HW1

Thomas Kwok

February 20, 2019

Question 15:

1. Prepare a stem-and-leaf plot for each of the predictor variables. Are any noteworthy features revealed by these plots?

Running a stem and leaf plot, the main noteworthy thing is the difference in size between Xi1, Xi2, and Xi3. Especially Xi3 which is so much smaller than everything else.

setwd("C:/Users/thoma/Desktop/Spring 2019/Linear Model 2")  
q15 <- read.csv('Ch6Q15.csv')  
attach(q15)  
stem(Xi1, scale = 1)

##   
## The decimal point is 1 digit(s) to the right of the |  
##   
## 2 | 23  
## 2 | 58899999  
## 3 | 012233344  
## 3 | 66678  
## 4 | 0012233344  
## 4 | 557779  
## 5 | 0233  
## 5 | 55

stem(Xi2, scale = .5)

##   
## The decimal point is 1 digit(s) to the right of the |  
##   
## 4 | 1234  
## 4 | 666678888899999  
## 5 | 000011111111223333444  
## 5 | 5678  
## 6 | 02

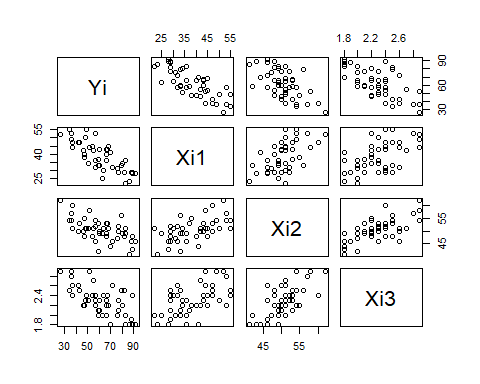
stem(Xi3, scale= .2)

##   
## The decimal point is at the |  
##   
## 1 | 888889  
## 2 | 0000011122222223333334444444555566678999

1. Obtain the scatter plot matrix and the correlation matrix. Interpret these and state your principal findings.

From the matrices it seems that there is a negative correlation between Yi and the three predictor variables. Also it seems that the predictor variables do have some correlation among each other also. This is especially interesting as I thought that Xi3 would not have any correlation with the other variables because it is so much smaller than the other variables in its values.

pairs(q15)



cor(q15)

## Yi Xi1 Xi2 Xi3  
## Yi 1.0000000 -0.7867555 -0.6029417 -0.6445910  
## Xi1 -0.7867555 1.0000000 0.5679505 0.5696775  
## Xi2 -0.6029417 0.5679505 1.0000000 0.6705287  
## Xi3 -0.6445910 0.5696775 0.6705287 1.0000000

1. Fit regression model (6.5) for three predictor variables to the data and state the estimated regression function. How is b2 interpreted here?

The regression function would be Yi = 158.4913 - 1.1416(Xi1) -.4420(Xi2) -13.4702(Xi3). The b2 which is -.4420 can be intrepreted as the difference in predicted value for Y for each unit difference in Xi2 if Xi1 and Xi3 remains constant.

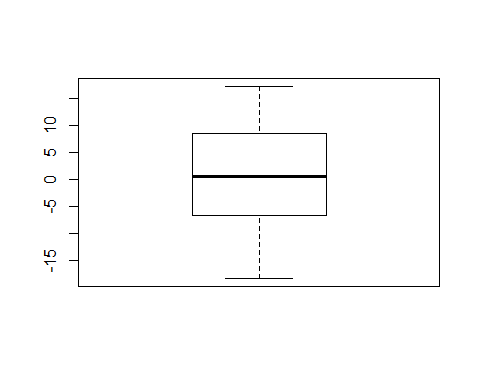
lm1 <- lm(Yi ~., data=q15)  
summary(lm1)

##   
## Call:  
## lm(formula = Yi ~ ., data = q15)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -18.3524 -6.4230 0.5196 8.3715 17.1601   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 158.4913 18.1259 8.744 5.26e-11 \*\*\*  
## Xi1 -1.1416 0.2148 -5.315 3.81e-06 \*\*\*  
## Xi2 -0.4420 0.4920 -0.898 0.3741   
## Xi3 -13.4702 7.0997 -1.897 0.0647 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 10.06 on 42 degrees of freedom  
## Multiple R-squared: 0.6822, Adjusted R-squared: 0.6595   
## F-statistic: 30.05 on 3 and 42 DF, p-value: 1.542e-10

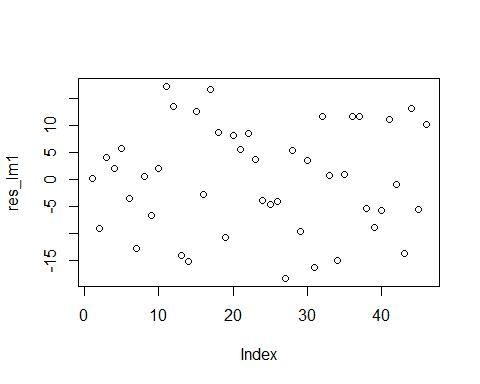
1. Obtain the residuals and prepare a box plot of the residuals. Do there appear to be any outliers?

According to the data there are definitely some outliers which are shown in the min and max. But according to the actual box plot, there doesn’t seem to be any outlie

lm1 <- lm(Yi ~., data=q15)  
res\_lm1 <- resid(lm1)  
boxplot(res\_lm1)



plot(res\_lm1)



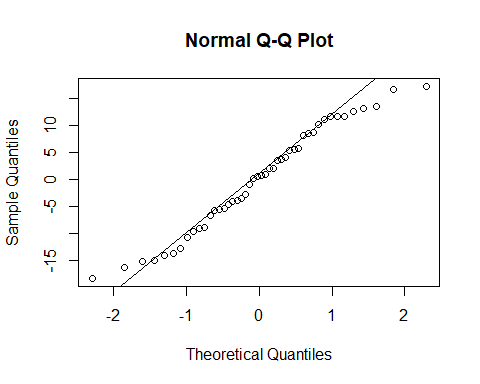
summary(res\_lm1)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -18.3524 -6.4230 0.5196 0.0000 8.3715 17.1601

1. Plot the residuals against Y, each of the predictor variables, and each two-factor interaction term on separate graphs. Also prepare a normal probability plot. Interpret your plots and summarize your findings.

Since the errors lie almost in a straight line, it means the errors may not be iid normal

lm1 <- lm(Yi ~., data=q15)  
res <- resid(lm1)  
res\_lm1 <- resid(lm1)  
qqnorm(res\_lm1)  
qqline(res\_lm1)



1. Can you conduct a formal test for lack of fit here?

No you can not. Or at least I do not know how to.

1. Conduct the Breusch-Pagan test for constancy of the error variance.

For the BP test, we want to have Ho be that the error variance is constant, and Ha to be that the error variance does change. With alpha as 0.1, it means that if the p-value is less than alpha, then the error variance is not constant. But from this test, we see that we can not come up with that conclusion from the test alone as the p-value is greater than alpha.

#install.packages("lmtest")  
library(lmtest)

## Warning: package 'lmtest' was built under R version 3.5.2

## Loading required package: zoo

## Warning: package 'zoo' was built under R version 3.5.2

##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

bptest(lm1)

##   
## studentized Breusch-Pagan test  
##   
## data: lm1  
## BP = 2.5583, df = 3, p-value = 0.4648

detach(q15)

Question 16:

1. Test whether there is a regression relation; use alpha = 0.1. State the alternative, decision, and conclusion. What does your test imply about betas? What is p-value?

The Ho in this case is that there is no relationship between the betas as B1 = B2 = B3. The alternative is that the betas are not equal and they do impact Yi differently. Since our p-value is 1.542e-10, it is less than the alpha of 0.1 which means that there is a difference between the betas in comparison to Y.

lm1 <- lm(Yi ~., data=q15)  
summary(lm1)

##   
## Call:  
## lm(formula = Yi ~ ., data = q15)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -18.3524 -6.4230 0.5196 8.3715 17.1601   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 158.4913 18.1259 8.744 5.26e-11 \*\*\*  
## Xi1 -1.1416 0.2148 -5.315 3.81e-06 \*\*\*  
## Xi2 -0.4420 0.4920 -0.898 0.3741   
## Xi3 -13.4702 7.0997 -1.897 0.0647 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 10.06 on 42 degrees of freedom  
## Multiple R-squared: 0.6822, Adjusted R-squared: 0.6595   
## F-statistic: 30.05 on 3 and 42 DF, p-value: 1.542e-10

1. Obtain joint interval estimates of B1, B2, B3 using a 90 percent family confidence coefficient. Interpret your results

Using a 90% confidence interval, it seems that each of the betas have a rather large difference between the two values for the confidence interval. This is especially true for Xi3 which makes sense with the data as it had numbers that were much smaller than the other two.

confint(lm1, level = 0.9, type = 'Bonferroni')

## 5 % 95 %  
## (Intercept) 128.004370 188.9781330  
## Xi1 -1.502893 -0.7803305  
## Xi2 -1.269467 0.3854587  
## Xi3 -25.411454 -1.5288719

1. Calculate the coefficient of multiple determination. What does it indicate here?

The R-square that I got was 0.6595 which means that approximately 66% of the data can be explained with a linear line of best fit according to this data.

summary(lm1)

##   
## Call:  
## lm(formula = Yi ~ ., data = q15)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -18.3524 -6.4230 0.5196 8.3715 17.1601   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 158.4913 18.1259 8.744 5.26e-11 \*\*\*  
## Xi1 -1.1416 0.2148 -5.315 3.81e-06 \*\*\*  
## Xi2 -0.4420 0.4920 -0.898 0.3741   
## Xi3 -13.4702 7.0997 -1.897 0.0647 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 10.06 on 42 degrees of freedom  
## Multiple R-squared: 0.6822, Adjusted R-squared: 0.6595   
## F-statistic: 30.05 on 3 and 42 DF, p-value: 1.542e-10

Question 17:

1. Obtain an interval estimate ofthe mean satisfaction when Xh1 = 35, Xh2 = 45, Xh3 = 2.2. Use a 90% confidence coefficient. Interpret your confidence interval.
2. Obtain a prediction interval for a new patient’s satisfaction when Xh1 = 35, Xh2 = 45, and Xh3 = 2.2. Use a 90 percent confidence coefficient. Interpret your prediction interval

Question 18:

1. Prepare a stem-and-leaf plot for each predictor variable. What information do these plots provide?

The stem and leaf plot show that the variables are all inconsistent with one another as Xi4 is much bigger than all of the other variables and Xi3 is much smaller than all of the variables. This is something of note when running them all to Yi in the future.

q18 <- read.csv('Ch6Q18.csv')  
attach(q18)  
stem(Xi1, scale = .5)

##   
## The decimal point is 1 digit(s) to the right of the |  
##   
## 0 | 01111111111111112222222222233333333333344444  
## 0 | 68  
## 1 | 112233344444  
## 1 | 555555556666666667888  
## 2 | 00

stem(Xi2, scale = .25)

##   
## The decimal point is 1 digit(s) to the right of the |  
##   
## 0 | 34  
## 0 | 5555556666677788888888899999999  
## 1 | 00000000011111111111111122222222222222333333333  
## 1 | 5

stem(Xi3, scale= .5)

##   
## The decimal point is 1 digit(s) to the left of the |  
##   
## 0 | 00000000000000000000000000000023333333333344444455555566788890234444  
## 2 | 12234773  
## 4 | 7  
## 6 | 03

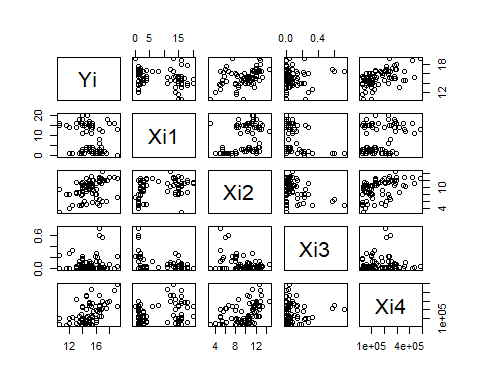
stem(Xi4, scale= 6)

##   
## The decimal point is 4 digit(s) to the right of the |  
##   
## 2 | 7  
## 3 | 012239  
## 4 | 011336  
## 5 | 0179  
## 6 | 000  
## 7 | 024  
## 8 | 125  
## 9 | 067  
## 10 | 11456  
## 11 | 00  
## 12 | 0238  
## 13 | 00006  
## 14 |   
## 15 | 4  
## 16 | 8  
## 17 |   
## 18 | 0459  
## 19 |   
## 20 | 289  
## 21 | 05  
## 22 | 0  
## 23 | 1466  
## 24 | 038  
## 25 | 15  
## 26 | 6  
## 27 | 66  
## 28 | 29  
## 29 | 179  
## 30 |   
## 31 | 5  
## 32 |   
## 33 |   
## 34 |   
## 35 | 0  
## 36 | 06  
## 37 |   
## 38 |   
## 39 |   
## 40 |   
## 41 |   
## 42 | 1  
## 43 | 5  
## 44 |   
## 45 |   
## 46 |   
## 47 |   
## 48 | 4

1. Obtain the scatter plot matrix and the correlation matrix. Interpret these and state your principal findings.

According to the plot and correlation matrix, it seems that Xi3 and Xi4 have no correlation with one another, which is something I thought was true from the stem and leaf plot, since one was so much bigger than the other and the other was much smaller. Also unlike Q15, it seems only Xi1 seems to have a negative correlation with Yi and it seems that Xi4 has the biggest correlation with Yi while Xi3 has the smallest correlation.

pairs(q18)



cor(q18)

## Yi Xi1 Xi2 Xi3 Xi4  
## Yi 1.00000000 -0.2502846 0.4137872 0.06652647 0.53526237  
## Xi1 -0.25028456 1.0000000 0.3888264 -0.25266347 0.28858350  
## Xi2 0.41378716 0.3888264 1.0000000 -0.37976174 0.44069713  
## Xi3 0.06652647 -0.2526635 -0.3797617 1.00000000 0.08061073  
## Xi4 0.53526237 0.2885835 0.4406971 0.08061073 1.00000000

1. Fit regression model (6.5) for four predictor variables to the data. State the estimated regression function.

The regression function would be Yi = 1.22e1 -1.42e-01Xi +2.82e-1X2 +6.193e-1X3 + 7.924e-6X4

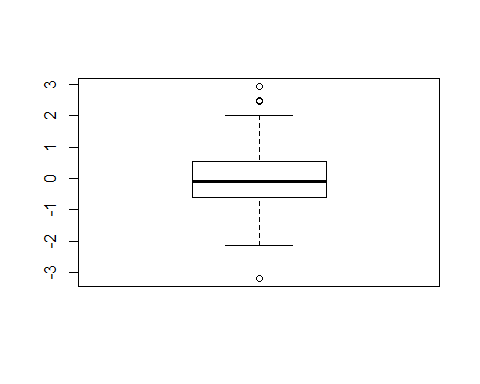
lm1 <- lm(Yi ~., data=q18)  
summary(lm1)

##   
## Call:  
## lm(formula = Yi ~ ., data = q18)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.1872 -0.5911 -0.0910 0.5579 2.9441   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.220e+01 5.780e-01 21.110 < 2e-16 \*\*\*  
## Xi1 -1.420e-01 2.134e-02 -6.655 3.89e-09 \*\*\*  
## Xi2 2.820e-01 6.317e-02 4.464 2.75e-05 \*\*\*  
## Xi3 6.193e-01 1.087e+00 0.570 0.57   
## Xi4 7.924e-06 1.385e-06 5.722 1.98e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.137 on 76 degrees of freedom  
## Multiple R-squared: 0.5847, Adjusted R-squared: 0.5629   
## F-statistic: 26.76 on 4 and 76 DF, p-value: 7.272e-14

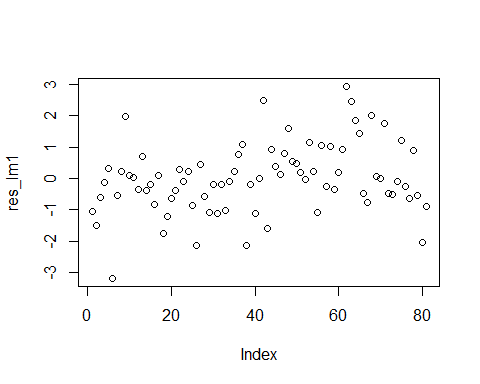
1. Obtain the residuals and prepare a box plot of the residuals. Does the distribution appear to be fairly symmetrical?

The box of the residuals seem to have a few outliers above and below the box plot but when I look at the plot by itself, it seems that most fit within 0,0.

lm1 <- lm(Yi ~., data=q18)  
res\_lm1 <- resid(lm1)  
boxplot(res\_lm1)



plot(res\_lm1)



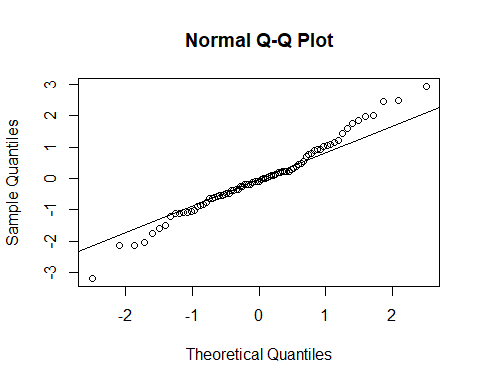
summary(res\_lm1)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -3.18718 -0.59105 -0.09095 0.00000 0.55794 2.94414

1. Plot the residuals against Y, each predictor variable, and each two-factor interaction term on separate graphs. Also prepare a normal probability plot. Analyze yom’ plots and summalize your findings.

As I discovered from part d, it seems that the residuals do fit the line here as it seems symmetrical. So most fall within the qqnormalized line as shown on the data.

lm1 <- lm(Yi ~., data=q18)  
res <- resid(lm1)  
res\_lm1 <- resid(lm1)  
qqnorm(res\_lm1)  
qqline(res\_lm1)



1. Can you conduct a fonnal test for lack of fit here?

For this specific data, it seems there is a fit here for most of the data points as it is symmetrical. So a lot do fit in the QQ plot line as shown from this result.

1. Divide the 81 cases into two groups. placing the 40 cases with the smallest fitted values Yi into group 1 and the remaining cases into group 2. Conduct the Brown-Forsythe test for constancy of the error variance, using ex = .05. State the decision rule and conclusion.

The decision rule would be that Ho represents if the error variance is consistent throughout the betas, and the alternative is if the error variance changes with the betas. According to my result, it seems that the data is not definitive enough to prove that the error variance changes with the differing beta as both are higher than our alpha of 0.05.

library(lmtest)  
library(caret)

## Loading required package: lattice

## Loading required package: ggplot2

set.seed(1234)  
sample <- createDataPartition(q18$Yi,  
 p = 0.5,  
 list = FALSE, groups = min(40, length(q18)))  
min <- q18[ sample ,]  
max <- q18[ -sample ,]  
  
lm1 <- lm(min)  
lm2 <- lm(max)  
  
bptest(lm1)

##   
## studentized Breusch-Pagan test  
##   
## data: lm1  
## BP = 8.853, df = 4, p-value = 0.06488

bptest(lm2)

##   
## studentized Breusch-Pagan test  
##   
## data: lm2  
## BP = 9.3118, df = 4, p-value = 0.05376

Question 19:

1. Test whether there is a regression relation using alpha = 0.05. State alternatives, decision rule, and conclusion.

The Ho in this case is that there is no relationship between the betas as B1 = B2 = B3. The alternative is that the betas are not equal and they do impact Yi differently. Since our p-value is 7.272e-14, it is less than the alpha of 0.05 which means that there is a difference between the betas in comparison to Y. So we reject Ho and we accept the alternative that the betas do impact Yi differently.

lm1 <- lm(Yi ~., data=q18)  
summary(lm1)

##   
## Call:  
## lm(formula = Yi ~ ., data = q18)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.1872 -0.5911 -0.0910 0.5579 2.9441   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.220e+01 5.780e-01 21.110 < 2e-16 \*\*\*  
## Xi1 -1.420e-01 2.134e-02 -6.655 3.89e-09 \*\*\*  
## Xi2 2.820e-01 6.317e-02 4.464 2.75e-05 \*\*\*  
## Xi3 6.193e-01 1.087e+00 0.570 0.57   
## Xi4 7.924e-06 1.385e-06 5.722 1.98e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.137 on 76 degrees of freedom  
## Multiple R-squared: 0.5847, Adjusted R-squared: 0.5629   
## F-statistic: 26.76 on 4 and 76 DF, p-value: 7.272e-14

1. Estimate the Betas by using Bonferroni using a 95% interval. Interpret your findings.

According to the Bonferroni 95% confidence interval, it seems that Xi4 has the biggest difference between the two ends while Xi2 and Xi3 do not have a big difference between the ends. Xi1 also does not seem to have a big difference between the two ends. This makes sense as Xi4 has the biggest numeric values.

confint(lm1, level = 0.95, type = 'bonferroni')

## 2.5 % 97.5 %  
## (Intercept) 1.104949e+01 1.335169e+01  
## Xi1 -1.845411e-01 -9.952615e-02  
## Xi2 1.561979e-01 4.078352e-01  
## Xi3 -1.545232e+00 2.783919e+00  
## Xi4 5.166283e-06 1.068232e-05

1. Calculate R-square and interpret this measure.

According to my linear model, the R-square is 0.5847 which means that 58.47% of the data can be explained by a linear line of best fit for the data.

summary(lm1)

##   
## Call:  
## lm(formula = Yi ~ ., data = q18)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.1872 -0.5911 -0.0910 0.5579 2.9441   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.220e+01 5.780e-01 21.110 < 2e-16 \*\*\*  
## Xi1 -1.420e-01 2.134e-02 -6.655 3.89e-09 \*\*\*  
## Xi2 2.820e-01 6.317e-02 4.464 2.75e-05 \*\*\*  
## Xi3 6.193e-01 1.087e+00 0.570 0.57   
## Xi4 7.924e-06 1.385e-06 5.722 1.98e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.137 on 76 degrees of freedom  
## Multiple R-squared: 0.5847, Adjusted R-squared: 0.5629   
## F-statistic: 26.76 on 4 and 76 DF, p-value: 7.272e-14

detach(q18)

Question 20:

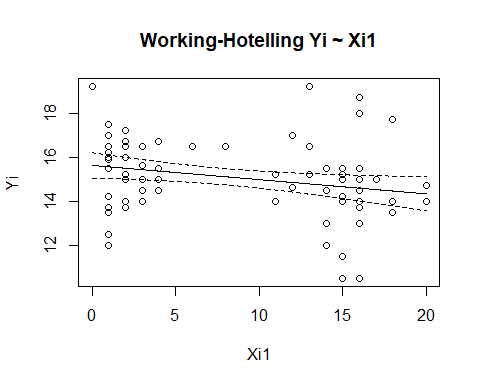
Obtain the family of estimates using a 95 percent family confidence coefficient. Employ the most efficient procedure.

The most efficient procedure in this case was the Working-Hotelling DelTime procedure.

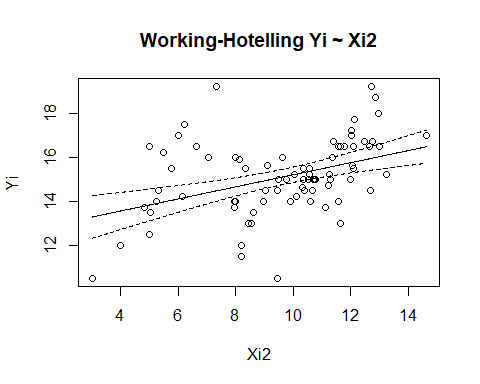
library(investr)

## Warning: package 'investr' was built under R version 3.5.2

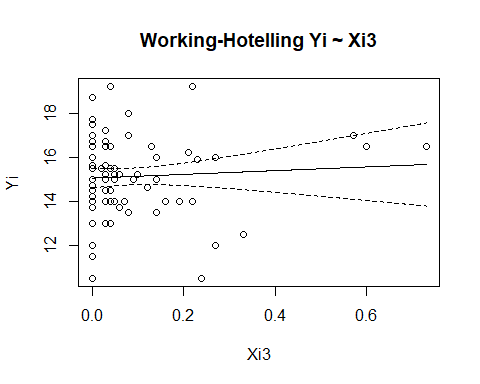
q20 <- read.csv('Ch6Q20.csv')  
lm1 <- lm(Yi~Xi1, data = q18)  
lm2 <- lm(Yi~Xi2, data = q18)  
lm3 <- lm(Yi~Xi3, data = q18)  
lm4 <- lm(Yi~Xi4, data = q18)  
  
plotFit((lm1), interval = 'confidence', k=.95, main = 'Working-Hotelling Yi ~ Xi1')



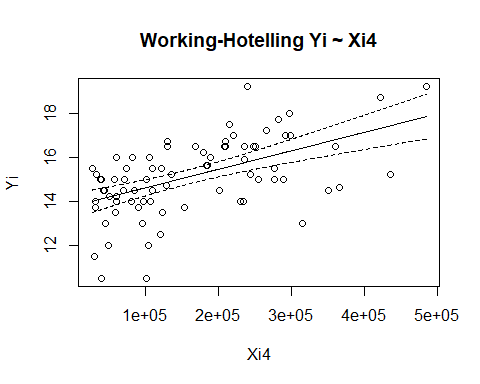
plotFit((lm2), interval = 'confidence', k=.95, main = 'Working-Hotelling Yi ~ Xi2')



plotFit((lm3), interval = 'confidence', k=.95, main = 'Working-Hotelling Yi ~ Xi3')



plotFit((lm4), interval = 'confidence', k=.95, main = 'Working-Hotelling Yi ~ Xi4')



Question 21:

Develop separate prediction intervals for the rental rates of these properties, using 95% statement confidence coefficient in each case. Can the rental rates of these three properties be predicted fairly precisely? What is the family confidence level for the set of three predictions?