homework six

Yi (Chris) Chen

November 10, 2017

Homework six

problem one: 3.14

(a) the F test of lack of fit

```
# read the data
setwd("C:/Users/cheny/Desktop/study/linear regression model/homework/homework record/homework
 six")
data1 <- read.table('1.22.txt',header = FALSE,col.names = c('hardness','time'))

reg1 <- lm(data1$hardness ~ data1$time)


library(alr3)
```

```
## Loading required package: car
```

```
pureErrorAnova(reg1)
```

```
## Analysis of Variance Table
##
## Response: data1$hardness
##               Df Sum Sq Mean Sq  F value    Pr(>F)
## data1$time     1 5297.5  5297.5 493.7487 4.067e-11 ***
## Residuals     14  146.4    10.5
##  Lack of fit   2   17.7     8.8   0.8237    0.4622
##  Pure Error   12  128.7    10.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
qf(0.99,2,12)
```

```
## [1] 6.926608
```

the answer of (a)

$h_0 : E(Y) = \beta_0 + \beta_1 * X$ $h_a : E(Y) \neq \beta_0 + \beta_1 * X$

Obviously, the F value for lack of fit is 0.8237, P value is 0.4622. The value of F_star is 6.926608. Thus, F value is samller than F_star value. Based on this, I can conclude that H0 is acceptable.

(b)

the answer of (b)

1. advantage: if we have equal number of replications at each of x levels. In this way, the precision of regression model can be inproved. Suppose that if we have a great number of replications when x level is low while very little number of replications when x level is high. The result of the whole regression line would heavily depend on the data when x is low. If we have the same number of replication for each x level, the importance(weight) of data for each x level is equal.
2. I think this would not have any disadvantage. For example the lack of fit test would no change. Because, we only care about the average value of replications for each x level in the reduced model.

(c)

the answer of (c)

According to lack of fit test, if the data is suitable for linear regression. For each x level, we would expect that the mean of replications should be one the regression line. Since we have assumed that the residuals follow the normal distribution with fixed variance and mean is zero.

If the we conclude that the regression function is not linear. We can not simply tell what regression function is better just based on F value. But we can see the difference between SSLF and SSPE at different x level to determine which part of data is linear which is not.

problem one: 7.7

(a)

```
data_2 <- read.table('6.18.txt',header = FALSE, col.names = c('Y','X1','X2','X3','X4'))
```

(b)

1. Model one: $Y = \beta_0 + \beta_4 * X_4$

```
reg_2_1 <- lm(data = data_2,Y~X4)
Anova(reg_2_1)
```

```
## Anova Table (Type II tests)
##
## Response: Y
##            Sum Sq Df F value    Pr(>F)
## X4         67.775  1  31.723 2.628e-07 ***
## Residuals 168.782 79
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Obviously $SSR(X_4) = 67.775$ and $SSE(X_4) = 168.782$

2. Model two:$ Y = _0 + _4 * X_4 + _1 * X_1$

```
reg_2_2 <- lm(data = data_2,Y~X4+X1)
Anova(reg_2_2)
```

```
## Anova Table (Type II tests)
##
## Response: Y
##             Sum Sq Df F value    Pr(>F)    
## X4          95.231  1  58.716 4.225e-11 ***
## X1          42.275  1  26.065 2.275e-06 ***
## Residuals 126.508 78                      
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Obviously $SSR(X_1|X_4) = SSE(X_4) - SSE(X_1, X_4) = |126.508 - 168.782| = 42.274$

3. Model three:$ Y = _0 + _4 * X_4 + _1 * X_1+_2*X_2$

```
reg_2_3 <- reg_2 <- lm(data = data_2,Y~X4+X1+X2)
Anova(reg_2_3)
```

```
## Anova Table (Type II tests)
##
## Response: Y
##            Sum Sq Df F value    Pr(>F)    
## X4         50.287  1  39.251 1.973e-08 ***
## X1         60.841  1  47.489 1.335e-09 ***
## X2         27.857  1  21.744 1.287e-05 ***
## Residuals 98.650 77                      
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Obviously $SSR(X_2|X_4, X_1) = SSE(X_1, X_2, X_4) - SSE(X_1, X_4) = |126.508 - 98.650| = 27.858$

4. Model four:$ Y = _0 + _4 * X_4 + _1 * X_1+_2*X_2+_3*X_3$

```
reg_2_4 <- lm(data = data_2,Y~.)
Anova(reg_2_4)
```

```
## Anova Table (Type II tests)
##
## Response: Y
##            Sum Sq Df F value    Pr(>F)    
## X1         57.243  1 44.2881 3.894e-09 ***
## X2         25.759  1 19.9294 2.747e-05 ***
## X3          0.420  1  0.3248    0.5704    
## X4         42.325  1 32.7464 1.976e-07 ***
## Residuals 98.231 76                       
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Obviously
$SSR(X_3|X_1, X_2, X_4) = SSE(X_1, X_2, X_3, X_4) - SSE(X_1, X_2, X_4) = |98.650 - 98.231| = 0.419$

To sum up

```
SSRX_4 = 67.775
SRRX_1.X_4=42.274
SSRX_2.X_4.X_1=27.858
SSRX_3.X_1.X_2.X_4=0.419
error <- 98.231
SSTO <- 168.782 + 67.775
regression <- SSTO-error
SS <- c(regression,SSRX_4,SRRX_1.X_4,SSRX_2.X_4.X_1,SSRX_3.X_1.X_2.X_4,error,SSTO)
df <- c(4,1,1,1,1,nrow(data_2)-5,nrow(data_2)-1)
MS <- SS/df

result <- data.frame(SS,df,MS,row.names = c("regression","X4","X1|X4","X2|X1,X4","X3|X1,X2,X
4",'ERROR','TOTAL'))

result
```

```
##                   SS df        MS
## regression  138.326  4 34.581500
## X4           67.775  1 67.775000
## X1|X4        42.274  1 42.274000
## X2|X1,X4     27.858  1 27.858000
## X3|X1,X2,X4   0.419  1  0.419000
## ERROR        98.231 76  1.292513
## TOTAL       236.557 80  2.956963
```

This problem can also be solved in this easy way. Since we include the new variables one by one. Thus, we can just use one anova to calculate all the information.

```
reg_2_5 <- lm(data=data_2,Y~X4+X1+X2+X3)
Anova(reg_2_5)
```

```
## Anova Table (Type II tests)
##
## Response: Y
##           Sum Sq Df F value    Pr(>F)
## X4        42.325  1 32.7464 1.976e-07 ***
## X1        57.243  1 44.2881 3.894e-09 ***
## X2        25.759  1 19.9294 2.747e-05 ***
## X3         0.420  1  0.3248    0.5704
## Residuals 98.231 76
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(b)

1. let’s make a t test first to determine whether we should keep X3 in the regression model.

```
reg_3 <- reg_2 <- lm(data = data_2,Y~.)
summary(reg_3)
```

```
## 
## Call:
## lm(formula = Y ~ ., data = data_2)
## 
## 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 ***
## X1          -1.420e-01  2.134e-02  -6.655 3.89e-09 ***
## X2           2.820e-01  6.317e-02   4.464 2.75e-05 ***
## X3           6.193e-01  1.087e+00   0.570     0.57    
## X4           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
```

clearly, as we can see for the t test for X3, the t-value is 0.570, and the p value is 0.57. Thus, we can conclude that we should not keep the X3 in the model.

2. Let’s do the partial F test again to determine whether we should keep X3 in the regression.

```
SSRX_3.X_1.X_2.X_4=0.419
df.X_3 <- 1
SSE.X_1.X_2.X_4 <- 98.650
df.X_1.X_2.X_4 <- nrow(data_2)-5

F_star <- (SSRX_3.X_1.X_2.X_4/df.X_3)/(SSE.X_1.X_2.X_4/df.X_1.X_2.X_4)
F_star
```

```
## [1] 0.3227978
```

```
F_key <- pf(0.95,df.X_3,df.X_1.X_2.X_4)

F_star > F_key
```

```
## [1] FALSE
```

clearly here we have $H0 : \beta_3 = 0$ and $Ha : \beta3 \neq 0$ And F_star > F_key, so we conclude that $\beta_3 = 0$ By the way, F_star is square value of t value in this test.

problem three: 7.10

```
data_3 <- read.table('6.18.txt',header = FALSE, col.names = c('Y','X1','X2','X3','X4'))
```

analysis: the model we have now is:

Full model:$Y = \beta_0 + \beta_1 X1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$

Since in this F test:

$H0 : \beta_1 = -0.1, \beta_2 = 0.4$ and $Ha : \beta_1 \neq -0.1, \beta_2 \neq 0.4$

We rewrite the model as:

$Y = \beta_0 + (-0.1) * X1 + (0.4) X_2 + \beta_3 X_3 + \beta_4 X_4$

Reduced Model:$Y + 0.1 * X_1 - 0.4 * X_2 = \beta_0 + \beta_3 X_3 + \beta_4 X_4$

```
# revalue the data
data_3$Y_NEW <- data_3$Y+0.1*data_3$X1-0.4*data_3$X2
#full model
reg_3_1 <- lm(data=data_3,Y~.-Y_NEW)
Anova(reg_3_1)
```

```
## Anova Table (Type II tests)
##
## Response: Y
##           Sum Sq Df F value    Pr(>F)
## X1        57.243  1 44.2881 3.894e-09 ***
## X2        25.759  1 19.9294 2.747e-05 ***
## X3         0.420  1  0.3248    0.5704
## X4        42.325  1 32.7464 1.976e-07 ***
## Residuals 98.231 76
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$SSE_F = 98.231$ and $df_F = 76$

```
reg_3_2 <- lm(data = data_3,Y_NEW~X3+X4)
Anova(reg_3_2)
```

```
## Anova Table (Type II tests)
##
## Response: Y_NEW
##            Sum Sq Df F value    Pr(>F)
## X3          6.600  1  4.6738   0.03369 *
## X4         31.872  1 22.5713 9.058e-06 ***
## Residuals 110.141 78
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
qf(0.99,2,76)
```

```
## [1] 4.89584
```

$SSE_R = 110.141$ and $df_R = 78$

analysis:

$F^* = \frac{SSE_R - SSE_F}{df_R - df_F} \div \frac{SSE_F}{df_f}$

$F^* = \frac{110.141 - 98.231}{78 - 76} \div \frac{98.231}{76} = 4.607$

$F(0.99, 2, 76) = 4.89584$

$F^* \leq F(0.99, 2, 76)$ thus, we conclude the H0

problem four: 7.16

(a)

```
#read the data
data_4 <- read.table('6.5.txt',header = FALSE,col.names = c('Y','X1','X2'))

# standardize the data
data_4 <- as.data.frame(scale(data_4, center=T,scale=T))
```

analysis

Here I just use the quick function in R to standardize the data. More details of how standardize the data are wroted as follow:

$$Y_i^* = \frac{1}{\sqrt{n-1}} * (\frac{Y_i-\overline{Y}}{S_Y})$$

$$X_{ik}^* = \frac{1}{\sqrt{n-1}} * (\frac{Y_{ik}-\overline{X_k}}{S_k})$$

```
#  regress the model
reg_4 <- lm(data=data_4,Y~X1+X2-1)
summary(reg_4)
```

```
## 
## Call:
## lm(formula = Y ~ X1 + X2 - 1, data = data_4)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.38423 -0.15391  0.00218  0.13863  0.36677 
## 
## Coefficients:
##    Estimate Std. Error t value Pr(>|t|)    
## X1  0.89239    0.05852  15.250 4.09e-10 ***
## X2  0.39458    0.05852   6.743 9.43e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.2266 on 14 degrees of freedom
## Multiple R-squared:  0.9521, Adjusted R-squared:  0.9452 
## F-statistic:   139 on 2 and 14 DF,  p-value: 5.82e-10
```

analysis

the result of the model: $\widehat{Y}^* = 0.89239 * X_1^* + 0.39458 * X_2^*$

(b)

As we can see from the model: when X2 is fixed, on average,the increase of one standard deviation of X1 would led to the increase of 0.89239 standard deviation in Y.

Similarly, when X1 is fixed, on average,the increase of one standard deviation of X2 would led to the increase of 0.39458 standard deviation in Y

(c)

the method of transform can be described as follow:

$$b_k = (\frac{S_Y}{S_k} b_k^*)$$

$$b_0 = \overline{Y} - b_1\overline{X_1} - b_2\overline{X_2}$$

```
b1_star <- 0.89239
b2_star <- 0.39458

data_4 <- read.table('6.5.txt',header = FALSE,col.names = c('Y','X1','X2'))

Sy <- sd(data_4$Y)
Sx1 <- sd(data_4$X1)
Sx2 <- sd(data_4$X2)

b1_expect <- (Sy/Sx1)*b1_star
b2_expect <- (Sy/Sx2)*b2_star
b0_expect <- mean(data_4$Y)-b1_expect*mean(data_4$X1)-b2_expect*mean(data_4$X2)
expect_value <- c(b0_expect,b1_expect,b2_expect)
names(expect_value) <- c('b0_expect','b1_expect','b2_expect')
reg_4_2 <- lm(data = data_4,Y~.)
expect_value;reg_4_2
```

```
## b0_expect b1_expect b2_expect
## 37.650124  4.424986  4.374992
```

```
##
## Call:
## lm(formula = Y ~ ., data = data_4)
##
## Coefficients:
## (Intercept)           X1           X2
##      37.650        4.425        4.375
```

As we can see they are the same

problem five: 7.24

(a)

```
# read the data
data_5 <- read.table('6.5.txt',header = FALSE,col.names = c('Y','X1','X2'))
# regress the simple model
reg_5_1 <- lm(data=data_5,Y~X1)
summary(reg_5_1)
```

```
##
## Call:
## lm(formula = Y ~ X1, data = data_5)
##
## Residuals:
##    Min     1Q Median     3Q    Max 
## -7.475 -4.688 -0.100  4.638  7.525 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   50.775      4.395  11.554 1.52e-08 ***
## X1             4.425      0.598   7.399 3.36e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.349 on 14 degrees of freedom
## Multiple R-squared:  0.7964, Adjusted R-squared:  0.7818 
## F-statistic: 54.75 on 1 and 14 DF,  p-value: 3.356e-06
```

analysis

the model we get: $Y_i = 50.775 + 4.425X_i + \varepsilon_i$

(b)

```
#regress the model
reg_5_2 <- lm(data=data_5,Y~.)
summary(reg_5_2)
```

```
##
## Call:
## lm(formula = Y ~ ., data = data_5)
##
## 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
```

the model we get: $Y_i = 37.6500 + 4.425X_{1i} + 4.3750X_{2i} + \varepsilon_i$

as we can see the estimate value of β_1 is same in both model

(c)

```
anova(reg_5_2);
```

```
## 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
```

```
anova(reg_5_1)
```

```
## 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
```

$SSR(X_1) = SSR(X_1|X_2) = 1566.45$

(D)

```
cor(data_5$X1,data_5$X2)
```

```
## [1] 0
```

as we can see the correlation between X1 and X2 is 0. This is also way the SSR(X_1)SSR(X_1|X_2) and we can see the estimate value of β_1 is same in both model.

problem six: 7.37

```
# read the data
data_6 <- read.table('c.2.txt',header = FALSE)
data_6 <- cbind(data_6$V8,data_6$V5,data_6$V16,data_6$V4,data_6$V7,data_6$V9,data_6$V10)
colnames(data_6) <- c('Y','X1','X2','X3','X4','X5','X6')
# regression between X1 and X2
data_6 <- as.data.frame(data_6)
reg_6_1 <- lm(data=data_6,Y~X1+X2)
```

(a)

```
# regression about X3
reg_6_2 <- lm(data=data_6,Y~X1+X2+X3)
Anova(reg_6_1);Anova(reg_6_2)
```

```
## Anova Table (Type II tests)
##
## Response: Y
##              Sum Sq  Df F value    Pr(>F)
## X1          1181173   1  3.6617   0.05633 .
## X2         22058054   1 68.3803 1.638e-15 ***
## Residuals 140967081 437
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Anova Table (Type II tests)
##
## Response: Y
##              Sum Sq  Df F value    Pr(>F)
## X1          2674327   1   8.517 0.0037005 **
## X2         15729788   1  50.095 5.889e-12 ***
## X3          4063370   1  12.941 0.0003583 ***
## Residuals 136903711 436
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$R^2_{Y3|12} = \frac{SSR(X_3|X1,,X2)}{SSE(X1,X2)} == \frac{SSE(X_1,X_2)-SSE(X_1,X_2,X_3)}{SSE(X_1,X_2)} = 1 - \frac{SSE(X_1,X_2,X_3)}{SSE(X_1,X_2)}$$

So:

$$R^2_{Y3|12} = 1 - 136903711/140967081 = 0.02882496$$

```
# regression about X4
reg_6_3 <- lm(data=data_6,Y~X1+X2+X4)
Anova(reg_6_1);Anova(reg_6_3)
```

```
## Anova Table (Type II tests)
##
## Response: Y
##              Sum Sq  Df F value    Pr(>F)
## X1          1181173   1  3.6617   0.05633 .
## X2         22058054   1 68.3803 1.638e-15 ***
## Residuals 140967081 437
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Anova Table (Type II tests)
##
## Response: Y
##              Sum Sq  Df F value    Pr(>F)
## X1          1245199   1  3.8662    0.0499 *
## X2         21777264   1 67.6152 2.303e-15 ***
## X4           541647   1  1.6817    0.1954
## Residuals 140425434 436
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$R^2_{Y4|12} = \frac{SSR(X_4|X1,,X2)}{SSE(X1,X2)} == \frac{SSE(X_1,X_2)-SSE(X_1,X_2,X_4)}{SSE(X_1,X_2)} = 1 - \frac{SSE(X_1,X_2,X_4)}{SSE(X_1,X_2)}$$

So:

$$R^2_{Y4|12} = 1 - 140425434/140967081 = 0.003842365$$

```
# regression about X5
reg_6_4 <- lm(data=data_6,Y~X1+X2+X5)
Anova(reg_6_1);Anova(reg_6_4)
```

```
## Anova Table (Type II tests)
##
## Response: Y
##              Sum Sq  Df F value    Pr(>F)
## X1          1181173   1  3.6617   0.05633 .
## X2         22058054   1 68.3803 1.638e-15 ***
## Residuals 140967081 437
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Anova Table (Type II tests)
##
## Response: Y
##             Sum Sq  Df F value    Pr(>F)
## X1        10822467   1  75.021 < 2.2e-16 ***
## X2        35802527   1 248.182 < 2.2e-16 ***
## X5        78070132   1 541.180 < 2.2e-16 ***
## Residuals 62896949 436
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$R^2_{Y5|12} = \frac{SSR(X_5|X1,,X2)}{SSE(X1,X2)} == \frac{SSE(X_1,X_2)-SSE(X_1,X_2,X_5)}{SSE(X_1,X_2)} = 1 - \frac{SSE(X_1,X_2,X_5)}{SSE(X_1,X_2)}$$

So:

$$R^2_{Y5|12} = 1 - 62896949/140967081 = 0.5538182$$

```
# regression about X6
reg_6_5 <- lm(data=data_6,Y~X1+X2+X6)
Anova(reg_6_1);Anova(reg_6_5)
```

```
## Anova Table (Type II tests)
##
## Response: Y
##              Sum Sq  Df F value    Pr(>F)
## X1          1181173   1  3.6617   0.05633 .
## X2         22058054   1 68.3803 1.638e-15 ***
## Residuals 140967081 437
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Anova Table (Type II tests)
## 
## Response: Y
##              Sum Sq  Df F value    Pr(>F)    
## X1            46727   1  0.1456   0.70297    
## X2         21996879   1 68.5365 1.537e-15 ***
## X6          1032359   1  3.2166   0.07359 .  
## Residuals 139934722 436                      
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$R^2_{Y5|12} = \frac{SSR(X_5|X1,,X2)}{SSE(X1,X2)} == \frac{SSE(X_1,X_2)-SSE(X_1,X_2,X_5)}{SSE(X_1,X_2)} = 1 - \frac{SSE(X_1,X_2,X_5)}{SSE(X_1,X_2)}$$

So:

$$R^2_{Y5|12} = 1 - 139934722/140967081 = 0.007323405$$

To sum up

```
value <- c(0.02882496,0.003842365,0.5538182,0.007323405)
value <- as.data.frame(value)
attributes(value)$row.names <- c('R_2_Y3.12','R_2_Y4.12','R_2_Y5.12','R_2_Y6.12')
value
```

```
##                 value
## R_2_Y3.12 0.028824960
## R_2_Y4.12 0.003842365
## R_2_Y5.12 0.553818200
## R_2_Y6.12 0.007323405
```

(b)

as we can see from the result in (a), X5 is much better than other variables.

```
value$sum_of_square <- c(4063370,541647,78070132,1032359)
value
```

```
##                 value sum_of_square
## R_2_Y3.12 0.028824960       4063370
## R_2_Y4.12 0.003842365        541647
## R_2_Y5.12 0.553818200      78070132
## R_2_Y6.12 0.007323405       1032359
```

And, yes obviously, X5 has much higher sum of square.

(c)

```
# full model
reg_6_6 <- lm(data=data_6,Y~X1+X2+X5)
Anova(reg_6_6)
```

```
## Anova Table (Type II tests)
##
## Response: Y
##             Sum Sq  Df F value    Pr(>F)
## X1        10822467   1  75.021 < 2.2e-16 ***
## X2        35802527   1 248.182 < 2.2e-16 ***
## X5        78070132   1 541.180 < 2.2e-16 ***
## Residuals 62896949 436
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# reduced model
reg_6_7 <- lm(data=data_6,Y~X1+X2)
Anova(reg_6_7)
```

```
## Anova Table (Type II tests)
##
## Response: Y
##              Sum Sq  Df F value    Pr(>F)
## X1          1181173   1  3.6617   0.05633 .
## X2         22058054   1 68.3803 1.638e-15 ***
## Residuals 140967081 437
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

analysis $H0 : \beta_5 = 0$ and $Ha : \beta_5 \neq 0$

$F^* = \frac{SSE_R - SSE_F}{df_R - df_F} \div \frac{SSE_F}{df_f}$

Here:

$F^* = \frac{140967081 - 62896949}{437 - 436} \div \frac{62896949}{436} = \frac{78070132}{144259.1} = 541.1799$

```
qf(0.99,1,137)
```

```
## [1] 6.823547
```

obviously $F^* > 6.823547$ conclude:Ha