivariate correlations. Why is this important? Because remember multicollinearity is a serious issue, especially when you have aggregate level data.

Dr. Bachman: So one quick way to examine whether or not multicollinearity causes a problem is just by looking at a simple bivariate correlation matrix. And I showed you how to do these last time under analyzed. You simply go under correlate and you have several options. We want to do a bivariate correlation, which is going to give us a matrix and I have put all of these. I'm going to add south because that's a new variable, that's going to put all of our variables, the IVs as well as the DV in a correlation matrix, and I'm going to leave it as two tails all the standard defaults for SPSS and click okay.

Dr. Bachman: And then I see in my output window this beautiful bivariate correlation matrix, and remember that the matrix is redundant. Everything on the top is identical just in a mirrored fashion to everything on the bottom. I always put my deepen it variable first, so I simply have to look at the first column to see the relationships between all my IVs and the DV. But again, the reason why we're doing this is to examine any correlations that may exist between any combination of our independent variables.

Dr. Bachman: A correlation, you say run at what is a correlation high enough to cause multicollinearity problems? The general rule is anything above .7, .75 could be problematic. In fact, probably we will be problematic. So you want to inspect the relationships between all of your independent variables to determine that you don't have correlations above .7. I see, I don't hear so I can assume multicollinearity is not going to be a problem.

Dr. Bachman: Now, we moved to the multiple regression equation. So I've assumed multicollinearity is not a problem, I've checked to make sure they're not bivariate outliers. Now we're gonna move on to our multiple regression model, predicting robbery rates in states using our independent variables. So again, we go to analyze and it's the same exact tab that we use for bivariate regression. We just put more independent variables into the box.

Dr. Bachman: We're looking at linear regression, that's what we're dealing with today or last regression, click on linear and you have a box to put your dependent variable, and then you have this other box to put all of your independent variables. Now, I've done screenshots on this lab, cut and
paste, and so you can see ... You can only see about three variables here, but if I scroll down, you see that I've got my poverty variable below.

Dr. Bachman: I'm worried about these percent moved as the census, US Census Bureau's variable that asks respondents have they moved within the last five years, percent rural is also from the census, south is just newly created variable that I dichotomized coded one for south, zero for non-south. And percent of families living below the social security administrations, poverty line is the final independent variable.

Dr. Bachman: So I'm not going to ask for anything special, simply going to click okay and I will have then in my output the regression output just like we got for the bivariate, but now we have several variables. If we scroll down in the final coefficient box, you see all of the independent variables here predicting the multiple regression equation.

Dr. Bachman: Now, I'm going to actually bring over my lab over here so you can see it if you don't happen to have it here, because I want to show you the multiple regression equation. This is the multiple regression equation that is predicting robbery rate and notice that it's as simple as extension of the bivariate case. Each of these unstandardized coefficients under the slope B column are the slope coefficient for each independent variable net of or after controlling for all of the other independent variables.

Dr. Bachman: And the constant 159 is our multivariate intercept. So let's look at the model summary first. This is R, the multiple R now, the size .7 is closer to one than it is to zero. I would say that's a moderately strong correlation between all of the independent variables here and the robbery rate. I always use the adjusted R squared when interpreting other coefficient of determination R square, just because it's a more conservative model that takes into account the number of IVs and their relative contribution to explaining the variation in robbery in this case.

Dr. Bachman: And I see that 49 percent of the variation in robbery can be explained by all of these variables. This F test here tells me that in fact, this is a significant amount of variation explained compared to just the constant alone, so I'm sitting good here. But notice that now I don't know about any specific relationship. I simply know that this model is good. I am predicting robbery relatively well here, at least significantly well. So let's go down to the regression equation, which is this now.
Dr. Bachman: And I'm going to interpret a couple of these separately so you get how to interpret this. This slope coefficient here B, four percent moved in the last five years of negative 3.3. I would say, this tells me specifically that for every one unit increase in the percent of the population in states that has moved in states, the robbery rate in states also goes down an average of 3.32 units. Even after controlling for morality region and poverty.

Dr. Bachman: So that's now what we can say. This is just not the effect alone. This is the effect even after controlling for these other things. Now, looking at the significance of this coefficient. I see that the significance is .17, meaning that we'd be wrong 17 percent of the time if we rejected. And a backup at the library correlation, remember it was a significant relationship. Now, it's no longer significant, so we can't conclude that mobility increases rates of robbery in states, net of poverty and region and percent rural.

Dr. Bachman: Now, let's look at percent rural and we see that that has a coefficient, again, negative. So there's a negative relationship between morality and robbery. This tells us that for every one unit increase in percent of the population that's rural, there's a corresponding 2.7 unit decrease in robbery rates even after controlling for the others and these two is significant. I mean, this is significant, unlike mobility.

Dr. Bachman: In fact, it's significant at the OOO level. I'm worried about SPSS, I don't know if I mentioned this last time, I'm going to click over here to the output. It only shows three decimal places, so I have students who say we're 100 percent certain that we're making the right decision. That's not correct. In probability theory and hypothesis testing, you can never be 100 percent sure, but SPSS only displays three decimal places.

Dr. Bachman: So if I go into the output and click on this, you see many, many, many decimals out there that there's a whole number out there that you don't see. Only the three decimal places are displayed, so what you can say however is you'd be less than .1 percent wrong. So it's a very significant relationship.

Dr. Bachman: So states residing states with higher percentages of rural population have significantly lower rates of robbery than states with high urban populations. Now, I just want to stick to the south because that's the dichotomy and it's a little trickier to interpret because it's not an interval ratio, you can't say for every one unit increase, you can simply say, "What
happens when this particular dichotomy goes from zero to one?" That's what this coefficient is telling us.

Dr. Bachman: So we remember that we recoded this zero equals states in the non-south and one equal states in the south. This is how I interpreted them. I start with compared to whatever zero was coded. In this case, compared to states in the south, compared to states in the non-south, forgive me, which are coated zero.

Dr. Bachman: When states reside in the south, there's a 39.28 unit increase in robbery rates. Net of the other three independent variables in the model. And that is also significant at the 0.01 level. So we can conclude that states in the south, have higher rates of robbery than states in the non-south. And looking finally at our percent poor families, poor variable that also remain significant at the 0.05 level. So after we've controlled for all of these, we see that mobility percent that has moved in the last five years, no, it was no longer significant, but the other three have maintained their relationships.

Dr. Bachman: Now, one of the things that you typically want to know is, "Okay, so all three are significant, which is the most important or the variable that most strongly is related to robbery in this case?" Some researchers or statisticians will have you look at this column, the standardized coefficient sometimes called betas. I never used those, because although most of the time they're accurate, sometimes they can lead you astray because they're calculated using the standard deviation.

Dr. Bachman: So you might have a variable that has extremely high standard deviation and it may look like it's stronger than another simply because of that. I simply use the size of the T values which correspond directly to the size of significance. The lower significance means the stronger the relationship, the highest the T value means a stronger relationship. The sign of the T values again reflect the sign of the slope coefficient, it has nothing to do with strength, it's just telling you that that coefficient was negative. So the T value is negative.

Dr. Bachman: So you look at absolute values of T to determine relative strength. Now, it would tell you generally the same thing as looking at significance, but sometimes as you see here, you have significance that is .000 and it only shows those three decimal places. Sometimes you get two variables that are both .000 in their significance to the dependent variable and you wouldn't know based on that without clicking on it in the SPSS output.
Dr. Bachman: So I just used the T value. So looking at these T values of those that are significantly related to robbery, I see that the T for the percent rural is the highest, which tells me that all the variables in the model predicting robbery percent rural is the most important, the most strongly related to robbery rates. I also want to show you about a cool way I typically present important variables like this.

Dr. Bachman: The great thing that you can do with the multiple regression equation is to use it for prediction, and a wonderful thing that you can do to illuminate the effect of that particularly strongly related IV and DV is to predict values of the dependent variable by changing a high and low value of that important variable while holding all other independent variables constant that they're mean or mode if it's a categorical variable. So remember, this is the regression equation here, and I've got just the basic access plugged in here. But I'm going to go down here and show you how I do that and it's so illuminating and great to present to an audience.

Dr. Bachman: Rurality was the most important variable here. So what I'm going to do is predict the value of robbery for a state in a non-south because that was the modal category that has an average rate of families living below the poverty, which is 9.9. An average rate of mobility, which was 15.2 and a very low percentage of rural population, which I'm going to set at five here. Okay? So all I have to do is plug in these values to their corresponding slope coefficients.

Dr. Bachman: So I have here 15.2, the first coefficient was for mobility. So I plugged that 15.2, that's the mean mobility rate beside that. The second coefficient negative 2.7 was for my rural population, so I plugged in five there, that's what I want. 39.28 was for the non-south and I'm plugging in zero there because I want states, I want this predicted value of robbery to be for non-southern states.

Dr. Bachman: And then I plugged in 9.9 here for my mean level of families in poverty, and if I just do the math, what I get here is a predicted robbery rate for this non-southern state with an average rate of poverty and mobility with a low percent rural. The predicted robbery rate is 149.48. I mean this is a low percent of rural population, right? Only five percent living in rural area. So it's a very, very urban state. Now, so what you always want to show an audience is compared to what, that's interesting but it means nothing and lets us compared to what.
Dr. Bachman: So now let's predict a high, the same non-southern average poverty, average mobility state. But now let's look at a state with 50 percent of its population rural. And we know that that's not the highest, there's a state with a high of 60 percent. So it's just a much higher rate than this five. And you plugged this in and then change this coefficient instead of five up here, we plug in 50 and we get a predicted robbery rate much, much lower, in fact, five times lower.

Dr. Bachman: So this is very illuminating when you're presenting regression output to an audience because especially when you want to highlight the effects of a very influential variable. It does so well and it allows you to say, "Look, this is what happens net of region and poverty and mobility." Keep in mind, even if you're like, this mobility variable is not significant, even if it's not significant you still have to set it at a constant.

Dr. Bachman: So you have to have this predicted rate equation hold constant every variable, regardless of whether it was significant or not. Okay? So that's cool. Now, I want to briefly move on to another research scenario. This time, I'm going to look at predicting sentence length for a sample of homicide defendants. This is again real data, it's a little dated, but it's real data. From 75 of the largest metropolitan areas and I'm going to toggle this back over here and go back to the data.

Dr. Bachman: Excuse me a second, I get this data up. This data again is homicide defendants. So the units of analysis are the actual homicide defendants and there are a number of variables that I have in this data set to dependent variable that I want to predict here is not all of these homicide defendants were convicted, but I want to predict prison time, the sentence, the incarceration term they received if they were convicted.

Dr. Bachman: So the dependent variable is going to include only those homicide defendants who were convicted, and I'm going to predict the severity of the incarceration sentence they got. I'm not going to include all of the variables that we think are important, I'm just going to give you an example of some of them. One of the things we know is race is always important to be controlled. Age of defendant, number of victims is important.

Dr. Bachman: And I'm not going to take you through the description here, I'm going to show them here on this lab that I've given you. Male gender of defendant is also important, prior convictions are important when you're looking at
severity of a sentence, particularly with mandatory minimums. So I have all of these variables. And I have everything coded agent defendant is interval ratio, it ranges from about a minimum of 14 years old to a maximum of 87, the mean was 29. All the others are dichotomies.

Dr. Bachman: The jury trial variable is coded one if it was a jury trial and zero if it was ... jury or a judge trial and zero if it was a plaid case. Male defendant, according to my custom, I called male because I coded one, I coded males one and zero for females. Prior conviction, some had several prior convictions, I simply coded this one if they had one or more priors and zero. Otherwise, this means tells me about 33 percent of these defendants had a prior. Number of victims here is coded one through five.

Dr. Bachman: Now, that tells me something could be problematic here. So I'm going to go and look at this actual distribution, so I'm going to go onto descriptives and give myself a frequency distribution of number of victims. And look what it actually tells me, so if I do that, my output tells me that almost 94 percent of the homicide defendants in their 1311 valid cases here killed single victims. The next 4.8 percent killed two victims and you know, and then it just kind of drags off. They have two homicide defendants who actually killed five people.

Dr. Bachman: But this is an extremely skewed distribution, it would not ... you cannot assume this is interval ratio even though it is because it is so extremely skewed, would give you a nonsense coefficient or not a valid one. When you have cases like this, just like the number of priors was in this case, there were most people had no priors, some had one, fewer had two and three and so on.

Dr. Bachman: So I recoded that already to zero and one. I'm going to go back and recode this variable. So that zero is going to be equal two, they just killed one victim and one is gonna be equal two, if they killed two or more victims. Again, I go under transform and again never recode into the same variable, I'm recoding into a different variable. Number of victims is my input variable and as my convention I'm going to call it more than one victim, more than one vic for short. And I'm going to give it a label more than one victim and then go into my old and new values.

Dr. Bachman: Again, system are using missing. First thing I do is put that in my data dictionary and in the original variable, one was the variable that I'm now going to code zero because that means they killed only one victim. And
all other values I could also do a range here, I could do range to the highest, I could say two to highest is going to be equal to one, but I'm just going to say all other values are equal to one.

Dr. Bachman: That is all I need for this, I'm going to click continue. SPSS says, "Are you sure?" I said, "Yep, change it, click okay," and notice down here I have this new variable now called more than one victim. But don't trust myself, I'm going to go back and get a frequency distribution for that new variable, more than one, and make sure it's coded correctly and sure enough, I have 94.1 percent coded zero now, which was my original one category.

Dr. Bachman: And see that a total of 5.9 percent of these homicide defendants killed two or more victims. That's going to be my new control variable to control for the severity of the defense, because we know that the heinousness and severity of offense also affects sentence. So I'm ready now because I'm sort of running out of time, I'm going to skip the bivariate cases here, it is in your handout not the scatter plots but the bivariate correlation. I'm going to go right to the multiple regression output, I go under analyze, regression, linear, and my dependent variable is going to be prison time which his sentence length and days, the independent variables is another handy dandy thing if you have variable names SPSS by default we'll give you the labels which is a pain in the butt because they are long and not telling you exactly what the variable is.

Dr. Bachman: For example, if you can right click and ask for variable names, the first variable I want to put in here is jury trial. If I'm on and I have variable names, I can just get in this thing and it's highlighted, I can type J-U-R-Y and we'll go to the first variable that starts with J-U-R-Y and that's easier than going through many, many, trying to find the variable you want.

Dr. Bachman: The next variable I want is defendant age, click that over. Then I want known victim. This is whether or not the control for whether or not the defendant knew their victim. I have male defendant here, is the next important. I have prior convictions, my zero in one variable, I have race which I've recoded into white versus non-white, white is coded one.

Dr. Bachman: I want to make sure I have everything here that I need. So I have jury trial defendant's age, whether the victim was known or not. All known victims are coded one, strangers are coded zero, male defendant prior convictions and a white defendant. So I think I'm set. So I simply click okay, and then I get my regression output and I talked about this last
time, you'll notice that the model summaries here are the R and R
squared are much smaller than they were for the aggregate case.

Dr. Bachman: I talk about this in the lab output. Note that you're never going to explain
typically, I've never seen more than 10 or 12 percent of the variation
explained in individual units of analysis, that's just typical. There's so
much variation in individual level phenomena, I sort of gave up predicting
individual behavior long ago. But nevertheless, we see that it's a rather
weak relationship, it explains all the variables in this model, explain eight
percent of the variation sentence length.

Dr. Bachman: The F test here tells us that that is a significant amount compared to the
constant alone. I'm going to bring over my lab sheet so I can show you.
This is the multiple regression equation from this output. Again, the
constant is the intercept A, and I've simply delineated all of the slope
coefficients, that are presented here. And I've given you my note on this.

Dr. Bachman: When you're at a conference, never say, excuse me, that's a very low R
squared to somebody presenting individual level data. I actually did that
when I was a grad student and I was told very clearly that that's typical.
You should know that. So where do the wise? So here we see that when
we look at the significance here of our variables, sort of the first thing I
do is proves before I look at the unstandardized slopes, what's significant
and what's not.

Dr. Bachman: We see that defendants age even though it's a negative relationship,
which would indicate older defendants got shorter sentence lengths. This
is not significant at all. Prior convictions are not significant, even though
this coefficient would tell us that those with one or more priors got about
a thousand 440 days longer sentence than those without priors. It's not
significant, race is not significant.

Dr. Bachman: The only significance variables we have here are whether they killed
someone they knew, this coefficient of 1977 says, compared to folks who
killed strangers, defendants who killed people they knew got about 1,977
more days in their sentence, even after controlling for these other
variables.

Dr. Bachman: The coefficient for gender male defendant tells me that compared to
female defendants, males got about 3,291 days added to their sentence
net of the other variables in the model. And that significant barely but at
the old five level, it just makes it. The one variable here that is the most
strongly related to sentence received on this output is rather different. I haven't done it the last variable here, but it is jury trial.

Dr. Bachman: Whether defendants pled their case or went to jury trial, compared this is coded zero for those who pled and one if they actually went to jury or judge trial. And we see that those who went to a trial compared to those who pled, those who went to a trial, had an increased sentence length and days of 8,873 days even after controlling for all of these other variables. That's if you divide that by 365 days, that's about 24 years longer in sentence that they got added just for going to a trial.

Dr. Bachman: So looking at the T values here, we see that that in fact is the most important predictor of sentence length here. In your little handout I've given you again a way to really convey this to an audience at that is such an important factor in looking at sentence length. I'm predicting here using this regression analysis, this multivariate regression analysis, predicting sentence length, I'm predicting holding everything else constant, the 29-year-old and that was my average age male, which was the modal category white, who had no prior convictions, who killed one victim and who pled rather than go to trial.

Dr. Bachman: That's what I'm predicting here. So I insert these values of X and the one coefficient here for whether or not they went to a jury trial. In this case the coefficient is 8,873. For this first prediction, I'm studying that at zero because that's the cases that were pled, and what I get is a predicted sentence length of about 10,193 days or 27.9 years.

Dr. Bachman: Now again, this is all very peachy but what you want to know and what you want to convey to an audience is compared to what? Well, let me show you what a similar 29-year-old white male who killed one person who they knew and had no prior convictions. This is what happens when they went to trial. They had from this equation, we would predict a sentence length received is 19,000 66 days, which translates into about 52.2 years.

Dr. Bachman: So with this, you can say even after controlling for gender, race, age, prior convictions and relationship, there is an added about 24 years onto their sentence length, just for going to a jury trial. Okay, so these are the things that you can do with multiple regression, and I hope I've conveyed both the importance of using multiple regression when you're displaying relationship information because when you display bivariate
relationships, even if you're using regression, it tells you nothing about the true causes and the true factors related to your dependent variables.

Dr. Bachman: If you're going to present something to a policy audience or through academic or research audience, it's always important to control for as many important independent variables as the literature say are required. Otherwise, you could be conveyed that there are relationships that are actually spurious. Like ice cream sales and violent crime. So with that I'm going to, how do I go back? How do I go back guys? Oh, here it is. I'm going to stop sharing and open it up to any questions.

Erin Farley: Okay? Sounds good.

Dr. Bachman: Okie dokie. Any questions from those attending?

Erin Farley: The audience. Give them a minute, there might be some typing. I just like to remind everybody that this document that I just put together will be available in conjunction with the webinar when it's posted on our website. So this will be, there's a link that was made available, but we can also, we'll also post it as well.

Erin Farley: So it doesn't look like there are any questions. So with that then I would just like to thank Ronet for your time and oh wait, we might have a question. Do we have access to the word document? Yeah. Yup, we got that. Okay, so what we're going to do then we're going to open up the poll so that all of those who are still attending the session.

Erin Farley: If you could take a couple minutes and complete the poll, we would greatly appreciate that. And again, we would like to thank everybody for attending and if you do have any questions that you think of, feel free to email us or email Ronet directly. I think that was okay, the last one.

Dr. Bachman: Yes, that's okay.

Erin Farley: Ronet@udel.edu. And I think that's it. So thank you Ronet-

Dr. Bachman: Thank you.

Erin Farley: Thank you everyone for attending and have a-

Dr. Bachman: Thank you, Erin and Jason.
Erin Farley: Yeah.

Dr. Bachman: Thanks so much.

Erin Farley: Oh no problem. Thank you.

Dr. Bachman: Happy holidays.

Erin Farley: Oh yes. Happy holidays.

Dr. Bachman: Right around the corner.


Dr. Bachman: Yeah.

Erin Farley: So we'll have a whole bunch of new webinars coming out and plan for 2017 so everybody keep an eye out for those. Thank you.