Assignment 4

YUNHEUM DAN SEOL

2017-12-06

Question 1

1.TestScores.csv contains data on standardized math test scores of 45 students from three Faculties in a university

(a)

```
library(readr)
```

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

```
Scores <- read_csv('http://www.math.mcgill.ca/yyang/regression/data/TestScores.csv')
```

```
## Parsed with column specification:
## cols(
##   Faculty = col_integer(),
##   Score = col_integer()
## )
```

```
str(Scores)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    45 obs. of  2 variables:
##  $ Faculty: int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Score  : int  44 40 44 39 25 37 31 40 22 34 ...
##  - attr(*, "spec")=List of 2
##   ..$ cols   :List of 2
##   .. ..$ Faculty: list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   .. ..$ Score  : list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   ..$ default: list()
##   .. ..- attr(*, "class")= chr  "collector_guess" "collector"
##   ..- attr(*, "class")= chr "col_spec"
```

```
head(Scores)
```

```
## # A tibble: 6 x 2
##   Faculty Score
##     <int> <int>
## 1       1    44
## 2       1    40
## 3       1    44
## 4       1    39
## 5       1    25
## 6       1    37
```

```
Scores$Faculty <- as.factor(Scores$Faculty)
str(Scores)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    45 obs. of  2 variables:
##  $ Faculty: Factor w/ 3 levels "1","2","3": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Score  : int  44 40 44 39 25 37 31 40 22 34 ...
##  - attr(*, "spec")=List of 2
##   ..$ cols   :List of 2
##   .. ..$ Faculty: list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   .. ..$ Score  : list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   ..$ default: list()
##   .. ..- attr(*, "class")= chr  "collector_guess" "collector"
##   ..- attr(*, "class")= chr "col_spec"
```

```r
fit.Scores.1 <- lm(Score~I(Faculty), data =Scores)
summary(fit.Scores.1)
```

```
##
## Call:
## lm(formula = Score ~ I(Faculty), data = Scores)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.800  -2.200   1.133   3.800   9.133
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 35.80000    1.59589  22.433  < 2e-16 ***
## I(Faculty)2  0.06667    2.25694   0.030    0.977
## I(Faculty)3 12.40000    2.25694   5.494 2.11e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.181 on 42 degrees of freedom
## Multiple R-squared:  0.488,  Adjusted R-squared:  0.4636
## F-statistic: 20.02 on 2 and 42 DF,  p-value: 7.843e-07
```

```r
anova(fit.Scores.1)
```

```
## Analysis of Variance Table
##
## Response: Score
##            Df Sum Sq Mean Sq F value    Pr(>F)
## I(Faculty)  2 1529.4  764.69  20.016 7.843e-07 ***
## Residuals  42 1604.5   38.20
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can conclude that there is a significant difference of scores between Faculties looking at the global fit from our F-test; moreover, we can further test between what specific faculties we can find evidence for the difference in scores and check whether our model actually holds by checking the residual plot. We do this in part (b).

(b)

```
library(lsmeans)
```

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

## The 'lsmeans' package is being deprecated.
## Users are encouraged to switch to 'emmeans'.
## See help('transition') for more information, including how
## to convert 'lsmeans' objects and scripts to work with 'emmeans'.
```

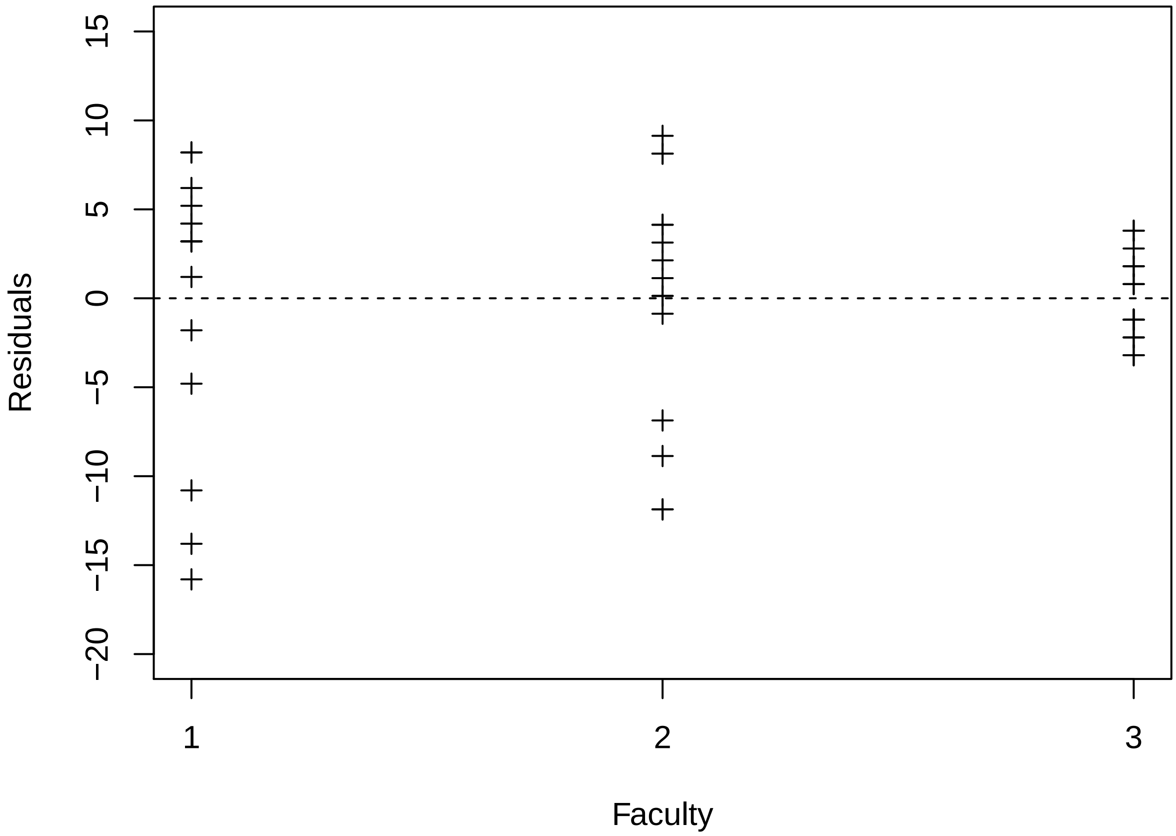
```
fit.Scores.2 <- lm(Score ~-1+I(Faculty), data = Scores)
summary(fit.Scores.2)
```

```
##
## Call:
## lm(formula = Score ~ -1 + I(Faculty), data = Scores)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.800  -2.200   1.133   3.800   9.133
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## I(Faculty)1   35.800      1.596   22.43   <2e-16 ***
## I(Faculty)2   35.867      1.596   22.47   <2e-16 ***
## I(Faculty)3   48.200      1.596   30.20   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.181 on 42 degrees of freedom
## Multiple R-squared:  0.9786, Adjusted R-squared:  0.9771
## F-statistic: 640.2 on 3 and 42 DF,  p-value: < 2.2e-16
```

```
#Acquiring means through lsmeans()
lsmeans(fit.Scores.2, ~I(Faculty))
```

```
##  Faculty   lsmean       SE df lower.CL upper.CL
##  1       35.80000 1.595894 42 32.57936 39.02064
##  2       35.86667 1.595894 42 32.64602 39.08731
##  3       48.20000 1.595894 42 44.97936 51.42064
##
## Confidence level used: 0.95
```

```
#Residual plot
residual.data<-data.frame(Residuals=residuals(fit.Scores.1),Faculty =Scores$Faculty)
par(mar=c(4,4,1,2))
stripchart(Residuals ~ Faculty,data = residual.data,pch=3,vertical=T,ylim=range(-20,15),
           xlab='Faculty') ;abline(h=0,lty=2)
```



The estimated mean score for faculty 1 would be 35.8, for faculty 2 would be $35.8000 + 0.0667 = 35.8667$, and for faculty 3 it would be $35.8 + 12.4 = 48.2$ with respective standard error 1.595894 for all faculties. A visual inspection of the residual plot suggests that the group variance for Faculty 3 might be smaller than that of others.

Question 2

```
#Reading in date on the noise emission level of 36 cars
Filter <- read_csv("http://www.math.mcgill.ca/yyang/regression/data/Filter.csv")
```

```
## Parsed with column specification:
## cols(
##   noise = col_integer(),
##   carsize = col_character(),
##   type = col_character()
## )
```

```
str(Filter)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    36 obs. of  3 variables:
##  $ noise  : int  810 820 820 840 840 845 785 790 785 835 ...
##  $ carsize: chr  "small car" "small car" "small car" "medium car" ...
##  $ type   : chr  "normal filter" "normal filter" "normal filter" "normal filter" ...
##  - attr(*, "spec")=List of 2
##   ..$ cols   :List of 3
##   .. ..$ noise  : list()
```

```
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   .. ..$ carsize: list()
##   .. .. ..- attr(*, "class")= chr  "collector_character" "collector"
##   .. ..$ type   : list()
##   .. .. ..- attr(*, "class")= chr  "collector_character" "collector"
##   ..$ default: list()
##   .. ..- attr(*, "class")= chr  "collector_guess" "collector"
##   ..- attr(*, "class")= chr "col_spec"
```

```r
fmodel1 <-lm(noise~1,data=Filter);summary(fmodel1)
```

```
##
## Call:
## lm(formula = noise ~ 1, data = Filter)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -50.139 -27.639   9.861  17.361  44.861
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  810.139      4.869   166.4   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29.22 on 35 degrees of freedom
```

```r
anova(fmodel1)
```

```
## Analysis of Variance Table
##
## Response: noise
##           Df Sum Sq Mean Sq F value Pr(>F)
## Residuals 35  29874  853.55
```

```r
#Adding the sum function to get the numeric data
SSres1<-sum(anova(fmodel1)[2])
```

```r
fmodel2 <-lm(noise~I(carsize),data=Filter);summary(fmodel2)
```

```
##
## Call:
## lm(formula = noise ~ I(carsize), data = Filter)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.7500  -8.7500  -0.8333  10.8333  21.2500
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            772.500      3.107  248.63  < 2e-16 ***
## I(carsize)medium car    61.250      4.394   13.94 2.20e-15 ***
## I(carsize)small car     51.667      4.394   11.76 2.42e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.76 on 33 degrees of freedom
```

```
## Multiple R-squared:  0.872,  Adjusted R-squared:  0.8643
## F-statistic: 112.4 on 2 and 33 DF,  p-value: 1.85e-15
```

```r
anova(fmodel2)
```

```
## Analysis of Variance Table
##
## Response: noise
##            Df  Sum Sq Mean Sq F value   Pr(>F)
## I(carsize)  2 26051.4 13025.7  112.44 1.85e-15 ***
## Residuals  33  3822.9   115.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
SSres2<-sum(anova(fmodel2)[2,2])
```

```r
fmodel3 <-lm(noise~I(type),data=Filter);summary(fmodel3)
```

```
##
## Call:
## lm(formula = noise ~ I(type), data = Filter)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -55.56 -29.72  15.28  20.28  39.44
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           815.556      6.862 118.849   <2e-16 ***
## I(type)Octel filter  -10.833      9.704  -1.116    0.272
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29.11 on 34 degrees of freedom
## Multiple R-squared:  0.03536,    Adjusted R-squared:  0.006985
## F-statistic: 1.246 on 1 and 34 DF,  p-value: 0.2721
```

```r
anova(fmodel3)
```

```
## Analysis of Variance Table
##
## Response: noise
##           Df  Sum Sq Mean Sq F value Pr(>F)
## I(type)    1  1056.2 1056.25  1.2462 0.2721
## Residuals 34 28818.1  847.59
```

```r
SSres3<-sum(anova(fmodel3)[2,2])
```

```r
fmodel4 <-lm(noise~I(carsize)+I(type),data=Filter);summary(fmodel4)
```

```
##
## Call:
## lm(formula = noise ~ I(carsize) + I(type), data = Filter)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.583  -7.292   1.250   6.250  15.833
##
```

```
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            777.917      3.099 250.987  < 2e-16 ***
## I(carsize)medium car    61.250      3.796  16.135  < 2e-16 ***
## I(carsize)small car     51.667      3.796  13.611  7.4e-15 ***
## I(type)Octel filter    -10.833      3.099  -3.495  0.00141 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.298 on 32 degrees of freedom
## Multiple R-squared:  0.9074, Adjusted R-squared:  0.8987
## F-statistic: 104.5 on 3 and 32 DF,  p-value: < 2.2e-16
```

```
anova(fmodel4)
```

```
## Analysis of Variance Table
##
## Response: noise
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## I(carsize)  2 26051.4 13025.7 150.659 < 2.2e-16 ***
## I(type)     1  1056.2  1056.2  12.217  0.001411 **
## Residuals  32  2766.7    86.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
SSres4<-sum(anova(fmodel4)[3,2])
```

```
fmodel5 <-lm(noise~I(carsize)*I(type),data=Filter);summary(fmodel5)
```

```
##
## Call:
## lm(formula = noise ~ I(carsize) * I(type), data = Filter)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.8333  -5.2083  -0.4167   5.0000  15.0000
##
## Coefficients:
##                                           Estimate Std. Error t value
## (Intercept)                                775.000      3.302 234.711
## I(carsize)medium car                        70.833      4.670  15.169
## I(carsize)small car                         50.833      4.670  10.886
## I(type)Octel filter                         -5.000      4.670  -1.071
## I(carsize)medium car:I(type)Octel filter   -19.167      6.604  -2.902
## I(carsize)small car:I(type)Octel filter      1.667      6.604   0.252
##                                           Pr(>|t|)
## (Intercept)                                < 2e-16 ***
## I(carsize)medium car                      1.30e-15 ***
## I(carsize)small car                       6.11e-12 ***
## I(type)Octel filter                        0.29282
## I(carsize)medium car:I(type)Octel filter   0.00688 **
## I(carsize)small car:I(type)Octel filter    0.80247
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.088 on 30 degrees of freedom
```

```
## Multiple R-squared:  0.9343, Adjusted R-squared:  0.9234
## F-statistic: 85.34 on 5 and 30 DF,  p-value: < 2.2e-16
```

```r
anova(fmodel5)
```

```
## Analysis of Variance Table
##
## Response: noise
##                    Df  Sum Sq Mean Sq  F value    Pr(>F)
## I(carsize)          2 26051.4 13025.7 199.1189 < 2.2e-16 ***
## I(type)             1  1056.2  1056.2  16.1465 0.0003631 ***
## I(carsize):I(type)  2   804.2   402.1   6.1465 0.0057915 **
## Residuals          30  1962.5    65.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
SSres5<-sum(anova(fmodel5)[4,2])
```

```r
t <- c("1", "1+X1", "1+X2", "1+X1+X2", "1+X1+X2+X1:X2")
v <- c(SSres1, SSres2, SSres3, SSres4, SSres5)
w <- c("Intercept", "Main Effect of Car Size",
       "Main Effect of Filter Type", "Main Effects only",
       "Main Effects plus interactions")
u <- c("lm(noise~1",
       "lm(noise~I(carsize))",
       "lm(noise~I(type))",
       "lm(noise~I(carsize)+I(type))",
       "lm(noise~I(carsize)*I(type))")
mtx <- cbind(t,w,round(v,3),u)
colnames(mtx) <- c("Model","Sum of Sq", "Description", "R")
mtx <- data.frame(mtx)
#table generated by kable(mtx, format = "latex") below
```

	Model	Sum.of.Sq	Description	R
1	1	Intercept	29874.306	lm(noise˜1)
2	1+X1	Main Effect of Car Size	3822.917	lm(noise˜I(carsize))
3	1+X2	Main Effect of Filter Type	28818.056	lm(noise˜I(type))
4	1+X1+X2	Main Effects only	2766.667	lm(noise˜I(carsize)+I(type))
5	1+X1+X2+X1:X2	Main Effects plus interactions	1962.5	lm(noise˜I(carsize)*I(type))

part (b)

```r
anova(fmodel2,fmodel5)
```

```
## Analysis of Variance Table
##
## Model 1: noise ~ I(carsize)
## Model 2: noise ~ I(carsize) * I(type)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     33 3822.9
## 2     30 1962.5  3    1860.4 9.4798 0.0001461 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
sum(anova(fmodel2, fmodel5)[2,3])
```

```
## [1] 3
```

```r
df1 <- sum(anova(fmodel2, fmodel5)[2,3]);df2<-sum(anova(fmodel2, fmodel5)[2,1])
c(df1, df2)
```

```
## [1]  3 30
```

```r
fstat <- sum(anova(fmodel2, fmodel5)[2,5]);fstat
```

```
## [1] 9.47983
```

```r
crit.value <- qf(0.95,df1,df2);crit.value
```

```
## [1] 2.922277
```

```r
pvalue<- 1-pf(fstat,df1,df2);pvalue
```

```
## [1] 0.0001460971
```

at

$$\alpha = 0.05$$

, we get the F critical value 2.922277, since our F statistic $9,4798$ is larger than the critical value, we would reject our null hypothesis in the test

$$H_0 = \mathrm{E}[Y_i|\mathbf{x_i}] = 1 + X_1 H_a = \mathrm{E}[Y_i|\mathbf{x_i}] = 1 + X_1 + X_2 + X_1 X_2$$

Where X1 = (the main effect of the size of the car to noise) and X2 = (the main effect of the filter type) and X1X2 being the interaction term

Question 3

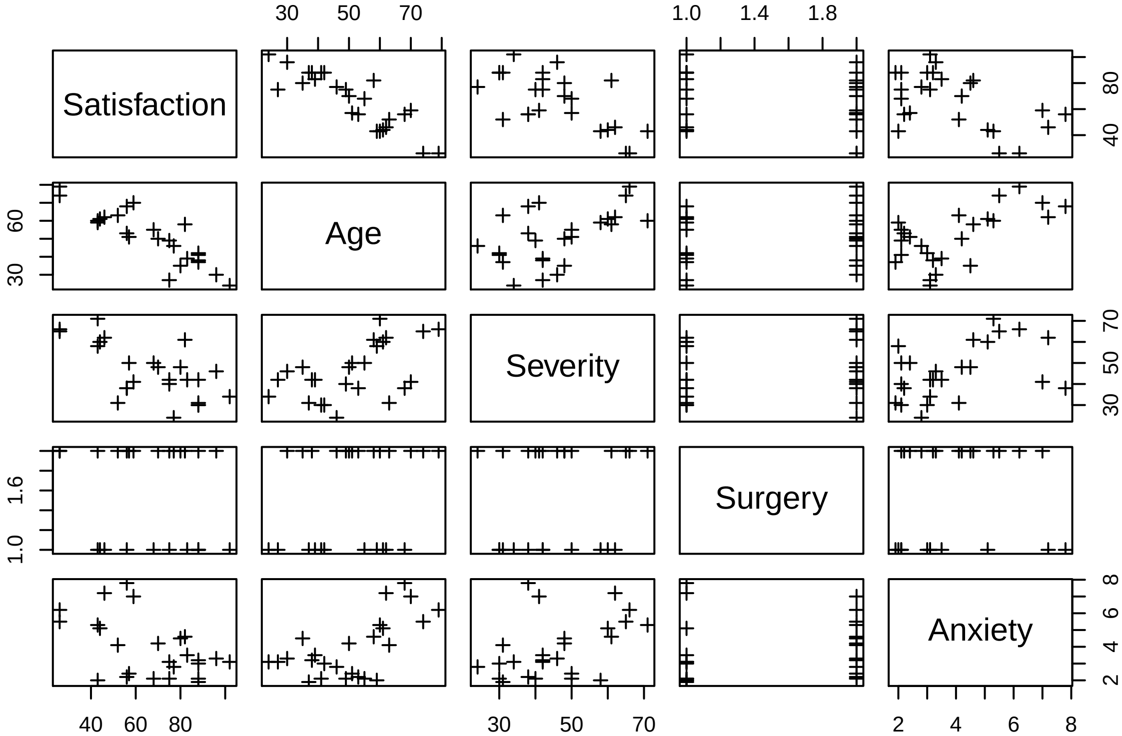
```r
#Reading in the information on patient satisfaction for 25 patients
#having undergone treatment at a hospital for the same condition
PS <- read.csv("http://www.math.mcgill.ca/yyang/regression/data/PatSat.csv", header=TRUE)
head(PS)
```

```
##   Satisfaction Age Severity Surgery Anxiety
## 1           68  55       50      No     2.1
## 2           77  46       24     Yes     2.8
## 3           96  30       46     Yes     3.3
## 4           80  35       48     Yes     4.5
## 5           43  59       58      No     2.0
## 6           44  61       60      No     5.1
```

```r
str(PS)
```

```
## 'data.frame':    25 obs. of  5 variables:
##  $ Satisfaction: int  68 77 96 80 43 44 26 88 75 57 ...
##  $ Age         : int  55 46 30 35 59 61 74 38 27 51 ...
##  $ Severity    : int  50 24 46 48 58 60 65 42 42 50 ...
##  $ Surgery     : Factor w/ 2 levels "No","Yes": 1 2 2 2 1 1 2 2 1 2 ...
##  $ Anxiety     : num  2.1 2.8 3.3 4.5 2 5.1 5.5 3.2 3.1 2.4 ...
```

```
#Renaming predictors and responses to make it simple
y_PS <- PS$Satisfaction
x1_PS <- PS$Surgery
x2_PS <- PS$Age
x3_PS <- PS$Severity
x4_PS <- PS$Anxiety
#Pairwise scatterplots
pairs(PS, pch=3)
```




```
#Fitting models
fit0_PS<-lm(y_PS~1)
fit1_PS<-lm(y_PS~I(x1_PS))
fit2_PS<-lm(y_PS~x2_PS)
fit3_PS<-lm(y_PS~x3_PS)
fit4_PS<-lm(y_PS~x4_PS)
fit12_PS<-lm(y_PS~I(x1_PS)+x2_PS)
fit13_PS<-lm(y_PS~I(x1_PS)+x3_PS)
fit14_PS<-lm(y_PS~I(x1_PS)+x4_PS)
fit23_PS<-lm(y_PS~x2_PS+x3_PS)
fit24_PS<-lm(y_PS~x2_PS+x4_PS)
fit34_PS<-lm(y_PS~x3_PS+x4_PS)
fit12i_PS<-lm(y_PS~I(x1_PS)*x2_PS)
fit13i_PS<-lm(y_PS~I(x1_PS)*x3_PS)
fit14i_PS<-lm(y_PS~I(x1_PS)*x4_PS)
fit23i_PS<-lm(y_PS~x2_PS*x3_PS)
fit24i_PS<-lm(y_PS~x2_PS*x4_PS)
fit34i_PS<-lm(y_PS~x3_PS*x4_PS)
fit123_PS<-lm(y_PS~I(x1_PS)+x2_PS+x3_PS)
fit124_PS<-lm(y_PS~I(x1_PS)+x2_PS+x4_PS)
fit134_PS<-lm(y_PS~I(x1_PS)+x3_PS+x4_PS)
```

```r
fit234_PS<-lm(y_PS~x2_PS+x3_PS+x3_PS)
fit123i_PS<-lm(y_PS~I(x1_PS)*x2_PS*x3_PS)
fit124i_PS<-lm(y_PS~I(x1_PS)*x2_PS*x4_PS)
fit134i_PS<-lm(y_PS~I(x1_PS)*x3_PS*x4_PS)
fit234i_PS<-lm(y_PS~x2_PS*x3_PS*x3_PS)
fit1234_PS<- lm(y_PS~I(x1_PS)+x2_PS+x3_PS+x4_PS)
fit1234i_PS<- lm(y_PS~I(x1_PS)*x2_PS*x3_PS*x4_PS)
```

```r
PS_numeric<-PS;PS_numeric$Surgery<-as.numeric(PS_numeric$Surgery)-1
cor(PS_numeric)
```

```
##              Satisfaction        Age   Severity    Surgery    Anxiety
## Satisfaction    1.0000000 -0.8707049 -0.6531434 -0.1822682 -0.5127287
## Age            -0.8707049  1.0000000  0.5290246  0.2456932  0.6212453
## Severity       -0.6531434  0.5290246  1.0000000  0.1775101  0.4471567
## Surgery        -0.1822682  0.2456932  0.1775101  1.0000000  0.1096486
## Anxiety        -0.5127287  0.6212453  0.4471567  0.1096486  1.0000000
```

It appears that there are some strong correlations among the predictors. Now, let's fit a model with surgery factor predictor only:

```r
summary(fit1_PS)
```

```
##
## Call:
## lm(formula = y_PS ~ I(x1_PS))
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -37.36 -15.00   4.00  17.00  32.64
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   71.000      6.433  11.036 1.15e-10 ***
## I(x1_PS)Yes   -7.643      8.597  -0.889    0.383
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.34 on 23 degrees of freedom
## Multiple R-squared:  0.03322,    Adjusted R-squared:  -0.008812
## F-statistic: 0.7904 on 1 and 23 DF,  p-value: 0.3832
```

It doesn't seem like as if having surgery significantly affected the patient satisfaction.

```r
#We begin the best model identification
#by examining the additive model that fits all predictors as main effects:

summary(fit1234_PS)
```

```
##
## Call:
## lm(formula = y_PS ~ I(x1_PS) + x2_PS + x3_PS + x4_PS)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.506  -5.096   1.306   4.738  28.722
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 140.1689     8.3191  16.849 2.77e-13 ***
## I(x1_PS)Yes   2.2259     4.1402   0.538   0.5968
## x2_PS        -1.1428     0.1904  -6.002 7.22e-06 ***
## x3_PS        -0.4699     0.1866  -2.518   0.0204 *
## x4_PS         1.2673     1.4922   0.849   0.4058
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.921 on 20 degrees of freedom
## Multiple R-squared:  0.8183, Adjusted R-squared:  0.7819
## F-statistic: 22.51 on 4 and 20 DF,  p-value: 3.611e-07
```

```r
drop1(fit1234_PS, test='F')
```

```
## Single term deletions
##
## Model:
## y_PS ~ I(x1_PS) + x2_PS + x3_PS + x4_PS
##          Df Sum of Sq    RSS    AIC F value   Pr(>F)
## <none>                1968.5 119.15
## I(x1_PS)  1      28.4 1997.0 117.51  0.2890  0.59677
## x2_PS     1    3545.1 5513.7 142.90 36.0182 7.22e-06 ***
## x3_PS     1     624.1 2592.6 124.04  6.3408  0.02043 *
## x4_PS     1      71.0 2039.5 118.04  0.7212  0.40579
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It seems that we can drop X1(Surgery) and X4(Anxiety) from the model.

```r
#Deleting Anxiety as a predictor
fita_PS <- update(fit1234_PS, ~.-I(x1_PS)-x4_PS)
anova(fita_PS, fit1234_PS, test='F')
```

```
## Analysis of Variance Table
##
## Model 1: y_PS ~ x2_PS + x3_PS
## Model 2: y_PS ~ I(x1_PS) + x2_PS + x3_PS + x4_PS
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     22 2062.3
## 2     20 1968.5  2    93.754 0.4763  0.628
```

```r
#
summary(fita_PS)
```

```
##
## Call:
## lm(formula = y_PS ~ x2_PS + x3_PS)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.3691  -5.9535   0.2975   4.0462  29.3439
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 139.9233     8.1002  17.274 2.78e-14 ***
## x2_PS        -1.0462     0.1573  -6.652 1.09e-06 ***
## x3_PS        -0.4359     0.1788  -2.439   0.0233 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.682 on 22 degrees of freedom
## Multiple R-squared:  0.8096, Adjusted R-squared:  0.7923
## F-statistic: 46.77 on 2 and 22 DF,  p-value: 1.193e-08
```

We can see that we fail to reject the null hypothesis that anxiety is a significant predictor for patient satisfaction. Let's see what roles the interactions might play.

```r
fitb_PS<-update(fita_PS, ~.+I(x1_PS)*x2_PS*x4_PS)
fitc_PS<-update(fitb_PS, ~.-I(x1_PS):x2_PS:x4_PS)
anova(fita_PS,fitc_PS,fitb_PS,test='F')
```

```
## Analysis of Variance Table
##
## Model 1: y_PS ~ x2_PS + x3_PS
## Model 2: y_PS ~ x2_PS + x3_PS + I(x1_PS) + x4_PS + x2_PS:I(x1_PS) + I(x1_PS):x4_PS +
##     x2_PS:x4_PS
## Model 3: y_PS ~ x2_PS + x3_PS + I(x1_PS) + x4_PS + x2_PS:I(x1_PS) + I(x1_PS):x4_PS +
##     x2_PS:x4_PS + x2_PS:I(x1_PS):x4_PS
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     22 2062.3
## 2     17 1739.6  5    322.72 0.5978 0.7023
## 3     16 1727.5  1     12.06 0.1117 0.7425
```

It seems that our model with Age and Severity only is still a better model than the others.

```r
fitd_PS<- update(fita_PS, ~.+x2_PS:x3_PS)
fite_PS<- update(fitd_PS, ~.+I(x1_PS))
anova(fita_PS, fitd_PS, fite_PS, test='F')
```

```
## Analysis of Variance Table
##
## Model 1: y_PS ~ x2_PS + x3_PS
## Model 2: y_PS ~ x2_PS + x3_PS + x2_PS:x3_PS
## Model 3: y_PS ~ x2_PS + x3_PS + I(x1_PS) + x2_PS:x3_PS
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     22 2062.3
## 2     21 2032.7  1    29.549 0.2929 0.5944
## 3     20 2017.8  1    14.967 0.1483 0.7042
```

Our model with Age and Severity only still remains to be a better model than the others.

Let's try some automated methods ###methods with step command

```r
fit.stepa_PS <-step(fit1234i_PS, k=2, trace=0)
print(summary(fit.stepa_PS), concise=T)
```

```
##
## Call:
## lm(formula = y_PS ~ I(x1_PS) + x2_PS + x3_PS + x4_PS + I(x1_PS):x2_PS +
##     I(x1_PS):x3_PS + x2_PS:x3_PS + I(x1_PS):x4_PS + x2_PS:x4_PS +
##     x3_PS:x4_PS + I(x1_PS):x2_PS:x4_PS + x2_PS:x3_PS:x4_PS)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.0565  -5.5923  -0.6494   3.0863  24.6576
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            239.72040  160.09378   1.497    0.160
## I(x1_PS)Yes            132.08818  107.10595   1.233    0.241
## x2_PS                   -2.01705    3.34228  -0.603    0.557
## x3_PS                   -5.34778    4.21765  -1.268    0.229
## x4_PS                  -25.67624   39.13060  -0.656    0.524
## I(x1_PS)Yes:x2_PS       -3.34196    2.14892  -1.555    0.146
## I(x1_PS)Yes:x3_PS        1.03425    0.60872   1.699    0.115
## x2_PS:x3_PS              0.06882    0.07897   0.872    0.401
## I(x1_PS)Yes:x4_PS      -40.96403   33.53151  -1.222    0.245
## x2_PS:x4_PS              0.41667    0.67947   0.613    0.551
## x3_PS:x4_PS              1.22388    0.95373   1.283    0.224
## I(x1_PS)Yes:x2_PS:x4_PS   0.72484    0.54702   1.325    0.210
## x2_PS:x3_PS:x4_PS       -0.02000    0.01639  -1.220    0.246
##
## Residual standard error: 10.36 on 12 degrees of freedom
## Multiple R-squared:  0.8811, Adjusted R-squared:  0.7623
## F-statistic: 7.412 on 12 and 12 DF,  p-value: 0.0007637
```

```
fit.stepb_PS <- update(fit.stepa_PS, ~.-I(x1_PS):x2_PS:x4_PS-x2_PS:x3_PS:x4_PS)
anova(fit.stepb_PS, fit.stepa_PS)
```

```
## Analysis of Variance Table
##
## Model 1: y_PS ~ I(x1_PS) + x2_PS + x3_PS + x4_PS + I(x1_PS):x2_PS + I(x1_PS):x3_PS +
##     x2_PS:x3_PS + I(x1_PS):x4_PS + x2_PS:x4_PS + x3_PS:x4_PS
## Model 2: y_PS ~ I(x1_PS) + x2_PS + x3_PS + x4_PS + I(x1_PS):x2_PS + I(x1_PS):x3_PS +
##     x2_PS:x3_PS + I(x1_PS):x4_PS + x2_PS:x4_PS + x3_PS:x4_PS +
##     I(x1_PS):x2_PS:x4_PS + x2_PS:x3_PS:x4_PS
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     14 1520.1
## 2     12 1287.5  2    232.56 1.0838 0.3693
```

```
fit.stepc_PS <- step(lm(y_PS ~(Surgery+x2_PS+x3_PS+x4_PS)^2, data=PS), k=2, trace=0)
summary(fit.stepc_PS)
```

```
##
## Call:
## lm(formula = y_PS ~ Surgery + x2_PS + x3_PS + Surgery:x2_PS +
##     Surgery:x3_PS, data = PS)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.508  -5.577  -1.272   3.764  26.465
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     145.1381    11.4691  12.655 1.05e-10 ***
## SurgeryYes       -5.2017    16.2534  -0.320  0.75243
## x2_PS            -0.7182     0.2575  -2.789  0.01169 *
```

```
## x3_PS              -0.9342     0.3122  -2.993  0.00748 **
## SurgeryYes:x2_PS  -0.4996     0.3260  -1.532  0.14194
## SurgeryYes:x3_PS   0.7085     0.3771   1.879  0.07571 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.443 on 19 degrees of freedom
## Multiple R-squared:  0.8436, Adjusted R-squared:  0.8024
## F-statistic: 20.49 on 5 and 19 DF,  p-value: 4.676e-07
```

This seems to be a relatively better model that includes the effect of surgery, but could we come up with a better way to compare models?

custom methods to compare all the measures : Used a method that was shown in one of the notes ("Factor Predictors - Examples")

```r
bigs.hat <- summary(fit1234i_PS)$sigma
criteria.eval <- function(fit.obj,nv,bigsig.hat){
  cvec <- rep(0,5)
  SSRes <- sum(residuals(fit.obj)^2)
  p <- length(coef(fit.obj))

  #R Squared
  cvec[1] <- summary(fit.obj)$r.squared
  #Adjusted R Squared
  cvec[2] <- summary(fit.obj)$adj.r.squared
  #Cp
  cvec[3] <- SSRes/bigsig.hat^2-nv+2*p
  #AIC in R computes
  #n*log(sum(residuals(fit.obj)^2)/n)+2*(length(coef(fit.obj)+1)+n*log(2*pi)+n)
  cvec[4] <- AIC(fit.obj)
  #BIC in R computes
  #n*log(sum(residuals(fit.obj)^2)/n)+2*length(coef(fit.obj))+1)+n*log(2*pi)+n
  cvec[5] <- BIC(fit.obj)
  return(cvec)}
```

```r
cvals <- matrix(0, nrow=28, ncol=5)

cvals[1, ] <-  criteria.eval(fit0_PS, 25, bigs.hat)
cvals[2, ] <-  criteria.eval(fit1_PS, 25, bigs.hat)
cvals[3, ] <-  criteria.eval(fit2_PS, 25, bigs.hat)
cvals[4, ] <-  criteria.eval(fit3_PS, 25, bigs.hat)
cvals[5, ] <-  criteria.eval(fit4_PS, 25, bigs.hat)
cvals[6, ] <-  criteria.eval(fit12_PS, 25, bigs.hat)
cvals[7, ] <-  criteria.eval(fit13_PS, 25, bigs.hat)
cvals[8, ] <-  criteria.eval(fit14_PS, 25, bigs.hat)
cvals[9, ] <-  criteria.eval(fit23_PS, 25, bigs.hat)
cvals[10, ] <-  criteria.eval(fit24_PS, 25, bigs.hat)
cvals[11, ] <-  criteria.eval(fit34_PS, 25, bigs.hat)
cvals[12, ] <-  criteria.eval(fit12i_PS, 25, bigs.hat)
cvals[13, ] <-  criteria.eval(fit13i_PS, 25, bigs.hat)
cvals[14, ] <-  criteria.eval(fit14i_PS, 25, bigs.hat)
cvals[15, ] <-  criteria.eval(fit23i_PS, 25, bigs.hat)
```

```r
cvals[16, ] <-  criteria.eval(fit24i_PS, 25, bigs.hat)
cvals[17, ] <-  criteria.eval(fit34i_PS, 25, bigs.hat)
cvals[18, ] <-  criteria.eval(fit123_PS, 25, bigs.hat)
cvals[19, ] <-  criteria.eval(fit124_PS, 25, bigs.hat)
cvals[20, ] <-  criteria.eval(fit134_PS, 25, bigs.hat)
cvals[21, ] <-  criteria.eval(fit234_PS, 25, bigs.hat)
cvals[22, ] <-  criteria.eval(fit123i_PS, 25, bigs.hat)
cvals[23, ] <-  criteria.eval(fit124i_PS, 25, bigs.hat)
cvals[24, ] <-  criteria.eval(fit134i_PS, 25, bigs.hat)
cvals[25, ] <-  criteria.eval(fit234i_PS, 25, bigs.hat)
cvals[26, ] <-  criteria.eval(fit1234_PS, 25, bigs.hat)
cvals[27, ] <-  criteria.eval(fit1234i_PS, 25, bigs.hat)
cvals[28,]  <- criteria.eval(fit.stepc_PS, 25, bigs.hat)

Criteria <- data.frame(cvals)

names(Criteria) <- c('Rsq', 'Adj.Rsq', 'Cp', 'AIC', 'BIC')
rownames(Criteria) <- c('Intercept',
                        'Surgery',
                        'Age',
                        'Severity',
                        'Anxiety',
                        'Surgery+Age',
                        'Surgery+Severity',
                        'Surgery+Anxiety',
                        'Age+Severity',
                        'Age+Anxiety',
                        'Severity+Anxiety',
                        'Surgery*Age',
                        'Surgery*Severity',
                        'Surgery*Anxiety',
                        'Age*Severity',
                        'Age*Anxiety',
                        'Severity*Anxiety',
                        'Surgery+Age+Severity',
                        'Surgery+Age+Anxiety',
                        'Surgery+Severity+Anxiety',
                        'Age+Severity+Anxiety',
                        'Surgery*Age*Severity',
                        'Surgery*Age*Anxiety',
                        'Surgery*Severity*Anxiety',
                        'Age*Severity*Anxiety',
                        'Surgery+Age+Severity+Anxiety',
                        'Surgery*Age*Severity*Anxiety',
                        'Step-selected model')
round(Criteria, 4)
```

```
##                                 Rsq Adj.Rsq      Cp      AIC      BIC
## Intercept                    0.0000  0.0000 62.0037 226.7293 229.1671
## Surgery                      0.0332 -0.0088 61.1798 227.8847 231.5413
## Age                          0.7581  0.7476 -0.4399 193.2457 196.9024
## Severity                     0.4266  0.4017 27.7415 214.8252 218.4818
## Anxiety                      0.2629  0.2308 41.6570 221.1038 224.7605
## Surgery+Age                  0.7592  0.7373  1.4694 195.1353 200.0108
```

```
## Surgery+Severity            0.4311  0.3794 29.3553 216.6263 221.5018
## Surgery+Anxiety             0.2790  0.2134 42.2901 222.5524 227.4279
## Age+Severity                0.8096  0.7923 -2.8148 189.2643 194.1398
## Age+Anxiety                 0.7594  0.7376  1.4501 195.1116 199.9871
## Severity+Anxiety            0.4875  0.4409 24.5676 214.0198 218.8953
## Surgery*Age                 0.7605  0.7262  3.3614 197.0029 203.0973
## Surgery*Severity            0.4743  0.3992 27.6883 216.6547 222.7491
## Surgery*Anxiety             0.2849  0.1827 43.7865 224.3461 230.4405
## Age*Severity                0.8123  0.7855 -1.0468 190.9035 196.9979
## Age*Anxiety                 0.7657  0.7322  2.9189 196.4536 202.5480
## Severity*Anxiety            0.4879  0.4147 26.5310 215.9988 222.0931
## Surgery+Age+Severity        0.8117  0.7848 -0.9935 190.9868 197.0812
## Surgery+Age+Anxiety         0.7606  0.7264  3.3474 196.9858 203.0802
## Surgery+Severity+Anxiety    0.4909  0.4182 26.2722 215.8497 221.9441
## Age+Severity+Anxiety        0.8096  0.7923 -2.8148 189.2643 194.1398
## Surgery*Age*Severity        0.8551  0.7954  3.3209 192.4444 203.4143
## Surgery*Age*Anxiety         0.7932  0.7080  8.5818 201.3336 212.3035
## Surgery*Severity*Anxiety    0.6447  0.4984 21.2016 214.8593 225.8292
## Age*Severity*Anxiety        0.8123  0.7855 -1.0468 190.9035 196.9979
## Surgery+Age+Severity+Anxiety 0.8183  0.7819  0.4494 192.1011 199.4144
## Surgery*Age*Severity*Anxiety 0.8941  0.7177 16.0000 200.5926 221.3134
## Step-selected model         0.8436  0.8024  0.2972 190.3508 198.8829
```

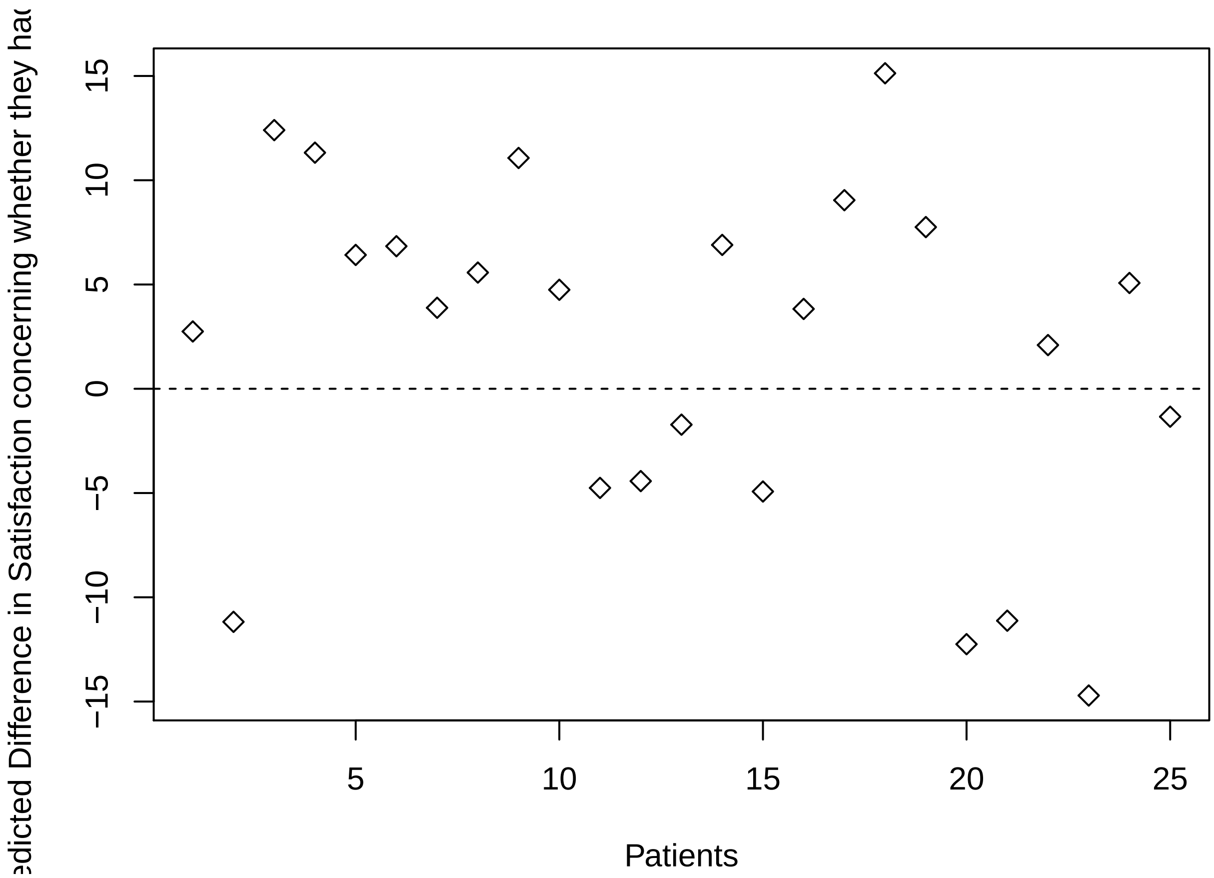
```r
mtx2 <-rbind(cvals[9,],cvals[15,],cvals[18,],cvals[21,],cvals[25,],cvals[28,])
rownames(mtx2) <- c("Age+Severity","Age*Severity",
                    "Surgery+Age+Severity","Age+Severitty+Anxiety",
                    "Age*Severity*Anxiety","Step-Selected model")
colnames(mtx2) <- c("Rsq","Adj.Rsq","Cp","AIC","BIC")
```

	Rsq	Adj.Rsq	Cp	AIC	BIC
Age+Severity	0.81	0.79	-2.81	189.26	194.14
Age*Severity	0.81	0.79	-1.05	190.90	197.00
Surgery+Age+Severity	0.81	0.78	-0.99	190.99	197.08
Age+Severitty+Anxiety	0.81	0.79	-2.81	189.26	194.14
Age*Severity*Anxiety	0.81	0.79	-1.05	190.90	197.00
Step-Selected model	0.84	0.80	0.30	190.35	198.88

Our conclusion is a rather not-so-neat one: it seems that surgery might have played a mild role in patient satisfaction, but we can't be so sure about it since we only have two models including surgery, and their AIC measures are not the lowest.

Let's further check the effect of predictor Surgery through a visual inspection:

```r
PS_no<-PS;PS_no$Surgery<- as.factor('No')
PS_yes<-PS;PS_yes$Surgery<- as.factor('Yes')
No_fit<-predict(fit.stepc_PS, newdata=PS_no)
Yes_fit<-predict(fit.stepc_PS, newdata=PS_yes)
par(mar=c(4,4,1,1))
plot(Yes_fit-No_fit,pch=5,xlab='Patients',ylab='Predicted Difference in Satisfaction concerning whether
abline(h=0,lty=2)
```



Since significant amount of data plots show there would be positive difference, we can estimate that the predictor Surgery does play a minor role.