300

400

Assn03Q1

# Sumanth Donthula 2022-10-08

Problem 1 1.a)

Yes, a linear relation is being observed in the data by scatter plot.

Data=read.table("AS3Q1Data.txt", header = FALSE, sep = "") Y = Data$V1

X = Data$V2

plot(X,Y,xlab="Sales",ylab="Sales Year")

# 0 2 4 6 8



Sales Year

100

200

Sales

lm1=lm(Y~X) lm1

##

## Call:

## lm(formula = Y ~ X) ##

## Coefficients:

## (Intercept) X

## 91.56 32.50

1.b)

By looking at the box cox plot a lambda of 0.5 is suggested. By SSE box cox plot it is evident that SSE is miniumum at lambda=0.5.

resid(lm1)

95%

20

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ## | 1 | 2 | 3 | 4 | 5 | 6 |
| ## | 6.4363636 | 10.9393939 | 5.4424242 | -11.0545455 | -0.5515152 | -22.0484848 |
| ## | 7 | 8 | 9 | 10 |  |  |
| ## | -3.5454545 | -19.0424242 | 22.4606061 | 10.9636364 |  |  |

library(MASS) boxcox(Y~X)

# −2 −1 0 1 2

log−Likelihood

0

5

10

15



boxcox(Y~X,seq(0,1,0.01))

18

19

# 0.0 0.2 0.4 0.6 0.8 1.0

95%

log−Likelihood

16

17



library('ALSM')

## Loading required package: leaps

## Loading required package: SuppDists ## Loading required package: car

## Loading required package: carData

boxcox.sse(X,Y)

SSE

30000

50000

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| −2 | −1 | 0 | 1 | 2 |
|  |  |  |  |  |

## lambda SSE

0 10000

|  |  |  |  |
| --- | --- | --- | --- |
| ## | 1 | -2.0 | 57788.3511 |
| ## | 2 | -1.9 | 50489.6939 |
| ## | 3 | -1.8 | 44054.1816 |
| ## | 4 | -1.7 | 38379.8733 |
| ## | 5 | -1.6 | 33377.2323 |
| ## | 6 | -1.5 | 28967.6103 |
| ## | 7 | -1.4 | 25081.9206 |
| ## | 8 | -1.3 | 21659.4777 |
| ## | 9 | -1.2 | 18646.9811 |
| ## | 10 | -1.1 | 15997.6263 |
| ## | 11 | -1.0 | 13670.3269 |
| ## | 12 | -0.9 | 11629.0334 |
| ## | 13 | -0.8 | 9842.1378 |
| ## | 14 | -0.7 | 8281.9520 |
| ## | 15 | -0.6 | 6924.2528 |
| ## | 16 | -0.5 | 5747.8831 |
| ## | 17 | -0.4 | 4734.4047 |
| ## | 18 | -0.3 | 3867.7951 |
| ## | 19 | -0.2 | 3134.1829 |
| ## | 20 | -0.1 | 2521.6190 |
| ## | 41 | 0.0 | 2019.8767 |
| ## | 21 | 0.1 | 1620.2804 |
| ## | 22 | 0.2 | 1315.5569 |
| ## | 23 | 0.3 | 1099.7093 |
| ## | 24 | 0.4 | 967.9088 |

|  |  |  |  |
| --- | --- | --- | --- |
| ## | 25 | 0.5 | 916.4048 |
| ## | 26 | 0.6 | 942.4498 |
| ## | 27 | 0.7 | 1044.2384 |
| ## | 28 | 0.8 | 1220.8598 |
| ## | 29 | 0.9 | 1472.2614 |
| ## | 30 | 1.0 | 1799.2242 |
| ## | 31 | 1.1 | 2203.3483 |
| ## | 32 | 1.2 | 2687.0483 |
| ## | 33 | 1.3 | 3253.5588 |
| ## | 34 | 1.4 | 3906.9485 |
| ## | 35 | 1.5 | 4652.1447 |
| ## | 36 | 1.6 | 5494.9660 |
| ## | 37 | 1.7 | 6442.1649 |
| ## | 38 | 1.8 | 7501.4808 |
| ## | 39 | 1.9 | 8681.7016 |
| ## | 40 | 2.0 | 9992.7371 |

1.c)

The linear relation function is : 10.26093+1.076X

Yroot=Yˆ0.5 lm2=lm(Yroot~X) lm2

##

## Call:

## lm(formula = Yroot ~ X) ##

## Coefficients:

## (Intercept) X

## 10.261 1.076

1.d)

Yes, the linear regression seems a good fit on the transformed data.

plot(X,Yroot)+abline(10.261,1.076,col=3)

16

18

20

# 0 2 4 6 8



Yroot

10

12

14

X

## integer(0)

1.e) Sum of residuals by looking at the residual plot is almost 0 which supports this transformation

From Qq plot, qq line does not line up perfectly but it appears to be in linear relation. So we conclude that residuals are normally distributed.

residual=lm2$residuals residual

## 1 2 3 4 5 6

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ## | -0.36143656 | 0.28172678 | 0.31440703 | -0.14814273 0.29997018 -0.41084412 |
| ## | 7 | 8 | 9 | 10 |
| ## | 0.10392174 | -0.47446579 | 0.46781397 | -0.07295049 |

predictors=lm2$fitted.values plot(predictors,residual)

0.0

0.2

0.4

# 10 12 14 16 18 20



residual

−0.4

−0.2

predictors

plot(lm2,which=c(1,2))

Residuals vs Fitted



9

6

8

Residuals

0.0

0.2

0.4

10 12 14 16 18 20

−0.4

Fitted values lm(Yroot ~ X)

Normal Q−Q



9

1

8

Standardized residuals

0.5 1.0 1.5

−1.5 −1.0 −0.5 0.0 0.5 1.0 1.5

−1.5

−0.5

1.f)

The estimated function in original units is: Ybar=(10.261+1.076X)ˆ2

Problem 2:

2.a)

The confidence intervals are

# Theoretical Quantiles lm(Yroot ~ X)

CI45 : 98.6309, 106.9691 CI55 : 88.11124, 93.68876 CI65 : 76.20837, 81.79163

Data2=read.table("AS32Data.txt", header = FALSE, sep = "") Y = Data2$V1

X = Data2$V2 n=length(X)

Xf =cbind(rep(1,n),X)

Ymat=as.matrix(Y) Xmat=as.matrix(Xf)

lm=lm(Y~X) lm

##

## Call:

## lm(formula = Y ~ X) ##

## Coefficients:

## (Intercept) X

## 156.35 -1.19

anova(lm)

## Analysis of Variance Table ##

## Response: Y

## Df Sum Sq Mean Sq F value Pr(>F)

## X 1 11627.5 11627.5 174.06 < 2.2e-16 \*\*\*

## Residuals 58 3874.4 66.8

## ---

## Signif. codes: 0 ’\*\*\*’ 0.001 ’\*\*’ 0.01 ’\*’ 0.05 ’.’ 0.1 ’ ’ 1

summary(lm)

##

## Call:

## lm(formula = Y ~ X) ##

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ## | Residuals: |  | | |
| ## | Min | 1Q Median | 3Q | Max |
| ## | -16.1368 -6.1968 -0.5969 6.7607 23.4731 | | | |
| ## |  | | | |
| ## | Coefficients: | | | |
| ## | Estimate Std. Error t value Pr(>|t|) | | | |
| ## | (Intercept) 156.3466 5.5123 28.36 <2e-16 \*\*\* | | | |
| ## | X -1.1900 0.0902 -13.19 <2e-16 \*\*\* | | | |
| ## | --- | | | |
| ## | Signif. codes: 0 ’\*\*\*’ 0.001 ’\*\*’ 0.01 ’\*’ 0.05 ’.’ 0.1 ’ ’ 1 | | | |
| ## |  | | | |
| ## | Residual standard error: 8.173 on 58 degrees of freedom | | | |
| ## | Multiple R-squared: 0.7501, Adjusted R-squared: 0.7458 | | | |
| ## | F-statistic: 174.1 on 1 and 58 DF, p-value: < 2.2e-16 | | | |

sse = sum((fitted(lm) - Y)ˆ2)

Sigmasquare=sse/n-2

*#Working hotelling method*

X1h=c(1,45) X2h=c(1,55) X3h=c(1,65)

Yhat1=156.35-1.19\*45 Yhat2=156.35-1.19\*55 Yhat3=156.35-1.19\*65

measure1=(Sigmasquare\*t(X1h)%\*%solve(t(Xmat)%\*%Xmat)%\*%X1h)ˆ0.5 measure2=(Sigmasquare\*t(X2h)%\*%solve(t(Xmat)%\*%Xmat)%\*%X2h)ˆ0.5 measure3=(Sigmasquare\*t(X3h)%\*%solve(t(Xmat)%\*%Xmat)%\*%X3h)ˆ0.5

W = sqrt(2 \* qf(p = 0.95, df1 = 2, df2 = n - 2)) W

## [1] 2.512342

conf1up=Yhat1+W\*measure1 conf1lo=Yhat1-W\*measure1 conf2up=Yhat2+W\*measure2 conf2lo=Yhat2-W\*measure2 conf3up=Yhat3+W\*measure3 conf3lo=Yhat3-W\*measure3

conf45=cbind(conf1lo,conf1up) conf45

## [,1] [,2]

## [1,] 98.6309 106.9691

conf55=cbind(conf2lo,conf2up) conf55

## [,1] [,2]

## [1,] 88.11124 93.68876

conf65=cbind(conf3lo,conf3up) conf65

## [,1] [,2]

## [1,] 76.20837 81.79163

2.b) NO the working hotel model is not the most efficient one as its range is wider compared to normal t distributions confidence interval.

For example here t is 1.67 where as w is 2.51 the band will be larger.

t = qt(0.95,nrow(Data2) - 2) t

## [1] 1.671553

2.c)

The confidence intervals are

CI48 : 95.62575, 102.8342 CI59 : 83.61339, 88.66661 CI74 : 64.3607, 72.2193

BX1h=c(1,48) BX2h=c(1,59) BX3h=c(1,74)

Yhat11=156.35-1.19\*48 Yhat21=156.35-1.19\*59 Yhat31=156.35-1.19\*74

Bmeasure1=(Sigmasquare\*t(BX1h)%\*%solve(t(Xmat)%\*%Xmat)%\*%BX1h)ˆ0.5 Bmeasure2=(Sigmasquare\*t(BX2h)%\*%solve(t(Xmat)%\*%Xmat)%\*%BX2h)ˆ0.5 Bmeasure3=(Sigmasquare\*t(BX3h)%\*%solve(t(Xmat)%\*%Xmat)%\*%BX3h)ˆ0.5 B = qt(1-0.05/(2 \* 3), n - 2)

conf11up=Yhat11+B\*Bmeasure1 conf11lo=Yhat11-B\*Bmeasure1 conf22up=Yhat21+B\*Bmeasure2 conf22lo=Yhat21-B\*Bmeasure2 conf33up=Yhat31+B\*Bmeasure3 conf33lo=Yhat31-B\*Bmeasure3 conf48=cbind(conf11lo,conf11up) conf48

## [,1] [,2]

## [1,] 95.62575 102.8342

conf59=cbind(conf22lo,conf22up) conf59

## [,1] [,2]

## [1,] 83.61339 88.66661

conf74=cbind(conf33lo,conf33up) conf74

## [,1] [,2]

## [1,] 64.3607 72.2193

Problem 3 3.a)

The observations from plots provided that there are no outliers and the distribution of each variable is normal.

Correlation matrix shows Y and X1 have significant positive correlation, Y and X2 are positively correlated, but less than Y and X1 and there’s no correlation between X1 and X2.

5

6

7

8

9

Data3=read.table("AS33Data.txt", header = FALSE, sep = "") Y = Data3$V1

X1= Data3$V2 X2=Data3$V3

pairs(~Y+X1+X2)

4 5 6 7 8 9 10

60 70 80 90

Y

60 70 80 90 100 2.0 2.5 3.0 3.5 4.0

X2

X1

4

2.0

3.0

4.0

colnames(Data3)=c("Y","X1","X2") cor(Data3)

## Y X1 X2 ## Y 1.0000000 0.8923929 0.3945807

## X1 0.8923929 1.0000000 0.0000000

## X2 0.3945807 0.0000000 1.0000000

3.b)The regression model is Y= 37.65 + 4.425X1 + 4.375X2. Holding the other variable constant, Increasing one unit of X1 leads to an increase in the brand liking by 4.425, and holding X1 constant, an one unit increase in X2 leads to an increase of the brand by 4.375.

Lmodel=lm(Y~X1+X2) summary(Lmodel)

##

## Call:

## lm(formula = Y ~ X1 + X2) ##

## 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) 37.6500 2.9961 12.566 1.20e-08 \*\*\*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ## | X1 | 4.4250 | 0.3011 14.695 | 1.78e-09 | \*\*\* |
| ## | X2 | 4.3750 | 0.6733 6.498 | 2.01e-05 | \*\*\* |
| ## | --- |  |  |  |  |

## Signif. codes: 0 ’\*\*\*’ 0.001 ’\*\*’ 0.01 ’\*’ 0.05 ’.’ 0.1 ’ ’ 1 ##

## Residual standard error: 2.693 on 13 degrees of freedom ## Multiple R-squared: 0.9521, Adjusted R-squared: 0.9447 ## F-statistic: 129.1 on 2 and 13 DF, p-value: 2.658e-09

*#Y=37.65+4.25X1+4.375X2*

3.c)

There are no outliers in the residuals and errors are normally distributed.

Lmodel$residual ## 1 2 3 4 5 6 7 8 9 10 11 12 13

|  |  |  |  |
| --- | --- | --- | --- |
| ## | -0.10 | 0.15 | -3.10 3.15 -0.95 -1.70 -1.95 1.30 1.20 -1.55 4.20 2.45 -2.65 |
| ## | 14 | 15 | 16 |
| ## | -4.40 | 3.35 | 0.60 |

boxplot(Lmodel$residual)

−2

0

2

4

3.d)

−4

Based on the plots, we observe that residuals are random and almost normally distributed with mean 0.

Yhat=37.65+4.25\*X1+4.375\*X2

plot(Yhat,Lmodel$residual)

0

2

4

# 65 70 75 80 85 90 95



Lmodel$residual

−4

−2

Yhat

plot(X1,Lmodel$residual)

0

2

4

# 4 5 6 7 8 9 10



Lmodel$residual

−4

−2

X1

plot(X2,Lmodel$residual)

0

2

4

# 2.0 2.5 3.0 3.5 4.0



Lmodel$residual

−4

−2

X2

plot(X1\*X2,Lmodel$residual)

0

2

4

# 10 15 20 25 30 35 40



Lmodel$residual

−4

−2

X1 \* X2

qqnorm(Lmodel$residual) qqline(Lmodel$residual)

**Normal Q−Q Plot**

0

2

4

# −2 −1 0 1 2



Sample Quantiles

−4

−2

Theoretical Quantiles

3.e)

Ho: Error variance is constant Ha: Error variance is not constant

The p value of the Breusch Pagan test is 0.3599 which is greater than alpha 0.05 so we reject the null hypothesis.

library(lmtest)

## Loading required package: zoo ##

## Attaching package: ’zoo’

## The following objects are masked from ’package:base’: ##

## as.Date, as.Date.numeric

Lmodel2=lm(log((Lmodel$residuals)ˆ2)~X1+X2) bptest(Lmodel2)

##

## studentized Breusch-Pagan test ##

## data: Lmodel2

## BP = 5.0338, df = 2, p-value = 0.08071

3.f)

Ho: Linear Model fits the Data(Y=b0+b1X1+b2X2) Ha: There is lack of fit in the model(Y<>b0+b1X1+b2X2) Since the P test value of X1 and X2 are greater than 0.01,so we reject the null hypothesis.

*#Lmodel=lm(Y~X1+X2)*

anova(Lmodel2,Lmodel)

## Warning in anova.lmlist(object, ...): models with response ’"Y"’ removed because ## response differs from model 1

## Analysis of Variance Table ##

## Response: log((Lmodel$residuals)^2)

## Df Sum Sq Mean Sq F value Pr(>F)

|  |  |  |
| --- | --- | --- |
| ## | X1 | 1 14.058 14.0580 3.1016 0.1017 |
| ## | X2 | 1 0.202 0.2015 0.0445 0.8363 |
| ## | Residuals | 13 58.923 4.5325 |

Problem 4 4.a)

Stem and leaf represents the histograms of quantitative data.

Data4=read.table("As34Data.txt", header = FALSE, sep = "") Y = Data4$V1

X1 = Data4$V2 X2 = Data4$V3 X3 = Data4$V4 X4 = Data4$V5

Xmat=as.matrix(cbind(rep(1,length(Y)),X1,X2,X3,X4)) stem(X1)

##

## The decimal point is at the | ##

|  |  |  |  |
| --- | --- | --- | --- |
| ## | 0 | | | 0000000000000000 |
| ## | 2 | | | 00000000000000000000000 |
| ## | 4 | | | 00000 |
| ## | 6 | | | 0 |
| ## | 8 | | | 0 |
| ## | 10 | | | 00 |
| ## | 12 | | | 00000 |
| ## | 14 | | | 0000000000000 |
| ## | 16 | | | 0000000000 |
| ## | 18 | | | 000 |
| ## | 20 | | | 00 |

stem(X2)

##

## The decimal point is at the | ##

|  |  |
| --- | --- |
| ## | 2 | 0 |
| ## | 4 | 080003358 |
| ## | 6 | 012613 |
| ## | 8 | 00001223456001555689 |
| ## | 10 | 013344566677778123344666668 |
| ## | 12 | 00011115777889002 |
| ## | 14 | 6 |

stem(X3) ##

## The decimal point is 1 digit(s) to the left of the | ##

## 0 | 0000000000000000000000000000002333333333334444445555556678889

## 1 | 023444469

## 2 | 1223477

## 3 | 3

## 4 |

## 5 | 7

## 6 | 0

## 7 | 3

stem(X4)

##

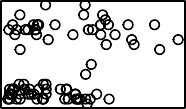
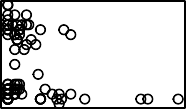
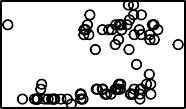
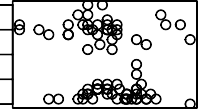
## The decimal point is 5 digit(s) to the right of the | ##

|  |  |  |  |
| --- | --- | --- | --- |
| ## | 0 | | | 333333444444 |
| ## | 0 | | | 555666667778899 |
| ## | 1 | | | 000001111222333334 |
| ## | 1 | | | 578889 |
| ## | 2 | | | 011122334444 |
| ## | 2 | | | 555788899 |
| ## | 3 | | | 002 |
| ## | 3 | | | 567 |
| ## | 4 | | | 23 |
| ## | 4 | | | 8 |

4.b)

From observing the correlation matrix we can see that there is no strong correlation between the predictor and response variables.

pairs(~Y+X1+X2+X3+X4)

0 5 10 15 20 0.0 0.4

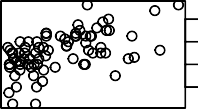
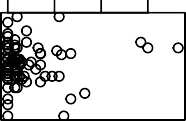
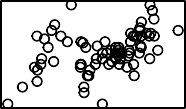
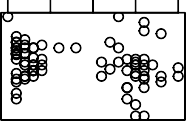
X1

0

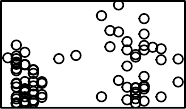
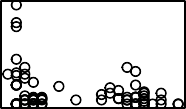
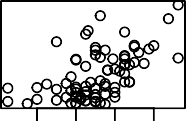
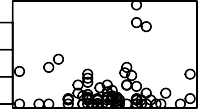
10

20

12 16

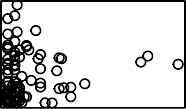
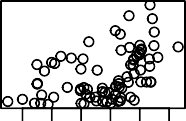
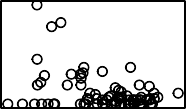


Y

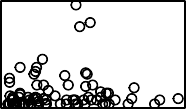
12 16

0.0

0.4

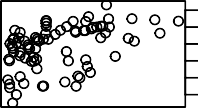
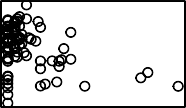
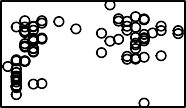
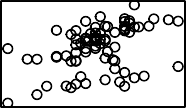
4 8 12

X3

1e+05 4e+05

X4

1e+05 5e+05



X2

4 8

14

colnames(Data4)=c("Y","X1","X2","X3","X4")

cor(Data4)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ## | Y | X1 | X2 | X3 | X4 |
| ## Y | 1.00000000 | -0.2502846 | 0.4137872 | 0.06652647 | 0.53526237 |
| ## X1 | -0.25028456 | 1.0000000 | 0.3888264 | -0.25266347 | 0.28858350 |
| ## X2 | 0.41378716 | 0.3888264 | 1.0000000 | -0.37976174 | 0.44069713 |
| ## X3 | 0.06652647 | -0.2526635 | -0.3797617 | 1.00000000 | 0.08061073 |
| ## X4 | 0.53526237 | 0.2885835 | 0.4406971 | 0.08061073 | 1.00000000 |

4.c)

The regression model is Y=12.22-0.14*X1+0.28* X2+0.61*X3+7.92e-06* X4

Lmod=lm(Y~X1+X2+X3+X4)

Lmod

##

## Call:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ## lm(formula = Y ~ X1 + X2 | + | X3 + X4) |  | |
| ## |  |  |
| ## Coefficients: |  |  |
| ## (Intercept) X1 |  | X2 | X3 | X4 |
| ## 1.220e+01 -1.420e-01 |  | 2.820e-01 | 6.193e-01 | 7.924e-06 |

4.d)

There are some outliers in the data and the plot is noy symmetrical and the assumption made on noise of regression is wrong.

Lmod$residuals

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ## | 1 | 2 | 3 | 4 | 5 | 6 |
| ## | -1.035672440 | -1.513806414 | -0.591053402 | -0.133568082 | 0.313283765 | -3.187185224 |
| ## | 7 | 8 | 9 | 10 | 11 | 12 |
| ## | -0.538356749 | 0.236302386 | 1.989220372 | 0.105829603 | 0.023124830 | -0.337070751 |
| ## | 13 | 14 | 15 | 16 | 17 | 18 |
| ## | 0.717869468 | -0.392411015 | -0.201019573 | -0.814937024 | 0.101690072 | -1.759131637 |
| ## | 19 | 20 | 21 | 22 | 23 | 24 |
| ## | -1.210114916 | -0.634341765 | -0.366004170 | 0.288596123 | -0.093200248 | 0.233884284 |
| ## | 25 | 26 | 27 | 28 | 29 | 30 |
| ## | -0.853339941 | -2.123934469 | 0.466014057 | -0.573974675 | -1.068826727 | -0.197717691 |
| ## | 31 | 32 | 33 | 34 | 35 | 36 |
| ## | -1.121737177 | -0.173906919 | -1.030125636 | -0.090953654 | 0.215053952 | 0.784804746 |
| ## | 37 | 38 | 39 | 40 | 41 | 42 |
| ## | 1.083920373 | -2.132451269 | -0.185470952 | -1.120385453 | -0.012771680 | 2.500938643 |
| ## | 43 | 44 | 45 | 46 | 47 | 48 |
| ## | -1.582833452 | 0.929599530 | 0.394236721 | 0.117200255 | 0.815339787 | 1.605896564 |
| ## | 49 | 50 | 51 | 52 | 53 | 54 |
| ## | 0.557941960 | 0.494737472 | 0.207611404 | -0.032045798 | 1.155796537 | 0.234272601 |
| ## | 55 | 56 | 57 | 58 | 59 | 60 |
| ## | -1.073489739 | 1.059646672 | -0.261711555 | 1.031651273 | -0.345957207 | 0.203372872 |
| ## | 61 | 62 | 63 | 64 | 65 | 66 |
| ## | 0.917961126 | 2.944144932 | 2.459696482 | 1.859088749 | 1.451807658 | -0.483857748 |
| ## | 67 | 68 | 69 | 70 | 71 | 72 |
| ## | -0.756250356 | 2.011402309 | 0.078550427 | 0.009892809 | 1.766898426 | -0.463930876 |
| ## | 73 | 74 | 75 | 76 | 77 | 78 |
| ## | -0.510410866 | -0.106354746 | 1.209427169 | -0.261085606 | -0.627547725 | 0.910085787 |
| ## | 79 | 80 | 81 |  |  |  |
| ## | -0.550846871 | -2.030180944 | -0.906819056 |  |  |  |

boxplot(Lmod$residuals)

−2

−1

0

1

2

3

4.e)

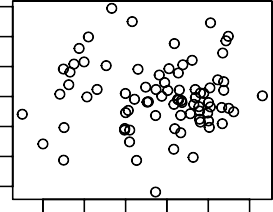
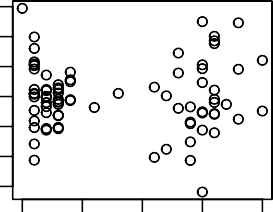
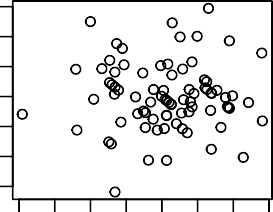


−3

Based on the residual plots we find that residuals are not uniform with mean 0.

par(mfrow=c(2,3)) plot(Lmod$fitted.values,Lmod$residual) plot(X1,Lmod$residual) plot(X2,Lmod$residual) plot(X3,Lmod$residual) plot(X4,Lmod$residual)

plot(X1,Lmod$residual)

11 13 15 17

Lmod$residual

−3

−1

1 2

3

Lmod$residual

−3

−1

1 2

3

Lmod$residual

−3

−1

1 2

3

0 5 10 15 20

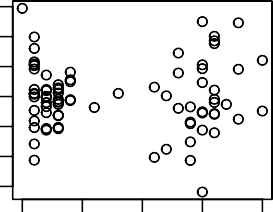
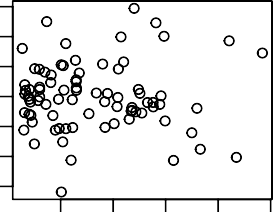
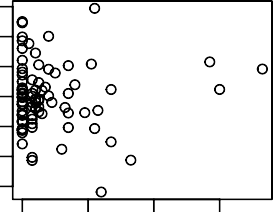
4 6 8 10 14

Lmod$fitted.values X1 X2

3

3

3

0.0 0.2 0.4 0.6

Lmod$residual

−3

−1

1 2

Lmod$residual

−3

−1

1 2

Lmod$residual

−3

−1

1 2

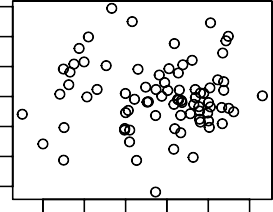
1e+05 3e+05 5e+05

0 5 10 15 20

X3 X4 X1

plot(X2,Lmod$residual) plot(X3,Lmod$residual) plot(X1\*X2,Lmod$residual) plot(X2\*X3,Lmod$residual) plot(X3\*X1,Lmod$residual)

plot(Lmod,which=2)

4 6 8 10 14

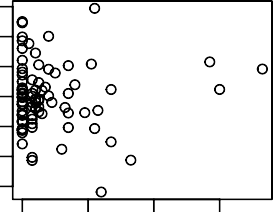
Lmod$residual

−3

−1

1 2

3

0.0 0.2 0.4 0.6

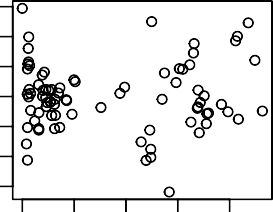
Lmod$residual

−3

−1

1 2

3

0 50 100 200

Lmod$residual

−3

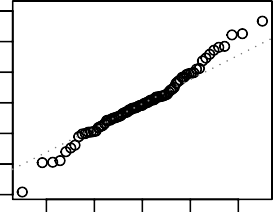
−1

1 2

3

X2 X3 X1 \* X2

# Normal Q−Q



62

42

6

Lmod$residual

1 2

3

Lmod$residual

1 2

3

Standardized residuals

1 2 3

0 1 2 3 4

−3

−1

−3

−1

−3

−1

0 1 2 3

−2 −1 0 1 2

X2 \* X3

X3 \* X1

Theoretical Quantiles

4.f)

From anova of Lmod we find that p value of coeficienr of X3 is less than 0.05 which implies X3 does not fit the model and X3=0.

The model can be Y~X1+X2+X4

anova(Lmod)*#Y~X1+X2+X3+X4*

## Analysis of Variance Table ##

## Response: Y

## Df Sum Sq Mean Sq F value Pr(>F)

## X1 1 14.819 14.819 11.4649 0.001125 \*\*

## X2 1 72.802 72.802 56.3262 9.699e-11 \*\*\*

## X3 1 8.381 8.381 6.4846 0.012904 \*

## X4 1 42.325 42.325 32.7464 1.976e-07 \*\*\*

## Residuals 76 98.231 1.293 ## ---

## Signif. codes: 0 ’\*\*\*’ 0.001 ’\*\*’ 0.01 ’\*’ 0.05 ’.’ 0.1 ’ ’ 1

Lmod2=lm(Y~X1+X2+X4)

anova(Lmod2,Lmod)

## Analysis of Variance Table

##

## Model 1: Y ~ X1 + X2 + X4

## Model 2: Y ~ X1 + X2 + X3 + X4

## Res.Df RSS Df Sum of Sq F Pr(>F) ## 1 77 98.650

## 2 76 98.231 1 0.41975 0.3248 0.5704

4.g)

H0: Error variance is constant Ha: Error variance is not constant

tstar>t so we reject null hypothesis. so the assumption error variance is constant is true.

Yhat=sort(fitted(Lmod2))

Y1=Yhat[1:40] Y2=Yhat[41:81]

e1=Y1-Y[1:40]

Med1=median(e1)

e2=Y2-Y[41:81]

Med2=median(e2)

n1=length(Y1) n2=length(Y2)

d1=e1-Med1 Mean1=mean(d1) d2=e2-Med2 Mean2=mean(d2)

s=sqrt(sum((d1-Mean1)ˆ2)+sum((d2-Mean2)ˆ2)/(length(Yhat)-2)) tstar=(Mean1-Mean2)/(s\*sqrt((1/n1)+(1/n2)))

tstar

## [1] 0.09588115

t=qt(0.05,df=n1+n2-2) t

## [1] -1.664371

Problem 5 5.a)

H0: b1=b2=b3=b4=0 (all the coeficients are 0) Ha: At least one of the coef is not 0

Since the p value of beta tests for X1, X2 , X3 and X4 are less than 0.05 we reject null hypothesis.

*#Q5*

anova(Lmod)

## Analysis of Variance Table ##

## Response: Y

## Df Sum Sq Mean Sq F value Pr(>F)

## X1 1 14.819 14.819 11.4649 0.001125 \*\*

## X2 1 72.802 72.802 56.3262 9.699e-11 \*\*\*

## X3 1 8.381 8.381 6.4846 0.012904 \*

## X4 1 42.325 42.325 32.7464 1.976e-07 \*\*\*

## Residuals 76 98.231 1.293 ## ---

## Signif. codes: 0 ’\*\*\*’ 0.001 ’\*\*’ 0.01 ’\*’ 0.05 ’.’ 0.1 ’ ’ 1

5.b)

The confidence intervals of betas are :

Beta1: (-0.19663959, -0.08742769) Beta2: (0.1203875, 0.4436456) Beta3: (-2.161312, 3.399999) Beta4: (4.381297e-06, 1.146731e-05)

n=length(Y) alpha=1 - 0.95 g=4

t=qt(1 - alpha/(2 \* g), n - 4 - 1) t

## [1] 2.558541

beta1 = coef(summary(Lmod))[,1][[2]]

beta2 = coef(summary(Lmod))[,1][[3]]

beta3 = coef(summary(Lmod))[,1][[4]]

beta4 = coef(summary(Lmod))[,1][[5]]

sebeta1 = coef(summary(Lmod))[,2][[2]]

sebeta2 = coef(summary(Lmod))[,2][[3]]

sebeta3 = coef(summary(Lmod))[,2][[4]]

sebeta4 = coef(summary(Lmod))[,2][[5]]

CIbeta1 = c(beta1 - t \* sebeta1, beta1 + t \* sebeta1) CIbeta2 = c(beta2 - t \* sebeta2, beta2 + t \* sebeta2) CIbeta3 = c(beta3 - t \* sebeta3, beta3 + t \* sebeta3) CIbeta4 = c(beta4 - t \* sebeta4, beta4 + t \* sebeta4)

CIbeta1

## [1] -0.19663959 -0.08742769

CIbeta2

## [1] 0.1203875 0.4436456

CIbeta3

## [1] -2.161312 3.399999

CIbeta4

## [1] 4.381297e-06 1.146731e-05

5.c)

The value of Rsquare is 0.58 the variation of model explains 58% variation in Y wrt X.

sse = sum((fitted(Lmod) - Y)ˆ2) sst=sum((Y-mean(Y))ˆ2)

Rsquare=1-(sse/sst) Rsquare

## [1] 0.5847496

Problem 6

The family of estimates of coeficients is CIb1: 145.7784, -126.7568 CIb2: 120.6433, -104.2315 CIb3: 4.627043,

-4.645127 CIb4:-4.645127, 4.627043

*#Q6*

Data42=read.table("As36.txt", header = FALSE, sep = "")

## Warning in read.table("As36.txt", header = FALSE, sep = ""): incomplete final ## line found by readTableHeader on ’As36.txt’

n=nrow(Data42) Xh1=t(cbind(rep(1,n),t(Data42$V1)))

## Warning in cbind(rep(1, n), t(Data42$V1)): number of rows of result is not a ## multiple of vector length (arg 1)

Xh2=t(cbind(rep(1,n),t(Data42$V2)))

## Warning in cbind(rep(1, n), t(Data42$V2)): number of rows of result is not a ## multiple of vector length (arg 1)

Xh3=t(cbind(rep(1,n),t(Data42$V3)))

## Warning in cbind(rep(1, n), t(Data42$V3)): number of rows of result is not a ## multiple of vector length (arg 1)

Xh4=t(cbind(rep(1,n),t(Data42$V4)))

## Warning in cbind(rep(1, n), t(Data42$V4)): number of rows of result is not a ## multiple of vector length (arg 1)

beta0=coef(summary(Lmod))[,1][[2]]

Bmat=cbind(beta0,beta1,beta2,beta3,beta4) Bmat=as.matrix(Bmat)

Yhat1=Bmat%\*%Xh1; Yhat2=Bmat%\*%Xh2; Yhat3=Bmat%\*%Xh3; Yhat4=Bmat%\*%Xh4;

sse = sum((fitted(Lmod) - Y)ˆ2)

Sigmasquare=sse/n-2 measure1=(Sigmasquare\*t(Xh1)%\*%solve(t(Xmat)%\*%Xmat)%\*%Xh1)ˆ0.5 measure2=(Sigmasquare\*t(Xh2)%\*%solve(t(Xmat)%\*%Xmat)%\*%Xh2)ˆ0.5 measure3=(Sigmasquare\*t(Xh3)%\*%solve(t(Xmat)%\*%Xmat)%\*%Xh3)ˆ0.5 measure4=(Sigmasquare\*t(Xh4)%\*%solve(t(Xmat)%\*%Xmat)%\*%Xh4)ˆ0.5 W = sqrt(2 \* qf(p = 0.95, df1 = 5, df2 = length(Y) - 5))

conf1up=Yhat1+W\*measure1 conf1lo=Yhat1-W\*measure1 conf2up=Yhat2+W\*measure2 conf2lo=Yhat2-W\*measure2 conf3up=Yhat3+W\*measure3 conf3lo=Yhat3-W\*measure3 conf4up=Yhat3+W\*measure3 conf4lo=Yhat3-W\*measure3 c(conf1lo,conf1up)

## [1] -126.7568 145.7784

c(conf2lo,conf2up) ## [1] -104.2315 120.6433

c(conf3lo,conf3up)

## [1] -4.645127 4.627043

c(conf4lo,conf4up)

## [1] -4.645127 4.627043

Problem 7 7.a)

Transforming the data and fitting the model

Ycor = sqrt(1/(length(Y)-1))\*((Y-mean(Y))/sd(Y))

X1cor = sqrt(1/(length(X1)-1))\*((X1-mean(X1))/sd(X1))

X2cor = sqrt(1/(length(X2)-1))\*((X2-mean(X2))/sd(X2))

X3cor = sqrt(1/(length(X3)-1))\*((X3-mean(X3))/sd(X3))

X4cor = sqrt(1/(length(X4)-1))\*((X4-mean(X4))/sd(X4))

Lmodel=lm(Ycor~-1+X1cor+X2cor+X3cor+X4cor) Lmodel

##

## Call:

## lm(formula = Ycor ~ -1 + X1cor + X2cor + X3cor + X4cor) ##

|  |  |  |  |
| --- | --- | --- | --- |
| ## | Coefficients: |  | |
| ## | X1cor X2cor | X3cor | X4cor |
| ## | -0.54785 0.42365 | 0.04846 | 0.50276 |

Lmod*#Y~X1+X2+X3+X4*

##

## Call:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ## lm(formula = Y ~ X1 + X2 | + | X3 + X4) |  | |
| ## |  |  |
| ## Coefficients: |  |  |
| ## (Intercept) X1 |  | X2 | X3 | X4 |
| ## 1.220e+01 -1.420e-01 |  | 2.820e-01 | 6.193e-01 | 7.924e-06 |

7.b)

The Standardization coef beta hat after transformation becomes : Betahat2=Sy/Sk X beta2hatstar

Betahat2=(sd(Y)/sd(X2))\*0.423*#value of beta from Lmodel*

Betahat2

## [1] 0.2815859

7.c)

The Standardization coef beta hat after transformatio becomes : Betahatk=Sy/Sk X betakhatstar

Betahat1=(sd(Y)/sd(X1))\*-0.547 Betahat2=(sd(Y)/sd(X2))\*0.423 Betahat3=(sd(Y)/sd(X3))\*0.048 Betahat4=(sd(Y)/sd(X4))\*0.502

Betahat1

## [1] -0.1418126

Betahat2

## [1] 0.2815859

Betahat3

## [1] 0.6134472

Betahat4

## [1] 7.912368e-06

Problem 8 8.a)

The regression model is Y=50.775 + 4.425X1

Y = Data3$Y X1= Data3$X1 X2=Data3$X2

Linearmod=lm(Y~X1) Linearmod

##

## Call:

## lm(formula = Y ~ X1) ##

## Coefficients:

## (Intercept) X1

## 50.775 4.425

8.b)

We have observed that the coefficient of moisture in model 6.5b is equal to that of the moisture coefficient in this model

lm(Y~X1+X2)*#model in 6.5b*

##

## Call:

## lm(formula = Y ~ X1 + X2) ##

## Coefficients:

## (Intercept) X1 X2 ## 37.650 4.425 4.375

8.c)

From anova table of the models #SSR(X1|X2) = SSR(X1)

anova(Linearmod)

## Analysis of Variance Table ##

## Response: Y

## Df Sum Sq Mean Sq F value Pr(>F)

## X1 1 1566.45 1566.45 54.751 3.356e-06 \*\*\*

## Residuals 14 400.55 28.61

## ---

## Signif. codes: 0 ’\*\*\*’ 0.001 ’\*\*’ 0.01 ’\*’ 0.05 ’.’ 0.1 ’ ’ 1

anova(lm(Y~X1+X2))

## Analysis of Variance Table ##

## Response: Y

## Df Sum Sq Mean Sq F value Pr(>F)

## X1 1 1566.45 1566.45 215.947 1.778e-09 \*\*\*

## X2 1 306.25 306.25 42.219 2.011e-05 \*\*\*

## Residuals 13 94.30 7.25

## ---

## Signif. codes: 0 ’\*\*\*’ 0.001 ’\*\*’ 0.01 ’\*’ 0.05 ’.’ 0.1 ’ ’ 1

*#from anova table*

Ssrx1\_x2=1566.45+306.25-306.25

Ssrx1\_x2

## [1] 1566.45

8.d)

Based on (b) and (c), and also the correlation matrix in Problem 6.5(a) confirms that there is a strong linear relationship between response variable and moisture content X1.

Problem 9 9.a)

By plotting the graph the relation did not appear the same for both the populations.

Data9=read.table("AS39Data.txt", header = FALSE, sep = "") Y = Data9$V1

X1 = Data9$V2 X2=Data9$V3

library(ggplot2)

ggplot(Data9, aes(X1, Y, colour = as.factor(X2))) + geom\_point()

90

## as.factor(X2)

80 0

Y

1

70

70.0 72.5 75.0 77.5

## X1

plot(Y,X1+X2)

76

78

80

# 65 70 75 80 85 90 95



X1 + X2

68

70

72

74

Y

9.b)

H0:All the coeficients are zero Ha : At least one of the coefficient is not zero since, Fstar > F-ratio i.e 18.65>3.15, therefore, we reject null hypothesis.

fit = lm(Y~X1+X2+X1\*X2) fit

##

## Call:

## lm(formula = Y ~ X1 + X2 + X1 \* X2) ##

## Coefficients:

|  |  |  |  |
| --- | --- | --- | --- |
| ## (Intercept) | X1 | X2 | X1:X2 |
| ## -126.905 | 2.776 | 76.022 | -1.107 |

summary(fit) ##

## Call:

## lm(formula = Y ~ X1 + X2 + X1 \* X2) ##

## Residuals:

## Min 1Q Median 3Q Max ## -10.8470 -2.1639 0.0913 1.9348 9.9836

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) -126.9052 14.7225 -8.620 4.33e-12 \*\*\*

## X1 2.7759 0.1963 14.142 < 2e-16 \*\*\*

## X2 76.0215 30.1314 2.523 0.01430 \*

## X1:X2 -1.1075 0.4055 -2.731 0.00828 \*\* ## ---

## Signif. codes: 0 ’\*\*\*’ 0.001 ’\*\*’ 0.01 ’\*’ 0.05 ’.’ 0.1 ’ ’ 1 ##

## Residual standard error: 3.893 on 60 degrees of freedom ## Multiple R-squared: 0.8233, Adjusted R-squared: 0.8145 ## F-statistic: 93.21 on 3 and 60 DF, p-value: < 2.2e-16

fit$coef

|  |  |  |  |
| --- | --- | --- | --- |
| ## (Intercept) | X1 | X2 | X1:X2 |
| ## -126.905171 | 2.775898 | 76.021532 | -1.107482 |

anova(fit) ## Analysis of Variance Table

##

## Response: Y

## Df Sum Sq Mean Sq F value Pr(>F)

## X1 1 3670.9 3670.9 242.2760 < 2.2e-16 \*\*\*

## X2 1 453.1 453.1 29.9073 9.282e-07 \*\*\*

## X1:X2 1 113.0 113.0 7.4578 0.008281 \*\*

## Residuals 60 909.1 15.2

## ---

## Signif. codes: 0 ’\*\*\*’ 0.001 ’\*\*’ 0.01 ’\*’ 0.05 ’.’ 0.1 ’ ’ 1

fit = lm(Y~X1) anova(fit)

## Analysis of Variance Table ##

## Response: Y

## Df Sum Sq Mean Sq F value Pr(>F)

## X1 1 3670.9 3670.9 154.28 < 2.2e-16 \*\*\*

## Residuals 62 1475.3 23.8

## ---

## Signif. codes: 0 ’\*\*\*’ 0.001 ’\*\*’ 0.01 ’\*’ 0.05 ’.’ 0.1 ’ ’ 1

fit1 = update(fit,.~.+X2+X1\*X2) anova(fit1)

## Analysis of Variance Table ##

## Response: Y

## Df Sum Sq Mean Sq F value Pr(>F)

## X1 1 3670.9 3670.9 242.2760 < 2.2e-16 \*\*\*

## X2 1 453.1 453.1 29.9073 9.282e-07 \*\*\*

## X1:X2 1 113.0 113.0 7.4578 0.008281 \*\*

## Residuals 60 909.1 15.2

## ---

## Signif. codes: 0 ’\*\*\*’ 0.001 ’\*\*’ 0.01 ’\*’ 0.05 ’.’ 0.1 ’ ’ 1

nrow(Data)

## [1] 10

*#From anova table*

*#SSR(X\_2, X\_1X\_2|X\_1) = SSR(X\_2, X\_1X\_2, X\_1) - SSR(X\_1) #=3670.9 + 453.1 + 113.0 - 3670.9*

SSR=3670.9 + 453.1 + 113.0 - 3670.9 MSEf=909.1/60

DofF=3 DofP=1

Fstar=(SSR/(DofF-DofP))/(909.1/60) Fstar

## [1] 18.68111

qf(0.95,2,60)

## [1] 3.150411

9.c)

The nature of difference between two models is linear that is Y=76.021+1.102X

Y11=Y[X2==1] Y12=Y[X2==0] X11=X1[X2==1] X10=X1[X2==0]

LinMod1=lm(Y11~X11) LinMod2=lm(Y12~X10)

LinMod1

##

## Call:

## lm(formula = Y11 ~ X11) ##

## Coefficients:

## (Intercept) X11

## -50.884 1.668

*#Y=-50.884+1.668X*

LinMod2

##

## Call:

## lm(formula = Y12 ~ X10) ##

## Coefficients:

## (Intercept) X10

## -126.905 2.776

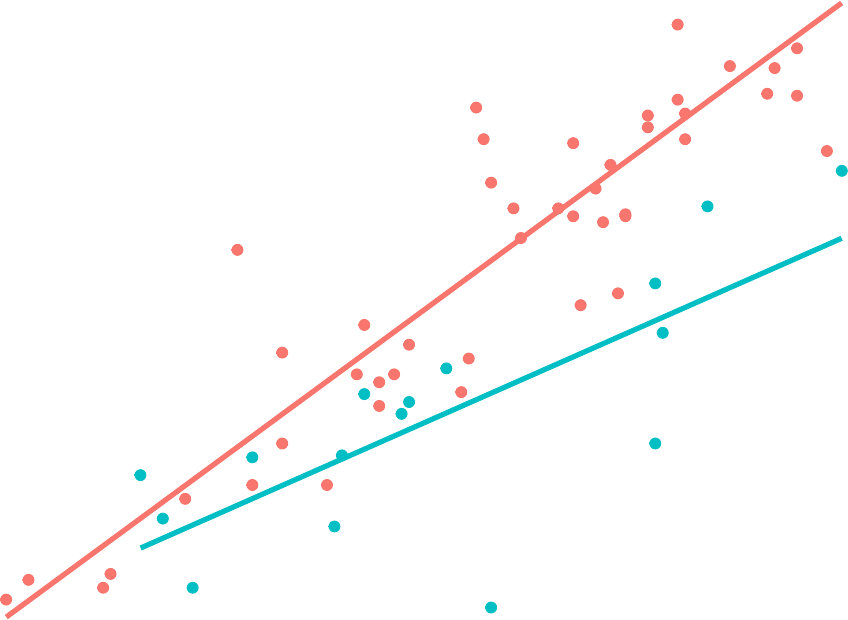
*#Y=-126.905+2.776X*

*#Difference in the Model #Y=76.021+1.102X*

ggplot(Data9, aes(x=X1, y=Y, col=as.factor(X2))) + geom\_point() + geom\_smooth(method="lm", se=FALSE)

## ‘geom\_smooth()‘ using formula ’y ~ x’

90



|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |

80

Y

70

70.0 72.5 75.0 77.5

## X1

as.factor(X2)

0

1