Chad Huntebrinker’s Homework 11

Chad Huntebrinker

2024-11-26

Problem 9.21 and 9.22

#Chad Huntebrinker  
  
library(leaps)

## Warning: package 'leaps' was built under R version 4.4.2

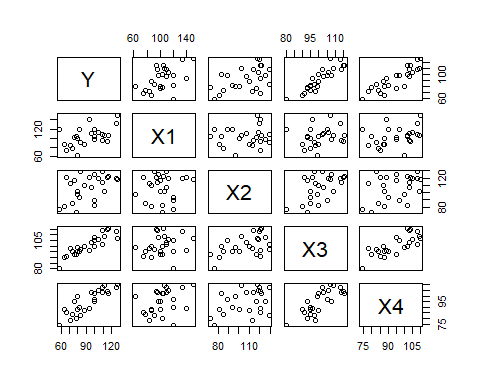
library(readxl)  
  
excel\_data <- read\_excel("Job\_Proficiency\_Data.xlsx")  
  
#Fit the model  
model\_1 <- lm(Y~X1 + X3 + X4, data=excel\_data)  
sum\_of\_model\_1 <- summary(model\_1)  
  
#Problem 9.21  
#Calculate the SSE and PRESS models  
model\_1\_SSE <- sum(model\_1$residuals^2)  
  
model\_1\_PRESS <- sum((model\_1$residuals / (1 - hatvalues(model\_1)))^2)  
  
model\_1\_PRESS

## [1] 471.452

model\_1\_SSE

## [1] 348.197

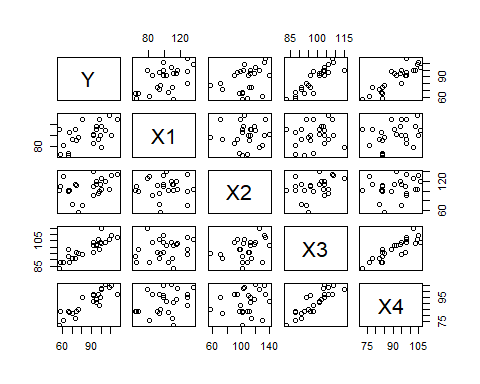
#Since PRESS is larger than the SSE, this suggest that the model might be overfitted and MSE is  
#not a reliable indicator.  
  
#Problem 9.22a  
excel\_data2 <- read\_excel("Job\_Proficiency\_Data2.xlsx")  
  
pairs(excel\_data)



cor(excel\_data[,-1])

## X1 X2 X3 X4  
## X1 1.0000000 0.1022689 0.1807692 0.3266632  
## X2 0.1022689 1.0000000 0.5190448 0.3967101  
## X3 0.1807692 0.5190448 1.0000000 0.7820385  
## X4 0.3266632 0.3967101 0.7820385 1.0000000

pairs(excel\_data2)



cor(excel\_data2[,-1])

## X1 X2 X3 X4  
## X1 1.00000000 0.01057088 0.1772891 0.3196395  
## X2 0.01057088 1.00000000 0.3437441 0.2207638  
## X3 0.17728907 0.34374413 1.0000000 0.8714466  
## X4 0.31963945 0.22076377 0.8714466 1.0000000

#Yes, they are reasonably similiar  
  
#Problem 9.22b  
model\_2 <- lm(Y~X1 + X3 + X4, data=excel\_data2)  
#Validation Model  
summary(model\_2)

##   
## Call:  
## lm(formula = Y ~ X1 + X3 + X4, data = excel\_data2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.4619 -2.3836 0.6834 2.1123 7.2394   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -122.76705 11.84783 -10.362 1.04e-09 \*\*\*  
## X1 0.31238 0.04729 6.605 1.54e-06 \*\*\*  
## X3 1.40676 0.23262 6.048 5.31e-06 \*\*\*  
## X4 0.42838 0.19749 2.169 0.0417 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.284 on 21 degrees of freedom  
## Multiple R-squared: 0.9489, Adjusted R-squared: 0.9416   
## F-statistic: 130 on 3 and 21 DF, p-value: 1.017e-13

summary(model\_2)$coefficients[, "Std. Error"]

## (Intercept) X1 X3 X4   
## 11.84783231 0.04729415 0.23261650 0.19749332

sum(model\_2$residuals^2)

## [1] 385.4536

anova(model\_2)["Residuals", "Mean Sq"]

## [1] 18.35493

#Y = 122.76705 + 0.31238X1 + 1.40676X3 + 0.42838X4  
  
#Regular Model  
sum\_of\_model\_1

##   
## Call:  
## lm(formula = Y ~ X1 + X3 + X4, data = excel\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.4579 -3.1563 -0.2057 1.8070 6.6083   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -124.20002 9.87406 -12.578 3.04e-11 \*\*\*  
## X1 0.29633 0.04368 6.784 1.04e-06 \*\*\*  
## X3 1.35697 0.15183 8.937 1.33e-08 \*\*\*  
## X4 0.51742 0.13105 3.948 0.000735 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.072 on 21 degrees of freedom  
## Multiple R-squared: 0.9615, Adjusted R-squared: 0.956   
## F-statistic: 175 on 3 and 21 DF, p-value: 5.16e-15

sum\_of\_model\_1$coefficients[, "Std. Error"]

## (Intercept) X1 X3 X4   
## 9.87405909 0.04367948 0.15183247 0.13105392

sum(model\_1$residuals^2)

## [1] 348.197

anova(model\_1)["Residuals", "Mean Sq"]

## [1] 16.58081

#Y = 124.20002 + 0.29633X1 + 1.35697X3 + 0.51742X4  
  
#Problem 9.22c  
model\_2\_prediction <- predict(model\_2)  
model\_2\_MSPE <- mean((excel\_data2$Y - model\_2\_prediction)^2)  
model\_2\_MSPE

## [1] 15.41814

model\_1\_MSE <- mean((excel\_data$Y - predict(model\_1))^2)  
model\_1\_MSE

## [1] 13.92788

#MSE is lower than MSPE but not significantly; thus, there isn't much  
#evidence of bias.  
  
#Problem 9.22d  
full\_data <- rbind(excel\_data, excel\_data2)  
model\_3 <- lm(Y~X1 + X3 + X4, data=full\_data)  
summary(model\_3)

##   
## Call:  
## lm(formula = Y ~ X1 + X3 + X4, data = full\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.7192 -2.7369 0.1278 2.0971 7.0657   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -123.44104 7.16508 -17.228 < 2e-16 \*\*\*  
## X1 0.30364 0.03072 9.886 5.86e-13 \*\*\*  
## X3 1.36906 0.12280 11.148 1.15e-14 \*\*\*  
## X4 0.48735 0.10475 4.652 2.79e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.006 on 46 degrees of freedom  
## Multiple R-squared: 0.9567, Adjusted R-squared: 0.9539   
## F-statistic: 338.9 on 3 and 46 DF, p-value: < 2.2e-16

summary(model\_1)

##   
## Call:  
## lm(formula = Y ~ X1 + X3 + X4, data = excel\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.4579 -3.1563 -0.2057 1.8070 6.6083   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -124.20002 9.87406 -12.578 3.04e-11 \*\*\*  
## X1 0.29633 0.04368 6.784 1.04e-06 \*\*\*  
## X3 1.35697 0.15183 8.937 1.33e-08 \*\*\*  
## X4 0.51742 0.13105 3.948 0.000735 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.072 on 21 degrees of freedom  
## Multiple R-squared: 0.9615, Adjusted R-squared: 0.956   
## F-statistic: 175 on 3 and 21 DF, p-value: 5.16e-15

#The intercept standard deviations are reduced, but not the other ones; in fact, they increased.

Problem 9.27

#Chad Huntebrinker  
  
library(leaps)  
library(readxl)  
  
excel\_data <- read\_excel("SENIC\_Data.xlsx")  
data\_part1 <- excel\_data[1:56,]  
data\_part2 <- excel\_data[57:113,]  
  
#Problem 9.27a  
#The best subset according to the Cp criterion is using age, check x-ray ratio, and census.  
#So we will use that as our regression model  
model\_1 <- lm(log(Length\_of\_Stay)~Age + `Routine\_Chest\_X-ray` + Average\_daily\_census, data = data\_part2)  
  
validation\_model <- lm(log(Length\_of\_Stay)~Age + `Routine\_Chest\_X-ray` + Average\_daily\_census, data = data\_part1)  
  
summary(model\_1)$coefficients[, "Std. Error"]

## (Intercept) Age `Routine\_Chest\_X-ray`   
## 0.2045023221 0.0037460580 0.0009643419   
## Average\_daily\_census   
## 0.0001049466

sum(model\_1$residuals^2)

## [1] 0.8579723

anova(model\_1)["Residuals", "Mean Sq"]

## [1] 0.01618816

summary(validation\_model)$coefficients[, "Std. Error"]

## (Intercept) Age `Routine\_Chest\_X-ray`   
## 0.2872850008 0.0048558381 0.0010067914   
## Average\_daily\_census   
## 0.0001431377

sum(validation\_model$residuals^2)

## [1] 1.16534

anova(validation\_model)["Residuals", "Mean Sq"]

## [1] 0.02241038

#The validation model seems similar to the model-building one.  
  
#Problem 9.27b  
model\_1\_MSE <- mean((data\_part2$Length\_of\_Stay - predict(model\_1))^2)  
validation\_model\_MSPE <- mean((data\_part1$Length\_of\_Stay - predict(validation\_model))^2)  
  
model\_1\_MSE

## [1] 53.88792

validation\_model\_MSPE

## [1] 62.3442

#The MSE and MSPE have a slight difference which wouldn't indicate any bias  
  
#Problem 9.27c  
model\_2 <- lm(log(Length\_of\_Stay)~Age + `Routine\_Chest\_X-ray` + Average\_daily\_census, data=excel\_data)  
summary(model\_2)$coefficients

## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 1.444129952 1.699914e-01 8.495313 1.122971e-13  
## Age 0.008117543 2.964507e-03 2.738244 7.216257e-03  
## `Routine\_Chest\_X-ray` 0.003303817 6.833786e-04 4.834534 4.403894e-06  
## Average\_daily\_census 0.000544456 8.617587e-05 6.317963 5.920449e-09

summary(model\_1)$coefficients

## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 1.4055759329 0.2045023221 6.873154 7.216907e-09  
## Age 0.0089342528 0.0037460580 2.384974 2.068909e-02  
## `Routine\_Chest\_X-ray` 0.0027050480 0.0009643419 2.805071 7.018153e-03  
## Average\_daily\_census 0.0006737651 0.0001049466 6.420078 3.859019e-08

#The coefficients and standard deviations are different from those of the model building set.  
#We would expect there to be a difference in the estimates due to new examples being introduced.

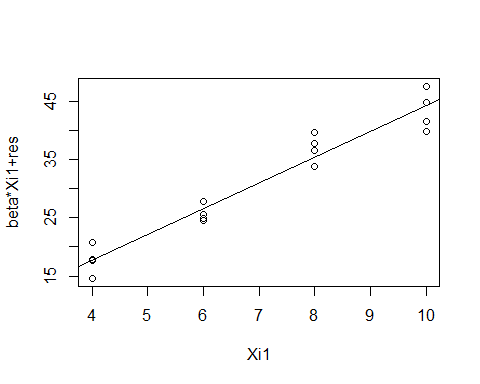
Problem 10.5

#Chad Huntebrinker  
  
library(readxl)  
require(faraway)

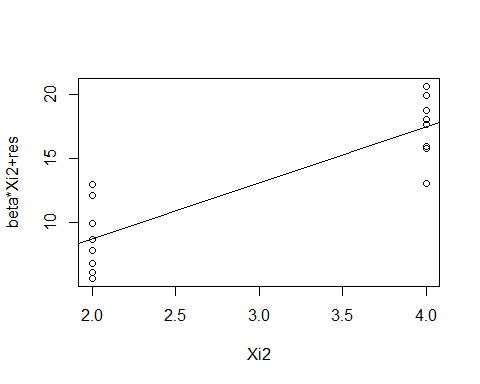
## Loading required package: faraway

## Warning: package 'faraway' was built under R version 4.4.2

excel\_data <- read\_excel("Brand\_preference\_data.xlsx")  
  
#Problem 10.5a  
model\_1 <- lm(Yi~Xi1+Xi2, data = excel\_data)  
  
prplot(model\_1,1)



prplot(model\_1,2)

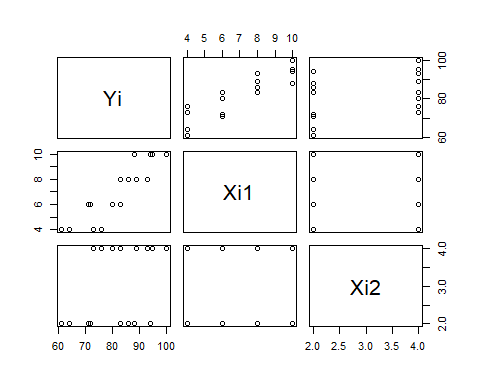


#Problem 10.5b  
#It looks like both Xi1 and Xi2 would contribute to the model due to both of them having  
#a linear relationship with the residuals of the model  
  
#Problem 10.5c  
model\_Yi\_X1 <- lm(Yi ~ Xi1, data = excel\_data)  
model\_X2\_X1 <- lm(Xi2 ~ Xi1, data = excel\_data)  
  
  
model\_2 <- lm(model\_Yi\_X1$residuals~model\_X2\_X1$residuals)  
  
summary(model\_2)

##   
## Call:  
## lm(formula = model\_Yi\_X1$residuals ~ model\_X2\_X1$residuals)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.400 -1.762 0.025 1.587 4.200   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.0000 0.6488 0.000 1   
## model\_X2\_X1$residuals 4.3750 0.6488 6.743 9.43e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.595 on 14 degrees of freedom  
## Multiple R-squared: 0.7646, Adjusted R-squared: 0.7478   
## F-statistic: 45.47 on 1 and 14 DF, p-value: 9.427e-06

Problem 10.15

#Chad Huntebrinker  
  
library(readxl)  
  
excel\_data <- read\_excel("Brand\_preference\_data.xlsx")  
model\_1 <- lm(Yi~Xi1+Xi2, data = excel\_data)  
  
#Problem 10.15a  
pairs(excel\_data)



cor(excel\_data[,-1])

## Xi1 Xi2  
## Xi1 1 0  
## Xi2 0 1

#It shows there is no association between X1 and X2  
  
#Problem 10.15b  
vif(model\_1)

## Xi1 Xi2   
## 1 1

#They are both equal to 1 because there is no correlation between the predictor variables

Problem 10.19

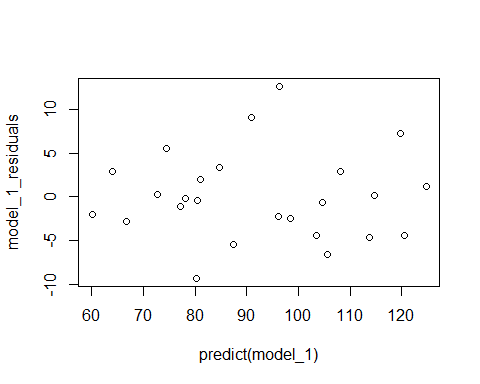
#Chad Huntebrinker  
  
library(readxl)  
library(dplyr)

##   
## Attaching package: 'dplyr'

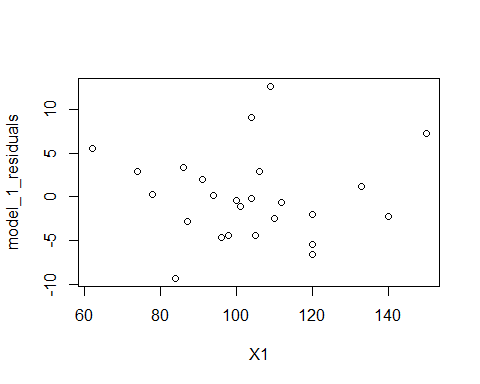
## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

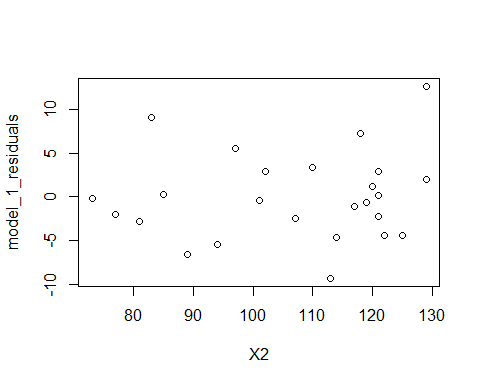
require(faraway)  
  
excel\_data <- read\_excel("Job\_Proficiency\_Data.xlsx")  
  
#Problem 10.19a  
model\_1 <- lm(Y~X1 + X3, data=excel\_data)  
model\_1\_residuals <- model\_1$residuals  
excel\_data <-excel\_data %>%   
 mutate(X1X3 = X1\*X3)  
  
plot(model\_1\_residuals~predict(model\_1))



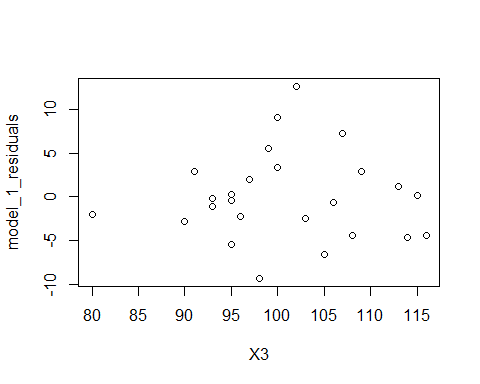
plot(model\_1\_residuals~X1, data = excel\_data)



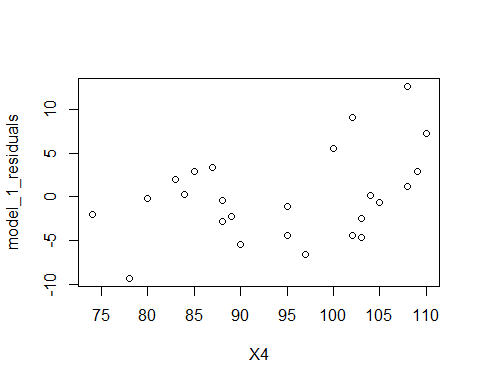
plot(model\_1\_residuals~X2, data = excel\_data)



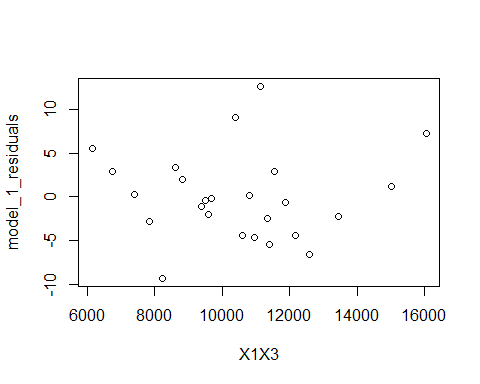
plot(model\_1\_residuals~X3, data = excel\_data)



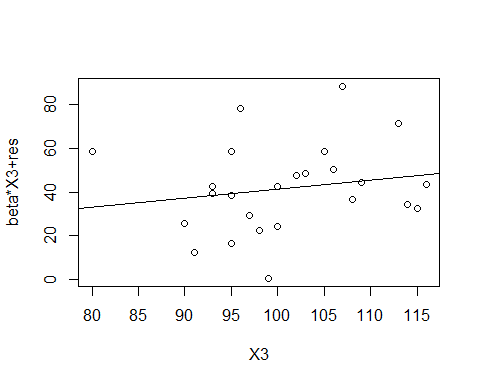
plot(model\_1\_residuals~X4, data = excel\_data)



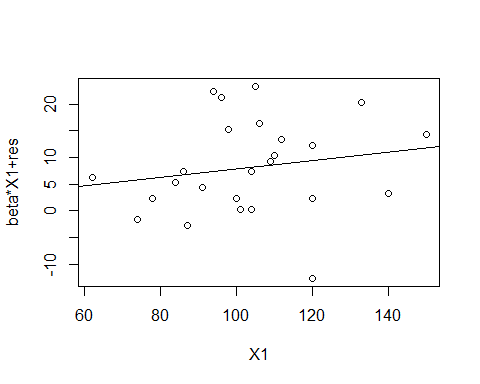
plot(model\_1\_residuals~X1X3, data = excel\_data)



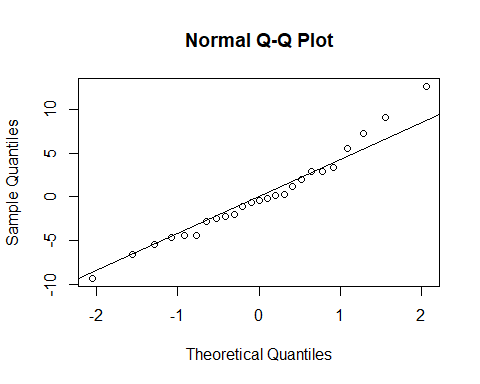
#The Yhat graph seems okay along with X1, X2, and X3 and X1X3.  
#X4 might have a curve (starts up, goes down, then starts going up again).  
#Overall, doesn't look like there's any modification needed.  
  
#Problem 10.19b  
model\_X1\_X3 <- lm(X1 ~ X3, data = excel\_data)  
model\_X3\_X1 <- lm(X3 ~ X1, data = excel\_data)  
  
prplot(model\_X1\_X3, 1)



prplot(model\_X3\_X1, 1)



#No modification seems to be needed.  
  
#Problem 10.19c  
qqnorm(model\_1$residuals)  
qqline(model\_1$residuals)



cor(model\_1$residuals, qqnorm(model\_1$residuals, plot.it = FALSE)$x)

## [1] 0.9840739

#The table has it listed between 0.957 and 0.96 and the correlation coefficient is 0.98  
#Thus, it is reasonable.  
  
#Problem 10.19d  
#Ha: If SDR > 2.508325, then it is an outlier  
#H1: If SDR <= 2.508325, then it is not an outlier  
  
model\_1\_SDR <- rstudent(model\_1)  
bon\_cutoff <- qt(1 - 0.5 / 50, df = model\_1$df.residual)  
outliers <- abs(model\_1\_SDR) > bon\_cutoff  
  
#The following are outliers  
which(outliers)

## 16   
## 16

model\_1\_SDR[16]

## 16   
## 2.826891

#Problem 10.19e  
which(hatvalues(model\_1)>2\*3/25)

## 7 18   
## 7 18

excel\_data[c(7,18),]

## # A tibble: 2 × 6  
## Y X1 X2 X3 X4 X1X3  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 58 120 77 80 74 9600  
## 2 127 150 118 107 110 16050

#Problem 10.19f  
dffits\_values <- dffits(model\_1)  
dfbetas\_values <- dfbetas(model\_1)  
cooks\_values <- cooks.distance(model\_1)  
  
dffits\_values[c(7, 16, 18)]

## 7 16 18   
## -0.3395422 0.6030725 0.9998970

#Case 7 and 16 are okay, case 18 indicates a possible influence as it is almost equal to 1  
  
dfbetas\_values[c(7, 16, 18)]

## [1] -0.24018835 -0.06867086 -0.46436969

#All the cases seem okay (they are less than 1)  
  
cooks\_values[c(7, 16, 18)]

## 7 16 18   
## 0.03982863 0.09199677 0.30812948

#Case 7 and 16 are okay, case 18 indicates a possible influence as it is greater than  
#the 10 or 20 percent cutoff.  
  
#As a result, case 7 and 16 do not seem to be outliers. Case 18 indicates some suspicion and  
#should be investigated deeper.  
  
#Problem 10.19g  
vif\_values <- vif(model\_1)  
vif\_values

## X1 X3   
## 1.033781 1.033781

#They indicate a basically no multicollinearity between the predictor variables.