**Data Cleaning and Missing Data**

Assignment 9

Note: Data cleaning varies based on each dataset and each researcher cleaning the data.

*Example Data Cleaning paragraph for a dissertation proposal:*

**Data Cleaning**

After the data has been collected, it will be screened for outliers and missing data. We will evaluate start times, and cases who complete the survey in less than 5 minutes will be excluded from the analyses. The internal reliability of all measures will be assessed. We will use an Expectation Maximization (EM) approach to account for missing data (Schafer & Graham, 2002); however, any case missing greater than 10% from any one measure will be deleted. The data will also be assessed for normality. We will assess skewness and kurtosis scores based on the guidelines set by Curran, West, and Finch, and values greater than three standard deviations from the mean will be removed (1996). In the event of non-normal data, we will also consider transforming the data for all variables except the LEC.

**Data Cleaning**

This dataset is a college student sample that completed an online survey related to trauma, growth, and well-being. The dataset is located on Canvas as:

“Growing+from+Stressful+Events+1”

1. Save a raw copy (Growing1\_Raw) and a clean copy (Growing1\_with\_Missing) of the dataset.

-Start a Data Cleaning Log in Word

2. Delete cases that did not complete the survey

3. Assess duration (Delete cases that completed the survey too quickly)

4. Delete all columns before ID

5. View > Freeze Panes > Freeze Top Row

6. Assess the data for data points that look out of place

-Either enter the correct value or change the values to missing

7. Change the headers

8. Delete Rows 2 and 3

9. Recode LEC

10. Replace remaining missing data points with a period.

-Record how many replacements there are in the dataset (Not including demographics)

-Divide total missing by total data points to add to your data cleaning paragraph

11. Assess the Age variable (Remove cases that the participant is under 18)

12. Use mean imputation for PANAS, MLQ, and GQ

13. Change the file type to Excel spreadsheet

14. Import the file into SPSS

-Save the dataset (Growing1\_Clean)

15. In variable view, make sure your variables are set up correctly.

-All Likert items should be Ordinal

-All Categorical items should be Nominal (ID, LEC, and Gender)

-Age should be Scale

-Enter values for Gender

16. Check the Min and Max of all items

17. Recode DASS and PCL

18. Missing data

-Little’s MCAR test

-Analyze > Missing Value Analysis

-Move all Ordinal variables to *Quantitative Variables*

-Select EM

-Paste and Run

-Impute missing values with Expectation Maximization

-One subscale at a time, move variables to *Quantitative Variables*

-Select EM

-In the EM tab, select *Save Completed Data* and enter the subscale name

-Paste and Run

-This will create a new dataset that you can copy into your full dataset.

19. Create total scores for PCL, DASS, and PTGI

-DASS has a total score and 3 subscales

20. Check skewness and kurtosis of the total scores

21. Check total scores for outliers

22. Determine whether total scores are different based on demographic variables

-Independent samples t-test for Gender

-Simple regression for Age

23. Check Cronbach’s alpha for PCL, DASS, and PTGI

24. Run a multiple regression with PCL, Depression, and Anxiety predicting PTGI

-Assess linearity, homoscedasticity, and multicollinearity

-Check for outliers

-Interpret the omnibus F and R2

-Interpret the Beta weights and Structure Coefficients

*Example Data Cleaning paragraph for a dissertation defense:*

**Data Cleaning**

Data collection occurred on Amazon’s MTURK on September 2, 2018. The initial dataset consisted of 258 respondents; however, several cases were excluded from analyses. Three cases were removed because they did not consent to participate in the study. Four cases were removed because they did not report any traumatic events on the LEC. Three cases were removed because they did not report their most stressful event. One case was removed because it failed the quality check questions, and three cases were removed because they contained greater than 10% missing data. No cases were removed due to length of time to complete the survey as all participants took >5 minutes with an average time of 22 minutes to complete the survey, leaving a final sample size of 244 participants (See Table 4 for demographic information).

Missing data was then assessed, indicating 140 total missing data points (.4%). The missing data was found to be missing completely at random (Little’s MCAR test *χ2*(6349) = 6049.81, *p* = .226), so an Expectation Maximization approach was used in in IBM SPSS Statistics 25 to estimate missing data, with each subscale estimated separately.

Skewness and kurtosis were within acceptable ranges (See Table 5; Curran, West, & Finch, 1996), and no significant outliers were detected (*Z* > 3.00). Standardized residual and standardized predicted value scatterplots indicated the assumptions of normality, linearity, and homoscedasticity for each regression analysis were all met. Multicollinearity for the multiple regression analyses was assessed. All regression models were found to have varying degrees of multicollinearity (Tolerance and Variance Inflation Factor (VIF) are reported for each hypothesis with multiple predictors, and squared structure coefficients are reported after each multiple regression analysis), but this was to be expected as all predictor and outcome variables were significantly correlated with each other (See Table 6).

Homework 9

**Scenario**

Using the **“Growing+from+Stressful+Events+2”**, clean the data and conduct a multiple regression analysis with PTSD symptoms, religiosity, and satisfaction with life predicting Posttraumatic Growth (PTG).

* Use EM for imputing missing values (Not mean imputation)
* Remember to fix the scoring of the PCL, DASS, PTGI
* Only obtain total scores for the PLC, RCI, SLS, and PTGI

Answer the following questions in Quiz 9 on Canvas:

1. Upload a Data Cleaning Log.
2. Upload an Excel spreadsheet with your clean dataset that still has missing values.
3. Upload your final dataset in SPSS format.
4. What percentage of your data were missing?
5. Report the results of Little’s MCAR test.
6. Report the Cronbach’s alphas for the PCL, RCI, SLS, and PTGI. Are they good?
7. Are there differences on the PTGI based on gender, age, or ethnicity?
   1. Report your statistics.
      1. Gender (use an independent samples t-test to look at differences between men and women)
      2. Age (use a simple regression to see if age predicts PTGI)
      3. Ethnicity (use a one-way ANOVA to look at group differences between ethnicities – Do not do post hocs. Just find the omnibus statistic)
8. Do the IVs predict the DV?
   1. Which predictors are strong according to the beta weights and structure coefficients?