Research Methods in International Relations and Political Science
Focus Questions for Unit 3 (Last updated: 11/20/2024)

Some ideas to review before you go to the rest of the guide

- As you read through the guide, you will not be familiar with many of the italicized concepts at first. Everything below will be covered *before* your test, and you will notice that many of the questions I want you to focus on also include an answer. **Any material that is changed after I initially post this guide will be bolded**.
- ***Don't let all of the specifics in this guide intimidate you.*** The sections below on linear and logistic regression focus primarily on application examples. They are very specific and ask you to answer questions based on hypothetical regression output. As you may recall from the last unit test, several multiple-choice items required you to identify which statistical method was appropriate for a given situation or to interpret a statistical finding. Some of the items below do the same.
- In the big picture: What I want you to understand best at the end of this unit is when, why, and how to use and interpret results that involve these statistical methods: (1) t-tests, **(2) Chi -squared tests,** (3) correlation, (4) linear regression, (5) statistical controls, (6) the interpretation of interval, dummy, and interaction variables, and (7) logistic regression
- What *must* you be able to do for the final test? You should assume that you will have required essay items that:
  1) give you SPSS output for a linear regression model and ask you to interpret the appropriate R-square statistic,
  2) ask you to do the same thing, but with a logistic regression model, and
  3) give you the summary of a study and ask you to identify the study's theory, hypotheses, dependent variable/s, independent variable/s, control variable/s, and an intervening variable if there is one.
- You will need to understand and be able to give examples of categorical, dummy, ordinal, and interval variables. Specific questions related to each type follow below.
- Most of this course unit examines how researchers use either *dummy* (sometimes called "binary" or "dichotomous") and *interval* (sometimes called "continuous") variables to examine and test the statistical significance of relationships between variables. While there are statistical methods designed specifically for *categorical* (often called "nominal") and *ordinal* variables, your textbook focuses exclusively on statistical methods involving dummy and interval variables. The textbook's focus reflects the fact that most researchers choose to convert categorical variables into dummy variables before any analyses that involve statistical tests. And researchers typically convert ordinal variables into dummy variables or treat the ordinal measures as interval variables, depending on their number of categories and frequency distributions. Be able to define and give an example of each of the four types of variables *italicized* above. Also, make sure you understand the key ideas communicated in this paragraph (i.e., this block of the study guide). This paragraph explains why this course focuses on a lot fewer statistics and methods than the typical social science methods course.
- What about the distribution of an ordinal variable—and even sometimes what a measure that seems to look like an interval variable—would make a researcher consider collapsing that variable into a dummy variable or perhaps two of them? Thinking about how household income and education are measured in the datasets we have worked with in class, why do researchers typically group and order the response categories for these types of interval variables so that each incremental gain/group doesn't have the same number of units (i.e., the number of years of education or range of income often is the not be the same for each unit)? Are these new variables still interval variables? (Hint: Yes, they better interval variables because have been modified so that each one-unit increase on the measure as a substantively similar effect).

- ***Testing for an association between two variables***

- There are numerous association measures that social scientists can use to analyze pairs of variables where *one or both measures are categorical (AKA nominal) or ordinal variables*. However, as explained above, political scientists and economists most often focus their statistical analyses on variables that have been transformed into dummy and interval variables (an exception is regression analysis designed to work with ordinal dependent variables, which is quite common, but that method is outside of the scope of this course).

  The only association tests that you will learn anything about in this course are correlation and *a chi-squared (x^2) test*. What kind of variable pairs are examined by a chi-squared (x^2) test? Hint, this is the most commonly used statistical test of whether *a pair of nominal (aka categorical) variables* are associated with one another, such that knowing something about a subject's category for one variable tells us something about what category they will belong to for the other variable. For example, we might want to know whether a person's preferred political party has a statistically significant association with that same individual's preferred color when given five options. Using crosstab results (or splitting the dataset by party and running a frequency on the color question), we could make a bar chart or frequency table to visually compare if there are differences in the share of Democrats, Republicans, independents, and third-party adherents who prefer each color. However, it is the chi-squared test results that will tell us whether there is a statistically meaningful association between the categorical variables that might not be found in repeated sampling (i.e., the probability of there being no consistent association is higher than .05). To calculated a *chi-squared test* in SPSS, we use Analyze -> Descriptive Statistics -> Crosstab (DV on the row) -> Statistics and check the appropriate test.

- What if you want to see whether belonging to a variable's category is statistically associated with a higher or lower average score on a dummy or interval variable when compared to the average score of another category of the same variable? An example of this would be if you want to know whether the typical level of education among Democrats is consistently higher/lower than Republicans.

  You can use an *independent sample t-test* to compare whether belonging to one or another of the groups measured by the same categorical variable is associated with the average value of an interval or dummy dependent variable. If you hypothesize that gender is a statistically significant predictor of differences in *income* (an interval variable) or identifying as *Democrat* (a dummy variable), you can answer this question with an independent samples t-test that compares two gender groups and tests for differences on the variable income and then Democrat.

- What method do you use if you want to know whether belonging to one group or another is associated with a different typical score, and the two groups you care about are measured by two different categorical variables? For example, we might want to know if the proportion of whites who identify as Republican is statistically different than the proportion of men who are Republican. Or we might want to know if identifying as racially white corresponds to a higher average household income than being a male.

  We can answer both questions if we split our dataset by gender and run separate *one-sample t-tests* on income and identifying a Republican, using white American's mean value for *Republican* and average income (that we have separately calculated) as the test values.

- When we want to examine the relationship between any pair of interval/dummy variables, *correlation* is the option most political scientists will choose, and it is the method discussed in your textbook. What does *Pearson's correlation (typically referred to as just "the correlation*" and denoted by the letter r) measure, and roughly what levels of a correlation statistic suggest a weak, moderate, or strong correlation?

- What else does a correlation statistic tell us? In addition to telling us (1) how much of an association there is, a correlation statistic is (2) signed to tell us which direction the association is (i.e., does an increase in one of the variables typically correspond to an increase or decrease (-.xxx) in the other. And (3) the squared value of a correlation can be interpreted as how much changes in one variable tell us about the value of the other variable. Finally, (4) a correlation statistic typically is reported with a p-

value or significance stars that tell us how probable it is that repeated sampling would find that the association is zero or runs in the other direction. If the p-value is higher than .05, there is not a statistically significant correlation.

- What does correlation NOT measure despite how the term is used in everyday language? A correlation statistic does *not* measure how much a change in the value of x will change y, on average. You need bivariate linear regression to do that. It also does *not* tell you whether x causes y, or if y causes x, or even if there is a causal relationship between the two. In short, only your theory or the temporal/logical ordering—rather than a statistical test—can help you to sort out if there is a causal relationship between two variables (for example, going to religious services more can't cause your gender).
- Sometimes, two variables have a correlation when one doesn't cause the other at all. What is a *spurious* correlation (hint: it is due to "omitted variable bias")? What, for example, explains the correlation between how much ice cream Americans eat in a month and crime?: https://www.foxweather.com/lifestyle/summer-crime-wave).
- What key assumptions about the nature of two variables' structure are made when one tests for a correlation between them? As a reminder, correlation works best with two continuous interval variables; however, social scientists widely use correlation analysis to analyze bounded variables (like dummy variables, percentages, Likert scale results, GPAs, >3-point indexes, 10-point scales, and so on). The only categorical variables suitable for correlation measures (including regression models) are dummy variables, which is why we code multi-category variables into a series of dummy variables before most statistical analyses. If we use correlation to analyze an ordinal variable, we are treating that variable as though it were an interval measure.
- What does correlation's (and regression's) *assumption of linearity* mean? What are some examples where two interval variables could have a logical, very consistent relationship and yet still have a very low or even correlation? You may find it helpful to think about the effect of (1) a higher income (the IV) on a person's happiness (DV), (2) the effect of each year of saving over time on the value of a person's investments, and (3) age on a person's level of physical independence. None of these relationships is linear because the effect of each additional unit of the IV doesn't consistently have the same influence on the dependent variable. However, all three situations clearly involve causal relationships from which we can make predictions (i.e., up to a certain point, making more money will provide you with more necessary resources and security, but at a certain point, earning more money has ittle influence on increased happiness).
- How can you deal with variables that have a non-linear relationship in statistical modeling (Hint: There are ways to quickly identify and address whether the relationship between an independent variable has a non-linear relationship with a dependent variable, but we will only consider—very briefly—how multiple dummy variables would allow you to analyze the three situations above). For example, we could code three dummies for age—1-20, 21-75, and older than 75. If we used a regression model to predict the level of a person's physical independence and used 21-75 as the reference category, we would expect to find that our dummy variables for young and old would both be significant and negatively signed. For what it's worth, these details won't be on any test you take, but I do want you to be aware of the fact that there are a variety of statistical techniques that we won't learn in this class that allow linear or logistic regression to work well when independent variables do not have a linear relationship with the dependent variable.

Statistical significance

- When we calculate statistics, our findings apply to the sample we are looking at. When we want to know whether the finding we are seeing would apply to the larger population, we need to look at the statistic's p-value. The p-value is the probability that the relationship we are seeing in the sample is non-existent or runs in the opposite direction. For example, if you flip a coin 10 times, you might see that 70% of your sample is heads. What is the probability that you would get more than 50% heads in repeated sampling? The p-value here will tell you that there is a very high probability that what you are seeing in

in your very small sample is due to chance.

- What is the difference between seeing a *statistically* significant association between two or more variables and reaching a *substantively* important finding? When are we most likely to find that a seemingly large and substantively important result in our sample is not statistically significant (answer: when you are looking at a sub-group for which there are few observations)?
- And when are you most likely to find a statistically significant finding that has little substantive value (answer: if you have a very large sample—the Cooperative Congressional Electoral Survey, for example, has tens of thousands of people in it—even a very slight difference in means or a tiny regression coefficient will generally be statistically significant)?
- T-test statistics, association measures like correlation, and regression coefficients are all reported with a corresponding *probability statistic* (i.e. a "*p-value*," which is often listed in SPSS output in the "sig." column). P values range from .000 to almost one. What exactly does the P-value measure for each type of statistic just listed? Hint: a one-sample T-test compares whether the means for two of a variable's sub-groups are the same for another variable (what's an example of that from your SPSS work?). A one-sample T-test can be used to compare a group or subgroup's mean to any value (what are some examples of test values against which we might test the mean of a group?).
- We interpret p-values to determine whether a finding we see for a sample is statistically significant. Sometimes SPSS labels p-values as "p-values" and sometimes, it labels them as "sig." Here are the kind of p-values we might see: 012, .453, .000, <.001, .002, .054, and .050. Hint: the easiest way to interpret a p-value is to think of it as a percentage. A p-value of .054 means there is over a five percent chance that a second, comparable sample would reach a finding that runs in the opposite direction from the one seen in the sample. So, let's say, we have a .101 p-value on a t-test that has found that women's average GPA is higher than men's. We would reject the hypothesis that women's grades are higher because over 10% of the time, we would expect that repeated sampling would find that men's grades are higher. A p-value of .001 means that there is less than .1% = $1/1000^{th}$ of a chance that a second, comparable sample would have a different finding.
- To make tables easier to read, it is a common practice to place significance stars next to statistics (e.g., a coefficient of .326***) to denote those that attain statistical significance instead of putting this information into an additional table common. Be able to interpret starred statistics assuming that * if p <.05, ** if p < .01 and ***if p < .001.

Regression with one independent and dependent variable

- What is the line of best fit (aka the regression line) measure for two variables, and how does that line relate to their correlation? Thinking about a scatterplot with a plotted regression line, which axis of the plot is the Y axis; which is the X axis? Specifically, by convention, which axis reports the value of the dependent variable? Answer: By convention, the dependent variable goes on the Y-Axis and the X-axis is where the independent variable is listed.
- *A bivariate linear regression* model calculates the *correlation* between a single independent and dependent variable AND provides other valuable information that is not available from just a correlation statistic. A regression model's results provide information about the correlation (both an *R* and *a R-square statistic*), the *Y-intercept* (aka, the model's "*constant*"), and the independent variable's *slope*. If we look at a regression table that includes the constant, how do we predict the value of a dependent variable for different values of the independent variable? The slope is reported as *an unstandardized regression coefficient*. What do all of these *italicized* terms mean? If you aren't sure, review textbook Chp. 9, the summary for linear regression in the schedule, and the handout of annotated SPSS output for regression.
- Assume we have a bivariate regression looking at the relationship between years of education (1-16 years) and income in $10k increments (1-15, with 15 representing >= $150k). The two variables are called Edu16 and Income15 The model includes years of education as the independent variable and

income as the dependent variable. The (adjusted) R-square statistic for the model is .54. The model's constant is -2.10, and the variable Edu16 has an unstandardized coefficient of 0.821. Its standardized coefficient (i.e., its Beta) is .425. The P-value for the Edu16 coefficient is .005. Be able to clearly explain these hypothetical findings for each of these underlined elements.

- What is a *scenario*, when talking about regression results? Selecting the relevant statistics from the hypothetical example just given, what would the predicted income be for someone with 6 years of education? (Hint: you can write an equation to calculate the first one and plug it into Google, which will solve it for you:
  -2.10 + (6*.821)
- If your regression model has only one dummy variable, what is the "reference" category? For example, if the variable Republican is in a regression model and there are no other partisan dummy variables, is the reference group Democrats or non-Republicans? Why are those not the same thing? Important to remember: the reference group is any group NOT in the regression model; if a Republican dummy variable is in your model, and it is the only partisan dummy variable, the reference group is non-Republicans, which includes Democrats and independents.

Multivariate (i.e., more than one independent variable) linear regression

- Assume that you also want to add multiple dummy categories for a single variable to the model you just examined. Specifically, you add two: the first indicates whether a person lives in the American South, and the second identifies respondents living in the West. What is the reference category for these two new variables? If the results returned statistically significant coefficients for West (say .492) and South (say -.103), how would you interpret these coefficients? Answer: The reference group for *both* of those new variables would be "people who don't live in the West or South."
- Staying with the same example used above (i.e., predicting a person's income bracket from their education level), assume that we decide to run a multivariate linear regression model by adding a second independent variable. The new variable, *Male*, is a dummy indicator. If the r-square statistic for the new model is .10 higher than our previous model's statistic, what does that tell us? It is possible that the coefficient for Edu16 would be smaller than before with the addition of a new variable? Answer, if a new variable adds valuable information, a regression model's adjusted r-square statistic will increase. When new variables are added, the value of the other variables' regression coefficients may increase or decrease as the model more accurately predicts the value of the outcome variable.
- What is *an interaction term*? Let's say our model of a person's income score (in the example, that's a 1-15 measure called *Income15*) includes *Edu16 Male* and *Edu16xMale*. Why would we put that last variable in the model (i.e., what is an example of a hypothesis that this interaction term can test)? What would it mean if our results showed it was statistically significant and positive? What if it was statistically significant and negative?
- What if the interaction term in the example wasn't statistically significant? What if it is statistically significant and its coefficient is signed negative or is positive? Answer: If the interaction term is not significant, the effect of education on income is the same for men and everyone else. If the interaction term is positive and significant, having more education increases the income of men more than everyone else. If the interaction term is negative, additional income has less of an effect for men than everyone else.
- Assume that the unstandardized coefficient for *Male* is 2.13 and the standardized coefficient is .214. In the revised model, the corresponding values for *Edu16* have been reduced to .623 (the unstandardized coefficient) and .340 (for the beta). The *constant* for the new model is: -1.015. Interpret the meaning of the unstandardized and standardized coefficients for the male dummy variable. What explains the reduction in the value of the standardized and beta coefficients for Edu16 in the new model? Looking at the standardized coefficients, which of the independent variables has more of an influence on *Income15*?
- Let's use those results to create *a predicted scenario for Income15 that includes a control.* If we know

that the mean for the variable Male in the sample is .521 (in other words, 52.5% of our sample is male), how would you calculate the estimated income bracket (i.e., a value somewhere between 1 and 10) for an individual with 3 years of education, controlling for the effect of gender by setting the value of that independent variable at its mean value? Answer: you would use this formula:
Constant + (x years of education times its regression coefficient) + (then mean for Male times its regression coefficient), which is:
-1.015 + (3x .623) + (.521 x 2.13)
If you plug that formula into Google, you will see that at those values, the expected income bracket (i.e. the value of *Income15*) is 1.96, or about $20-30K (remembering that a 2 *Income15* corresponds to that income range. If you do nothing but change the 3 years of education to 16, you see that a college-educated individual has an estimated income bracket score of 10.06, which corresponds mostly closely with $100-100K bracket for *Income15*.

- What is *multicollinearity*, and why is it sometimes a problem when using multivariate regression? Why would you, perhaps, need to be worried about multicollinearity if you had a regression model that tried to predict an outcome with the 7-point variable of party identification (typically coded so that 1=Strong Democrat thru 7=Strong Republican) and a 10-point measure of how conservative someone is? A very common use of correlation in social science research involves a "correlation matrix" measuring the correlation between each of the independent and control variables in multivariate regression models. How does a correlation matrix with all of the independent and control variables help us to identify and deal with collinearity that is so high that we may have a difficult time telling whether or how much a given independent variable is correlated with our dependent variable?

Logistic Regression

- What kind of regression do you use when we are dealing with a dependent variable that is a dummy variable? (*logistic regression* is the most common technique)
- Why is it often preferable to use logistic regression to analyze ordinal variables (especially Likert- scale items) that have been recoded into binary outcomes rather than using linear regression with the original variables? If you had a variable that originally asked respondents to rate the current US president's performance as: Very good, good, typical, bad, or very bad, what would be a logical way to code a dummy variable for his performance?
- Assume that you are using logistic regression to predict whether a person voted. Your model includes *Male* and *Education16* (as defined above). In your results, you have multiple pseudo R-square measures. Hint: You can report and interpret either the Nagelkerke's R-Square or the Cox & Snell R-Square (but not both), since these measures are both widely used. What does this statistic tell you and how do you interpret it?
- Continuing with our logistic example, why will you want to report data from the SPSS output column that lists odds ratios (i.e., the statistics in the column exp(B)) rather than the unstandardized coefficients that are so useful with linear regression? (Note, you will see tables in published articles that report b instead of exp(b), but this is usually the case only if the body of the paper analyzes odds ratios or predicted probabilities similar to what you see in the sample Setzler and Yanus paper on characteristics that predict attitudes about gender equality.
- Assume that our logistic model predicting voting finds that *Male* has an exp(B) of .875 and *Education16* has an exp(B) of 1.25. The Wald score for *Male* is 4.2 and for the education measure, it is 7.2. Interpret these odds ratios and explain why we know education is the more powerful predictor.
- We probably will learn how to predict scenarios from logistic regression models. Although it requires a little more work, scenarios are by far the best way to explain logistic regression results to an audience that does not regularly work with statistics. Be able to describe in very basic terms how we predict a logistic regression scenario and provide examples.