Spring 2019
Dr. Hakan Gogtas
Student: Yinan Kang

CSCI E-106: Data Modeling

Assignment 5

Due: March, 25 2019 at 11:59 pm EST

Instructions: Students should submit their reports on Canvas. The report needs to clearly state what question is being solved, step-by-step walk-through solutions, and final answers clearly indicated. Please solve by hand where appropriate.

Please submit two files: (1) a R Markdown file (.Rmd extension) and (2) a PDF document generated using knitr for the .Rmd file submitted in (1) where appropriate. Please, use RStudio Cloud for your solutions.

All questions are coming from Kutner, M. *et al*: Applied Linear Statistical Models, Fifth Edition.

1. (4.9)
2. (4.19)
3. (4.23) 4.23 Hand-written, scanned in back. Rest on Knitr PDF.
4. (4.26)
5. (4.27)

YK_Assignment5

Yinan Kang

3/23/2019

Problem 4.9 - PLASTIC HARDNESS

(a)

```
# Import plastic hardness data

colnames <- c("y","x")
df <- read.table(url("http://users.stat.ufl.edu/~rrandles/sta4210/Rclassnotes/data/textdatasets/KutnerDa
n <- nrow(df)

# (a)
# Bonferroni confidence intervals for X = 20, 30, 40 hours
fitreg <- lm(y~x, df)
Xh <- c(20,30,40)
pred <- predict.lm(fitreg,data.frame(x = Xh), se.fit = T, level = 0.9)
B = rep(qt(1-0.1/(2*length(Xh)),n-2),length(Xh))
a.results <- cbind(pred$fit-B*pred$se.fit, pred$fit+B*pred$se.fit)
colnames(a.results) <- c("lwr","upr")

print(a.results)
```

```
##        lwr      upr
## 1 206.7277 211.8473
## 2 227.6762 231.5863
## 3 246.7824 253.1676
```

What is the meaning of the family confidence coefficient here?

Response: Family confidence coefficient at 90% confidence means, when sampled repeatedly, 90% of the time the entire family of estimates would be valid.

Here specifically, if we sampled the below Xh values repeatedly, 90% of the time all three Yh values would exist in their respective ranges:
Xh = 20 (hours), the Yh (hardness) value between [206.7277428, 211.8472572]
Xh = 30, the Yh value between [227.6762032, 231.5862968]
Xh = 40, the Yh value between [246.7824219, 253.1675781]

(b)

```
# Checking Working-Hotelling interval - if it produces a tighter prediction interval, it'd be considere

W <- rep(sqrt(2*qf(0.9,2,n-2)),length(Xh))
b.results <- cbind(pred$fit-W*pred$se.fit, pred$fit + W*pred$se.fit)
colnames(b.results) <- c("lwr","upr")
print(b.results)
```

```
##        lwr      upr
## 1 206.7545 211.8205
## 2 227.6966 231.5659
## 3 246.8158 253.1342
```

Is the Bonferroni procedure the most efficient one to be employed here?

Response: No, the Working-Hotelling procedure produced tighter intervals among the Yh predictions shown thus far. Therefore, the Working-Hotelling procedure is the more efficient one here.

(c)

```
# Predict Yh for Xh = 30,40 using Working-Hotelling at 90% confidence

Xh <- c(30,40)
pred <- predict.lm(fitreg,data.frame(x = Xh), se.fit = T, level = 0.9)
W <- rep(sqrt(2*qf(0.9,2,n-2)),length(Xh))
c.results <- cbind(pred$fit-W*pred$se.fit, pred$fit + W*pred$se.fit)
colnames(c.results) <- c("lwr","upr")
print(c.results)
```

```
##        lwr      upr
## 1 227.6966 231.5659
## 2 246.8158 253.1342
```

Predictions of Yh for Xh = 30,40 with 90% Confidence:

When Xh = 30, Yh is between 227.6966397 and 231.5658603
When Xh = 40, Yh is between 246.8157946 and 253.1342054

Problem 4.19 - GPA

(a)

```
# Import Data
rm(list=ls())
colnames <- c("y","x")
df <- read.table(url("http://users.stat.ufl.edu/~rrandles/sta4210/Rclassnotes/data/textdatasets/KutnerDa
n <- nrow(df)

# Create linear model
model <- lm(y~x, df)

# Collecting figures needed for s2{predX}
mse <- mean(model$residuals^2)
b0 <- 2.11405 # From model
b1 <- 0.03883 # From model
xbar <- mean(df$x)
sxx <- 0
for (i in 1:nrow(df)){
  sxx.temp <- (df$x[i] - xbar)^2
  sxx <- sxx+sxx.temp
}
```

```r
# Calculating Xh.new
Yh.new <- 3.4
Xh.new <- (Yh.new-b0)/b1

# Calculating s{predx} for Xh = 3.4
s2.predx <- (mse/(b1^2)) * (1+(1/n) + ((Xh.new-xbar)^2)/sxx)
s.predx <- sqrt(s2.predx)

# Calculating Interval at 90%
t <- qt(1-0.1/2,n-2)

Xh.new.low <- Xh.new-t*s.predx
Xh.new.hi <- Xh.new + t*s.predx

print(Xh.new.low)
```

```
## [1] 6.239732
```

```r
print(Xh.new.hi)
```

```
## [1] 59.99514
```

Results: For a student with GPA = 3.4, we are 90% confident that their ACT score would've been between 6.2397325 and 59.9951375.

(b)

```r
# Checking criterion 4.33
crit <- ( (t^2) * mse ) / ((b1^2) * sxx)
print(crit)
```

```
## [1] 0.2924515
```

The guideline given for criterion (4.33) is it should be small, < 0.1. Here, the value is 0.29, therefore the criterion for appropriateness is NOT MET.

Problem 4.26 - CDI

(a) Obtain joint Bonferroni confidence intervals at 95%

```r
# Importing Data
rm(list=ls())
cdi.df <- read.csv("CDI.csv")
n <- nrow(cdi.df)
attach(cdi.df)

# Generating linear regression model
model <- lm(Number.Active.Physicians ~ Total.Population)
b0 <- -1.106e+02   # from model
b1 <- 2.795e-03   # from model
```

```r
# Calculating B
B <- qt(1-.05/4,n-2)

# Collecting values for s.b0 and s.b1 (to calculate confidence intervals)
mse <- mean(model$residuals^2)
xbar <- mean(Total.Population)
sxx <- 0
for (i in 1:nrow(cdi.df)) {
  sxx.temp <- (Total.Population[i] - xbar)^2
  sxx <- sxx + sxx.temp
}
s2.b0 <- mse * ((1/n) + (xbar^2)/sxx)
s.b0 <- sqrt(s2.b0)
s2.b1 <- mse/sxx
s.b1 <- sqrt(s2.b1)

# Calculating confidence intervals

b0.low <- b0 - B*s.b0
b0.hi <- b0 + B*s.b0
b1.low <- b1 - B*s.b1
b1.hi <- b1 + B*s.b1
```

Bonferroni Joint Confidence Intervals:
b0: $[-188.5706792, -32.6293208]$
b1: $[0.0026865, 0.0029035]$

(b)

Do the joint confidence intervals from (a) support the view that B0 should be -100 and B1 should be .0028?

Response: YES, according to the confidence intervals for b0 and b1 from (a), the researcher's views on beta0 and beta1 can be supported.

(c)

Will the Working-Hotelling or Bonferroni procedure be more efficient?

Reponse: Based on the reasoning provided in the textbook, for smaller families of prediction with fewer statements, the Bonferroni will give the more efficient and tighter interval. This is the case here (3 prediction intervals), so BONFERRONI would likely be more efficient.

(d)

```r
# Generating Bonferroni 90% confidence intervals
Xh <- c(500*1000, 1000*1000, 5000*1000)
pred <- predict.lm(model, data.frame(Total.Population=Xh),se.fit=T,level=0.9)

B <- rep(qt(1-0.1/(2*length(Xh)),n-2),length(Xh))

d.results <- cbind(pred$fit-B*pred$se.fit, pred$fit+B*pred$se.fit)
```

```r
colnames(d.results) <- c("lwr","upr")
print(d.results)
```

```
##         lwr       upr
## 1  1224.013  1350.142
## 2  2596.566  2773.014
## 3 13386.746 14346.233
```

Interpretation:

When repeatedly sampled, 90% of the time areas with the following 'populations' will have 'number of physicians' in the given ranges:
For populations of 500 thousand, the mean number of physicians is in range: [1224.0131191, 1350.1422048]
For populations of 1000 thousand, the mean number of physicians is in range: [2596.5662999, 2773.0139024]
For populations of 5000 thousand, the mean number of physicians is in range: $[1.3386746 \times 10^4, 1.4346233 \times 10^4]$

Problem 4.27 - SENIC

(a)

```r
rm(list=ls())
# Import Data
senic <- read.table("/cloud/project/APC1.DAT", quote="\"", comment.char="")
names(senic) <- c("ID", "length.of.stay", "age" , "infection.risk", "routine.culturing.ratio", "routine
n <- nrow(senic)

attach(senic)

# Generating model
model <- lm(length.of.stay ~ infection.risk)

# Bonferroni join confidence intervals for B0 and B1, 90% family confidence
bonf <- confint(model,level=1-0.1/2)
print(bonf)
```

```
##                    2.5 %    97.5 %
## (Intercept)    5.3038443 7.3697288
## infection.risk 0.5336442 0.9871976
```

Joint confidence interval, 90% Bonferroni:
B0: [5.3038443, 7.3697288]
B1: [0.5336442, 0.9871976]

(b)

Does joint confidence interval in (a) support researcher's belief that B0 = 7, B1 = 1 ?

Response: No, the joint confidence interval in (a) DOES NOT support the researcher's belief, as B1 = 1 is not in the joint confidence interval range for B1.

(c)

Will the Working-Hotelling or Bonferroni procedure be more efficient for creating CIs for new Xh values?

Response: Not sure here, as Xh has size 4, which is not tiny, thus not guaranteeing Bonferroni would be more efficient. Thus, I'll try both in (d) and see for myself.

(d)

```
Xh <- c(2,3,4,5)
pred <- predict.lm(model,data.frame(infection.risk=Xh),se.fit=T,level=.95)    # 95% confidence interval

# Creating Bonferroni
B <- rep(qt(1-.05/(2*length(Xh)),n-2),length(Xh))
b.results <- cbind(pred$fit-B*pred$se.fit, pred$fit+B*pred$se.fit)

# Creating Working-Hotelling
W <- rep(sqrt(2*qf(0.95,2,n-2)),length(Xh))
w.results <- cbind(pred$fit-W*pred$se.fit, pred$fit+W*pred$se.fit)

# Compare Bonferroni vs. Working-Hotelling
print(b.results)
```

```
##       [,1]      [,2]
## 1 7.071051  8.644206
## 2 8.065356  9.170743
## 3 8.977088  9.779852
## 4 9.708059 10.569723
```

```
print(w.results)
```

```
##       [,1]      [,2]
## 1 7.088991  8.626266
## 2 8.077961  9.158137
## 3 8.986242  9.770698
## 4 9.717885 10.559897
```

Conclusion: WORKING-HOTELLING was the more efficient choice here. This makes sense, as the family size = 4, which is large enough for the Working-Hotelling to be tighter than the Bonferroni.

Interpretation: When repeatedly sampled, 95% of the time the following family of estimates will be valid:

Infection Risk = 2, Avg. Length of Hospital Stay = [7.0889906, 8.626266]
Infection Risk = 3, Avg. Length of Hospital Stay = [8.0779611, 9.1581374]
Infection Risk = 4, Avg. Length of Hospital Stay = [8.9862423, 9.7706979]
Infection Risk = 5, Avg. Length of Hospital Stay = [9.7178849, 10.5598971]

Problem 4.23 on next page.

(4.23) Show that for $\hat{Y}_i = b_1 X_i$,

$$\Sigma X_i e_i = 0$$

According to the textbook,

$\hat{Y}_i = b_1 X_i$ (4.15)

relies on the estimator of b_1, $b_1 = \frac{\Sigma X_i Y_i}{\Sigma X_i^2}$ (4.14),

which is the result of the normal equation:

$\Sigma X_i (Y_i - b_1 X_i) = 0$ (4.13)

For regression through origin, $e_i = Y_i - b_1 X_i$ (4.16)

Putting these together,

$$\Sigma X_i e_i = \Sigma X_i (Y_i - b_1 X_i) = 0$$ end of sol'n