Psy 524 Lab #2

Data Screening

Write any answers and paste any figures into this document after the appropriate question and upload on Canvas.

**Using ARC** (download and install the ARC program)

1. Create a .lsp file out of the **forclass.sav** data (download the creating a .lsp file document on the class webpage).
   1. Save **forclass.sav** as a tab delimited file (title it **forclass.dat**).
   2. Open arcinstruction.txt in NOTEPAD and make sure to insert actual information in any line that has a ! in front of it and delete the first two lines of directions.
   3. Open up forclass.dat in notepad as well (you may need to open up notepad separately so that both are open at the same time). Delete the variable names from the top of the data file and use them in the .lsp syntax. Copy all of the columns in the forclass.dat data and insert in between the parentheses. “Save as” **forclass.lsp**.
2. Looking at graphical relationship among variables.
   1. Open Arc and load **forclass.lsp** (you’ll need to find it wherever you saved it).
   2. Create a Multipanel plot of SOI (fixed axis on horizontal) versus N, E, A, O, C, ego and SOS, marking by gender (assume females are 1). To do this go to **Graph and Fit**, move SOI into **Fixed Axis** and the others (n to sos) into **changing axis**, move gender into **mark by** and then hit **OK**.
   3. Put OLS = 1 and use the bottom slide bar to change through predictors. Which variables appear to have relatively strong relationships with SOI? Copy and paste the graph with the strongest relationship below.

HIGHLIGHT HERE AND PASTE THE GRAPH FOR #2C

1. Looking at the same multiple panel plot from put OLS back at 0 and select “Remove Linear Trend”, do any of the variables seem to be heteroscedastic?

HIGHLIGHT HERE AND ANSWER THE QUESTION FOR #3

1. Examine the histograms for each of the variables (except gender, ethn and case numbers) and identify any variables with outliers (disconnection) or skewness (use 10 bins in each histogram).
   1. Go to **graph and fit**, **plot of** and move each variable in one at a time.
   2. If there is a variable with outlier copy and paste it below and then delete the outliers from the distribution.

HIGHLIGHT HERE AND PASTE THE GRAPH FOR #4B

* 1. If there is a variable that is skewed copy and paste it below.

HIGHLIGHT HERE AND PASTE THE GRAPH FOR #4C

1. If there was a variable from #3 that is skewed transform it. Go to forclass button (if this is what you put in the “dataset =” part of the .lsp syntax). Click on transform, move the variable over, change it to log transform, put the number 3 in the “c” box (this is a constant your adding to take it out of a scale that includes 0), and hit OK. Plot a histogram of this newly transformed variable and paste it below.

HIGHLIGHT HERE AND PASTE THE GRAPH FOR #5

**In SPSS**

1. Open the “**salary.sav**” data set using SPSS
   1. Create a casenum variable by using compute 🡪 casenum = $casenum.
   2. Calculate mahalanobis distances. **Analyze** 🡪 **Regression** 🡪 **Linear** and move **casenum** into dependent and everything else except “female” into Independent(s). Hit “**Save**” and click on “**Mahalanobis**”. **Continue** 🡪 **OK**.
   3. Close the output because you don’t need it. What is the cutoff value for multivariate outliers given these predictors? And are there any cases that qualify as multivariate outliers? If so which case(s)?

HIGHLIGHT HERE AND ANSWER 6C.

* 1. Split the file by **Data** 🡪 **Split File** click on “**compare groups**” and move “**female**” into “**Groups based on**”. What function does this serve? Repeat “b” above. Did the values change? Are there any multivariate outliers this time? What did this exercise demonstrate?

HIGHLIGHT HERE AND ANSWER 6D.

1. Dealing with Missing Value. Open “**social.sav**”.
   1. **Identifying Cases with no responses**:
      1. Transform 🡪 Compute and enter num=nvalid(ciccomp to supcomp) 🡪 OK.
      2. Data 🡪 Select Cases 🡪 If condition satisfies 🡪If… then enter num <> 0 into the window 🡪 continue, select “unselected cases are” DELETED.
   2. **Using Missing Value Analysis**:
      1. Analyze 🡪 Missing Value Analysis, include everything except gender, order and num into **Quantitative Variables**.
      2. Click on “Descriptives” and select “t-test with groups…” and “Include probabilities in table”, continue.
      3. Select “EM”, then hit new “EM” button 🡪 “Save completed data” 🡪 “File” and title it **social\_em.sav**, click on continue and then hit **OK**.

HIGHLIGHT HERE AND PASTE THE OUTPUT FOR #7Biii

* + 1. Interpret the output and pick the variable causing the largest dependency on the other missing values. What is the probability of Little and Ruben’s MCAR analysis? Repeat the steps above removing the variable causing biggest problem. What is the MCAR probability now? Pretend it’s OK and continue.

HIGHLIGHT HERE AND ANSWER THE QUESTIONS IN #7Biv

* 1. **Using Multiple Imputation:**
     1. Using the “**social.sav**” data set go to Analyze 🡪 Multiple Imputation 🡪 Impute Missing Data Values…, include everything except gender, order and num into **Variables in Model**. In the “Create a New Dataset” box enter **social\_imp.**
     2. Under **Method tab** leave it as **Automatic**.
     3. Under **Constraints Tab** click on the **Scan/Rescan** button and visually review the values to see that they make sense. In the **Define Constraints** box enter 1 in the rounding column for each variable.
     4. Under the Output tab make sure that both **Imputation Model** and **Descriptive** **Statistics for Variables with Imputed Values** are selected. Click on **OK**.

HIGHLIGHT HERE AND PASTE THE OUTPUT FOR #7Biii

* + 1. Randomly select one of the datasets (1-5) and use **Data** 🡪 **Select Cases** to save that dataset to a file titled “**social\_imp.sav**”.

1. Explore “**social\_imp.sav**” separately by gender. You should:
   1. Split file by **gender** (follow same steps as 6D above).
   2. Test for skewness and univariate outliers using **Analyze** 🡪 **Descriptive Statistics** 🡪 **Explore…**
      1. Calculate Z for skewness for each variable and tell me which variables violate this test.
      2. Take care of any outliers as you see fit (paste graphs with outliers, tell me why it’s an outlier and explain what you did to fix it (delete or change, etc…).

HIGHLIGHT HERE AND ANSWER 8Ai.

* 1. Do a further test of normality by asking for a P-P plot (make sure and split the file first). **Analyze** 🡪 **Descriptive Statistics** 🡪 **P-P plot…**, move the variables you want over and hit continue. Paste the graph that seems to show the worst violation and interpret it (tell me what it means, refer to T&F book for help).

HIGHLIGHT HERE AND PASTE THE GRAPH AND ANSWER 8B.

* 1. Perform square root transformations using **Transform** 🡪 **Compute Variable** and the T&F book’s table titled “Syntax for Common Data Transformations” (I would use “Paste” and run it from the syntax, hint hint). Run explore again and tell me which variables this worked for (normalized) if any by looking at the histograms and Z for skewness.

HIGHLIGHT HERE AND ANSWER 8C.

* 1. Do the same thing as in C for any variables square root did not normalize, but using a log10 transformation of the original variables (not the square rooted ones). Does this help any? For which variables?

HIGHLIGHT HERE AND ANSWER 8D.