# Psy 524 Lab #2

# Data Screening

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

## SPSS

### Mahalanobis Distance

1. Open the “**salary.sav**” data set using SPSS

1.a. Create a casenum variable by using compute 🡪 casenum = $casenum.

1.b. 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**.

1.c. 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 1.c.

1.d. 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 1.d.

### Missing Values Analysis (MVA) and Expectation Maximization

1. Open “**social.sav**”.

2.a. Identifying Cases with no responses:

2.a.i. **Transform** 🡪 **Compute** and enter num=nvalid(ciccomp to supcomp) 🡪 OK.

2.a.ii. **Data** 🡪 **Select Cases** 🡪 **If condition satisfies** 🡪**If…** then enter num <> 0 into the window 🡪 continue, select “unselected cases are” DELETED.

2.b. Using Missing Value Analysis:

2.b.i. **Analyze** 🡪 **Missing Value Analysis**, include everything except gender, order and num into Quantitative Variables.

2.b.ii. Click on “Descriptives” and select “t-test with groups…” and “Include probabilities in table”, continue.

2.b.iii. 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 #2.b.iii.

2.b.iv. 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 #2.b.iv.

### Missing Values and Multiple Imputation

1. Using the “**social.sav**” data set

3.a. **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.**

3.b.Under **Method tab** leave it as **Automatic**.

3.c. 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.

3.d. Under the Output tab make sure that both **Imputation Model** and **Descriptive** **Statistics for Variables with Imputed Values** are selected. Click on **OK**.

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

HIGHLIGHT HERE AND PASTE THE OUTPUT FOR #3.d.

### Normality and Transformations

1. Explore “**social\_imp.sav**” separately by gender. You should:

4.a. Split file by **gender** (follow same steps as 6D above).

4.b. Test for skewness and univariate outliers using **Analyze** 🡪 **Descriptive Statistics** 🡪 **Explore…**

HIGHLIGHT HERE AND ANSWER 4.a.i.

4.b.i. Calculate Z for skewness (by hand) for each variable and tell me which variables violate this test.

4.b.ii. 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 4.b.ii.

4.c. 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 4.c.

4.d. 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 4.d.

4.e. 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 4.e.

## R-studio

### Setting up a Notebook file, Installing/Loading Packages, and Loading Data

1. Open a new R notebook

5.a. **File** 🡪 **New File** 🡪 **R Notebook**

5.b. At the top edit the file to:

---

title: "Psy524 Lab #2 R code"

author: "Your Name"

output:

word\_document: default

html\_notebook: default

---

5.c. Delete the section that reads:

“This is an [R Markdown](http://rmarkdown.rstudio.com) Notebook. When you execute code within the notebook, the results appear beneath the code.

Try executing this chunk by clicking the \*Run\* button within the chunk or by placing your cursor inside it and pressing \*Ctrl+Shift+Enter\*.”

And replace it with:

# Psy524 Lab #2 Question 5

And delete everything below that.

5.d. Copy and paste the following chunk into the notebook and run it.

```{r Packages}

install.packages("foreign")

library(foreign)

```

5.e. Copy and paste the following chunk into the notebook and run it.

```{r Loading Data}

forclass <-read.spss("https://github.com/AndrewAinsworth/psy524/raw/master/LabAssignments/Lab02/forclass.sav",

use.value.label=TRUE,

to.data.frame=TRUE)

```

### Mahalanobis Distances

1. Computing and testing mahalanobis distances in R
   1. Copy and paste the following chunk into the notebook and run it.

# Computing Mahalanobis Distances and Checking for Multivariate outliers

```{r Mahalanobis}

mahal <- mahalanobis(forclass[ , -c(1:3)],

colMeans(forclass[ ,-c(1:3)], na.rm = TRUE),

cov(forclass[ , -c(1:3)], use = "pairwise.complete.obs"))

cutoff <- qchisq(.999, ncol(forclass[ , -c(1:2)]))

summary(mahal < cutoff)

```

HIGHLIGHT HERE AND PASTE OUTPUT FROM #6.1

* 1. What does the output mean? Explain.

### Normality and Simple linearity

1. Simple regression plots in R
   1. Copy and paste the following chunk into the notebook and run it.

# Simple normality and linearity

```{r Normality and Simple linearity}

#Panel Plot of histrograms for the first 4 variables

distribution1 <- par(mfrow=c(2,2))

distribution1 <- hist(forclass$sos)

distribution1 <- hist(forclass$ego)

distribution1 <- hist(forclass$n)

distribution1 <- hist(forclass$e)

#Panel Plot of histrograms for the first 4 variables

distribution2 <- par(mfrow=c(2,2))

distribution2 <- hist(forclass$o)

distribution2 <- hist(forclass$a)

distribution2 <- hist(forclass$c)

distribution2 <- hist(forclass$soitot)

#Simple plot separated by gender predicting SOITOT by A

plot(forclass$sos, forclass$soitot,

col = forclass$gender,

xlab = "SOS",

ylab = "SOITOT score",

main="Prediction of SOITOT by SOS")

abline(lm(soitot ~ sos, data = forclass), col="black")

```

HIGHLIGHT HERE AND PASTE OUTPUT FROM #7.1

* 1. Based on the histograms do any of the distributions look skewed? Do there appear to be any outliers? Explain.
  2. Based on the scatterplot does there appear to be any issues with linearity or homoskedasticity? Explain.

### Multicollinearity

1. Testing for multicollinearity with correlations in R
   1. Copy and paste the following chunk into the notebook and run it.

# Testing for Multicollinearity

```{r Multicollinearity}

#We can look at the correlations

correlation <- cor(forclass[ , -c(1:2)], use = "pairwise.complete.obs")

View(correlation)

symnum(correlation)

```

HIGHLIGHT HERE AND PASTE OUTPUT FROM #8.1

* 1. Does there appear to be issues with multicollinearity? Explain.

### More Linearity and Homoskedasticity

1. Using LM function, Q-Q plot and histograms in R
   1. Copy and paste the following chunk into the notebook and run it.

# More Linearity and Homoskedasticity

```{r Linearity and Homoskedasticity}

#In order to check normality in regression we have to run a model first

model = lm(soitot ~ sos + ego + n + e + o + a + c, data = forclass)

fitted = scale(model$fitted.values)

standardized = rstudent(model)

#Q-Q Plot of standardized values

qqnorm(standardized)

abline(0,1)

#Histogram of standardized fitted values

hist(standardized)

#Removing Linear Trend to check for homoskedasticity

plot(fitted, standardized)

abline(0,0)

abline(v = 0)

```

HIGHLIGHT HERE AND PASTE OUTPUT FROM #9.1

* 1. Do you see any issues with normality and homoskedasticity in the plots above? Explain.