***Downloading a Pew Dataset and Using SPSS for Analysis: Part Three***

**ANALYZING DATA IN SPSS: UNIVARIATE ANALYSIS**

Once you have recoded all your variables, you will want to begin doing some basic univariate analyses of your data. There are three main purposes for univariate analysis:

* The first purpose is to see whether you have made any obvious mistakes while coding your data. Creating a “descriptive statistics” table that lists the sample size, mean, and minimum and maximum values of all of your variables will help you to see if any variable has a large number of missing observations or some types of coding errors (look to see if the minimum, maximum, and means look right).
* Univariate statistics also summarize the “central tendency” of a study’s independent and dependent variables. *Means, medians,* and *modes* are measures of central tendency. When you want to talk about the typical respondent, you usually will cite either a mean or a median.
* Univariate stats also are used to show how a given quality is distributed, e.g., we could look at the standard deviation associated with our survey’s mean income statistic to see whether the incomes of most people tend to cluster around the average income for the sample or if incomes are more broadly distributed. We also could look at a statistic of skewness to see if the distribution deviates from a bell curve (e.g. income is skewed towards…e.g. the sharp tail points at…the wealthy in that lots of people are quite poor and comparatively few people are rich). Since standard deviation statistics and measurements of skewness are difficult and aren’t something that most people are familiar with, we might choose to instead summarize the distribution and skew of income in our sample by reporting means for people at each quartile, quintile, or decile. Alternatively, we might report frequencies in a table or figure showing the percentage of individuals at each level of income.

**Creating a codebook and a descriptive statistics table.** Assuming you have already coded all of your variables properly (as explained in the earlier parts of these handouts), you can use SPSS to generate a descriptive statistics table that will give you a big-picture look at all of the variables you are working with. You also should create a codebook where you will see the frequency distribution of each variable and the number with which it is coded:

1. If you are running univariate or other statistics on a variable for the first time, take a few seconds first to run CODEBOOK command to examine the variables you are using as well as any original variables you may have used to create them. You need to verify that the coding looks right. In SPSS’s click-and-point interface, you create a codebook by: **Analyze -> Reports -> Codebook.**Any time, you are working in syntax, you can quickly verify one or more variables’ coding and distribution with:
**CODEBOOK var1 var2.**
2. Running a descriptives statistics table any time you are working with a set of variables—whether the original dataset version or ones you have (re)coded–is a must! You want to look carefully at the results table to see if any variable has an unusually large number of missing observations. This will alert you if there is something miscoded or when a survey question may have only been asked to part of the sample.

To generate a descriptives statistics table for one or more variables In SPSS, use: **Analyze -> Descriptive Statistics ->Descriptives.** Then, select the all of variables you want and use the checkboxes in the "options" window to generate results for only the mean, standard deviation, min. value, and max value. These four statistics generally are reported in the order just listed in a table with each variable in a separate row and the four statistics each in a separate column. It is very common in social science research to include a table with all of a study’s variables at the end of the methods section where the author has just described how each of the study’s variables is coded.
3. For statistical research papers and theses, if you have categorical (aka, nominal) variables (for example, a variable coded 1=Protestant, 2=Catholic, etc), you should create a dummy variable for each value of these variables before putting them in your descriptives statistics table. The mean or standard deviation for a 7-category religious denomination variable (i.e., 1=Catholic, 2= mainline Protestant, etc.) is meaningless, but descriptives statistics for each of the individual dummy variables will tell you what percentage of the sample is identified in each of the seven categories. For example, if you have a variable coded 1=Protestant, 0=other denomination or no faith, a mean of .326 indicates that 32.6% of the sample is Protestant.

**Some helpful suggestions as you begin to generate descriptives and other kinds of statistics**

* **When using SPSS’s point-and-click interface, you can select multiple variables at a time to add or remove from the list of variables** **to be analyzed.** If you are using a PC, you just need to hold the control button down as you select each variable. And you can select a continuous range of variables by holding the shift key down as you select variables. Mac users, depending on your trackpad settings, usually need to press the control button in combination with the shift key to select a range of variables.
* To make it faster to find the specific variables you want to analyze, **when you are using SPSS’s point-and-click interface to run statistics, you can change the view of the variable list so that you see the short variable names rather than labels. You can also sort labels to alphabetically.** To do so with a PC, hover over the variable list, use a right-hand mouse click, and select the option to see the variables by their names. Repeat this step to order them alphabetically. For a Mac, depending on your trackpad settings, what is a right-hand mouse click on a PC typically involves pressing the control button while clicking your trackpad.
* I**t is so fast and easy to sort and find variables in the descriptives statistics window that this usually is the best way to create a list of your study's variables any time you need one**. Just point-click-and-the-paste a descriptives command. In syntax, you can move the variables around if you think you are going to want them in a different order.

**Generating frequencies tables/figures or calculating central measures of tendency not available in the Descriptives output.** To calculate measures of central tendency other than means or if you want to create a table or figure showing the percentage of respondents in each category across a particular variable (this is commonly done with a key independent variable or dependent variables), do the following:

* + - 1. Staying on your syntax page, select **“Analyze” 🡪“Descriptive Statistics” 🡪 “Frequencies”**.
			2. **Select the variable (or several if you want) that you are interested in** analyzing and click the 🡪 button.
			3. To find out additional information on this variable**, select “statistics”** and then click on “mean,” median,” or mode depending on what makes the most sense for what you want to know. **Click “continue”** when finished in that dialogue box.
			4. You may want to **look at your data in a frequency figure**; if so, **select “charts”** and your preference of chart type, then click “continue.” If you are dealing with a variable that has nominal categories, bar charts are usually the easiest to read. For variables that are continuous and have lots of values (age in years or raw income for example), a line chart or histogram is your best option. Whatever chart you do, make sure to check the options to **display your data in percentages rather than counts** (e.g., if comparing men and women, you want to see percentages rather than the number of respondents in each category). Any time you intend to use an SPSS figure in any type of paper or presentation, you need to have titles in the appropriate places so that your work makes sense to the reader.

			Pro-tip: While you can edit SPSS figures, you will save yourself a lot of time and be much less likely to inaccurately describe something if you have properly **labeled all of your dataset’s variables and the value labels corresponding to each response category.** Assuming that you have done your coding in syntax (see the other handouts in this series), you can always modestly amend your syntax to improve the labels if you find that they are clunky in the figures you are creating). To change a variable’s label, just rerun the command in your syntax. If the variable Liberal10 was labeled “How liberal respondent is 10-point,” and you want a shorter label in your SPSS output, you might just run this line of syntax:
			VARIABLE LABELS Liberal10 “Liberal, 10 pts”.

			Another pro-tip: As will be explained below, **bar charts look a lot better when created in Excel, and more often than not, you will want to merge multiple frequency figures into one rather than giving presentations that have multiple SPSS output figures** (e.g., you can easily combine data about men and women or younger, middle-aged, and older Americans into a single figure). This should be done with Excel.
			5. Once you have finished selecting your various options for frequencies, hit the OK button to run your statistics. If you want to see the SPSS coding for these procedures, once again click “paste” and this will paste the command code in the syntax file. Highlight the command code and once again select “Run” and “Selection”…and presto…you will have your chart and frequencies. In all cases, carefully at the results in your output file to make sure that something doesn’t look strange as a consequence of a variable coding mistake.

**ANALYZING DATA IN SPSS: BIVARIATE ANALYSES**
Before social scientists start to look at the effects of multiple variables at a time in their analyses, they usually start by seeing if there is a statistically significant relationship between the study’s independent and dependent variables**.**

**Generating means or frequencies of different sub-groups.** The fastest way to take a quick look at whether different values of an independent variable may influence the value of a dependent variable is to “split” the independent variable and run descriptive statistics on your dependent variable.

For example, let’s say you have the variable Race5 (Af Amer, white, Latino, AAIP, and other race/multiple). If you think that race has a relationship to who identifies as a Democrat (assume that you have a dummy variable coded 1=Democrat, 0=Not), you could **produce** **univariate stats for each of the Race5 subgroups**.

To do so, you can use the command **Data🡪Split File and check the “compare groups” option.** If you split your dataset on Race5, you could quickly create a descriptives statistics table listing the means for the variable Democrat for the five groups. If you wanted to generate and compare descriptive statistics for the variable Democrat for males and non-males for each race, you could split the file by both of the relevant variables in the “compare groups” window.

**After you are done with any subgroup analyses, you need to navigate back to the split command to tell SPSS to resume analyzing “all cases.”**

Keep in mind that this method is used to visually show bivariate associations; you will not be “testing” any relationship between two or more variables if you do this kind of work because you will not be using a statistical test to determine whether the differences you visually observe in your sample will be found in repeated sampling.

If you are thinking about using the split file option for the purposes of creating frequency tables or a figure listing several different groups, keep in mind that you will end up with a separate table or figure for each group if you try to use SPSS’s chart’s setting. Your work will look much better in a presentation or paper if you combine frequencies/discriptives results for the multiple groups into a single table or figure in Excel. I explain how to do this in this screencast: <https://www.youtube.com/watch?v=T6kHpZ2oReQ>.

**Comparing the means of different groups to see if they really are different (i.e., if what you see in your sample would be predictably replicated in the larger population): T-tests.**

**Note, senior seminar theses rarely use T-tests because these projects typically use multiple regression even when examining the relationship between two variables because multiple regression allows a researcher to isolate a bivariate relationship from the effects of other variables by employing controls. If you are taking PSC 4099, check in with your instructor before you read this section on T-tests.**

T-tests frequently appear in social science research. Sometimes, we want to see if different groups have statistically different means (e.g., Do Republicans, on average, make more or less money than non-Republicans? Do they differ for their average level of agreement with the statement: you can trust the government most of the time?).

If you are comparing groups to see if there are *statistically* significant differences among them, you can’t just split your dataset and run descriptive statistics because that approach will not test whether the means for your subgroups are statistically different. In other words, you don’t know how likely it is that what you are finding in your sample would be found in repeated sampling of the larger population. For this, you need to use a T-Test

1. If you are looking at whether **two groups coded on the same independent variable have statistically different means for a second variable**, you will **use the Independent Samples T test**: Analyze🡪Compare Means🡪Ind. Samples T Test. This is an appropriate statistical test, for example, if you want to compare the average incomes for men and women using a gender variable that is coded 0/1.

When running this test, the dichotomous interval, ordinal (if you are considering it to be an interval variable for this test) outcome variable you want to examine goes into the window marked “dependent variable.” If your categorical independent variable happens to contain three or more subgroups (e.g., Democrats =1, Republicans=2, and others=3), you can compare the means for just two of them at a time. To do so, you need to use the “Define groups” option to indicate which two values of the variable should be compared (e.g., groups 1 and 2 if you want to compare the means for Dems and Repubs, where they are coded as noted above).
2. We also use T-Tests to test whether the mean of a variable is equal to a specified value.This type of test, **a single sample T-Test, is what you use to test whether the mean value for a group on a dependent variable is different than the mean value for another group (or groups) that are measured by other variables.**

For example, let’s say you have a survey question asking people to place themselves on a scale where 1 indicates that respondents said God is "not at all important" in their life and a 10 indicates they believe God is very important. In a recent sample using this question, the typical American (i.e. the survey mean) had a response value of 7.17 on a 10-point scale. Is the sample's mean value for this dependent variable different than the mean for Democrats? How about Republicans? How about men? How about women?

We can answer these questions if we split our data on the relevant independent variables and then run a single sample T-Test (Analyze-> Compare Means -> One Sample T-test). For the tests, we would enter the average for the sample as a whole, 7.17, into the place in SPSS that asks for the "Test value." In the survey I mentioned above, the mean score for Democrats on the 10-point scale of how important God is in their life was 6.60, and the T-Tests **two-sided significance test** (this is the one you want to use) says that the average for Democrats was different than the test value with a significance statistic of .001 level. We are almost certain that we would find in repeated samples that the mean for Democrats is going to be lower than the national average for this variable. Incidentally, using the same data and test shows that the average Republican value for the importance-of-God measures is 8.25, which is significant at the <.000 level. To see if men or women also are different from the national average, we would need to go back to Data-> split file -> compare groups and swap out party with a gender variable. From there, we would once again run a single sample t-test and use the test value of 7.17.

**Correlation.** This is the most commonly reported measure of association used in political science work because multivariate regression analysis builds on it. It can be used to explore relationships among any two variables as long as they have been previously coded into either dummy (that is 1/0 dichotomous) or interval variables. When ordinal variables have five or more categories—as is almost always the case with Likert scales—it is very common to treat these variables as though they are interval variables. It also is very common to recode multi-category variables into one or more dummy variables.

Running a correlation matrix—where you enter all of your independent variables at a time—is also an important step if you going to do multivariate regression analysis (discussed in the next section). When planning to use multivariate regression, you need to create a correlation matrix to see how much each of your independent variables is correlated with each of the others so that you know if you will need to combine or drop any independent variables for your analysis due to ***multicollinearity.***

What’s multicollinearity? Imagine that you think that one of two men who are close friends is committing a certain type of pretty unique crime in the evenings. You are sure that your suspect is not the only one committing this type of crime because sometimes there have been incidents when neither your suspect nor his friend has been present. Your suspect and his friend hang out together a lot, and when they've been in an area at night, there frequently is evidence the next day that the criminal act has occurred. You would like to pin the blame on your main suspect, but you have a problem. While you believe that your main suspect has been doing the all of crimes and his friend hasn’t committing any of them, it's also possible that both men have been doing these crimes, taking turns. Or they both could be doing the crimes together. Or, it may be possible that you've identified the wrong guy entirely and it's the second man doing all of the crime that happens when these two men are together. To reliably test your hypothesis that it is just your suspect committing crimes, you ideally will have lots of instances where only one of the guys was in an area on a given night, so you can see if a crime happened that night. However, if they are such good friends that you don't have very many of these observations, you may have to determine who is guilty by using just a few observations (something we try to avoid with statistics) or by splitting responsibility and assigning the crime to both of them equally whenever both were present (which could be flat out wrong). That’s essentially what multicollinearity is, and it is why we look at how correlated our independent variables are with each other. If we find that an increase in one of our variables consistently predicts a change in the other, we will need to rethink how we approach our study.

Beforeyou think about using any of the techniques described below, the first quick step in bivariate analysis usually is to take a look at a correlation matrix that includes all of the variables you are interested in. To do this: **Analyze🡪correlation🡪bivariate correlation**. Check just the option to return just **the Pearson’s correlation coefficient** and then go to the left window and double-click on every variable in which you are interested—including both dependent and independent variables. This will give you a quick view of which variables are probably significantly correlated (i.e., those starred by SPSS or that carry a significance statistic that is <.05). The Pearson’s measurement also gives you some idea of how strong a negative or positive relationship is. If the coefficient is really small and insignificant, it is unlikely that the more precise tests described below will add much information to your story.

**As a very general guideline for thinking about correlation values (and other association statistics), <.10 means that there is a very weak or no association between the variables; .20 can be interpreted as a meaningful but modest association; .30 is a moderate association, and >.40 is a strong association, although in every case, you need to put these findings into context** (e.g., a .40 association between being Republican and being conservative would be a much weaker finding than you would anticipate, so it wouldn't make sense to refer to this scenario as being evidence of a very strong association). Most of the association statistics range from -1 to 1, and any negative statistic means that increased values in one variable is associated with declines in the value of the other variable.

Is correlation the only way to determine whether two variables are statistically associated with one another? No, but it is the technique most often used in political science, which frequently treats ordinal variables as though they were interval measures when running descriptive statistics. Similarly, as noted above, most political science research converts nominal variables into a series of dummy variables, which can be analyzed with correlation analysis.

When you are looking at the association between two variables that are structured as ordinal or categorical measures, there are measures of association specifically designed to address these variables. To access these tests, you need to use the **crosstabs** command: Analyze🡪 descriptive statistics🡪 cross-tabs. SPSS can guide you to the most appropriate tests for your specific variable type. **Typically, we do not review association tests other than correlation in class (and you will not be expected to be familiar with them) because they are not frequently used in political science.**

**What exactly do we mean by a statistically significant “relationship” (aka “association”) between two variables?** Depending on the specific “test of association” you use, you will be using SPSS to tell you up to five things:

* First, you can run some type of procedure **to learn if two or more variables in your dataset are associated**, which is to say tend to vary together, as if they were causally connected in some way. Most association tests run from 0 (no association) to +/- 1 (strongest positive/negative relationship).
* Second, if the association test coefficient for a statistical method is not zero, **you want to know if this seeming association is due purely to chance.** The standard in political science is that for you to say that something is associated or correlated with something else, you want to be confident that if you were to redo a survey 100 times, with similar samples drawn from the same larger population at the same time, would you find at least 95% of the time that these two variables vary together and in the same direction indicated by the test coefficient. If there is a 5% chance or greater that the association we see in a sample could be zero or in the opposite direction, it is the norm in the social sciences to say that there is no statistically significant relationship.

When you look at social science journals and books, you often will see that the statics reported in tables have one, two or three astericks next to them. These refer to the probability (\* p=.05, \*\* p=.01, and \*\*\* p= .001) that the result being reported could be 0 or in the other direction in repeated sampling. An association statistic of +.21\*\*\* between two variables, for example, means that we think there is less than a one-in-one thousand chance that repeated sampling would uncover an example where the two variables were negatively or not at all associated.
* Third, you want to know **how “strong” (aka consistent) any association is**. That is, how consistently do the two variables co-vary? It is really important to note: These association tests only tell us how often an increase in one of the variables corresponds to an increase/decrease in the other, not how much a change in one changes the other. Thus, a correlation statistic telling us that higher levels of exercise are strongly correlated (say +.80) with more years of life, would tell us only that people who exercise more consistently live longer and not how much longer.
* Fourth, if the two variables have some internal order (e.g., the range in value from less to more), you want to know **“the direction of the relationship”** by which we mean whether an increase in the value of one of the variables appears to co-vary positively or negatively with the value of the other variable. A positive association indicates that a higher level on one variable typically corresponds to a higher level on the other. A negative association shows that a higher level for one variable usually corresponds to a lower value for the other.
* Finally, some association tests are designed to measure the association between two variables, assuming that one of them is the cause (independent variable) and the other is the outcome (effect). We will not look at any of these tests closely.

**ANALYZING DATA IN SPSS: MULTIVARIATE ANALYSIS**

Many of you will not be satisfied with using your research project to explore how just one or two variables influence a given outcome. If you need to know how much each of several variables influences an outcome, which variables are most influential, and how much an outcome your variables as a collective can explain, you will want to use multivariate regression analysis.

**Why Regression?** Have you ever wondered why most research published or widely publicized in political science, natural science, and medical journals uses “regression” models even though most Americans have no idea what a regression model is or how to interpret its output? Most social science research today uses somewhat more sophisticated methods than crosstabs and association measures for four reasons:

* First, life is complicated, and explaining human attitudes and behaviors is especially so. To figure out what is going on, we often need to consider how changes in each one of many different independent variables will influence an outcome in which we are interested. Controlling for more variables makes it clearer how each of our variables of interest *independently* shapes the outcome we are studying. So, rather than looking at how race impacts voting, how gender impacts voting, how education impacts voting, etc., social scientists frequently want to know how differences in race, gender, and education all interact to simultaneously influence voting behavior.
* Second “regression” analytical techniques also allow us to answer the question of whether there is a relationship between an independent variable of special interest and our dependent variable/s, once other relevant factors are taken into account**.** Sometimes, it looks like there is a relationship, when there isn’t much of one. For example, it typically looks like there is a relationship between gender and support for Democratic presidents, with women being more supportive. However, women are also more likely to be Democrats. Regression analysis shows us that partisianship explains support for Democratic presidents rather than the “spurious” relationship between gender and support.
* Third, regression models tell us how much variables collectively explain an outcome and how much different variables contribute to that outcome.
* While bivarariate correlation analysis tells us how consistently a variable influences an outcome, regression (even bivariate regression) tells us how much and how consistently variables influence an outcome.

As an example of how we can use regression models to answer a question that would be impossible to get at with bivariate methods, consider a study that wants to know if Mexican Americans vote at the same rate as other Americans. Let’s assume that previous research has suggested that 1) Mexican Americans vote in presidential elections at lower rates when we just compare mean rates of turnout, 2) Mexican Americans, on average, have lower socioeconomic indicators, and 3) people with lower socioeconomic indicators vote at lower rates. Thus our research might want to see whether Mexicans still vote at a lower rate than other Americans once we “take into account,” that is “control” for various indicators of socio-economics that impact voting. In other words, we want to know whether there is something about being a Mexican American that leads to lower voting rather than other qualities that disproportionately apply to this group. To answer this question, we can use a statistical program like SPSS to work out all of the permutations necessary to compare the voter turnout of Mexicans and other Americans across varying levels of income, educational attainment, occupational status, etc. Our results might allow us to say something like, “While the odds that the typical Mexican-American will vote in a presidential election is approximately half those of other Americans, once the effects on voting of the socio-economic differences between the two groups are taken into account, the odds that a non-Mexican will vote in a presidential election is only 20 percent higher than the odds for a Mexican-American.”

**Linear Regression (aka OLS).** The vast majority of research in political science reports the findings of either linear regression or some variation of logistic regression. Linear regressionis an extension of correlation analysis and is appropriate when you are looking at an outcome that is measured as an interval variable. This could be the right choice, for example, if we wanted to see how education, gender, age, and religiosity influence a person’s support for torture and our survey has a 7-point aggregate measure of support for torture based on responses to 7 yes-no questions (e.g., to get information that might prevent a future crime). Here is how we calculate and interpret OLS regression models in SPSS:

**(1) How to estimate a linear regression model.** When you have an interval-dependent variable and want to use regression to predict how changes in an independent variable increase or decrease the value of that outcome variable, you use this method:
**SPSS: Analyze -> Regression ->Linear.** You will need to identify your dependent variables and independent variables.

**(2) Really important: If your model is going to include dummy independent variables, carefully think through what your reference categories will be as you get ready to run the model!** If you put in a dummy variable for males, the reference category in your output statistics will be non-males. If you put in dummy variables for both Republicans and Democrats, your reference category will be people who are neither Republican nor Democrat, which may not be the comparison you are looking for if you have hypothesized that Republicans will be the most supportive of torture. If you are making that hypothesis, your model should include dummy variables for Democrats and independents so that your results will show you how different each of these groups is from the reference category (Republicans) and whether that difference is statistically significant.

Once you have your results, you will want to focus on two sections of the output:

**(1) First, use your SPSS output to figure out how well the variables in the model collectively explain variation in the dependent variable**. To assess a model's "fit,” use SPSS’s adjusted R-squared statistic from the output. An example of interpretation: If a model's adjusted R-square is .365 and we are predicting how much a person supports torturing terrorism suspects on a 1-7 point measurement, we would way that, "the variables in this model collectively account for about 37 percent of the variation in how much people support torture." Alternatively, depending on what hypotheses we are testing, we might interpret the same statistic by saying, "Most of what predicts the level of support for torture--that is nearly two-thirds of the explanation--for the dependent variable is due to factors not considered by this model" [1 - .365 = .635].

**(2) Then, go to the coefficients table to look at how increases in each independent variable change the estimated value of the dependent variable when all other variables are held at a constant influence.**As part of this step, we first need to verify whether each coefficient showing an effect is significantly different from zero. The model's unstandardized coefficients are listed in your output’s "coefficients" table, specifically in the column labeled "B." For the dependent variable, the column “Sig,” lists the probability that a given coefficient is not different from zero or that the relationship is signed in the wrong direction. By convention, we want the probability (aka, “the p-value”) of the coefficient to be meaningless or wrong to be less than .05. As explained earlier in the this handout, if a variable’s coefficient is greater than .05, we will not interpret that statistic because we aren’t sure if there is a real effect on the dependent variable.

Next, we will look at the unstandardized coefficients (they are listed in the “B” column), so that we can interpret each of those that are statistically significant. Say we are trying to explain what variables predict a person's level of support for torturing terrorism suspects on the seven-point measure. If we had unstandardized coefficients of .451\*\* for our dummy variable Republican (assuming it is the only partisan variable in the model) and -.134\* for edu4 (measured as a four-unit, interval variable), a suitable interpretation of these results would be:

"Compared to other non-Republicans, Republicans' level of support for torture was around a half of a point higher on the seven-point scale, controlling for the influence of other variables. On the other hand, each increase in education modestly decreased support for torture. Respondents with the highest levels of education had around a half a point lower level of support when compared to the least educated" (i.e., each level of education reduced a person's score by -.137; thus going up three units--from 1 to 4--is equal to 3 x -.134 = -.402). **Remember, we need to look carefully at what the reference category is when using dummy variables. If our model had included dummies for both Republicans and Democrats, the comparison here would be to independents, since they were the only partisan group not in the model.**

**(4) Next, we will want to consider how the predictors rank in their influence on the independent variable.** In the same table containing the unstandardized coefficients, the "Beta" column lists the model's "standardized" coefficients. These each measure how many standard deviations the dependent variable increases or decreases with each one-standard-deviation unit increase in the applicable independent variable (e.g., going from having an average level of education to the 84th percentile). Most often, these statistics will be less than one, suggesting that one standard unit increase in a given variable corresponds to less than a one-standard-deviation increase/decrease in the dependent variable's value.

The main use for "betas" is that they help us to figure out which variables have the most powerful influence on the value of the dependent variable. For example, when predicting support for torture--as the model does in the screencast--a one-standard-deviation increase in racism causes a much larger increase in support than a one-standard-deviation increase in religiosity in education. In interpreting a model that had these models, we might say, "The standardized coefficients for the model indicate that the most powerful predictor of how much a person supports torturing terrorists is that individual's level of racial animosity."

 **(6) Finally, we can make our results come alive by using scenarios.** For linear regression, it is very straightforward to calculate specific scenarios from SPSS output. For example, if we want to estimate the degree to which a 30-year old female Democrat who is a college graduate thinks torturing terrorism suspects is necessary (i.e. what her score would be on a hypothetical 7-point index that assesses how justified someone thinks torture is for suspects of terrorism), we would do the following:

(1) Take the constant value reported in the regression model (i.e., what a person’s expected score on the index is if all of the independent variables are at zero, even if that isn’t a possibility given how they are coded),

(2) and then add the influence of gender (if you have the variable "male" in model, this value is 0),

(3) and then add the effect of the dummy CollegeGrad variable times its unstandardized regression coefficient,

(4) and finally, add the effect of age (if AgeInYears is your variable, this would be 30 times the value of unstandardized coefficient reported for variable AgeInYears).

You can do this kind of math very quickly in an Excel spreadsheet and then perhaps compare the openess-to-torture score for our 30-year old female Democrat who is a college graduate against the expected score of a 55-year old male Republican without a college degree.

**Logistic regression**. Most HPU students doing statistical projects in our PSC classes will want to use logistic regression. If you are fortunate and your dependent variable of interest is an interval variable with a wide range of values, you will be able to use the most easy-to-interpret type of regression: ordinary least squares (aka linear regression). If, however, you have a dependent variable that takes the form of a count (e.g., how many times have you voted in the last year?) or an ordinal ranking (e.g., do you support the torture of suspected terrorists “often,” “some of the time,” “rarely,” or “never”?), the interpretation of the best type regression model for these kinds of variables is sufficiently complicated that we will ask you to convert your dependent variable into a dichotomous variable so that you can use logistic regression.

In the example below I discuss the results from a model used in a paper I co-authored about who intended to vote for Donald Trump a month out of the 2016 Presidential election:

**(1) How to run this type of regression.** When you have a *dichotomous* dependent variable and want to use regression to predict how changes in an independent variable increases/decreases the likelihood of an outcome, you use this method:
**SPSS: Analyze -> Regression ->Binary Logistic**

You enter variables for this type of regression just like you do with OLS regression, although the look of the window is a little different. See my note of caution above about being very careful to **think about what the reference categories will be for each dummy variable you include in the model and whether you are making the best choice for a reference category to test your hypotheses**.

**(2) In the output, you will want to ignore most of it, going straight to the “pseudo” R-squared statistics first. These will allow you to figure out how well the variables in the model as a whole explain the "likelihood" of the outcome you are predicting**. Report just one of the two R-squared statistics that SPSS lists in model results. Researchers have not yet settled on one measure as being the best, but the Nagelkerke pseudo-R-square is the most like OLS regression’s R-square. Note that researchers typically label this statistic "pseudo-R-square" when reporting the results of logistic regression because it is not actually the mathematical square of a Pearson's R statistic the way it is for linear regression. Here's an example of how to interpret a pseudo-R-square:

Let's say the model of who intended to vote for Donald Trump (yes or no), gave us SPSS results including output with a Nagelkerke R-square of .142. In writing up our findings, we might say:

"The model's pseudo-R-square (Nagelkerke) indicates that the predictors [i.e., the independent variables] in the model collectively account for just over 14% of the variation in whether or not someone intended to vote for then-candidate Donald Trump." Another way to interpret the same results would be: "While numerous studies have suggested that each of the characteristics in the model were important predictors of who voted for Trump in 2016, the model's R-square (Nagelkerke) statistic indicates that over 85% [i.e., 1.0 - .142  is > .15] of the factors that led only some individuals to vote for Trump lie beyond the indicators examined here."

**(3) Then look at the odds ratios to determine how much a one-unit increase in each independent variable changes the "likelihood" of the outcome you are predicting. when all other variables are held at a constant influence.** Odds ratios are listed in the Ex(B) column of our SPSS output. An odds ratio of higher than 1.0 indicates that increases in the value of that independent variable increase the likelihood of the outcome predicted by the model.

Here is an example of three factors that were significantly correlated with supporting Donald Trump in 2016; however, I am making up their coefficients to review how odds ratios falling into three different value ranges are interpreted:  Let's say we were predicting whether someone intended to vote for Trump and our results had an odds ratio [Ex(B)] of 12.321 for a dummy independent variable identifying *Republican* respondents (assuming it is the only partisanship variable in the model). For our second variable, our output reports an odds ratio of  1.137 for *edu4*, which is a four-point measure of educational attainment (Here, we assume that it is an interval variable; if we didn't want to assume that each one-unit increase in education will the same effect on the dependent variable, we would need to recode that variable into a series of dummy indicators, say "more than high school," "college degree or more," and "advanced degree," leaving "high school or less" as our reference category). Finally, we have a third variable--*Torture7--*created from a 7-point item asking how much respondents agreed with a statement that "torture should be used to obtain information from suspected terrorists. For this variable, our output reports an odds-ratio of  .911.

To recap, a truncated version of our output for this example would read:

Indep Variables ***Ex(B) Note that this is not the first column in SPSS output!***
Republican              12.321
Edu4                         1.137
Torture 7                  .911

A suitable interpretation of our odds-ratio results would be:

"Compared to all other respondents, the typical Republican was over 12 times as likely to vote for Trump, after taking the influence of other variables into account. Moreover, each increase in education on a four-point scale increased the likelihood of voting for Trump by about 14% (i.e., 1.137 - 1.0 = 13.7). Conversely, being more supportive of torture made a person less likely to support Trump. On its seven-point scale, each additional level of support for using torture corresponded to a 9 percent decrease in the likelihood of voting for Trump (i.e., 1 - .911 = .089)."

In this example, I am showing the very basic math involved in interpreting odds ratios with a value of 1-to-2 (a positive relationship that effect that can be stated as an "x percentage increase in the likelihood) and for odds ratios between 0-1, which indicate a negative effect. Normally, you do not show this math.

**(4) Remember: odds ratios that are lower than 1.0 indicate that increases in the value of an independent variable decrease the likelihood of the outcome.** These results generally are reported as percentages, which often are calculated by subtracting the odds ratio from one). An example: Assume we are predicting whether someone intends to vote for Trump and our results have odds ratios of .101 for the dummy variable *Democrat* (assuming it is the only partisanship variable in the model) and .913 for edu4 (as defined above). A suitable interpretation of these results would be:

"Compared to other respondents, Democrats were about 90% less likely to vote for Trump [i.e., 1.0 - .101= .899] after taking the influence of other variables into account. Increases in education also decreased the likelihood of supporting him. Respondents with the highest level of education were approximately 24% less likely to vote for Trump than persons with only a high school degree or less schooling.”

For the *Democrat* variable, another way to say the same thing, but use the odds ratio of .101 differently would be to say:

"Compared to other respondents, Democrats were approximately 90 percent less likely to vote for Donald Trump, controlling for the other predictors in the model."

To calculate the statistic we make in the statement about education’s influence, we first need to figure out what the first unit of change does (going from Edu4 = 1 to Edu4 = 2) and then compound that effect to take into account the two additional units of change that distinguish people with the lowest and highest levels of education. In other words, we need to solve this problem: (1-.913 =.087) to the third power. If you aren't math-inclined, Google will do this math for you, just do this search: "(1-.087) to the third power". The answer is 76.10. So, 100 percent - 76 percent = 24 percent, which is the statistic noted above.

How do we get to this statistic again? Going from 1 to 2 units of edu4 makes a respondent 91.3% as likely to vote for Trump. Then, those with a score of 3 on edu4 are just 91.3% as likely to vote for Trump as those with a score of 2 (i.e. .913 x .913 = .834). Finally, those with a score of 4 on edu4 are just 91.3 percent as individuals with a score of 3: 91.3% of 83.4% as likely (i.e.: .913 x .834 = .761). Of course, this is the same result that you get by calculating (1-.087)^3.

**(5) If appropriate for your hypotheses, consider how the predictors rank in how consistent their influence is on the independent variable.** And here's something that is not covered in the video but may be useful to think about. There is some debate about how to best rank the relative influence of each variable in a logistic regression model, but one common approach is to compare their "Wald scores" much like we do with "betas" for linear regression. For logistic regression, this approach tells us which variables most consistently predict the outcome, but not the magnitude of their influence, the way a standardized coefficient does in linear regression.

**(6) As with linear regression, interpreting logistic regression analysis in an interesting way is best done with scenarios.** For linear regression, creating scenarios is a very straightforward process because it is easy to use Excel to calculate specific scenarios from SPSS output (see the example above).

Unfortunately, it is more challenging to create scenarios with logistic regression. With logistic regression, scenarios are reported as the predicted probability of doing or believing something at a given value of the independent variable, with the effect of all other variables held constant at their mean. A in-plain-English example of this can be seen in thinking about the number of injuries to key players on a basketball team and its likelihood of winning a game, controlling for whatever effect playing at home or away may have. Let's assume that having one, or two, or even three injuries may modestly impact whether the team will lose. On the other hand, having a lot of injuries virtually ensures they will lose, even if they sustain additional injuries beyond what they already have. But, there also is a middle area where each additional injury will have a large effect on their probability of winning. In short, the effect of each additional injury on winning is not always the same. This is why logistic regression coefficients are reported as odds ratios. If a team has lots of injuries and virtually no chance of winning, what does it matter that having one less injury would double its odds of winning (perhaps going from 1 to 2 percent). At the other end of the injury count, if a team with one injury already has a 90% chance of winning, doubling its odds of winning by reducing its injuries to none doesn’t actually do much.

In order to address the fact that each additional injury does not have the same effect on the dependent variable (as it would in linear regression), we need to do some additional math before we can create scenarios similar to those I describe above for linear regression. For example, if we want to predict the probability that a 30-year old female Republican supports (yes or no) the torture of individuals suspected of terrorism, we first need to mathematically transform the values of our regression output (specifically the constant and the unstandardized logistic coefficients).

I have put together a screencast (<https://www.youtube.com/watch?v=OtnjEfXB930>) and an Excel sheet (see me for this if you need the spreadsheet and it is not in your course materials) that shows you how to do generate. If you have the spreadsheet and follow the directions in the screencast, it is a pretty straightforward process

**CONCLUDING COMMENTS:**

A lot of information is covered in the three parts of this document. Keep in mind that you will probably only use a fraction of the methods covered, depending on what kinds of variables you use in your project). More importantly, this is a resource that you will use in combination with screencasts, other assigned materials, various readings, and lots of conversations with your instructor.