MIS453 – Business Intelligence
HOA# 3: Regression in R

Fall 2020
due October 6 before class

Follow the below instructions to run the R code for Regression Analysis

Getting Started

Charts for associations

- Bar charts

```
# Creating charts for group distributions

# Load data file about Google searches by state

#data set that is based on Google searches by state. The idea here is that the Google search data is showing how many standard deviations above or below the national average each state is in their relative interest in a search term.
#Variables: State, that's the name of the state, state_code, that's like CA for California. Then we have their relative interest in data visualization; so, how often do they search for that relative to their other searches? Then we also have searches for Facebook, searches for NBA, and for fun, to put down whether that state had an NBA team. Also, the percentage of people in that state with a college degree, whether that state had a K-12 curriculum for statistics, and the region of the country.

google <- read.csv("google_correlate.csv", header = T) names(google)
str(google)

# Does interest in data visualization vary by region? # Split data by region, create new data frame viz.reg.dist
<- split(google$data_viz, google$region)

# Draw boxplots by region

boxplot(viz.reg.dist, col = "lavender")

# To draw barplot with means

viz.reg.mean <- sapply(viz.reg.dist, mean)

# Run next two together (or sequentially)
```

```
barplot(viz.reg.mean, col = "beige",

main = "Average Google Search Share of\n\"Data Visualization\" by Region of US") abline(h = 0)

# Install and load "psych" package to print means, etc.

install.packages("psych")
library("psych") describeBy(google$data_viz, google$region)
```

• Scatterplots

```
# Load data file about Google searches by state

# Is there an association between the percentage of people in a state with college degrees and # interest in
data visualization?
#plot(x,y)
plot(google$degree, google$data_viz)

# Add title, labels, change circles to points

plot(google$degree, google$data_viz,
main = "Interest in Data Visualization Searches\nby Percent of Population with College Degrees", xlab =
"Population with College Degrees",
ylab = "Searches for \"Data Visualization\"",

pch = 20,

col = "grey") # Add fit lines

# Linear regression line (y ~ x)

abline(lm(google$data_viz ~ google$degree), col="red")
```

Statistics for Associations

- **Calculating Correlations**

```
google <- read.csv("google_correlate.csv", header = T) names(google)

# Create data frame with only quantitative variables

g.quant <- google[c(3, 7, 4, 5)]
# Correlation matrix for data frame

cor(g.quant)

# Can test one pair of variables at a time
# Gives r, hypothesis test, and confidence interval cor.test(g.quant$data_viz, g.quant$degree)

# Install "Hmisc" package to get p-values for matrix

install.packages("Hmisc") library("Hmisc")
```

```
# Need to coerce g.quant from data frame to matrix # to get correlation matrix and p-values
rcorr(as.matrix(g.quant))
```

- **Computing regression**

```
google <- read.csv("google_correlate.csv", header = T)

names(google)

reg1 <- lm(data_viz ~
degree + stats_ed + facebook + nba + has_nba + region, data = google)

summary(reg1)
```