PREDICT 401 Introduction to Statistical Analysis Data Analysis Assignment #4

Introduction and Objectives:

There are three objectives to this assignment. First, contingency table analysis may be used as part of EDA to uncover associations between variables. This assignment will demonstrate how this may be done. Second, the classical two-sample t-test is commonly used to compare different populations. Classical t confidence intervals assume a symmetric underlying sampling distribution for the t statistic, and do not adjust for asymmetry or skewness when it is present in the sampling distribution. Large sample sizes are required in this situation. Bootstrap methods empirically adjust for skewness. This can be important when comparing samples from populations that have materially different variances. Using samples from asymmetric distributions, the performance of the classical procedure will be compared to confidence intervals based on bootstrapping. While there are instances in which the three methods agree closely, it is a good practice to use bootstrapping as a check on the classical t procedure. With unequal sample sizes and skewness, bootstrap t confidence intervals are preferred. Third, with more than two populations a commonly used procedure is the analysis of variance. This method will be demonstrated along with a procedure for simultaneous comparisons. Examples of linear regression will also be given.

Overview:

This assignment is comprised of three parts. The first part is an EDA which includes contingency table analysis and chi square tests. The second part addresses two-sample t-tests and bootstrapping. The third part deals with one-way analysis of variance, the Tukey Honest Significance Differences test and linear regression. Instructions and examples are provided to guide the investigations. Use the code supplied as instructed. Follow these instructions carefully. Execute the code and enter information into the tables presented in the instructions. Upon completion be sure to take the associated quiz.

Preliminaries:

The language R will be required throughout the assignment. Two R packages will also be required: "moments" and "ggplot2." Both are available from the Comprehensive R Archive Network (CRAN). The data sets "Hospital.csv" and "schools.csv" must be downloaded and be available.

Ebook Version:

Assignment instructions reference a chapter, page and problem number. The ebook version of *Business Statistics* does not have page numbers. The data dictionary and problem references can be located in WileyPlus by using Chapter Resources. For each chapter navigate to the item *Analyzing the Databases* located on the left hand side of the screen. The data dictionary is located there under Chapter 1. The problems referenced by chapter are located therein for each specified chapter.

Supplemental References:

1) Chihara and Hesterberg, *Mathematical Statistics with Resampling and R* Chapter 3 pages 58-64, and Chapter 7 pages 199-200. (Located in course reserves on the course site.)
2) Verzani, Using R for Introductory Statistics, pages 85, 132-137, 146-148 and 408-410.

Part 1 Exploratory Data Analysis (EDA)

This part of the assignment will use the data in "Hospital.csv". Refer to Black *Business Statistics* Chapter 10 page 410 problem 3. Use the code in Appendix A for Part 1. Complete the following.

1) Use the results from summary() to determine if extreme outliers are present. Use these results to fill in the following table. Refer to Section 3.4 of Black page 85 (Figure 3.13).

Variable	IQR	Q_3+3.0*IQR	Maximum Value	Extreme Outlier Yes or No
admissions				
beds				
expenditures				

2) Evaluate the various displays. Evaluate the histograms for skewness.
3) Review the boxplots of admissions, beds and expenditures and compare Psychiatric hospitals versus General hospitals. Compare the boxplots to the corresponding scatter plots.
4) Note how table() is used in conjunction with chisq.test(). This portion of the program will demonstrate the effect of Yates' Continuity Correction on the chi square test of independence. Part of this involves using the user supplied chi() function. Fill in the following table.

Comparison of chisq.test and chi	chisq.test()		chi()	
	X-squared	p-value	X-squared	p-value
service_beds				
service_admissions				
service_exp				

5) Evaluate the table showing the geographic distribution of hospitals.

Part 2 Two-Sample t tests and Bootstrapping

Refer to Black *Business Statistics* Chapter 10 page 410 problem 3. Also consult the references cited earlier in this assignment. Of particular importance is Chihara and Hesterberg, *Mathematical Statistics with Resampling an d R* Chapter 7 pages 199-200 (course reserves). Follow the steps as shown in the code supplied in Appendix B. The first portion of the program deals with admissions.

6) Using the code supplied in Appendix B of this assignment, subset the data into two data.frames, one for general medical hospitals and the other for psychiatric hospitals. Do an initial EDA using the code supplied taking note of the distributions shapes and locations.
7) Execute the resampling code supplied and examine the resampling distributions presented for the mean difference and the t statistic. Take note of how close the vertical lines are in the plot of the sampling distribution for the t statistic. These mark where the quantiles are located.
8) Continue with the program and calculate: 1) the traditional two-sample t confidence interval, 2) the percentile bootstrap confidence interval, and 3) the bootstrap t confidence interval. Do this as shown at the 90% confidence interval. Compare results.
9) Note the tail area comparison between the theoretical t distribution and the bootstrap t resampling distribution. Ideally the two values would both equal 0.05. If both values are within 0.025 and 0.075, the theoretical distribution may be an acceptable match to the resampling results. There is no hard and fast rule for making this judgment.
10) Repeat the preceding steps for expenditures and beds. For beds, you will need to adapt the code supplied to that variable. After completing these steps, enter the results in the table below. Round off to a single decimal place. (The last column is the conclusion reached (at 90% confidence) from testing the null hypothesis of no difference between the population means versus the alternative hypothesis that there is a difference.)

	Confidence Intervals Based on Three Methods			Different from zero or not?
Variable	Traditional t	Bootstrap percentile	Bootstrap t statistic	
admissions				
expenditures				
beds				

Part 3 One-way Analysis of Variance and Linear Regression

This part of the assignment will use the data in "schools.csv". There is no reference to Black *Business Statistics*. The file schools.csv contains educational expenditures from three consecutive years. Each year fifty schools are selected at random across the nation. These schools are identified by the geographic region from which they were selected. The nation is divided into four regions. Variables Y and X are not annual numbers. We will ignore the timeframe associated with X and Y for this exercise.

The definition of the variables is:

Y Per capita expenditure on public education
X Per capita personal income
region A: Northeast, B: North Central, C: South, D: West
year "1", "2", "3"

The assignment starts with EDA, followed by AOV and linear regression. Assume the variances are equal and that an analysis of variance can be performed. Use the code supplied in Appendix C. Complete the following.

11) In the overview table that is constructed, take note of which region has the highest average expenditure per capita on education, and how average expenditures compare to average income.
12) Compare the Pearson Correlation Coefficient with the R-squared result from the simple linear regression. The square of the correlation coefficient equals the R-squared value of 0.4.
13) Review the next series of boxplots prior to completing the AOV. Ask yourself how the AOV results might turn out. The AOV is sufficiently robust to handle the differences in variability.
14) Perform the two AOVs relating Y to year and to region. The F value is the test statistic in an AOV. With a one-way analysis of variance, only when this statistic produces a small p-value can the results be deemed statistically significant and TukeyHSD() used.
15) Examine the results from TukeyHSD() and determine which comparisons resulted in significant differences. Then adapt the code and duplicate the analyses only this time use X instead of Y. Summarize the results of the TukeyHSD() procedure in the table below as shown.

Regions Compared	Y as dependent variable		X as dependent variable	
	diff	p adj	diff	p adj
B-A	-3.17	0.9735		
C-A	-26.24	0.0014		
D-A	25.09	0.0041		
C-B	-23.08	0.0024		
D-B	28.25	0.0003		
D-C	51.33	0.0000		

16) Compare to the boxplots produced at (13).
17) Execute the two-way AOVs and take note of which factor(s) turn out to be statistically significant. This could include the interaction term. Would you expect this result?

The next portion of the assignment uses linear regression to combine two variables which appear to be important for predicting Y, expenditures per capita. They are region and X, income per capita.

18) Execute the multiple linear regression and note which if any terms in the model turn out to be statistically significant. (This is an example of what is referred to as a parallel lines regression.)
19) Compare the regression results with the associated ggplot. Note how the term in the regression model for region D reflects the positioning of those data points relative the other data points.
20) Evaluate how well the residuals conform to a normal distribution. The fit is reasonable.

Save all your computational results and program(s) for the quiz.

```
# Predict 401 Data Analysis Assignment #4
# Assignment #4
#-------------------------------------------------------------------------
# Appendix A Code
# Part 1
#-------------------------------------------------------------------------
# Analyzing the Databases Problem 3 page 410

hospital <- read.csv(file.path("c:/RBlack/","Hospital.csv"),sep=",")
# hospital <- read.csv("hospital.csv", sep = ",")
str(hospital)
library(moments)

admissions <- hospital$Admissions
beds <- hospital$Beds
hospital$Tot..Exp. <- hospital$Tot..Exp./1000   # Expenditures in thousands of dollars
exp <- hospital$Tot..Exp.

# Define factors which will be used for displays and contingency tables.
# Label as indicated S for South, N for Northeast, M for Midwest, SW for
# Southwest, R for Rocky Mountain, C for California and NW for Northwest.

service <- factor(hospital$Service, labels = c("General", "Psychiatric"))
region <- factor(hospital$Geog..Region, labels = c("S","N","M","SW","R", "C", "NW"))
hospital <- data.frame(hospital, region, service, exp)

str(hospital)

# Initial EDA

summary(admissions)
summary(beds)
summary(exp)

# Summary statistics provide information to evaluate if extreme outliers are present.
# An extreme outlier is defined as an observation located outside either the third or
# the first quartile by more than 3.0*IQR.  (Black Section 3.4 Figure 3.14 page 85.)

hist(admissions, col = "red")
hist(beds, col = "blue")
hist(exp, col = "green", xlab = "Expenditures")

# Compare the summary information to the boxplots for admissions, beds and exp.
# To find out more about the arguments in boxplot(), type help(boxplot) in the console.

boxplot(admissions, notch = T, range = 1.5, main = "Admissions", col = "red")
boxplot(beds, notch = T, range = 1.5, main = "Beds", col = "blue")
boxplot(exp, notch = T, range = 1.5, main = "Expenditures", col = "green")

# Scatter plots can be helpful in visualizing relationships between variables.
# They are also revealing in terms of comparative hospital operations.
```

```
with(hospital, plot(hospital$Beds,hospital$Admissions,
  col=c("red","green3")[service], main = "Admissions, by number of beds"))
legend("bottomright", c("General", "Psychiatric"), col = c("red", "green3"), pch = 1)
with(hospital, plot(hospital$Admissions,hospital$Tot..Exp.,
  col=c("red","green3")[service], main = "Total Expenditures, by Admissions"))
legend("bottomright", c("General", "Psychiatric"), col = c("red", "green3"), pch = 1)
with(hospital, plot(hospital$Beds,hospital$Tot..Exp.,
  col=c("red","green3")[service], main = "Total Expenditures, by Beds"))
legend("bottomright", c("General", "Psychiatric"), col = c("red", "green3"), pch = 1)

# Boxplots help reveal how the distributions compare across service.
# Take note of how the distributions differ for General and Psychiatric hospitals.
# With skewed distributions, the F test should not be used to compare sample variances.
# There is no general consensus on the best way to test for equality in this situation.
# Given the visual displays the variances will be assumed to be different.

boxplot(Tot..Exp.~service,data=hospital,main="Exp by Service", col = "red", range = 1.5)
boxplot(Beds~service,data=hospital,main="Beds by Service", col = "blue", range = 1.5)
boxplot(Admissions~service,data=hospital,main="Admission by Service", col = "green", range = 1.5)

# Dichotomize the data. This is necessary for simple contingency table analysis on counts.
# A convenient way to start is using a logical condition which generates a TRUE or a
# FALSE result and labeling accordingly.  Having the right order in the labels argument is
# critical to maintaining a correct outcome.

beds_median <- factor(beds > median(beds), labels = c("below","above"))
admissions_median <- factor(admissions > median(admissions), labels = c("below","above"))
exp_median <- factor(exp > median(exp), labels = c("below","above"))

hospital <- data.frame(hospital,exp_median,beds_median,admissions_median)

# Comparison of the resulting contingency tables to the bivariate plots presented
# previously is informative.  EDA on the tables suggests a number of insights.
# Expenditures are proportionately lower for Psychiatric hospitals than General hospitals.
# Admissions are proportionately lower for Psychiatric hospitals than General hospitals.
# Regarding beds it looks like a comparable split.

service_beds <- table(service ,beds_median)
addmargins(service_beds)
service_exp <- table(service, exp_median)
addmargins(service_exp)
service_admissions <- table(service, admissions_median)
addmargins(service_admissions)

# This is confirmed by chi square tests on the contingency tables.  The chi square test
# tests the hypothesis of independence of the factors involved.

chisq.test(service_exp)
chisq.test(service_beds)
chisq.test(service_admissions)
```

```
# Refer to pages 58-60 of Chihara and Hesterberg Mathematical Statistics
# Chapter 3.5 is available through Course Reserves.  Yates'continuity correction
# may overcorrect.  An example where it is not used is shown for beds.

# This is the numerical calculation of Pearson's Chi square without Yates' Correction.

(84-168*101/200)^2/(168*101/200)+(84-99*168/200)^2/(99*168/200)+
  (17-101*32/200)^2/(101*32/200)+(15-99*32/200)^2/(99*32/200)

# The user supplied function chi() does the same thing starting with a table.

chi <- function(x){
 # To be used with 2x2 contingency tables that have margins added.
 # Expected values are calculated.
 e11 <- x[3,1]*x[1,3]/x[3,3]
 e12 <- x[3,2]*x[1,3]/x[3,3]
 e21 <- x[3,1]*x[2,3]/x[3,3]
 e22 <- x[3,2]*x[2,3]/x[3,3]
 # Value of chi square statistic is calculated.
 result <- (x[1,1]-e11)^2/e11+(x[1,2]-e12)^2/e12+(x[2,1]-e21)^2/e21+(x[2,2]-e22)^2/e22
 result
}

# Example of using the function chi() and obtaining a p-value.
service_beds <- addmargins(service_beds)
q <- chi(service_beds)
q # [1] 0.1050105
pchisq(q, 1, lower.tail = FALSE) # [1] 0.7458977

service_exp <- addmargins(service_exp)
q <- chi(service_exp)
q # [1] 12.05357
pchisq(q, 1, lower.tail = FALSE) # [1] 0.0005169325

service_admissions <- addmargins(service_admissions)
q <- chi(service_admissions)
q # [1] 38.09524
pchisq(q, 1, lower.tail = FALSE) # [1] 6.737436e-10

# Record the chi square results from the chisq.test().  Also use the function chi()
# on service_beds, service_exp and service_admissions.  Record the results.

# Create a contingency table of service by regions.  It is apparent the distribution
# of Psychiatric hospitals differs from General hospitals across geographic regions.
# Because of the number of zeros in the table,a Chi square test is not advised.

addmargins(table(hospital$service,hospital$region))

#------------------------------------------------------------------------
# Appendix B Code
```

```
# Part 2
#------------------------------------------------------------------------

# Two sample bootstrap study of psy and gen expenses.  Refer to Chihara and
# Hesterberg pages 199-200 for details on two-sample bootstrapping.

#------------------------------------------------------------------------
# Two sample bootstrap study of psy and gen using Admissions.
# First define the two samples for testing.

gen <- hospital[service == "General", ]
psy <- hospital[service == "Psychiatric", ]

psy.A <- psy$Admissions  # These are the samples for comparison.
gen.A <- gen$Admissions

# The aggregate function provides a convenient way to compare the samples.

aggregate(Admissions~service,data=hospital,summary)

# Investigate the distributions.

par(mfrow = c(2,2))
hist(psy.A, col = "red", xlab = "Admissions")
hist(gen.A, col = "green", xlab = "Admissions")
boxplot(psy.A, col = "red", ylab = "Admissions")
boxplot(gen.A, col = "green", ylab = "Admissions")
par(mfrow = c(1,1))

# We will assume the data constitutes random samples from larger populations.
# Bootstrap resampling will be conducted on these data.  N pairs of samples
# will be drawn and statistics calculated on each.  A classical two-sample
# t confidence interval is also prepared. Each time this section is run, use
# the same set.seed value for comparable results.

np <- length(psy.A)
ng <- length(gen.A)
mu <- mean(gen.A)-mean(psy.A)
std.s <- sqrt((var(psy.A)/np)+(var(gen.A)/ng))  # Sample standard deviation of difference.

N <- 10^4
diff.mean <- numeric(N)
diff.t <- numeric(N)
set.seed(123)   # Keep this set.seed the same to assure comparable results.

for (i in 1:N)
{
  psy.sample <- sample(psy.A, np, replace=TRUE)
  gen.sample <- sample(gen.A, ng, replace=TRUE)
  x <- mean(gen.sample)-mean(psy.sample)                # Calculate the mean difference.
  diff.mean[i] <- x
```

```
  std <- sqrt((var(psy.sample)/np) + (var(gen.sample)/ng))
  diff.t[i] <- ((x-mu)/std)                    # Calculate the t statistic.
}

# Evaluate the bootstrapping sampling distributions.

x <- seq(-4*std.s+mu, mu+4*std.s, 0.1)
hist(diff.mean, main = "Resampling distribution of the mean differences", col = "red", prob = T)
abline(v=mean(diff.mean), col="blue", lty = 2, lwd=2)
curve(dnorm(x,mean = mu, sd = std.s ),add=TRUE, col= "green", lwd = 2, type = "l")
legend("topright", legend = c("skewness = 0.06", "kurtosis = 3.03"))
abline(v= quantile(diff.mean, probs = 0.05), col = "blue", lwd = 2, lty = 2)
abline(v= quantile(diff.mean, probs = 0.95), col = "blue", lwd = 2, lty = 2)
skewness(diff.mean) # [1] 0.06415897
kurtosis(diff.mean) # [1] 3.031802

hist(diff.t, main = "Resampling Distribution of t-statistic df = n=ng+np-2", breaks = "Sturges",
    ylim = c(0, 0.45), prob = TRUE, col = "green", xlab = "t-statistic values")
curve(dt(x, df=ng+np-2),add=TRUE, col= "darkred", lwd = 2)
legend("topright", legend = c("skewness = -0.20", "kurtosis = 3.12"))
abline(v= quantile(diff.t, probs = 0.05), col = "blue", lwd = 2, lty = 2)
abline(v= quantile(diff.t, probs = 0.95), col = "blue", lwd = 2, lty = 2)
abline(v= qt(c(0.05,0.95),ng+np-2,lower.tail=T),col="darkred",lwd=2,lty=2)
skewness(diff.t) # [1] -0.201206
kurtosis(diff.t) # [1] 3.120071

#-------------------------------------------------------------------------------------------
# Traditional confidence interval using t-statistic.
t.test(gen.A, psy.A, var.equal=FALSE,conf.level=0.9)

#         Welch Two Sample t-test
#
# data:  gen.A and psy.A
# t = 11.5348, df = 196.103, p-value < 2.2e-16
# alternative hypothesis: true difference in means is not equal to 0
# 90 percent confidence interval:
#  5501.413 7341.497
# sample estimates:
# mean of x mean of y
#  7859.268  1437.812

# Determine two-sided confidence interval using mean bootstrap distribution.
round(quantile(diff.mean, probs=c(0.05,0.95)), digits = 2)

#      5%     95%
# 5505.79 7343.27

# Bootstrap confidence interval based on the t-statistic.
round((mu - quantile(diff.t, probs = 0.95, names = FALSE)*std.s), digits = 2)
# [1] 5549.68
```

```
round((mu - quantile(diff.t, probs = 0.05, names = FALSE)*std.s), digits = 2)
# [1] 7423.47

# Check the tail area of the bootstrap t distribution against the theoretical
# t value with degrees of freedom equal to 35 and 198.  Ideally both values
# would be 0.05 which reflects With good convergence.

sum(diff.t > 1.658)/N # [1] 0.042
sum(diff.t < -1.658)/N # [1] 0.0622

#-----------------------------------------------------------------------------------------------
#-----------------------------------------------------------------------------------------------
# Two sample bootstrap study of psy and gen using Tot..Exp.
# First, split the data set into two subsets and examine.

gen <- hospital[service == "General", ]
psy <- hospital[service == "Psychiatric", ]

psy.E <- psy$Tot..Exp.  # These are the samples for comparison.
gen.E <- gen$Tot..Exp.

# The aggregate function provides a convenient way to compare the samples.

aggregate(exp~service,data=hospital,summary)

# Investigate the distributions.

par(mfrow = c(2,2))
hist(psy.E, col = "red", xlab = "Expenditures in Thousands of Dollars")
hist(gen.E, col = "green", xlab = "Expenditures in Thousands of Dollars")
boxplot(psy.E, col = "red", ylab = "Thousands of Dollars")
boxplot(gen.E, col = "green", ylab = "Thousands of Dollars")
par(mfrow = c(1,1))

# We will assume the data constitutes random samples from larger populations.
# Bootstrap resampling will be conducted on these data.  N pairs of samples
# will be drawn and statistics calculated on each.  A classical two-sample
# t confidence interval is also prepared. Each time this section is run, use
# the same set.seed value for comparable results.

np <- length(psy.E)
ng <- length(gen.E)
mu <- mean(gen.E)-mean(psy.E)
std.s <- sqrt((var(psy.E)/np)+(var(gen.E)/ng))  # Sample standard deviation of difference.

N <- 10^4
diff.mean <- numeric(N)
diff.t <- numeric(N)

set.seed(123)  # Keep this set.seed the same to assure comparable results.
```

```
for (i in 1:N)
{
  psy.sample <- sample(psy.E, np, replace=TRUE)
  gen.sample <- sample(gen.E, ng, replace=TRUE)
  x <- mean(gen.sample)-mean(psy.sample)                # Calculation of the mean difference.
  diff.mean[i] <- x
  std <- sqrt((var(psy.sample)/np) + (var(gen.sample)/ng))
  diff.t[i] <- ((x-mu)/std)                             # Calculation of the t statistic.
}

# Evaluate the bootstrapping sampling distributions.

x <- seq(-4*std.s+mu, mu+4*std.s, 0.1)
hist(diff.mean, main = "Resampling distribution of the mean differences", col = "red", prob = T)
abline(v=mean(diff.mean), col="blue", lty = 2, lwd=2)
legend("topright", legend = c("skewness = -0.01", "kurtosis = 3.09"))
curve(dnorm(x,mean = mu, sd = std.s ),add=TRUE, col= "green", lwd = 2, type = "l")
abline(v= quantile(diff.mean, probs = 0.05), col = "blue", lwd = 2, lty = 2)
abline(v= quantile(diff.mean, probs = 0.95), col = "blue", lwd = 2, lty = 2)
skewness(diff.mean)
kurtosis(diff.mean)

hist(diff.t, main = "Resampling Distribution of t-statistic df = n=ng+np-2", breaks = "Sturges",
    ylim = c(0, 0.45), prob = TRUE, col = "green", xlab = "t-statistic values")
curve(dt(x, df=ng+np-2),add=TRUE, col= "darkred", lwd = 2)
legend("topright", legend = c("skewness = 0.03", "kurtosis = 3.12"))
abline(v= quantile(diff.t, probs = 0.05), col = "blue", lwd = 2, lty = 2)
abline(v= quantile(diff.t, probs = 0.95), col = "blue", lwd = 2, lty = 2)
abline(v= qt(c(0.05,0.95),ng+np-2,lower.tail=T),col="darkred",lwd=2,lty=2)
skewness(diff.t)
kurtosis(diff.t)

#-------------------------------------------------------------------------------------
# Traditional confidence interval using t-statistic.
t.test(gen$Tot..Exp.,psy$Tot..Exp.,var.equal=FALSE,conf.level=0.9)

# Determine two-sided confidence interval using the percentile bootstrap method
# on the resampled distribution for the mean difference.
round(quantile(diff.mean, probs=c(0.05,0.95)), digits = 2)

# Bootstrap confidence interval based on the t-statistic.
round((mu - quantile(diff.t, probs = 0.95, names = FALSE)*std.s), digits = 2)
round((mu - quantile(diff.t, probs = 0.05, names = FALSE)*std.s), digits = 2)

# Check the tail areas of the bootstrap t distribution against the theoretical
# t value with degrees of freedom equal to 198. Ideally, both values would
# be 0.05 which reflects With good convergence.

sum(diff.t > 1.658)/N # [1] 0.0509
sum(diff.t < -1.658)/N # [1] 0.0461
#---------------------------------------------------------------------
```

```
#-----------------------------------------------------------------------
# Two sample bootstrap study of psy and gen using Beds

gen <- hospital[service == "General", ]
psy <- hospital[service == "Psychiatric", ]

psy.B <- psy$Beds
gen.B <- gen$Beds

# The aggregate function provides a convenient way to compare the samples.

aggregate(Beds~service,data=hospital,summary)

# Adapt the coding examples shown previously to calculate confidence intervals.
#-----------------------------------------------------------------------------


#-----------------------------------------------------------------------------
# Traditional confidence interval using t-statistic.


# Determine two-sided confidence interval using mean bootstrap distribution.


# Bootstrap confidence interval based on the t-statistic.


# Check the tail area of the bootstrap t distribution against the theoretical
# t value with degrees of freedom equal to 198.  Ideally, both values would
# be 0.05 which reflects With good convergence.


#-----------------------------------------------------------------------------
# Appendix C Code
# Part 3  One-way Analysis of Variance
#-----------------------------------------------------------------------------
# schools.csv contains educational expenditures over three consecutive years.
# Each year fifty schools are selected at random across the nation. The nation is
# divided into four regions.  Variables Y and X are not annual numbers.  We will
# ignore the timeframe for this exercise.  The definition of the variables is:
# Y Per capita expenditure on public education
# X Per capita personal income
# region A: Northeast, B: North Central, C: South, D: West
# year "1", "2", "3"

# Perform initial EDA
schools <- read.csv(file.path("c:/R401/","schools.csv"),sep=",")
# schools <- read.csv("schools.csv", sep = ",")
schools$year <- factor(schools$year, labels = c("1","2","3"))
str(schools)
```

```
# Check summary statistics and visual plots.
summary(schools)

# Sometimes it is useful to form an overview table.
my <- aggregate(schools$Y~schools$region, data = schools, mean)
mx <- aggregate(schools$X~schools$region, data = schools, mean)
mx <- mx[,2]
overview <- cbind(my,mx)
colnames(overview) <- c("region","expenditures", "income")
overview

#   region expenditures   income
# 1      A     85.55556 2260.556
# 2      B     82.38889 2090.917
# 3      C     59.31250 1572.792
# 4      D    110.64103 2118.359

# Evaluate distributions.
par(mfrow = c(1,2))
hist(schools$Y, col = "red", main = "Distribution of Schools by Expenditure")
hist(schools$X, col = "blue", main = "Distribution of Schools by Income")
par(mfrow = c(1,1))

# A bivariate plot is a useful way to visualize data.
plot(schools$X, schools$Y,main = "Expenditures versus Personal Income",
    xlab = "Per capita personal income", ylab = "Per capita expenditure on public education",
    col = "red", pch = 16)

# Evaluate correlation between Y and X.  This is the Pearson Product Moment Correlation
# Coefficient sometimes referred to as the linear correlation coefficient.
cor.test(schools[,1],schools[,2])

# This suggests a regression analysis may be used.
result <- lm(Y~X,data=schools)
summary(result)

# Note that the correlation coefficient from cor.test when squared equals
# the multiple R-squared value of 0.4 when rounded.  However, this is only
# a preliminary model.  Other factors should be considered.

par(mfrow = c(2,2))
boxplot(Y~year, data = schools, col = "red", main = "Expenditures by Year")
boxplot(Y~region, data = schools, col = "red", main = "Expenditures by Region")
boxplot(X~year, data = schools, col = "blue", main = "Income by Year")
boxplot(X~region, data = schools, col = "blue", main = "Income by Region")
par(mfrow = c(1,1))

# Perform initial one-way analyses of variance.

aov.year <- aov(Y~year, schools)
summary(aov.year)
```

```
# No significant difference is found.

aov.region <- aov(Y~region, schools)
summary(aov.region)
# Significant difference is found.  Perform TukeyHSD.  Compare to the boxplots.
TukeyHSD(aov.region)

#-------------------------------------------------------------------------------
# Repeat the one-way analysis of variance using Income as the dependent variable.


#-------------------------------------------------------------------------------
# This next section shows the efficiency which can be gained with a two-way AOV.

# Combine factors and perform a two-way analysis of variance.
result <- aov(Y~region+year+region*year,schools)
summary(result)

result <- aov(X~region+year+region*year,schools)
summary(result)

#-------------------------------------------------------------------------------
# The AOV results points to region as an important factor for a multiple regresson model.

# Using ggplot2 it is possible to visualize the role played by region.

result <- lm(Y~X+region,schools)
summary(result)

library(ggplot2)
ggplot(schools, aes(x = X, y = Y))+geom_point(aes(color = region), size = 3)+
  ggtitle("Plot of Expenditures versus Income Colored by Region")

# A multiple regression model needs to be evaluated.  Using the residuals is one way.
# It is highly desirable for the residuals to conform to a normal distribution with
# few to no outliers.

r <- residuals(result)

par(mfrow = c(1,2))
hist(r, col = "red", main = "Histogram of Residuals", xlab = "Residual")
boxplot(r, col = "red", main = "Boxplot Residuals", ylab = "Residual")
par(mfrow = c(1,1))

qqnorm(r, col = "red", pch = 16, main = "QQ Plot of Residuals")
qqline(r, col = "green", lty = 2, lwd = 2)
skewness(r) # [1] -0.002500969
kurtosis(r) # [1] 3.251458
# The residuals indicate the regression model is a reasonable fit to the data despite a few outliers.
```