Regression Design Essentials

Everyone, I wanna welcome you to our research methodology group webinars.

Uh, we are excited to have Dr.

Jim Rice here with us to talk about regression design essentials.

Uh, he is a masterful, I know you're in for a treat, um, as he discusses regression with you and with us.

And if, uh, there are students here, uh, feel free to ask your questions or, um, either during the presentation or at the end of the presentation.

But please take advantage of the knowledge that we have here with Dr. Rice.

So, without further ado, I'd like to hand it over to him to conduct the webinar.

Uh, Dr. Smith, thank you so much for that, uh, very kind introduction, but just set the bar a little too high for me there.

So, uh, well, I'm looking forward to, uh, sharing some thoughts about regression design tonight.

Um, it's interesting.

This is a, uh, probably an underused design.

I think there's a lot of opportunities to use this, and I don't see it used very often, so I kind of hope after tonight we'll have a couple more, uh, uh, proposals show up with a regression analysis proposed.

Um, so my name is Dr. Jim Rice.

Um, I'm a faculty here at University of Phoenix.

I'm actually a chair, so I'm not a use, I'm not a research methodologist.

So I will bow to the wisdom of all the methodologists that we have here on the call.

Um, but I do enjoy, uh, quantitative, uh, research.

So, um, I enjoy talking about it.

I enjoy, uh, sharing ideas about it, and I enjoy helping students, uh, wrap their dissertation proposals around quantitative research.

So, I hope we have some, uh, opportunities to have some good discussions and some good conversation tonight.

Um, gonna talk a little bit about what regression is and what regression isn't.

Um, you know, I get a lot of conversations with students who confuse regression analysis from with correlation analysis, and they have a lot of similarities.

Uh, but regression is very different.

Regression is really about, um, identifying an opportunity to forecast or predict, uh, an outcome as opposed to correlation.

That's just looking at, looking for patterns, uh, in the numbers.

So I think it'll be a little more ev evident as we go through.

But I always like to start out by just highlighting that regression is different than correlation, but we use a lot of the same inferential statistics as we go through the process.

And a lot of the same information and insights that you may have about correlation designs or about, um, variable types and sampling and population, uh, analysis are very, very similar.

And so we'll talk about those as we go through this.

This presentation, by the way, is longer, and I'm going to then, uh, anybody could rightfully go through in an hour.

Dr. Smith is laughing at me, but I intentionally put this together.

Um, everybody on this call will get a link to this presentation as well as this video afterwards.

And I'm happy to, uh, dive into more detail in any of these topic areas during an office hour.

Uh, so if, if there's something that you have follow up questions and we don't get them covered while we talk today, um, I'm always happy to chat with folks about it.

Uh, but I will skim through particularly some of those areas that are addressed in more detail in the correlation or in the sampling webinars that, uh, are on the website already.

So those that have worked with me before have seen my roadmaps.

Uh, as we walk through dissertation journeys, as we talk about, um, research design, I often cast it in the context of a dissertation proposal, uh, 'cause many students are interested in how do I apply this design to my dissertation research.

So we're gonna talk about the different aspects of a dissertation, um, the, the problem and the purpose, and the design elements, and the variables, and the things that are unique to, um, regression research as you work through the different aspects of a research proposal.

And, uh, hopefully will, it'll become a little more clear.

And, uh, the light bulb will come on about why you would use this in a proposal as we go through this.

And then we will wrap this up with the actual discussion and regression analysis.

So I'll be going through the first parts of this fairly quickly, um, but in a way that I hope will provide some context for the analysis that's helped.

Um, so what's the goal of regression? Uh, what makes regression unique, um, regression is really about, um, trying to predict an outcome.

So when we're looking at sets of data, when we're looking at variables, uh, variable pairs, um, what we're doing is trying to figure

out if, based on a, an independent variable, if we can predict what the value would be of a dependent variable, uh, in that dataset.

And for, uh, simple data pairs, where we've got two, two sets of two variables in our dataset.

It tends to be fairly simple, but, um, regression analysis can be performed on multiple combinations of variables.

It's called multi regression analysis.

And that's where it gets very interesting.

And we see some really interesting insights in the process.

And, and so always think about regression not as an understanding of the pattern in the data or the, the, the shape of the data or the relationships in the data, but think about regression as a, a way to build a model that will help, uh, predict a, uh, depend the value of dependent variable based on the independent variables.

Uh, we really do wanna look at the strength and direction, because we may build models that are more accurate and more predictive, and models that are less predictive.

Um, uh, and it's really about testing hypotheses.

Now, what's different in, uh, regression analysis is the intense in, uh, insights that we're trying to find in the variable relationships.

So we'll talk about different kinds of variables, particularly about confounding variables and codependent variables, et cetera.

So we'll talk about those in a little bit more detail.

But, um, regression is about looking at those relationships between those independent variables and how they can be combined in a model that predicts the outcome.

So again, ultimately, if we're trying to figure out, for example, what is the, um, likelihood that a student will have a large student debt after their master's program, we can look at all kinds of variables like their family income, their age, their, uh, uh, area of study.

There's a lot of different variables we can put in there, and we can look at independently how those variables relate to the outcome of student debt.

But we can also look at them in combination and see how we can combine them to be more predictive of that outcome.

There's all kinds of wonderful examples, uh, of this type of prediction.

Uh, for example, if anybody's ever been in a hospital room and you've ever been, uh, awakened by a nurse in the middle of the night to take your blood pressure and your pulse, and, uh, you're, you know, just looking you in the eye to see if you can, if you're loosened, what they're doing is they're collecting variables that are combined in a model that predict how, how likely you are to become a critical care patient in the next six hours.

So that model was developed through a regression analysis.

And so there's a lot of very practical applications to regression analysis that have a, a great deal of, uh, value to our life and our, and our daily living.

So, uh, let's dive into this a little bit.

So let's start with a problem statement.

You know, there's, there's, we had always talked about different topics, but, um, you know, the problem is where most students, uh, stumble.

And a problem statement in, in research is always, is always about figuring out what we don't know and that we need to know to mitigate an issue.

Um, a problem is not an issue statement.

A problem is about what we don't know, and it always has these five elements in it.

But for a, uh, regression analysis, there's a few things I always look for in a student's, uh, or in a researcher's, uh, problem statement.

I'm always looking for the variables.

What is it that they're trying to understand, uh, both from an independent standpoint, and if in a well constructed problem statement, what's the dependent variable? And the independent variables are often separated by some keyword that, um, indicates that there's an influence or an a, uh, prediction, predictable capability, um, uh, that we can discern, uh, about that dependent variable from the independent variable.

So in this particular case, we've got tuition costs, family income, financial aid, availability, academic performance, as all the independent variables in this proposed study.

And we're looking to see how they can influence or predict the amount of student debt that's incurred by North American graduate students.

And again, every good problem statement identifies and to limits the population as well that we're collecting the data from.

So look for that, um, design indicator in your problem statement influence prediction, something that indicates that it's more than just a correlation, that there's actually a causal, um, uh, effect in this, uh, in this proposed study that we're looking for.

So that's the problem statement.

Let's talk a little bit about the purpose.

Why, why do we do regression research? Um, and so when we dive into the purpose statement, we really wanna make sure we understand, uh, what the study is about.

But it's, what's it trying to understand in this particular example of a purpose statement, we're looking at the purpose of this quantitative regression study, uh, to examine the factors or the independent variables that influence the amount of student debt incurred, uh, by college students in the United States.

This is kind of a restatement of our problem statement, you know, what are we trying to, uh, get out of this? And then we're using mul, in this case, we're using multiple regressions.

So there's multiple variables in our case, tuition cost, family income, financial aid, availability, academic performance, um, and that a purpose section in a dissertation should always have objectives.

What are we gonna do with this research? And a prediction research or, or regression research modeling of the data, uh, should always have, uh, some kind of purpose.

We are all practitioners, and we are in a practitioner, scholar, uh, leader program.

And so we wanna make sure that our objectives that we state, uh, identify what a stakeholder or a practitioner may do with the findings of our study to help mitigate the issue.

So our objectives have to be clear.

So in this particular example, the objective of the study is to provide insight that can guide policy makers and educational institutions in developing strategies to reduce the financial burden on students, and for more, more equitable, equitable access to higher education.

We're looking at can a policy maker use the insight in this model to better understand how to influence these contributing factors to student debt? So regression analysis can be, um, have a substantial value on, uh, social policy on the way practitioners, um, uh, contribute to their field.

Um, this can be very powerful research.

So it's always something to, uh, consider when you're, uh, looking at various topics.

So, uh, problem statement, purpose statement, uh, research questions.

Um, so again, research questions should have some kind of design indicator.

So in here, we, uh, you know, what factors predict the amount of student debt incurred by college graduates in the United States? You notice a pattern here.

We're looking at the, the effects and factors that affect an independent variable.

And in the hypotheses we're gonna go look at, we're very often gonna go look at the relationship and the model that's associated with each of the, um, independent variables.

So in this case, um, higher tuition costs higher, yeah, higher tuition costs are positively associated with greater amounts of student debt.

That was one of our, uh, independent variables.

Family income is positively, but we're not positively in the negative, uh, associated with greater amounts of student debt.

So you'll often see a research question with a set of hypotheses that are directly tied to the various independent variables that we're including in our study.

Dr. Rice, uh, Dr has a question.

Oh, certainly. Sorry to interrupt. Um, thought here.

But, so these hypothesis say positively associated, so I get the correlation aspect, but is there a predictive aspect that goes into the hypothesis? Um, yes.

You, you could, and again, we're the, the basic research question includes the prediction aspect.

So since the hypothesis is directly associated with the research question, the predictive aspect or the influential aspect, um, is in the research question, I'm kind of switching back and forth.

These are just examples.

So I'm, I'm switching back and forth between predict and influence and, um, I should probably be consistent.

But I, I wanna do, give examples of different kinds of words that would be appropriate.

The, um, the hypotheses have to be specifically testable.

And so what I'm looking for is a, is a positive association, negative association, no association, and the hypotheses, um, that make, ensures that there's a correlation there.

But these are also giving me an opportunity to, uh, determine whether or not these factors are, uh, predictive of the, that the research question means that they're predictive, which largely comes out of the analysis phase of the, the modeling that we do.

Okay. So, okay. I understand. Thank you.

Okay. Did that help? Yeah, that did. Thank you.

Yeah. Yeah.

And we'll talk a little bit about that more when we actually talk about the regression analysis, uh, a little bit later.

I'm flying through this awfully quickly.

So I'm trying to touch all the major pieces of a dissertation proposal, and obviously there would be more hypotheses if there's more variables, and there could even potentially be more research questions if I've got multiple independent variables.

But that's actually kind of rare.

Uh, typically for a student regression analysis, we'll have a, a single research question with multiple hypotheses.

Let's talk a little bit about the variables.

Did anybody have questions about that before I go on? Okay. Let's talk a little bit about the variables.

And this is, you're gonna see a lot of similarities with correlation research here, but this is where we really start to diverge in the language between correlation and regression.

Um, you know, up to this point, and as we just discussed with the research questions, we talk, we use a lot of the same words when we talk about, uh, uh, positively associated, negatively associated, um, there's a, a relationship.

Um, there's a, there's a causal, there's a predictive effect.

Um, but what we're gonna start talking about here are the variables and the insight about the predictability, uh, that we're seeking to understand really comes from an understanding of the data.

And so, as we, I've used the term independent variables, independent variables for a great deal.

Um, the, the independent variables, um, just like they are in a correlation, uh, and the dependent variables, just like they are a correlation, are the two variables that we're looking to see if they cova, if they are, um, uh, varying or co varying at roughly a predictable rate.

Um, but what we're also looking for in, um, regression analysis are three other kinds of variables that may participate in our model.

One is called a confounding variable.

Um, and we'll talk about that more in a little bit.

We'll talk about moderating variables and mediator variables.

So all regression analysis must have some kind of discussion about these types of variables, um, where we're really not doing much more than correl or, uh, correlation analysis.

Uh, 'cause what we're looking for is contributors to the models that we're gonna build, the equations that we're gonna build that describe the relationship between these variables.

And so let's talk about those a little bit more.

Um, so the, the confounding variables, now, this is probably the most difficult concept for most, uh, students to understand, but confounding variables, independent variables that are, um, associated with the independent and dependent variables.

So a confounding variable may be a variable that changes, and as it changes, it modifies the independent variable and the dependent variable.

So there's actually a, a relationship that exists.

So as a confounding variable changes, it will change the independent variable, uh, meaning that directly or indirectly, it will, uh, affect the dependent variable.

Um, and a confounding variable may directly affect an inde, the dependent variable, and appear to, uh, cause a, uh, uh, Covance with the independent variable.

So, confounding variables are, are a little bit challenging to, um, identify, but they're pretty easy to control for.

Um, so if, for example, uh, in our student study, if we are interested in understanding if, um, uh, parental education levels have an effect on, uh, student income and or, or parental socioeconomic status has an effect on student socioeconomic status and their student debt, we could look at parental socioeconomic status as a confounding variable.

And typically what we would do is we would do a couple of things.

We'd rather randomize our population to spread out these compounding variables more uniformly.

But if it's something that we can identify and even potentially measure, uh, we can start to do sample stratification and start to bucket these, uh, group subgroups of our, um, sample into different confounding stratums stratus.

Uh, this allows us to continue to study the d the vari, the variation or Covance, or the relationship between the independent variable and the dependent variable.

Um, within different con, uh, populations of data that, uh, are stratified by the confounder.

Um, there's also some other statistic patrols.

We can also do some matching to make sure we've got a uniform distribution of confounders in different subgroups of our population to make sure that the confounder isn't having an undue influence.

But whenever you're doing a regression analysis, there's an opportunity for the researcher to sit back and speculate about potential confounding variables that may exist in the population that they want to control for.

And so, anytime I'm looking at a regression study, I'm looking for controlling factors or controlling or controls within the study to help mitigate the impact of compounding or confounding variables.

Uh, if I don't see that, uh, then we have a conversation.

We figure out how to, uh, strengthen the study by putting those controls in place.

So always look for compounding variables.

And as a researcher, it's an opportunity to sit back and, and think about what are some other variables variability in this population that I need to control for that could potentially influence that relationship, either exaggerated, um, uh, lead to a false understanding of the association, or, um, suppress the relationship that exists between the variables.

So, confounding variables or important thing to look for.

There's two others that are a little easier to understand.

Um, and generally I'm looking for a description of those, uh, when they exist, or identification of those if possible.

Uh, but they typically have a little less influence, direct influence on the analysis.

One is called a mediating variable.

Um, mediating variable is a variable that exists between, um, the independent, the dependent variable.

So in my example earlier where I talked about, um, the socioeconomic status of parents as opposed to the socioeconomic status of, of students, uh, as a measure of, uh, student debt.

Um, if we were measuring parental socioeconomic status as an independent variable, the, uh, student socioeconomic status might be a mediating variable, indirectly influenced by the parental status, uh, and, and having an effect on the dependent variable or their, their debt load.

Uh, so sitting back and thinking about, um, your variable relationships and thinking about whether or not there might be a mediator, uh, in between our independent variable, independent variable, um, another type of variable would be considered a moderating variable.

Moderating variable is a variable that is influenced by the independent variable.

Uh, but the independent variable is still doing a direct influence on the, uh, dependent variable.

So, for example, um, you might have, uh, an independent variable being, um, the students at socioeconomic status and their debt load, but you also may have a moderator in there that is, uh, related to the, um, job income of the student.

You know, is the student working, how much do they make? What kind of career are they in that, um, their socioeconomic status may influence that moderating variable, which could have a, an alternate, uh, separate influence on the in or the dependent variable, uh, thus, uh, biasing or exaggerating or, um, minimizing the relationship between the independent variables, independent variables, again, the confounding variables, the mediating variables and moderating variables are an important part of the discussion about the population and the analysis that you'd like to see, uh, in any research proposal that includes a regression analysis.

'cause you're really looking at, um, the strength of relationship between the independent dependent variables and identification of confounders mediators and moderators is an important part of that, uh, uh, understanding of the population that you're, um, uh, seeking to study.

Of course, the types of variables are all still, you know, these are the same as you would have with a correlation analysis.

Um, you know, we've got ratios, intervals, ordinals, nominals, discrete, continuous.

Um, the type of variable will often influence the type of regression analysis that we're doing.

Uh, so for example, if you're doing a, a linear regression analysis, the dependent variable has to be a continuous variable or a ratio variable.

You can't have an ordinal variable and a linear, uh, linear relationship.

So knowing your variable type will have a direct influence on the, uh, regression type that you choose.

And we'll talk about regression types a little bit later.

Variables must be defined. Yeah, go ahead.

This is Pam. Sorry, can you go back a, a slide? Sure. You know, you just said about the certain variables, do certain things.

Do you have a slide that you put, uh, ordinals and say that this is the only way, this is the only variable that can go with that.

Do you have anything with that? You know, I don't have that in here right now because I'm lost.

That's an excellent addition here. Um, yes, please.

The, uh, I will tell you the vast majority of regression research that students do yeah.

Um, is on, uh, ratio or continuous variables.

And so linear, and most of them will do linear regressions.

Yeah. Um, if they start to do, if they start to include ordinal variables or nominal variables, um, and they've got multiples of those in their population, um, and they've got an output that's, uh, ordinal or interval, have 'em call me.

We'll, we'll talk through the, we'll talk through the different, uh, okay.

Types of regression analysis. Thank you.

This, this slide deck could be 200 pages long, I'm sure.

Thank you. Thank you.

So, um, happy to talk to talk to them, but as I said, the, the vast majority of them are gonna be simple linear regressions, or even multiple linear regressions.

And those are very, uh, straightforward to do.

Um, uh, typically you do them in Excel spreadsheet.

If you've got the data load in Excel, you could do a, a linear regression in five minutes and, and, uh, begin to do your analysis on the, on the regression.

Even even a multiple regression, if they've never done it before and they pull YouTube up, it might take 'em 25 minutes to do it.

It's, they're not that hard to do.

The hardest part is knowing what to do, not how to do it.

So, um, and I can certainly help them identify what type of regression.

I'll talk a little bit about different types of regression later, but I'm not gonna go into detail about what, uh, what the requirements are for each of them.

Maybe in a future version of this slide deck, I'll, I'll add that in.

But, um, that's a little bit beyond what I wanted to do today.

Does that help? Yes, sir. Okay.

I just needed a, a little, yeah. Thank you.

You know, people get really afraid of, um, quantitative research because this numbers and math and statistics, you betcha.

And the tools that we've got at our, our fingertips today make it so easy to do.

Um, you know, it takes longer to prepare the data than it does to actually do the analysis.

Um, but the trick is knowing what the analysis is telling you.

So you kind of have to understand what it's telling you more than knowing how to do it.

So, um, so we've talked through right now, um, all the red stuff up top, the, uh, uh, the pur problem purpose, a little bit of the methodology, design alignment variables.

But, um, I'm gonna blow through the sample description and sampling, uh, or the, the population and sample and sampling

techniques very quickly here.

I'm just gonna blow past them because, um, they're exactly the same as the correlation, uh, uh, language.

And we've got another webinar that we did on correlation, and I wanna make sure we stay on time.

'cause I think there'll be a few questions on the analysis.

So, um, just in chapter one, you're gonna talk about the population.

Always encourage students to make sure that they're clearly describing their population to limitation.

Uh, they clearly talk about the delimited size, the demographic data, so they can make sure that they're, the data they collected is representative of their, um, overall population.

And also ensure that, um, uh, it is, um, well or is, is properly sampled.

So, for example, if, if I'm studying a population of students, and, and even if I'm looking at just their financial debt, uh, and 50% of the students in the overpopulation are the overall population are female, and, and 50% of them are male, I, it is, I need to have a sampling control to ensure that my sample has roughly that same ratio.

Or I can't claim that my sample is representative, even if I randomly sampled it.

I have to make sure that I've got sampling controls to make sure that my sample roughly approximates the demographics of my overall population.

So I can assert representation.

The sampling technique also has to make sure that it, it identifies the minimum size, and we've got good determination on when we've sampled enough to have a substantive sample.

All the same, um, sampling controls, how to del limit, how to understand an overall population, how to understand the delimited population.

Um, how to understand samples is, it's all the same for most any quantitative research.

So the slides are here.

Um, you're welcome to go through this.

And there's another, an entire, another webinar just on population and sampling that we did a couple months ago that, uh, I would encourage people that are interested in learning more about this section.

Uh, go, go and listen to separately.

'cause that was a whole hour just on sampling.

So, um, I'm not in the tools for doing that.

Again, defining minimum sample size and statistical significance sounds very scary.

It usually means plugging three numbers into a, a dialogue box and getting the output number that that's how hard it's, you have to

understand what to put in there, but actually doing the calculation isn't that hard.

Um, so again, population sample sampling techniques.

Um, purpose of sampling, snowball sampling, convenience sampling, random sampling, case study sampling, stratified sampling, maximum variation, theoretical sampling.

All these have to do more about with your population.

Then the, the, uh, the design.

You could be doing correlation.

You could be doing expos facto, you could be doing, um, uh, regression.

You could be doing, uh, many different quantitative designs.

And, uh, depending on your population, you may choose almost any of these sampling techniques.

They've got a different, each technique has a different purpose, a different value, and they'll all result in an appropriate sample if, uh, executed properly.

So I just blew through sampling, see how fast we got through the blue stuff.

Let's, uh, talk a little bit about the descriptive statistics here, and then we'll dive into the regression.

Um, so again, uh, descriptive statistics.

Uh, we always wanna make sure that our data that we've collected is representative, that it is a good, uh, has a good central tendency.

So there's good variability in the patterns, uh, that we've got with, so we can be assured that it represents the population.

And so, taking a look at measures of center, center of tendency, variability and distribution, uh, the mean median and mode, uh, are some of the most common ways of taking a look at those and assessing any assumptions that are necessary.

Uh, so for example, in a correlation analysis, it often relies on certain assumptions about the data.

The same is true with, uh, regression and, uh, uh, in regression analysis.

We're often looking at the, uh, the type of regression.

And we, we choose as dependent on the, uh, type of dependent variable.

And we'll talk about that more in a minute.

Um, and we may still wanna do some data cleaning.

So if we see some significant outliers in the data, we may want to, um, reject those.

We may have some, uh, variable comparisons.

We may find some variables that are redundant.

We wanna discard those. Um, so data cleaning is important.

Um, but while all this math looks funny for doing mean media and mode, uh, the reality is most of the time students are looking at the, the scatter chart for their, uh, variables.

So take a look at the, the scatter plot for your different variables.

Look for these significant outliers that may exist.

Throw 'em out, you know, look at them, understand why they're outliers.

They may just be a collect data collection anomaly.

It might be somebody that filled out a survey and just filled the number one in, in every column.

Um, we'll almost always find an outlier in the data.

And the larger the data sets, the less significant the outliers are.

But many of our students are collecting, uh, fairly modest data sets, you know, 600, uh, under 600, or even under, uh, 200, uh, data points.

And some of these outliers can, uh, can significantly alias their data.

Uh, it's also, uh, worth taking a look at the shape of the scatter graft to get a level of confidence that it looks like it might be linear.

It looks like it might have a correlation like this one does this, this looks like it's a relatively linear, um, relationship that has a, uh, linear and positive relationship that gives us a level of confidence that if we do a linear regression, we're gonna get a good output and we're gonna have a relatively high confidence interval or r squared of the, uh, the model that we built.

So I always encourage students, once you've collected your data, produce the scatter plots.

Take a look at those, visually inspect them.

There's obviously statistical ways of testing this.

Uh, but I'm perfectly comfortable with any, uh, student that, uh, produces a scatter plot, describes what they're seeing, uh, because that's, that's a better description that they under, that's a better, uh, indication to me that they understand what they're, um, looking at.

Uh, rather than just plugging data into a, into a tool.

Um, yeah, you know, is the data skewed? Is it ketosis? Do we need to do, um, uh, transformations of the data? Again, we talked about that, uh, at great, at great length in the, uh, correlation, uh, presentation.

So I'll let you go and look at that one.

Let's spend a little bit time taking a look at the actual regression analysis itself.

Uh, this is what you guys all came to listen to today anyway, it was the actual regression.

And this is where we get a significant, uh, divergence from something like a correlation or an expo factor was the analysis itself.

Um, regression analysis, it's all about building a model, and it's building a mathematical equation that models the data.

And that equation, um, will result in a line that's get plotted through that scatter plot.

And we wanna measure, uh, we need to determine how to calculate that line, what that model needs to look like, what that equation needs to look like.

And then we also have to assess how that fits the data.

Uh, that's the r squared fit.

Uh, and then we have to interpret the results.

So these are the, these are the steps I'm always looking for.

When in chapter three, when a student is going through and describing their data analysis.

This is, these are kind of the steps I'm looking for.

I wanna know, um, how are they identifying the variables? Why did they identify the variables they did? How did they collect the data? Um, you know, so that we've got a degree of confidence in the data.

That's often in the data collection section, in the data analysis section.

We start with that exploration. Did they plot it? Did they prepare it? Did they remove the outliers? Uh, did they, um, do a transform on any of the data to make sure it was linear? Um, and be able to, uh, uh, basically later under be able to describe that it's not the analysis was linear, but the, the relationship was exponential, or the relationship was quadratic or, or multinomial.

Um, then we have to turn, once we've, once we've prepared that data, we have to do a, uh, a regression.

And this is where that, um, variable type, uh, comes in.

And I think, uh, uh, whoever asked the question, what the dependent variable type was, uh, this is really what determines the type of regression we do.

So I would say 99% of, uh, the research there's students are gonna do, will ultimately be a linear regression, which means their output variable, their dependent variable, uh, has to be a continuous, uh, variable.

Uh, it can't be, uh, binary. It can't be categorical.

It must, for a linear regression, it must be a continuous dependent variable.

But if they have a binary output or a, um, if they've got a categorical output, uh, or they have a bucketed output, um, there are other types of regressions, uh, logistic regression, what's on regression of multinomial regression of the most common, um, for the different types of output or dependent variables that, um, uh, students may have as a result of their regression model.

Um, once they determine the type of regression, uh, they have to start building the model.

They have to define what the equation looks like.

Um, they have to select that, uh, regression model and the regression model to type determines the type of equation.

Um, I will tell you, most of the time, the tools you use will determine the equation for you.

So you'll, you'll look at your data, you will determine that you wanna do a linear regression, and you'll go tell Microsoft Excel.

You wanna do a linear regression, and it will pop out the model, uh, and it will pop out the model with, um, all the, uh, uh, various modifiers on it necessary.

Uh, but you also have to assess the model fit.

How well does that model, or how well does that equation fit that, uh, data set? Uh, so if you have a really loose scatter plot in the line drawn through the middle of it, that r-squared value is gonna be, um, you know, to one extreme, if your, your data relationships are, um, almost linear, one for one, uh, then that r-squared value is gonna be gonna indicate that the model is a very tight fit.

And, uh, we'll look at that, we'll look at some examples here in just a minute.

And then the most important thing I'm, I'm always looking for is, do they understand enough about what their model tells 'em to be able to analyze, or more importantly, interpret the results to be able to describe what that model is telling us, uh, about their data so that they can make an informed prediction about the relationships and those variables going forward.

And that, that, that prediction is valid.

So let's look at a few examples here.

So, Dr. Ryan, uh, Dr. Baron Had a question, sorry to sorry to interrupt you again.

Um, do you tend to point students towards SPSS or do you pretty much count on Excel to do the analysis? Um, You know, I personally, I use SaaS quite a bit, and I, I do have an SPSS license that I've used it if students are using, using SPSS for that.

Um, but increasingly, the, for most of the regressions we're doing with students, um, they're, they're looking for relatively simple regressions.

And so we just use Excel.

And that way they're not having to lay out any, uh, money for, uh, tools.

Um, if they wanna use a statistical tool or if they are familiar with the statistical tool, I will certainly help them with that.

There aren't too many out there that I haven't used, um, uh, SaaS.

Many students like to use SaaS because SaaS offers their licenses for free for students, right? SPSS, the students get a student rate on the tool.

Um, all of those tools do, in fact, the, the graph that you're looking here, this, uh, Kline graph is actually an SPSS output.

Um, so that was a screenshot from an SPSS run.

The one on the left is actually an Excel run.

Do you think there's any benefit one way or the other? Uh, from a research standpoint, no.

Um, 'cause we're really not teaching the tool.

We're teaching, um, how to do the research and the analysis.

And as long as they understand what the tool is doing, I don't care what tool they're using.

Got it. Okay. Thank you. Yeah.

Um, I tend to be a really cheap, uh, chair.

So, um, I'll point them to the tool that is the cheapest, that they're the most comfortable using.

I'm with you. I, I don't know too many students that have an excess amount of funds that can lay out several a hundred dollars for SPSS.

Um, but it's, uh, but SaaS is all web-based and free for them.

But it's, there's no graphic interface, uh, for this, uh, or the SaaS.

The graphic interface for the SaaS student license is pretty limited.

So, um, yeah, it's whatever tool they're most comfortable with.

So, like I said, the actual regression analysis that I'm about to show you, um, only takes a couple minutes.

So if they've collected the data and they pulled it up and they run a scatter plot, um, what, however they do it, they can do it, draw it by hand if they want, if they wanna do the scatterplot by hand, uh, and look at the, uh, data elements and see if they think there's a fit.

So, for example, this linear example here, um, if it were my data set, I would probably be inclined to throw out that number in the upper left.

'cause it's probably an outlier.

It's probably somebody that misentered the data, um, before I did the modeling, and I'd keep the rest of it.

Um, I actually did it on this data set, and I found it didn't really matter much.

But, you know, that would be me, the one on the right, that Kbo linear graph.

You know, you look at that scatterplot and you go, oh my goodness, how did they ever put a line in there? And, uh, there's actually, uh, multiple, uh, variable types in that data, uh, weighted variables.

And, uh, so that, uh, that's a multi-variant analysis.

And so we'll see how that was, uh, calculated here in a minute.

But, um, that wound up being a non-linear or a curve linear, uh, analysis.

Um, and that's what, and that was done by SPSS.

Now, I would say S-S-P-S-S, um, is far more powerful.

You know, if you give it 60 different variables and tell it to, tell it to identify the outliers and throw them out and tell it to, um, identify

the type of best fit curve by running through all the analysis, it will do that for you.

Um, and honestly, if a student knows how to do that in SPSS, I'll, I'll accept that because they know how, they know at least know to ask the question.

Um, SPSS is a much more powerful tool.

SaaS is a much more powerful tool.

R is a much more powerful tool.

Um, uh, power BI is a very powerful tool.

Um, there's a lot of incredibly powerful tools, uh, available to our students today.

They don't cost anything.

Um, and if they understand the basics of doing regression analysis, they can pick up those tools pretty quickly and go spend 10 minutes with YouTube and figure out what the commands are to actually run their regression and their data types.

So this is, um, this is again, the, the first step.

This picture really here is an indication of what I'm really looking for a student to do.

First, this is map your data, become familiar and comfortable with your data, create the scatter plot and really stare at it and look at it before you put a line on it and, and take a look and say, does it look like there's a relationship here? And if so, what type of relationship is it? And it's, it's that basic understanding of looking at this collection of data and being able to see that line in there before you, you calculate it, is, uh, probably one of the most empower.

One of the most powerful lessons that a student will take away from doing a regression analysis is that they're looking for a way to predict what, um, one, the occurrence of one variable, the value of one variable will say about the prediction of the, the dependent variable.

And being able to figure out, can I model that? Can I create an equation that can predict that? Um, and the good thing is that equation predict that creation build, while it looks really complicated there on the bottom, uh, is all done by the tools.

These days. You don't, very seldom do you have to do that by hand anymore.

I don't think I've done a hand regression in 25 years.

So, um, different types of regressions.

Um, we talked a little bit about some of these, uh, linear regression is probably the most common, uh, or multiline regression where you do multiple regressions, um, between different variables and bring them together.

Um, but there are other regressions, um, quadratic cubic, exponential logarithmic, I would say.

Um, you, while you can do exponential logarithmic, um, um, uh, regressions, you can also often transform the data, um, with expert with an exponential or log with a transform to make it linear.

And then just do the linear analysis.

And just remember when you're doing your analysis of the data, that that relationship's actually logarithmic or that relationship is actually exponential.

A little harder to do with cubic and quadric.

Uh, those aren't really transformable data, but exponential logarithmic are.

Um, and like I said, there's a, there's a plethora of tools out there today that are used for regression.

And you will actually find, if a student comes to you and says, I wanna do a regression analysis, they're gonna often come with their tool of choice.

It usually means they're working with that tool at work, and they want to use the tool they've got at their fingertips that they're either comfortable with or they get for free.

Um, Microsoft Excel is probably the most common that I've had with students.

Um, SPSS, because we, we often have used that in our statistical classes.

Um, SaaS is required by the US Federal Government.

Uh, if you're doing any statistical analysis and submitting a regression model to the federal government, it has to be in a SaaS formula or a SaaS model.

Um, so there's a lot of, uh, statisticians that are very comfortable with SaaS, but increasingly power bi, uh, Tableau, um, Python, uh, are, are, uh, pretty common.

MATLAB's a little dated. We don't see a lot of, I haven't seen MATLAB in a long time.

My daughter came to me with a MATLAB model.

I went, who are you and what to do with my daughter? Um, and, uh, sta uh, you know, I've seen a couple of times pop up, so I put it on the list, but don't be afraid of the tool.

'cause the, the tools all do the same thing.

Um, you give the tool a set of data or a set of data pairs, and you tell it to do a regression, and it will come back, uh, in most, and sometimes you have to tell it to do a linear regression, and it will come back with the, the line fit through this, the, uh, scatterplot.

It will give you the equation.

It will even tell you what the, um, confidence interval is, or the best fit is the r squared value, uh, for that equation.

And so the actual number crunching gets done by the tools.

Um, so just as an example here, this is a, this is a very simple linear regression.

You can see there's a tuition cost and a student debt ratio.

And we just said, okay, what's the scatterplot? And I, I selected that in Excel and I said, insert plot, put it here, insert scatter plot.

And it built a scatter plot for me.

And then I went into the attributes of the table, I said, add the re linear regression equation.

And it put the equation on the chart for me.

And I checked the little box that said, give me the best fit or the confidence I ever wanted.

And it put the r sport value on it that took, you know, that simple regression took, you know, two minutes maybe.

And that's because I was formatting the table to make sure I got rid of the grid line.

So it was nice and easy to read.

This is probably as complicated as most of our students do.

They're looking at a relationship between two variables, and they wanna know if one variable is predictive of the other.

Um, and so this is probably the most common, uh, dissertations I've seen.

And again, I'm not that concerned about the complexity for me as a chair.

I'm not that concerned about the complexity of the regression.

But did they do the analysis to look at confounding variables? Did they do the analysis to look for, uh, moderators and mediators? Um, do they understand the tenets of regression? If they do, then I'm perfectly happy with a relatively simple model.

And then the model here is y equals x plus 51.

It's pretty straightforward. Um, uh, I would say the students that get a little fancier, uh, are gonna look at, uh, multi multivariate regressions, where they've got multiple variables.

It sounds like a fancy term. Does it? Multi-variant regression just means there's multiple variable pairs.

We're looking at, uh, student debt in comparison to student tuition costs.

We're looking at student debt in relationship to family income.

We're looking at student debt in relationship to financial aid.

And you select that whole table.

And in Excel you go up and you say, if you haven't already included the stats module, you check the little box and it adds the extra icon on your data ribbon.

And you check regression and you drop down and say, I wanna do a linear regression.

And you select that table and it'll kick out that summary output.

And the summary output will tell you what the regression values are, the residual values, and it will build the equation for you.

Uh, that is the multi-variant equation.

And so you can see, um, the dependent variable y with, um, the offset and the, uh, contribution of X one, which is student cost X two, which is family and income X three, which is financial aid to that outcome.

So it built the model for me.

I didn't, I, I should have put the R value on here.

I didn't, uh, 'cause it kicked that out too, and I just never did add it to the slide.

Now that I'm staring at it, I'm going, why Kicking myself again, I didn't put it on the slide.

Um, but it will actually even help you with the analysis.

It'll do the independent graphs down here.

Uh, you can see, you can scroll through here, see the independent graphs.

Now I, I laugh at this particular example because, um, I put it on here and someday I'm gonna go fix this one.

And you'll see these pound numb exclamation points.

And it's because, um, when I ran this particular, uh, regression for this, this, uh, screenshot, i, it, I let it select the income variable type and said family income was a bucketed variable and not a ratio variable.

And it thought financial aid was the same thing.

And so it chose the wrong variable type.

So I had, I had to, when I actually ran this for real, I had to go back and change the variable type on all of them to make sure they properly represented the data.

And that, that's what you have to do.

You have to understand what your variables are, what types of variables they are, you have to be able to interpret the data.

But the actual calculation itself, uh, is very simple and straightforward.

This multi regression analysis, it took me longer to enter the data than it did to actually run the regression.

So, um, and what does this tell me? It tells me that there is a actually positive relationship, uh, between all of these factors and all of these factors contribute to student debt.

And, uh, and they all contribute in different ways.

Uh, so if you look, there's two of these variables that have roughly the same, uh, contribution.

And one of 'em has a much more modest contribution.

So family income had less substantial, uh, contribution and tuition, debt and financial aid.

And you can see that from the model.

Just looking at the model, you can see that variation.

Excel will do linear regressions, multivariate litter, linear, uh, linear regressions.

It'll do exponential regressions, it'll do logarithmic regressions.

Um, uh, if you go beyond that, you're better off going over to SPSS or or SaaS to do the, uh, regression.

'cause it'll, it'll get a little too complicated Regression analysis, it's not that hard.

It's about building a model and then being able to take a look at that model and interpret it, uh, to be able to tell us in the real world what the relationship is between those factors that are contributing to our dependent variable or to the outcome that we're looking for.

Uh, in that relationship, sometimes we find no relationship at all.

Sometimes we find very tight relationships.

Uh, we very often find nonlinear relationships.

Uh, and I will tell you, when I find nonlinear relationships, that's when I'm looking more closely for confounding variables.

Uh, 'cause there's usually another factor in there we're not taking into account, but, um, they're not hard to do.

Um, but what I encourage students to do is spend a little bit of time, um, actually learning what regression tells us, when do you use it, and what does it tell us? And from there, back into, back in from that understanding into the actual processes, and it's a pretty straightforward design to execute.

So, um, we have five minutes left, so I think that was my last slide.

So any questions I threw, I flew through a whole lot of stuff very fast, and hopefully I didn't scare people because it's, it's pretty straightforward.

Um, the hardest part is just under understanding and not being afraid of the numbers.

Jim, this is Karen Johnson. Hey, Karen. Hey, how are you? I'm doing wonderfully. How are you doing today? Sure, same. Same.

Um, we all know that we're dealing with, uh, students using AI in our classes.

Mm-hmm. Do you see any role for AI in, in doing quantitative analysis? You know, I do. Um, uh, and I have done this with a couple of students already that are interested in, uh, quantitative, uh, analysis, correlation, x plus facto.

You can ask ai, tell me how to do this in Excel.

Tell me what are the variable types and how do I describe them? And it will do a really nice job of teaching the student how to do it.

It'll write terrible text and it'll create lists of things that are almost unintelligible, but it's unintelligible from an a PA standpoint.

But it does a really nice job of providing a step-by-step process that if the student is interested in doing, for example, a regression analysis, and they say, I have three variables.

This is what they are, they're ratio variables, and I want to use Excel to do a re uh, regression analysis.

Tell me what the steps are, I have to follow.

And it will do a really nice job.

It'll make a couple of mistakes, and the student may have to catch those, or we don't have to catch those.

But, um, it'll do a really nice job of helping the student not miss major steps in the process that they have to follow.

Um, it, it can do a nice job of making our job our life a little bit easier, because when they actually go to execute and do the work, it won't do the, it won't do the analysis for them.

At least it won't do it well, but, um, it'll definitely do a good job of showing them how it should be done.

You can almost think of it as a YouTube on steroids, uh, when it comes to quantitative research.

Right, right. I've often said it's just Google on steroids for, yeah, I, I did a interesting experiment the other day.

I gave it a table of data.

I gave, uh, chat GPTA table of data.

I said, identify the variable types.

Um, and it did a really good job.

It not only did a good job, did a better job than I did, I had misidentified one of them.

Um, and I won't even say why I did that.

I just mislabeled it.

But, um, it caught it and, uh, helped me correct my good, helped me correct my, uh, interpretation of my variables.

Um, I've used it, um, I've used chat g PT to do, um, and, uh, co-pilot both to do assessments of, um, data collection instruments.

So when students will build a survey instrument with a bunch of questions, I will have it go, tell me how long will it take a normal person to take the survey and when it comes back and says 32 hours.

And, you know, I will encourage the student to shorten their survey.

Um, but you can also ask it, is there bias in the survey? Is the survey, um, guiding a, um, a participant in a, in a particular path, uh, resulting in confirmation bias? And, um, I will tell you, check GPT does a great job of assessing instruments. Have you tried, Have, have you tried having it generate the actual scatterplot? Um, I have and it does it poorly.

Okay. So, Um, uh, copilot, I think that's one of my concerns because I don't think this, the typical student would know whether or not it was done correctly.

Yeah, true. Yeah. But I will, when they submit it, Well, that's true.

Right? Right. Yeah.

And I will ask them to go fix it and figure out why it's wrong.

Okay. Yeah.

Um, you know, and then we'll say, but, but, but AI told me this is the way it is.

I said, oh, so you don't know how it how it got here, huh? Okay. You know, AI is not a, and I, and I tell my students this all the time, I don't care if they use ai, they are still obligated to understand what they are communicating.

And, um, so if they want to use AI to check their work, if they wanted to make recommendations, if they wanna use it to teach them how to do something, I'm okay with that.

They're still obligated to ensure that their, um, proposals and their dissertations are formatted correctly, that their content is accurate, and that they understand the content that they're presenting as their own work. Yeah, I agree.

Thank you.

Any other questions? I think we're at time.

Dr. Smith. I ended on time.

Good. Great job. Thank you so much.

I have been posting in the chat the link to our feedback survey.

So please go in and share your thoughts about this webinar.

This is a research design webinar.

There'll be a question where it'll ask you the type, so if you can select that you'll get the proper, uh, questions for, for the webinar.

Um, also if you are interested, the, and I know you all are the PowerPoint and the recordings will be available on the research methodology, uh, team site.

Uh, and the, I think I saw a question asking about the correlation slides.

Those are already, they should already be there.

I'll check and see, 'cause I've gotta remember if we did that webinar this year or last year.

But I will go in there and pull 'em from their other resting, uh, or, or the other place where they're stored if they're not there, and get them onto the research methodology.

Um, group site. Um, correlation was wanna, The correlation was last year the, um, uh, population, the sample size sample, yeah.

Okay. I'll go pull that 'cause and put the, I will put it in the channel and then if you're a member of the research methodology group team site, you'll get an announcement when it gets posted and you'll know where, where it's at.

So I, right now I'm posting, um, how to join the methodology group.

So it's this really long teams link, but if you click that link, um, ask to join, I will add you because it's a private group and you'll have access to all of the resources that we have on the site.

Um, and years of webinars about different methods that, um, hopefully will be helpful for you as you're working on your research or supporting your students, um, as they're doing research.

And then finally putting one more thing in the chat, um, if you're interested in knowing when other events are, um, all of our events along with events from other, oh, it's not there, along with events from other groups are on the research hub, uh, event calendar.

And so if you go there, you'll see events from all, uh, sorts of, are all areas within the college of, uh, doctoral studies.

So I am going to go get that link and post it in the chat as well.

Yeah. And Dr. Smith is probably worth noting. Yeah.

Um, there's some great folks in this group and that'd love to, uh, share their experience with, uh, with the designs.

Um, I love talking about quantitative research design, so I've got a Calendly link that, uh, Dr. Smith can share if anybody wants to schedule time during one of my office hours.

So any final questions before we wrap up for tonight?