MATH 423 - Assignment 3

YUNHEUM DAN SEOL

2017-11-15

```r
#Reading the data
cigs <- read.csv("http://www.math.mcgill.ca/yyang/regression/data/cigs.csv")
summary(cigs)
```

```
##       TAR           NICOTINE          WEIGHT             CO
##  Min.   : 1.00   Min.   :0.1300   Min.   :0.7851   Min.   : 1.50
##  1st Qu.: 8.60   1st Qu.:0.6900   1st Qu.:0.9225   1st Qu.:10.00
##  Median :12.80   Median :0.9000   Median :0.9573   Median :13.00
##  Mean   :12.22   Mean   :0.8764   Mean   :0.9703   Mean   :12.53
##  3rd Qu.:15.10   3rd Qu.:1.0200   3rd Qu.:1.0070   3rd Qu.:15.40
##  Max.   :29.80   Max.   :2.0300   Max.   :1.1650   Max.   :23.50
```

```r
#Defining the variables
y <- cigs$CO
x1 <- cigs$TAR
x2 <- cigs$NICOTINE
x3 <- cigs$WEIGHT
#Constructing the models
fit.intercept <- lm(y~1)
fit.cigs123 <- lm(y~x1+x2+x3)
fit.cigs12 <- lm(y~x1+x2)
fit.cigs1 <- lm(y~x1)
fit.cigs23 <-lm(y~x2+x3)
fit.cigs13 <- lm(y~x1+x3)

#Summaries of each model

summary(fit.intercept)
```

```
##
## Call:
## lm(formula = y ~ 1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.028  -2.528   0.472   2.872  10.972
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  12.5280     0.9479   13.22 1.65e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.74 on 24 degrees of freedom
```

```r
summary(fit.cigs1)
```

```
##
## Call:
```

```
## lm(formula = y ~ x1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1124 -0.7167 -0.3754  1.0091  2.5450
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.74328    0.67521   4.063 0.000481 ***
## x1           0.80098    0.05032  15.918 6.55e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.397 on 23 degrees of freedom
## Multiple R-squared:  0.9168, Adjusted R-squared:  0.9132
## F-statistic: 253.4 on 1 and 23 DF,  p-value: 6.552e-14
```

```r
summary(fit.cigs12)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.89941 -0.78470 -0.00144  0.91585  2.43064
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.0896     0.8438   3.662 0.001371 **
## x1            0.9625     0.2367   4.067 0.000512 ***
## x2           -2.6463     3.7872  -0.699 0.492035
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.413 on 22 degrees of freedom
## Multiple R-squared:  0.9186, Adjusted R-squared:  0.9112
## F-statistic: 124.1 on 2 and 22 DF,  p-value: 1.042e-12
```

```r
summary(fit.cigs123)
```

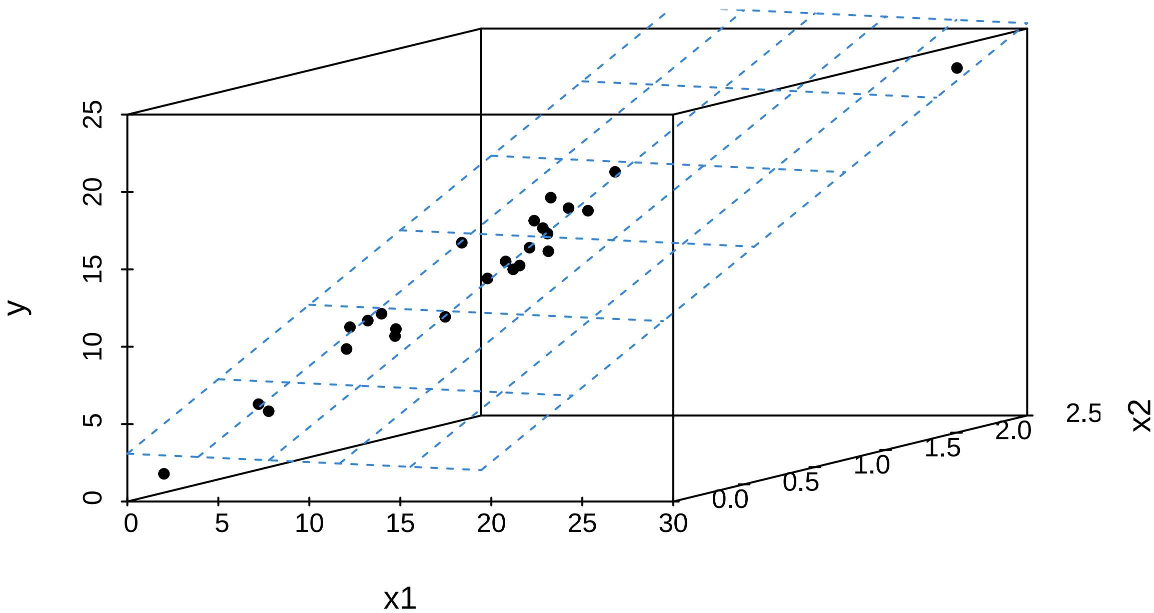
```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.89261 -0.78269  0.00428  0.92891  2.45082
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.2022     3.4618   0.925 0.365464
## x1            0.9626     0.2422   3.974 0.000692 ***
## x2           -2.6317     3.9006  -0.675 0.507234
## x3           -0.1305     3.8853  -0.034 0.973527
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.446 on 21 degrees of freedom
## Multiple R-squared:  0.9186, Adjusted R-squared:  0.907
## F-statistic: 78.98 on 3 and 21 DF,  p-value: 1.329e-11
```

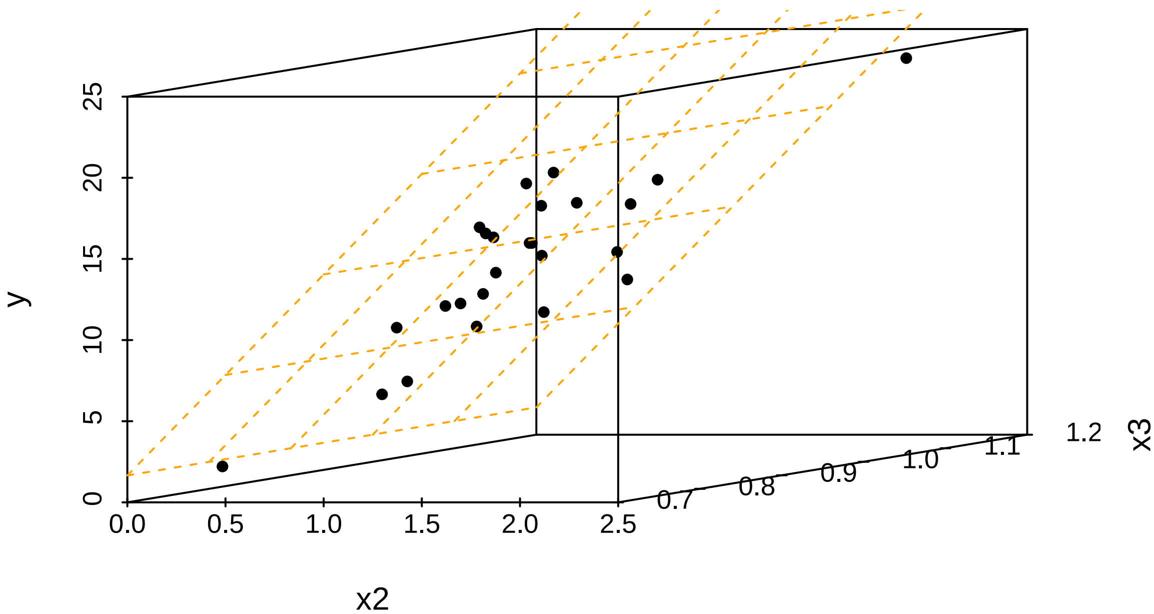
```r
library(scatterplot3d)
```

```
## Warning: package 'scatterplot3d' was built under R version 3.3.2
```

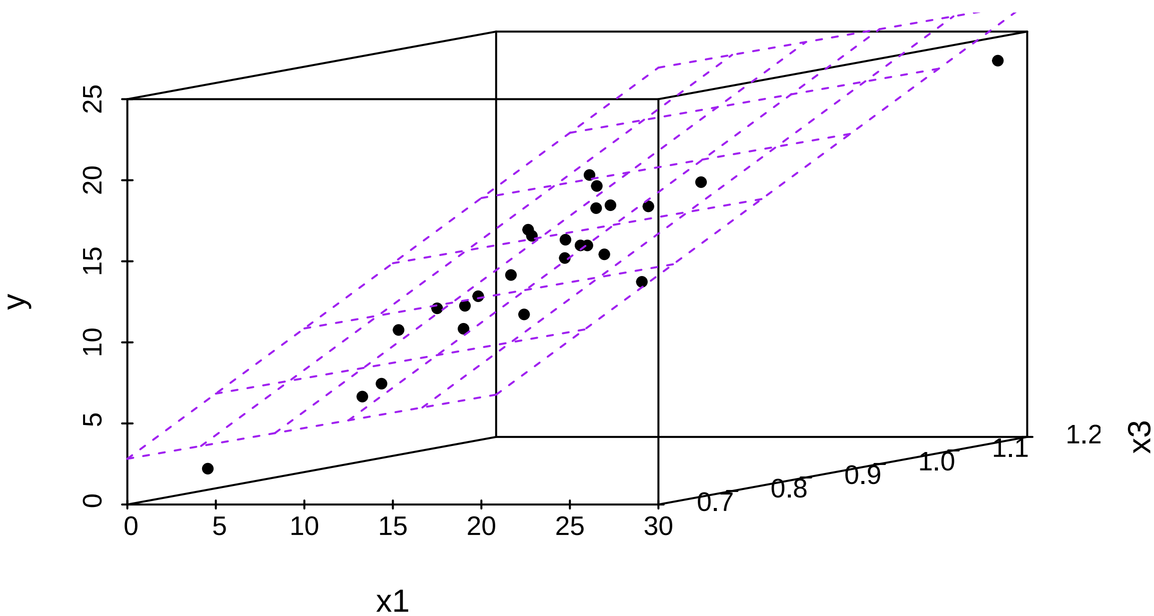
```r
#3d plot for fit.cigs12
s3d<-scatterplot3d(x1,x2,y,pch=20,grid=FALSE,angle=20)
s3d$plane3d(fit.cigs12,col="#3388dd")
```




```r
#3d plot for fit.cigs23
s3d<-scatterplot3d(x2,x3,y,pch=20,grid=FALSE,angle=15)
s3d$plane3d(fit.cigs23,col="orange")
```




```
#3d plot for fit.cigs123
s3d<-scatterplot3d(x1,x3,y,pch=20,grid=FALSE,angle=15)
s3d$plane3d(fit.cigs13,col="purple")
```



(a)

From the notes we know

$$SS_{res} := (n-p)\hat{\sigma}^2 = (n-p) * \hat{\sigma}^2$$

so for $SS_{res}(\beta_0, \beta_1, \beta_2, \beta_3)$

$$SS_{res}(\beta_0, \beta_1, \beta_2, \beta_3) = (25-4) * 1.445726^2 = 43.89259$$

#(b) and

$$SS_{res}(\beta_0, \beta_1, \beta_2) = (25-3) * (1.412524)^2 = 43.89494$$

```r
#Calculating SS_res for models

SS_res_beta123 <- summary(fit.cigs123)$df[2]*summary(fit.cigs123)$sigma^2
SS_res_beta123
```

```
## [1] 43.89259
```

```r
SS_res_beta12 <- summary(fit.cigs12)$df[2]*summary(fit.cigs12)$sigma^2
SS_res_beta12
```

```
## [1] 43.89494
```

Let us check whether those values coincide with the definition

$$SS_{res} = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$$

```r
fitted.beta123 <- fitted(fit.cigs123)
SSresBeta123 <- sum((y-fitted.beta123)^2)
SSresBeta123
```

```
## [1] 43.89259
```

```r
fitted.beta12 <- fitted(fit.cigs12)
SSresBeta12 <- sum((y-fitted.beta12)^2)
SSresBeta12
```

```
## [1] 43.89494
```

```r
SS_res_beta123-SS_res_beta12
```

```
## [1] -0.002357293
```

So our computation is correct. We should remark that the difference between $SS_{res}(\beta_0, \beta_1, \beta_2, \beta_3)$ and $SS_{res}(\beta_0, \beta_1, \beta_2)$ is very small. This might suggest the variation in predictor x_{i3} might not explain the variation in y very #(c) The F test statistic for comparing the two models

$$\mathrm{E}_{Y|X}[Y_i|\mathrm{x}_i] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} \mathrm{E}_{Y|X}[Y_i|\mathrm{x}_i] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}$$

can be found as

$$F = \frac{(SS_{Res}(\beta_0, \beta_1, \beta_2) - SS_{Res}(\beta_0, \beta_1, \beta_2, \beta_3))/r}{SS_{Res}(\beta_0, \beta_1, \beta_2, \beta_3)/(n-p)} = \frac{(SS_{Res}(\beta_0, \beta_1, \beta_2) - SS_{Res}(\beta_0, \beta_1, \beta_2, \beta_3))/1}{SS_{Res}(\beta_0, \beta_1, \beta_2, \beta_3)/(21)}$$

$$\frac{(SS_{Res}(\beta_0, \beta_1, \beta_2) - SS_{Res}(\beta_0, \beta_1, \beta_2, \beta_3))/1}{SS_{Res}(\beta_0, \beta_1, \beta_2, \beta_3)/(21)} = \frac{(43.89494 - 43.89259)/1}{43.89259/21} = 0.001127825$$

```r
partial.F.12<- anova(fit.cigs12)
partial.F.123 <- anova(fit.cigs123)

partial.F.123
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq  F value    Pr(>F)
## x1         1 494.28  494.28 236.4843 6.651e-13 ***
## x2         1   0.97    0.97   0.4661    0.5023
## x3         1   0.00    0.00   0.0011    0.9735
```

```
## Residuals 21  43.89    2.09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
partial.F.12
```

```
## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq  F value    Pr(>F)
## x1         1 494.28  494.28 247.7322 1.858e-13 ***
## x2         1   0.97    0.97   0.4882     0.492
## Residuals 22  43.89    2.00
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
SSres012 <- partial.F.12[3,2]
SSres123 <- partial.F.123[4,2]
MSres123 <- partial.F.123[4,3]
#Calculating F Statistic in two different ways
F_1a <-(SSres012-SSres123)/(MSres123)

#Comparing it with direct location of F Statistic
F_1b <- partial.F.123[3,4]
#Check whether they match
F_1a
```

```
## [1] 0.001127825
```

```r
F_1b
```

```
## [1] 0.001127825
```

```r
#HYPOTHESIS TESTING AT ALPHA = 0.05
qf(0.95, 1, 21) < F_1a
```

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

We can remark that our F-Statistic value is very small. If you perform a test of statistical hypothesis with hypotheses

$$H_0 : \beta_3 = 0 H_a : \beta_3 \neq 0$$

We would fail to reject our null hypothesis at $\alpha = 0.05$, so we have no sufficient evidence to claim that x_{i3} is influential. So we would say the reduced model

$$\mathrm{E}_{\mathrm{Y|X}}[Y_i|\mathrm{x}_i] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$$

gives a better explanation than the full model. #(d) Our sums-of-squares decomposition would be done as below:

$$\bar{SS}_R(\beta_1, \beta_2, \beta_3|\beta_0)_{(1)} = \bar{SS}_R(\beta_3|\beta_0)_{(2)} + \bar{SS}_R(\beta_2|\beta_3, \beta_0)_{(3)} + \hat{SS}_R(\beta_1|\beta_2, \beta_3, \beta_0)_{(4)}$$

1:

$$\bar{SS}_R(\beta_1, \beta_2, \beta_3|\beta_0) = \bar{SS}_R(\beta_1, \beta_2, \beta_3, \beta_0) - \bar{SS}_R(\beta_0)$$

2:

$$\bar{SS}_R(\beta_3|\beta_0) = \bar{SS}_R(\beta_3, \beta_0) - \bar{SS}_R(\beta_0)$$

3:>

$$\bar{SS}_R(\beta_2|\beta_3, \beta_0) = \bar{SS}_R(\beta_2, \beta_3, \beta_0) - \bar{SS}_R(\beta_3, \beta_0)$$

4:>

$$\bar{SS}_R(\beta_1|\beta_2,\beta_3,\beta_0) = \bar{SS}_R(\beta_1,\beta_2,\beta_3,\beta_0) - \bar{SS}_R(\beta_2,\beta_3,\beta_0)$$

```r
partial.F.321 <- anova(lm(y~x3+x2+x1))
fitted.F.321 <- fitted(lm(y~x3+x2+x1))
fitted.F.3 <- fitted(lm(y~x3))
fitted.F.23 <- fitted(lm(y~x2+x3))
fitted.F.0 <- fitted(lm(y~1))

#Computing sums-of-squares using two different ways
bar_SS_reg_321 <- sum(fitted.F.321^2)-sum(fitted.F.0^2)
bar_SS_reg_3 <- sum(fitted.F.3^2) - sum(fitted.F.0^2)
bar_SS_reg_23 <- sum(fitted.F.23^2) - sum(fitted.F.3^2)
bar_SS_reg_132 <- sum(fitted.F.321^2)-sum(fitted.F.23^2)
partial.F.321
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## x3         1 116.06  116.06  55.526 2.522e-07 ***
## x2         1 346.20  346.20 165.636 1.982e-11 ***
## x1         1  33.00   33.00  15.789 0.0006921 ***
## Residuals 21  43.89    2.09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
barSSreg1032 <- partial.F.321[3,2]
barSSreg203 <- partial.F.321[2,2]
barSSreg30 <- partial.F.321[1,2]
barSSreg123 <- sum(c(barSSreg1032,barSSreg203, barSSreg30))

#Check whether the answers match!
c(bar_SS_reg_321, bar_SS_reg_3, bar_SS_reg_23, bar_SS_reg_132)
```

```
## [1] 495.25781 116.05651 346.19988  33.00142
```

```r
c(barSSreg123, barSSreg30, barSSreg203, barSSreg1032)
```

```
## [1] 495.25781 116.05651 346.19988  33.00142
```

We have computed that

$$S\bar{S}_R(\beta_1,\beta_2,\beta_3|\beta_0) = 495.25781$$
$$S\bar{S}_R(\beta_3|\beta_0) = 116.05651$$
$$S\bar{S}_R(\beta_2|\beta_0,\beta_3) = 346.19988$$
$$S\bar{S}_R(\beta_1|\beta_0,\beta_3,\beta_2) = 33.00142$$

We can notice including β_3 and β_1 to the reduced model model

$$\mathrm{E}_{Y|X}[Y_i|\mathrm{x}_i] = \beta_0 + \beta_3 x_{i3}$$

can explain the variation in response y better.

(e)

$$\bar{SS}_R(\beta_1,\beta_2|\beta_0)_{(a)} = \bar{SS}_R(\beta_1|\beta_0)_{(b)} + \bar{SS}_R(\beta_2|\beta_0,\beta_1)_{(c)}$$

a:

$$\bar{SS}_R(\beta_1,\beta_2|\beta_0) = \bar{SS}_R(\beta_1,\beta_2,\beta_0) - \bar{SS}_R(\beta_0)$$

b:

$$\bar{SS}_R(\beta_1|\beta_0) = \bar{SS}_R(\beta_0,\beta_1) - \bar{SS}_R(\beta_0)$$

c:

$$\bar{SS}_R(\beta_2|\beta_0,\beta_1) = \bar{SS}_R(\beta_1,\beta_2,\beta_0) - \bar{SS}_R(\beta_0,\beta_1)$$

```
partial.F.12
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq  F value    Pr(>F)
## x1         1 494.28  494.28 247.7322 1.858e-13 ***
## x2         1   0.97    0.97   0.4882     0.492
## Residuals 22  43.89    2.00
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
bar_SS_reg12 <- sum(fitted(lm(y~x1+x2))^2) - sum(fitted.F.0^2)
bar_SS_reg1 <- sum(fitted(lm(y~x1))^2) - sum(fitted.F.0^2)
bar_SS_reg2 <- sum(fitted(lm(y~x1+x2))^2) - sum(fitted(lm(y~x1))^2)
barSSreg10 <- partial.F.12[1,2]
barSSreg201 <- partial.F.12[2,2]
barSSreg120 <- barSSreg10+barSSreg201
c(bar_SS_reg12, bar_SS_reg1, bar_SS_reg2)
```

```
## [1] 495.2554571 494.2813099   0.9741472
```

```
c(barSSreg120,barSSreg10, barSSreg201)
```

```
## [1] 495.2554571 494.2813099   0.9741472
```

$$S\bar{S}_R(\beta_1,\beta_2|\beta_0) = 495.2554571$$
$$S\bar{S}_R(\beta_1|\beta_0) = 494.2813099$$
$$S\bar{S}_R(\beta_2|\beta_0,\beta_1) = 0.9741472$$

#(f)

```
partial.F.1 <- anova(fit.cigs1)
partial.F.1
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## x1         1 494.28  494.28  253.37 6.552e-14 ***
## Residuals 23  44.87    1.95
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
partial.F.12
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq  F value    Pr(>F)
## x1         1 494.28  494.28 247.7322 1.858e-13 ***
## x2         1   0.97    0.97   0.4882     0.492
## Residuals 22  43.89    2.00
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
SSres01 <- partial.F.1[2,2]
MSres012 <- partial.F.12[3,3]

F_2a <- (SSres01-SSres012)/MSres012
F_2b <- partial.F.12[2,4]
F_2a
```

```
## [1] 0.4882394
```

```
F_2b
```

```
## [1] 0.4882394
```

```
qf(0.95, 1, 22) < F_2a
```

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

We know that x_{i3} is not influential. Given a test of hypotheses

$$H_0 : \beta_2 = 0 H_a : \beta_2 \neq 0$$

,under $\alpha = 0.05$ we can notice that our F value is smaller than $100 * (1 - \alpha)\%$ quantile with degrees of freedom 1 and 22, so we can conclude that there is no sufficient evidence to reject our null hypothesis. So we would say the reduced model

$$\mathrm{E}_{\mathrm{Y|X}}[Y_i|\mathrm{x}_i] = \beta_0 + \beta_1 x_{i1}$$

is a better model than

$$\mathrm{E}_{\mathrm{Y|X}}[Y_i|\mathrm{x}_i] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$$

(g)

```
partial.F.0 <- anova(fit.intercept)
partial.F.0
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value Pr(>F)
## Residuals 24 539.15  22.465
```

```
partial.F.12
```

```
## Analysis of Variance Table
##
## Response: y
```

```
##            Df Sum Sq Mean Sq  F value    Pr(>F)    
## x1          1 494.28  494.28 247.7322 1.858e-13 ***
## x2          1   0.97    0.97   0.4882     0.492    
## Residuals 22  43.89    2.00                       
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
SSres0 <- partial.F.0[1,2]

F3a <- (SSres0-SSres012)/MSres012
F3b <- partial.F.12[1,4]+partial.F.12[2,4]
F3a
```

```
## [1] 248.2204
```

```r
F3b
```

```
## [1] 248.2204
```

```r
qf(0.95, 1, 22) < F3a
```

```
## [1] TRUE
```

Given hypothesis

$$H_0 : \beta_1 = \beta_2 = 0 H_a : \text{At least one of } \beta_{\text{j}} \text{ is not zero }_{j=1,2}$$

under $\alpha = 0.05$, we have sufficient evidence to reject our null hypothesis. Therefore we say the "full" model

$$\mathrm{E}_{\mathrm{Y}|\mathrm{X}}[Y_i|\mathrm{x}_i] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$$

is better than the reduced model

$$\mathrm{E}_{\mathrm{Y}|\mathrm{X}}[Y_i|\mathrm{x}_i] = \beta_0$$

.