**Supplemental Information about These Analyses**

*To help you understand why this is important and help you interpret each statistic for the project*

**Data Cleaning**

Summary: Sometimes survey respondents provide low-quality data. This can happen for a number of reasons such as respondents not paying attention, being bored or tired, or not wanting to complete the survey. Regardless of reason, it is our responsibility to “clean” the data, identifying and screening out respondents who provide low-quality data. Higher-quality data helps us make better and more trustworthy decisions.

Related Concepts:

* Social desirability: Faking good (selecting answers that make the respondent look good)
* Social undesirability: Faking bad (selecting undesirable responses)
* Acquiescence: Tendency to agree (i.e., select the “yes” responses)
* Disacquiescence: Tendency to disagree (i.e., select the “no” responses)
* Careless responding: Answering without sufficient thought or effort

Statistics and Techniques:

* “Use me” question: Including a survey question asking the respondent whether or not we should trust and use his or her data
* Bogus items: Including survey questions with a clearly “correct” answer
* Instructed items: Asking the respondent to select a specific answer
* Longstring: Examining how many consecutive invariant responses a respondent selects
* Response time: Measuring how much time a respondent spends on the survey
* Outliers: Identifying extreme or “illegal” (i.e., impossible) responses

Examples:

* “Use me” question: “Did you respond thoughtfully and with sufficient effort?” “Do you believe we should use your results?”
* Bogus item: “I currently live on the planet Earth.” “I was born on February 30.”
* Instructed item: “Please select “agree” for this item.” “Please leave this item blank.”
* Longstring: A response pattern of 5,5,5,5,5,5,5,5,5,5 has a longstring of 10 (ten “5s” in a row). A response pattern of 5,5,5,5,4,4,4,5,5,5, has a longstring of 4 (four “5s” in a row).
* Response time: Having the computer or website record how much time a respondent spends on a given item, page, or survey.
* Outliers: Illegal values (e.g., a score of “2” for sex when females are coded as 0 and males are coded as 1). Impossible responses (e.g., a GPA of 15.34).

**Scales and Scoring**

Summary: When we give people surveys, we often ask multiple questions about a given characteristic (e.g., a 10-item extraversion scale, a 50-question intelligence test). We do this to get a better picture of the characteristic we are trying to measure because, generally speaking, having more information or more data points is better than relying on a single piece of information.

However, when we analyze data, we generally want to have a single quantitative estimate for each characteristic. To obtain this estimate, we have to combine information from each question.

Related Concepts:

* Number of items: The number of items contained in a given test
* Item score: Score for a single item on a test
* Scale score: “Overall” score on the entire scale/test (usually a sum or average)

Statistics and Techniques:

* Sum: The result of adding all item scores
* Average: A statistic used to summarize scores on multiple items ()

Examples: Consider a five-item test (number of items = 5) where respondent #1 answered 1,3,3,4,2 and respondent #2 answered 5,3,4,2,4.

* Sum:
  + Respondent #1: 1+3+3+5+2 = 14
  + Respondent #2: 5+3+4+2+4 = 18
* Average:
  + Respondent #1:
  + Respondent #2:

**Correlations**

Summary: A correlation is a measure of association between two things. Correlations are typically computed to measure the association between two variables. As a result, they are sometimes called bivariate correlations (“bi” = two, “variate” = variables). Correlations are often denoted by the letter “r” with subscripts indicating the two variables. For example, a correlation between variable X and variable Y may be written as rXY. Correlations range from -1 to +1. There are two key pieces of information about a correlation: the *direction* and the *magnitude*.

Related Concepts:

* Direction: The *direction* (sign, either positive or negative) of a correlation tells you how the two variables are related. Positive numbers indicate a positive association such that when one variable increases, the other variable increases. High levels of X are associated with high levels of Y and low levels of X are associated with low levels of Y.
  + A correlation of -1 indicates a perfect negative relationship (i.e., when X increases, we know exactly how much Y will decrease).
  + A correlation of +1 indicates a perfect positive relationship (i.e., when X increases, we know exactly how much Y will increase).
  + A correlation of 0 indicates no relationship (i.e., when X increases, Y might increase or it might decrease).
* Magnitude: The *magnitude* (absolute value) of a correlation indicates the strength of the relationship between the two variables.
  + Values between 0 and -1 are imperfect negative relationships while values between 0 and +1 are imperfect positive relationships. Imperfect relationships are common, and simply mean that there is some “error” that limits our predictive ability. While we still know the general tendency (e.g., what happens to Y when X increases or decreases), the presence of error means that we do not know exactly how (or by how much) Y will change as X changes.
  + Absolute value- if you can predict something going up as going down you are equally able to tell the correlation; -.8 = .8

Examples:

* Example 1:
  + Scores on variable X: 5,4,2,5,4,4,7,7,7,6
  + Scores on variable Y: 7,6,4,2,4,3,7,5,5,4
  + rXY = .34
  + Interpretation: This is a positive correlation, so higher values of X are associated with higher values of Y (as X increases, Y increases).
* Example 2:
  + Scores on variable X: 3,4,4,1,1,3,3,4,4,4
  + Scores on variable Y: 4,1,5,3,7,4,7,3,6,6
  + rXY = -.17
  + Interpretation: This is a negative correlation, so higher values of X are associated with lower values of Y (as X increases, Y decreases).
    - The magnitude in Example 2 (.17) is lower than the magnitude in Example 1 (.34), meaning there is more error in Example 2, or that our ability to predict Y using X is stronger in Example 1.

**Multiple Regression**

Summary: Regression is related to correlation. Multiple regression is a technique used to predict values of one variable (the “criterion”) using a set of other variables (the “predictors”). While correlation can provide a good summary of bivariate relationships, we often use multiple regression when we are interested in using multiple variables in our predictive models.

Related Concepts:

* Criterion (or “dependent variable”): The variable we want to predict; the outcome of interest. The criterion is sometimes referred to as “Y.”
* Predictors (or “independent variables”): The variables we are using to predict the criterion variable. The predictors are sometimes referred to as numbered “X” variables (e.g., if we have 4 predictors, we might call then X1, X2, X3, and X4)
* R2: A measure of the overall predictive power of our model. Technically, R2 indicates the amount of variance in Y we can explain using our set of predictor variables. Practically, R2 ranges from 0 (no explanation) to 1 (perfect explanation). Higher R2 values mean that that a model is better at predicting Y (while lower values mean that a model is worse).
  + We can often increase R2 by adding additional predictors to our model.
  + We can often increase R2 by using better predictor (i.e., predictors that are more conceptually and/or empirically related to our criterion variable).
* Regression equation: A summary of the relationship by the predictors and criterion, usually in a form such as Y = B0 + B1X1 + B2X2 + B3X3 + B4X4 + e
* Intercept (B0): The value of Y when all predictors are equal to zero.
* Regression coefficient (Bk, where k is the number corresponding to the same X variable (e.g., B2 corresponds to predictor X2): A number indicating the relationship between a specific predictor and the criterion variable. Positive values mean that as the predictor increases, Y tends to also increase. Negative values mean that as the predictor increases, Y tends to also decrease.
* Error (e): The amount of uncertainty in our model, often quantified by (1 – R2).

Example:

* Scores on X1: 3,7,6,1,4,4,6,1,1,5
* Scores on X2: 3,5,6,2,5,2,7,2,2,1
* Scores on X3: 1,1,4,2,2,1,7,6,5,6
* Scores on Y: 4,14,13,16,9,9,16,8,16,6
* R2 = .25, so 25% of the variance in Y is explained by X1, X2, and X3
* Regression equation: Y = 8.31 + (-0.74)X1 + (1.36)X2 + (0.23)X3 + e
  + B0 (intercept): Y = 8.31 when X1, X2, and X3 all equal 0.
  + B1: -0.74. As X1 increases, Y tends to decrease
  + B2: 1.36. As X2 increases, Y tends to increase
    - This regression coefficient has the highest magnitude, so X2 has the strongest effect on Y of the three predictors
  + B3: 0.23. As X3 increases, Y tends to increase
    - This regression coefficient has the lowest magnitude, so X3 has the weakest effect on Y of the three predictors