Math484_MidtermExam

Kevin Tchouate Mouofo

10/9/2020

Problem 1

```r
x <- c( 1, 0, 2, 0, 3, 1, 0, 1, 2, 0)
y <- c(16, 9, 17, 12, 22, 13, 8, 15, 19, 11)
sum_Xi <- 10; sum_Xi_2 <- 20; sum_Yi <- 120; sum_Yi_2 <- 2194
sum_XiYi <- 182; t8_.5alpha <- 2.31
n <- length(y)
```

1. Regression function:

```r
Sxy <- sum((x-mean(x))*(y-mean(y)))
Sxx <- sum((x-mean(x))^2)

beta1_hat <-  Sxy/Sxx
beta0_hat <- mean(y) - beta1_hat * mean(x)

cat("y_hat =",beta0_hat,"+",beta1_hat,"* x")
```

```
## y_hat = 10.2 + 4 * x
```

2. Points estimates:

```r
y_hat_i <- function(xi){
  return(beta0_hat + beta1_hat * xi)
}

xi <- c(0,1,2,3)
ans <- sapply(xi, FUN=function(a){
  cat("For xi =",a, ", Y_hat_i =",y_hat_i(a),"\n")
})
```

```
## For xi = 0 , Y_hat_i = 10.2
## For xi = 1 , Y_hat_i = 14.2
## For xi = 2 , Y_hat_i = 18.2
## For xi = 3 , Y_hat_i = 22.2
```

```r
SSE <- sum((y - y_hat_i(x))^2)
MSE <- SSE/(n-2)
cat("SSE =",SSE,", MSE =", MSE)
```

```
## SSE = 17.6 , MSE = 2.2
```

3. Point estimate and 95% confidence interval for X=1:

```r
t <- qt(1-(.05/2),n-2)
xh <- 1
yh =y_hat_i(xh)
s_yh <- sqrt(MSE*((1/n)+((xh-mean(x))^2/Sxx)))
cat("yh(1) =",yh,", 95% CI = {", yh-(t*s_yh),",",yh+(t*s_yh),"}")
```

```
## yh(1) = 14.2 , 95% CI = { 13.11839 , 15.28161 }
```

4. Point estimate and 95% confidence interval for X=3:

```r
t <- qt(1-(.05/2),n-2)
xh <- 3
yh =y_hat_i(xh)
s_yh <- sqrt(MSE*((1/n)+((xh-mean(x))^2/Sxx)))
cat("yh(3) =",yh,", 95% CI = {", yh-(t*s_yh),",",yh+(t*s_yh),"}")
```

```
## yh(3) = 22.2 , 95% CI = { 19.78144 , 24.61856 }
```

5. Estimate of $\beta 1$ and hypothesis testing:

H0: $\beta 1 = 0$, Ha: $\beta 1 \neq 0$

Test statistic : $t* = \frac{b1}{s\{b1\}}$, if $|t*| \leq t(1-\alpha/2; n-2)$ conclude H0,

otherwise, conclude Ha.

```r
t <- qt(1-(.05/2),n-2)
b1 <- beta1_hat
s_b1 <- sqrt(MSE/Sxx)
t_star <- b1/s_b1
if(t_star<=t){
  cat("b1 =", b1, ", t* =", t_star, ", t(1-alpha/2, n-2) =", t, ", conclude H0")
}else{
  cat("b1 =", b1, ", t* =", t_star, ", t(1-alpha/2, n-2) =", t, ", conclude Ha")
}
```

```
## b1 = 4 , t* = 8.528029 , t(1-alpha/2, n-2) = 2.306004 , conclude Ha
```

6. F test:

H0: $\beta 1 = 0$, Ha: $\beta 1 \neq 0$

Test statistic : $F* = \frac{MSR}{MSE}$, if $F* \leq F(1-\alpha; 1, n-2)$ conclude H0,

otherwise, conclude Ha.

```r
f <- qf(1-.05,1,n-2)
MSR <- sum(sapply(x, FUN = function(x) (y_hat_i(x) - mean(y))^2))
f_star <- MSR/MSE
if(f_star<=f){
  cat("F* =", f_star, ", F(1-alpha,1, n-2) =", f, ", conclude H0")
}else{
  cat("F* =", f_star, ", F(1-alpha,1, n-2) =", f, ", conclude Ha")
}
```

```
## F* = 72.72727 , F(1-alpha,1, n-2) = 5.317655 , conclude Ha
```

7. F* and t* relation:

```r
cat("F* = ", f_star, ", (t*)^2 = ", t_star^2,
    "\nF(1-alpha,1, n-2) =",f,", [t(1-alpha/2, n-2)]^2 =",t^2)
```

```
## F* =  72.72727 , (t*)^2 =  72.72727
## F(1-alpha,1, n-2) = 5.317655 , [t(1-alpha/2, n-2)]^2 = 5.317655
```

We have $t^2* = F*$, and $[t(1-\alpha/2; n-2)]^2 = F(1-\alpha; 1, n-2)$ we can conclude that the t test is equivalent to the F test.

Lets prove it, we have :

$SSR = \sum_{i=1}^{n}(\hat{y} - \bar{y})^2$

$= \sum_{i=1}^{n}(\hat{\beta 1}xi + \hat{\beta 0} - \bar{y})^2$

$= \sum_{i=1}^{n}(\hat{\beta 1}xi + \bar{y} - \hat{\beta 1}\bar{x} - \bar{y})^2$

$= \sum_{i=1}^{n}(\hat{\beta 1}xi - \hat{\beta 1}\bar{x})^2$

$= \sum_{i=1}^{n}(\hat{\beta 1})^2(xi - \bar{x})^2$

$= \hat{\beta 1}^2 \sum_{i=1}^{n}(xi - \bar{x})^2$

$SSR = b1^2 \sum(xi - \bar{x})^2$, hence,

$F* = \frac{MSR}{MSE}$

$= \frac{\frac{SSR}{1}}{MSE}$

$= \frac{b1^2 \sum(X_i - \bar{X})^2}{MSE}$

$= \frac{b1^2}{\frac{MSE}{\sum(X_i - \bar{X})^2}}$

$= \frac{b1^2}{s^2\{b1\}}$

$= (t*)^2$

Problem 2

1. t-statistic for treat81:

```r
coef <- -19.952
SE_coef <- 3.006
cat("t-statistic =", coef/SE_coef)
```

```
## t-statistic = -6.637392
```

2. Comment on the effect of thorax :

One unit increase in body length (thorax) result in an increase of male longevity by 135.82 days, when every other predictors are held constant,

3. Explanation:

We observe that the coefficient for the treatment group of 8 pregnant females is larger than the one of 1 pregnant female. This is counter-intuitive since we expected the males kept with more females to be more sexually active and hence, to die earlier. The anomaly comes from the fact that the females fruit flies placed with males are inseminated, there are some deviations in behaviour to take into account now, (some flies only mate once, other are less attracted to recently mated females). Hence, it is not the same case as that with receptive females.

"Second, male D. melanogaster experienced with courting recently mated, unreceptive females learn to selectively avoid recently mated females but not receptive virgin females" (Dukas R, 2005. Experience improves courtship in male fruit flies. Anim Behav69:1203–1209.).

4. Explanation:

We observe that the coefficient for the treatment group of 8 receptive virgin females is smaller than the one of 1 receptive virgin female. This is consistent with my commun sense of animal behavior since male animals tend to be more sexually active if there are more receptive females, consequently in our case, their longevity should tend to decrease as the number of virgin receptive females increases. This has been proved staticsally:

"The insemination rate declined from approximately 7 females/day at age one week to just under 2/day at age eight weeks in the males supplied with eight virgin females per day, and from just under 1/day at age one week to approximately 0.6/day at age eight weeks in the males supplied with one virgin female per day." (James A. Hanley and Stanley H. Shapiro, http://jse.amstat.org/datasets/fruitfly.txt).

5. Anova table:

```r
dfE <-  124 - 5
SSE <- 38252.8-25108.1
MSR <- 25108.1/5
MSE <- SSE/dfE
F_star <- MSR/MSE
p_value <- 1-pf(F_star, df1=4, df2=119)
cat(" Df Error =" , dfE, "\n", "SS Error =" , SSE, "\n", "MS Regression =" , MSR, "\n",
    "MS Error =" , MSE, "\n", "F* =", F_star, "\n P_value =", p_value)
```

```
##  Df Error = 119
##  SS Error = 13144.7
##  MS Regression = 5021.62
##  MS Error = 110.4597
##  F* = 45.46112
##  P_value = 0
```

6. The estimate of model variance is MSE.

```r
cat("MSE =", MSE)
```

```
## MSE = 110.4597
```

7. The unit of $\hat{\sigma^2}$ is days.
8. R^2 and R^2_{adj}:

```r
SSTO <- 38252.8
SSR <- 25108.1
R_squared <- SSR/SSTO
R_squared_adj <- 1 - (124/119)*(SSE/SSTO)
cat("R^2 =", R_squared, ", R^2_adj =", R_squared_adj)
```

```
## R^2 = 0.6563729 , R^2_adj = 0.6419348
```

9. Assuming the same treatment, we expect a fly with a thorax greater by 0.2mm to live 27.164 days longer

```r
0.2*135.82
```

```
## [1] 27.164
```

10. It won't be okay to do a one way ANOVA ignoring the thorax length because the variable thorax length has a strong effect on survival, it is important to take it into account to increase the precision of between-group contrasts, even though it is distributed similarly across groups.

Problem 3

Considering $y = (\frac{x}{k0+k1.x+k2.x^2})^2$

We could use a transformation similar to the box-cox transformation for $\lambda = -0.5$ and have:

$y^\lambda = \frac{1}{\sqrt{y}}$

That way, the problem will be reduced to build the linear regression model of the equation :

$y' = k0.x^{-1} + k1 + k2.x$ where $y' = \frac{1}{\sqrt{y}}$

A model could then be fitted with $y' = \frac{1}{\sqrt{y}}$, $X1 = x^{-1}$ and $X2 = x$, we would obtain the linear regression equation:

$y' = \beta 0 + \beta 1.X1 + \beta 2.X2$

We will then have:

$k0 = \beta 1$

$k1 = \beta 0$

$k2 = \beta 2$

Problem 4

proof that

(i) $H = X(X'X)^{-1}X'$ is symmetric:

We know that the square matrix $X'X$ is symmetric.

We have:

$H' = [X(X'X)^{-1}X']'$

$= (X')'[(X'X)^{-1}]'X'$

$= X[X'(X')']^{-1}X'$

$= X(X'X)^{-1}X'$

$= H$

(ii) H and $I - H$ are idempotent:

We have:

$HH = [X(X'X)^{-1}X'][X(X'X)^{-1}X']$

$= X(X'X)^{-1}(X'X)(X'X)^{-1}X'$

$= X(X'X)^{-1}X'$

$= H$

$(I - H)(I - H) = I - 2H + HH = I - 2H + H = (I - H)$

(iii) $trace(H) = p$ and $trace(I - H) = n - p$:

$trace(H) = trace(X(X'X)^{-1}X')$

$= trace(X'X(X'X)^{-1})$

$= trace(I_p)$

$= p$

$trace(I_n - H) = trace(I_n) - trace(H) = n - p$

(iv) $HX = X$ and $(I - H)X = 0$:

$HX = X(X'X)^{-1}X'X$

$= X(X'X)^{-1}(X'X)$

$= X$

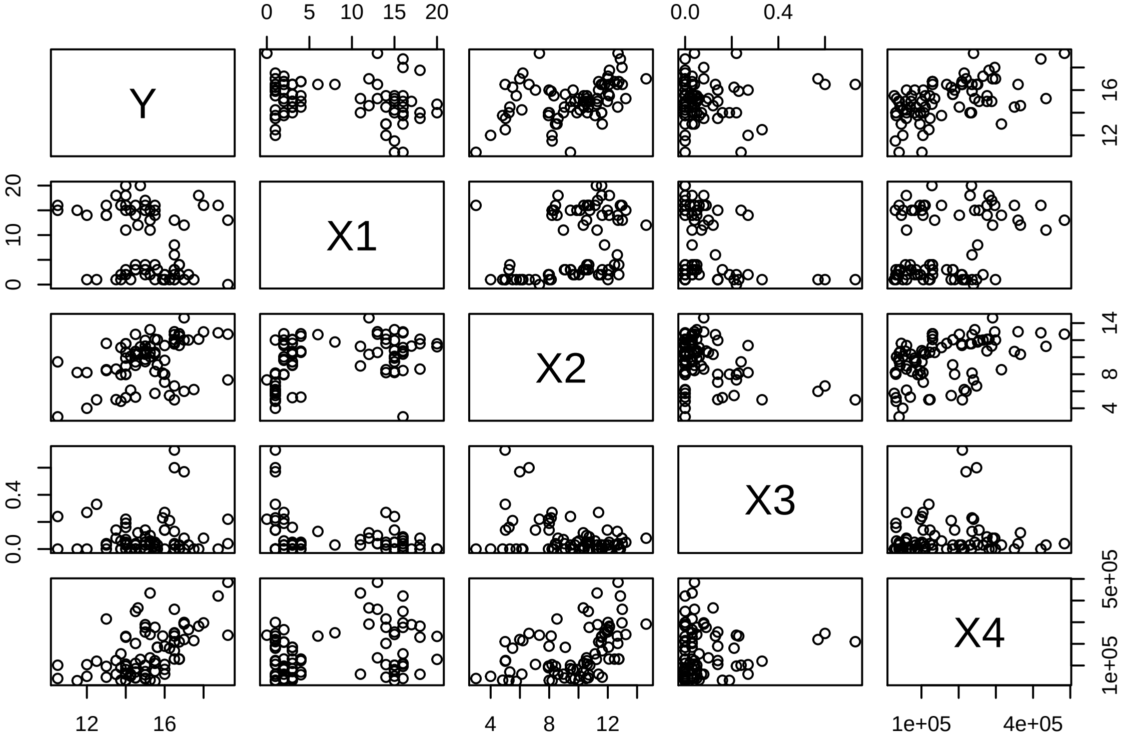
$(I - H)X = (I - X(X'X)^{-1}X')X = X - X = 0$

Problem 5 Commercial Properties.

```
#Load the data
data <- read.csv("/Users/kevinmouofo/Desktop/MATH 564/Midterm/commercial.txt",
                 sep="", header = TRUE )
n <- nrow(data)
```

1. scatter plot matrix :

```
plot(data)
```



```
cor(data)
```

```
##               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
```

From the scatter plot, we can see that there is no strong linear correlation among the variables, however the correlation between Y and X2, and also the correlation between Y and X4, are relatively high as confirmed by the correlation matrix.

2. Model 1:

```
model1 <- lm(Y~X1+X2+X3+X4,data=data)
model1
```

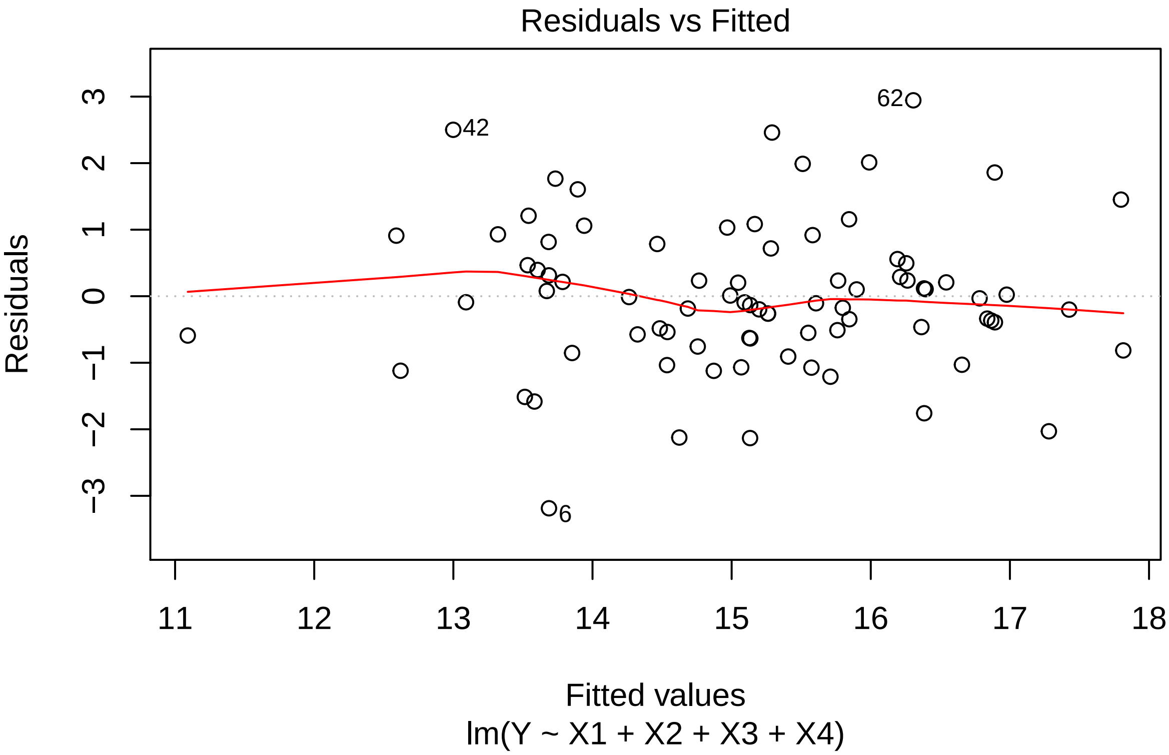
```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3 + X4, data = data)
##
## Coefficients:
## (Intercept)           X1           X2           X3           X4
##   1.220e+01   -1.420e-01    2.820e-01    6.193e-01    7.924e-06
```

We have the estimated regression function:

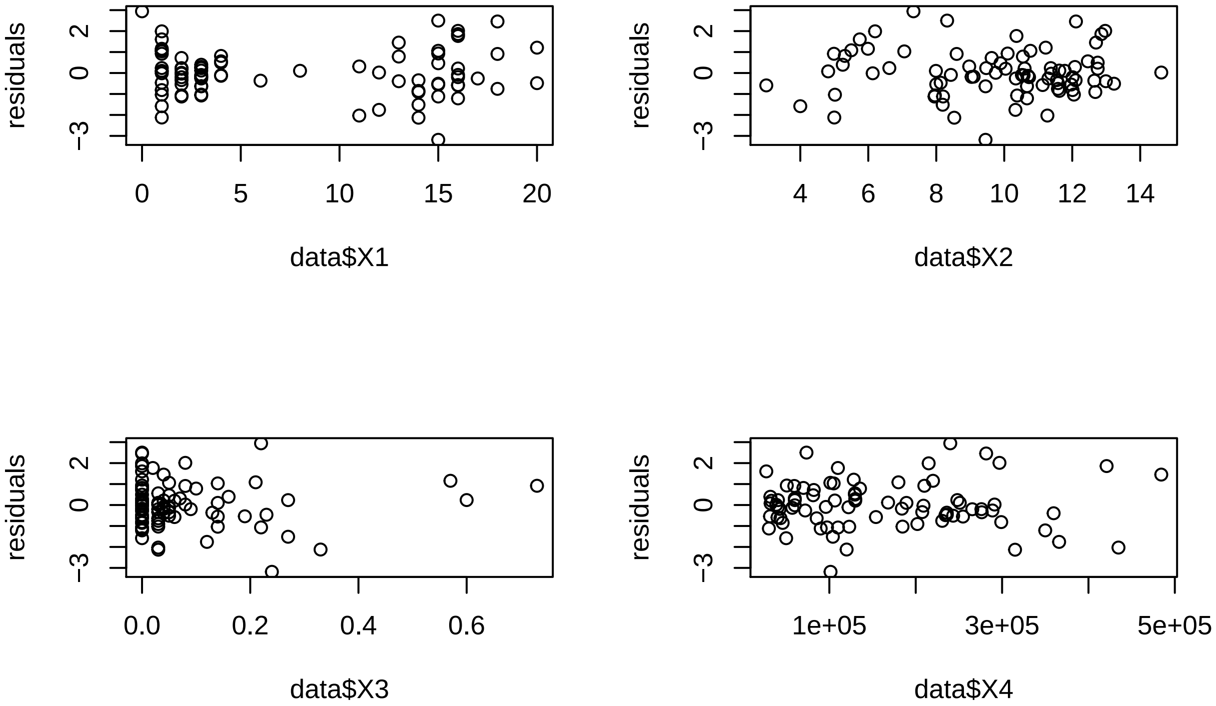
$$\hat{y} = (1.220e + 01) - (1.420e - 01)X1 + (2.820e - 01)X2 + (6.193e - 01)X3 + (7.924e - 06)X4$$

3. Residuals study:

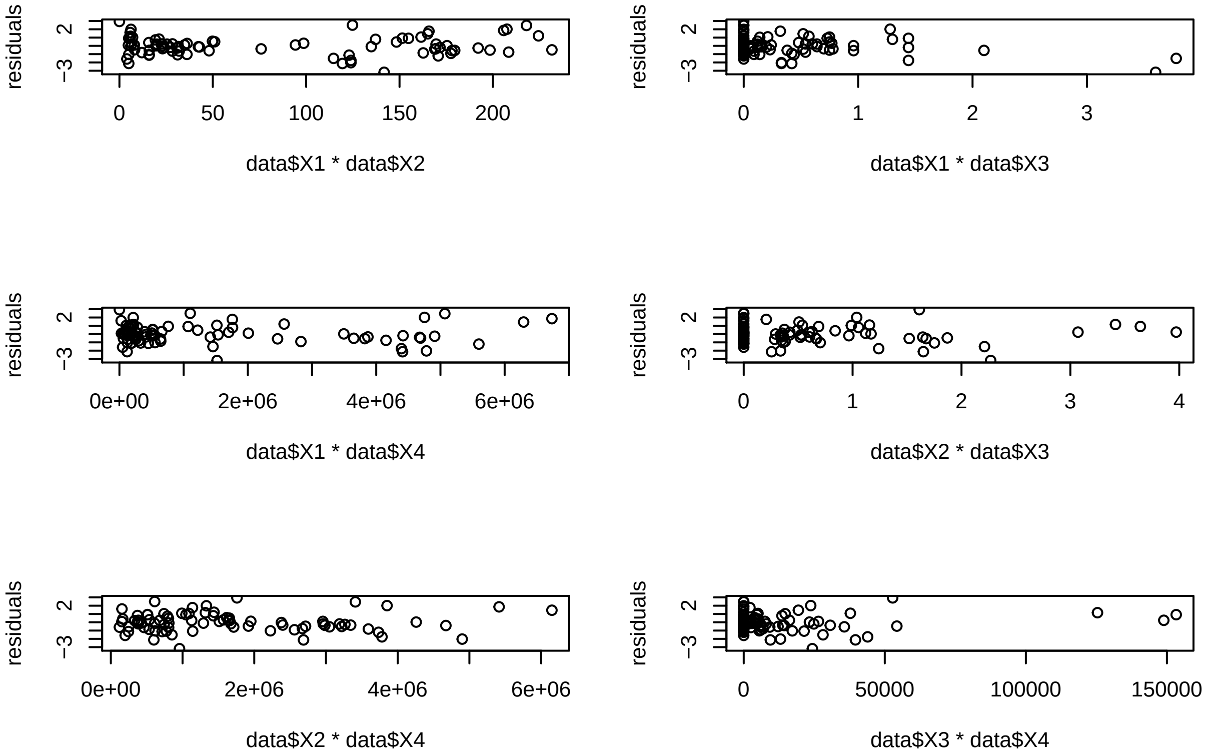
```
residuals <- model1$residuals
plot(model1, which=1)
```



```
#Residuals vs predicators
par(mfrow=c(2,2))
plot(residuals~data$X1)
plot(residuals~data$X2)
plot(residuals~data$X3)
plot(residuals~data$X4)
```



```r
#Residuals vs each 2 predicators
par(mfrow=c(3,2))
plot(data$X1*data$X2, residuals)
plot(data$X1*data$X3, residuals)
plot(data$X1*data$X4, residuals)
plot(data$X2*data$X3, residuals)
plot(data$X2*data$X4, residuals)
plot(data$X3*data$X4, residuals)
```



Looking at the plots of the residuals against X1, X2, X3 and their combinations, we observe that the errors are getting more concentrate around 0 as the predicator increases. The plot Y against $\hat{Y}$ suggest that the error variance is not constant since we observe a curvature.

4. Anova table:

```
anova(model1)
```

```
## 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
```

We can see that F is significant.

```
MSR <- sum(14.819,72.802,8.381,42.325)/4
MSE <- 1.293
F_star_model1 <- MSR/MSE
if(F_star_model1<=qf(.95,4,76)){
  cat("F-ratio =", F_star_model1, ", F(1-alpha,p-1, n-p) = ",qf(.95,4,76), "F-ratio not significant" )
}else{
  cat("F-ratio =", F_star_model1, ", F(1-alpha,p-1, n-p) = ",qf(.95,4,76), ", F-ratio significant" )
}
```

```
## F-ratio = 26.74536 , F(1-alpha,p-1, n-p) =  2.492049 , F-ratio significant
```

5. R^2 and R^2_{adj}:

```
SSE <- 98.231
SSR <- sum(14.819+72.802+8.381+42.325)
SSTO <- sum(14.819+72.802+8.381+42.325+98.231)
MSE <- 1.293
R_squared <- SSR/SSTO
R_squared_adj <- 1 - (80/76)*(SSE/SSTO)
cat("R^2 = ", R_squared, ", R^2_adj = ", R_squared_adj)
```

```
## R^2 =  0.5847488 , R^2_adj =  0.5628934
```

6. Mean estimate, CI, and prediction interval:

```
X <- cbind(rep(1, nrow(data)), data$X1, data$X2,data$X3, data$X4)
Xt <- t(X)
Xp <- Xt%*%X
Xp.inv <- solve(Xp)

#Load the data
Xnew <- read.csv("/Users/kevinmouofo/Desktop/MATH 564/Midterm/new.txt",
                 sep="", header = TRUE )
alpha <- .05
Ynew <- predict(model1, Xnew)

t <- qt(1-(alpha/2),76)

ans <- sapply(1:nrow(Xnew), FUN=function(x){
  xh <- cbind(1, Xnew$X1[x], Xnew$X2[x], Xnew$X3[x],Xnew$X4[x])
  xht <- t(xh)
  s_yh <- sqrt(MSE*(xh%*%Xp.inv%*%xht))
  s_pred <- sqrt(MSE*(1 + xh%*%Xp.inv%*%xht))
  cat("For i = ", x, ", yh_hat =", Ynew[x], ", CI : {", Ynew[x]- t*s_yh, ",",
      Ynew[x] + t*s_yh, "}", ", PI : {", Ynew[x]- t*s_pred, ",",
      Ynew[x] + t*s_pred, "}\n")
})
```

```
## For i =  1 , yh_hat = 15.1485 , CI : { 14.76822 , 15.52877 } , PI : { 12.85206 , 17.44493 }
## For i =  2 , yh_hat = 15.54249 , CI : { 15.15358 , 15.9314 } , PI : { 13.24461 , 17.84037 }
## For i =  3 , yh_hat = 16.91384 , CI : { 16.18344 , 17.64424 } , PI : { 14.53424 , 19.29344 }
```

7. Model2 and partial F test:

```
model2 <- lm(Y~X1+X2+X4, data = data)
summary(model2)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X4, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0620 -0.6437 -0.1013  0.5672  2.9583
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.237e+01  4.928e-01  25.100  < 2e-16 ***
## X1          -1.442e-01  2.092e-02  -6.891 1.33e-09 ***
## X2           2.672e-01  5.729e-02   4.663 1.29e-05 ***
## X4           8.178e-06  1.305e-06   6.265 1.97e-08 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.132 on 77 degrees of freedom
## Multiple R-squared:  0.583,  Adjusted R-squared:  0.5667
## F-statistic: 35.88 on 3 and 77 DF,  p-value: 1.295e-14
```

```
anova(model2)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## X1         1 14.819  14.819  11.566  0.001067 **
## X2         1 72.802  72.802  56.825 7.841e-11 ***
## X4         1 50.287  50.287  39.251 1.973e-08 ***
## Residuals 77 98.650   1.281
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

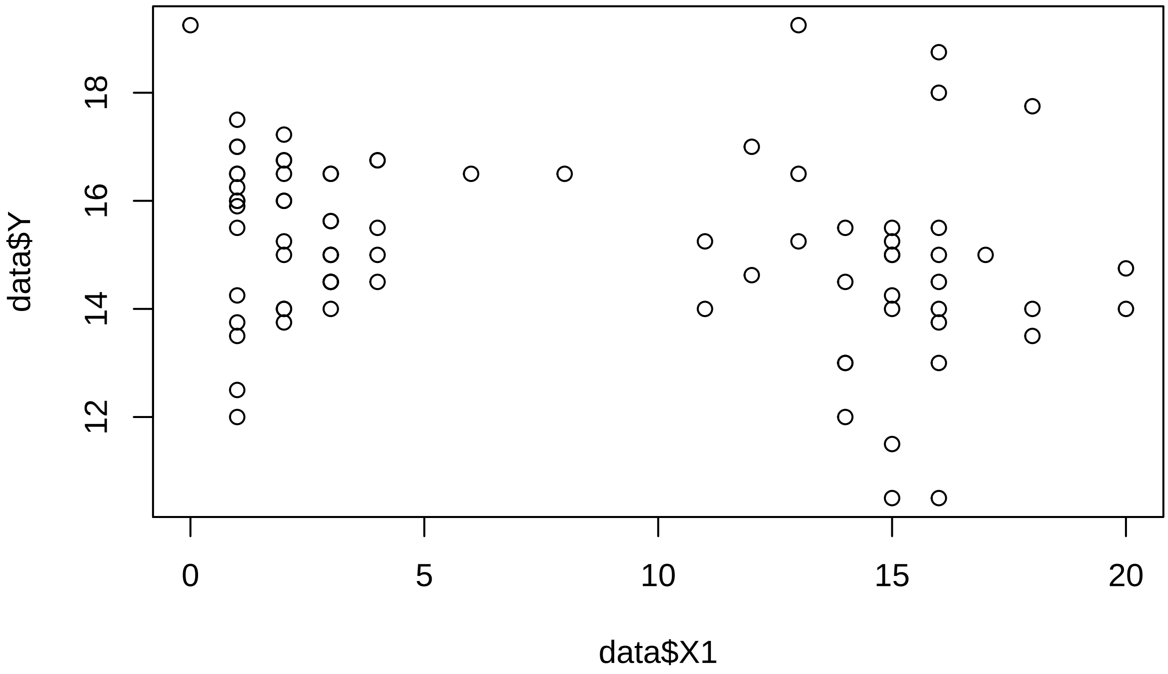
We have the model 2 $\hat{Y} = (1.237e+01) - (1.442e-01)X1 + (2.672e-01)X2 + (8.178e-06)X4$

```
SSE_model2 <- 98.650
F_star <- (SSE_model2 - SSE) / (SSE/76)
if (F_star <= qf(.95, 1, 76)){
  cat("F* = ", F_star, "F(1-alpha, 1, n-5) = ", qf(.95, 1, 76),
      "\n Conclude X3 can be dropped from regression model that already have X1, X2, X4")
}else {
  cat("F* = ", F_star, "F(1-alpha, 1, n-5) = ", qf(.95, 1, 76), ", conclude can not be dropped")
}
```

```
## F* =  0.3241746 F(1-alpha, 1, n-5) =  3.96676
##  Conclude X3 can be dropped from regression model that already have X1, X2, X4
```

8. Y against X1:

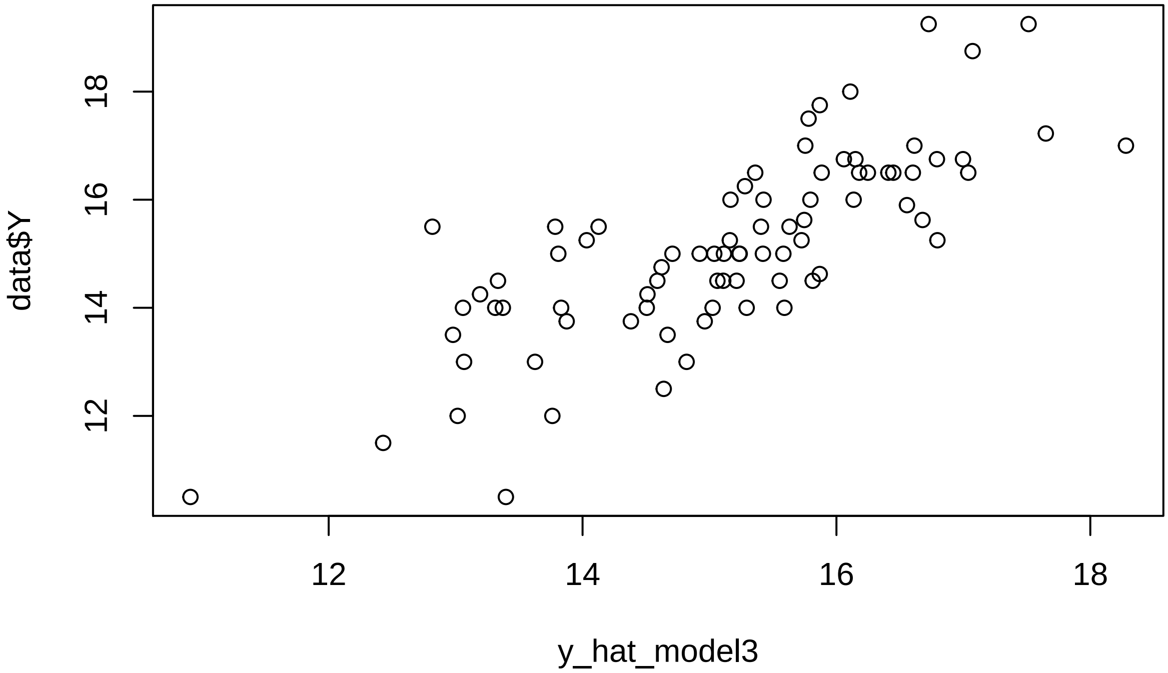
```
plot(data$Y~data$X1)
```



We observe a curviture around the middle.

9. Model 3:

```
X1_squared <- (data$X1)^2
data$X1_squared <- X1_squared
model3 <- lm(Y~X1+X2+X4+X1_squared, data = data)
dat <- data.frame(X1 = data$X1, X2 = data$X2, X4 = data$X4, X1_squared = data$X1_squared )
y_hat_model3 = predict(model3, dat)
plot(data$Y~y_hat_model3)
```



We have the regression equation :

$\hat{Y} = (1.249e + 01) - (4.043e - 01)X1 + (3.140e - 01)X2 + (8.046e - 06)X4 + (1.415e - 02)X1^2$

The plot Y against $\hat{Y}$ suggest that this model is a good enough fit.

```
summary(model3)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X4 + X1_squared, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.89596 -0.62547 -0.08907  0.62793  2.68309
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.249e+01  4.805e-01  26.000  < 2e-16 ***
## X1          -4.043e-01  1.089e-01  -3.712  0.00039 ***
## X2           3.140e-01  5.880e-02   5.340 9.33e-07 ***
## X4           8.046e-06  1.267e-06   6.351 1.42e-08 ***
## X1_squared   1.415e-02  5.821e-03   2.431  0.01743 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.097 on 76 degrees of freedom
## Multiple R-squared:  0.6131, Adjusted R-squared:  0.5927
## F-statistic:  30.1 on 4 and 76 DF,  p-value: 5.203e-15
```

```
anova(model3)
```

```
## Analysis of Variance Table
##
## Response: Y
##            Df Sum Sq Mean Sq F value    Pr(>F)
## X1          1 14.819  14.819 12.3036 0.0007627 ***
## X2          1 72.802  72.802 60.4463 2.968e-11 ***
## X4          1 50.287  50.287 41.7522 8.907e-09 ***
## X1_squared  1  7.115   7.115  5.9078 0.0174321 *
## Residuals  76 91.535   1.204
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

10. Comparison model2 and model3 :

```
SSE_model3 <- 91.535
F_star_2 <- (SSE_model2-SSE_model3) / (SSE_model3/76)
if (F_star_2 <= qf(.95, 1, 76)){
  cat("F* = ", F_star_2, "F(1-alpha, 1, n-5) = ", qf(.95, 1, 76), " Conclude beta4 = 0")
}else {
  cat("F* = ", F_star_2, "F(1-alpha, 1, n-5) = ", qf(.95, 1, 76), ", conclude beta4 is not 0")
}
```

```
## F* =  5.907467 F(1-alpha, 1, n-5) =  3.96676 , conclude beta4 is not 0
```

We have a statistical evidence that $X1^2$ is significant.