Hw3

Stone Cai

2024-02-18

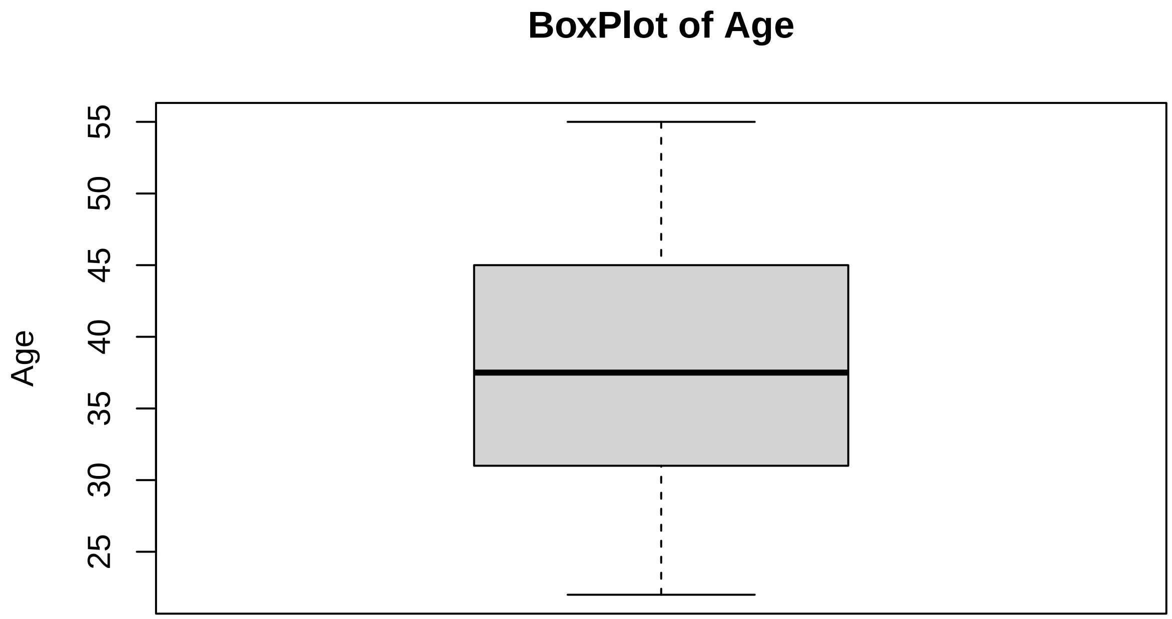
Problem 1. Although R^2 cannot decrease the more predictors you add, there is a heavy emphasis in model building that if two models have similar variance explained, the simpler model is better. We would only want to add additional predictors to the model if we can demonstrate a statistically significant coefficient and thus R^2 improves significantly if the predictor is added. As an alternative, adjusted R^2 can often be used for determining variance explained by a model as it tends to penalize the addition of predictors if it doesn't meaningfully improve the variance explained in a model.

Inputting Data

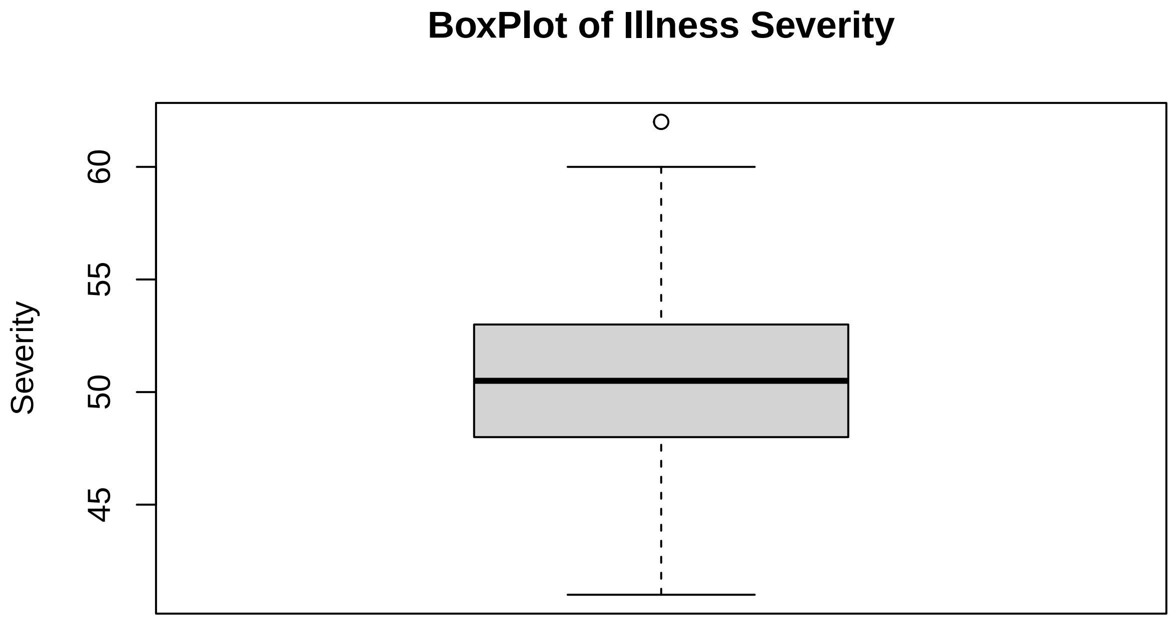
```r
hospital <- read.csv("/Users/stone/Documents/Stat 408/Homework Data Files/Hw3/hospital_satisfaction.csv

summary(hospital)
```

```
##   satisfaction        age           severity        anxiety     
##  Min.   :26.00   Min.   :22.00   Min.   :41.00   Min.   :1.800  
##  1st Qu.:48.25   1st Qu.:31.25   1st Qu.:48.00   1st Qu.:2.100  
##  Median :60.00   Median :37.50   Median :50.50   Median :2.300  
##  Mean   :61.57   Mean   :38.39   Mean   :50.43   Mean   :2.287  
##  3rd Qu.:76.75   3rd Qu.:44.75   3rd Qu.:53.00   3rd Qu.:2.475  
##  Max.   :92.00   Max.   :55.00   Max.   :62.00   Max.   :2.900  
```

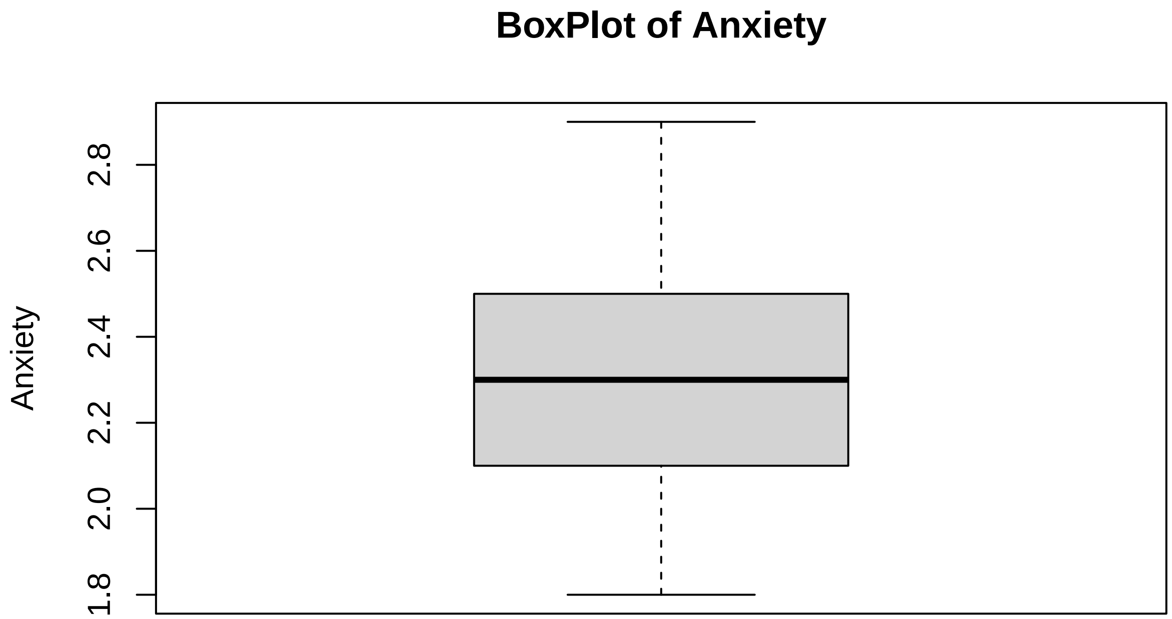
```r
boxplot(hospital$age, main="BoxPlot of Age", ylab="Age")
```



```
boxplot(hospital$severity, main="BoxPlot of Illness Severity", ylab="Severity")
```



```r
boxplot(hospital$anxiety, main="BoxPlot of Anxiety", ylab="Anxiety")
```



Problem 2a) After plotting the individual predictors' boxplots, I don't notice anything that particularly stands out. Only the box plot of severity has one outlier observation which is the max of 62.Otherwise, the distribution of individual predictors do not heavily skewed. The boxplot of anxiety may have a small right skew.

```
library(ggplot2)
```

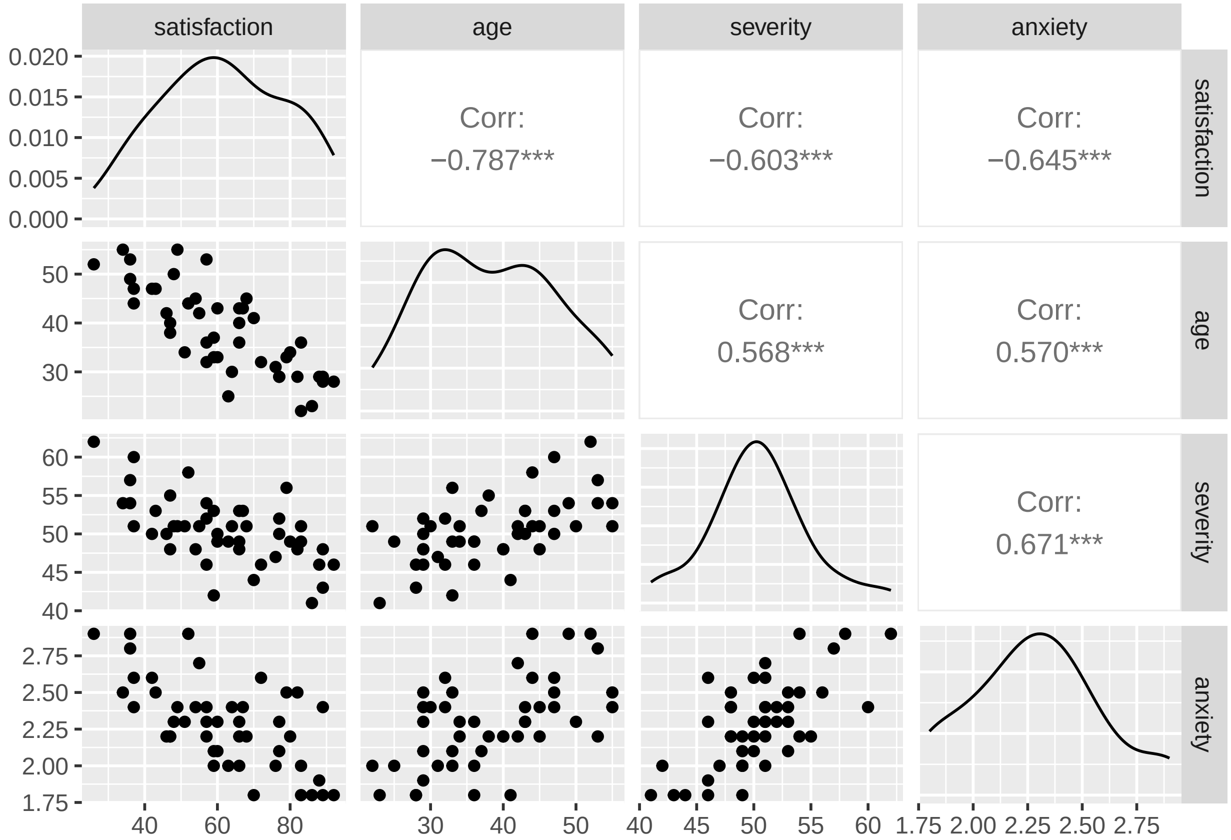
```
## Warning: package 'ggplot2' was built under R version 4.3.2
```

```
library(GGally)
```

```
## Warning: package 'GGally' was built under R version 4.3.2
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
ggpairs(hospital)
```



b) From the above plot we obtain both the scatter plot matrix and the coefficient matrix. When looking at the the scatter plots between Y and the 3 predictors X1,X2,X3 individually, we see that there is a negative slope relationship between them. We confirm with negative correlations from the correlation matrix. However, we also notice a potential problem as the relationships between the predictors are all positively correlated, but the relationship between response and predictors have negative correlation. This potentially indicates a problem of multi-collinearity. There is a relatively strong correlation between the predictors themselves.

```
lm_hosp<- lm(satisfaction~age+ severity+anxiety,data=hospital)
summary(lm_hosp)
```

```
##
## Call:
## lm(formula = satisfaction ~ age + severity + anxiety, data = hospital)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.3524  -6.4230   0.5196   8.3715  17.1601
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 158.4913    18.1259   8.744 5.26e-11 ***
## age          -1.1416     0.2148  -5.315 3.81e-06 ***
## severity     -0.4420     0.4920  -0.898   0.3741
## anxiety     -13.4702     7.0997  -1.897   0.0647 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.06 on 42 degrees of freedom
## Multiple R-squared:  0.6822, Adjusted R-squared:  0.6595
## F-statistic: 30.05 on 3 and 42 DF,  p-value: 1.542e-10
```

c) I fit a multiple linear regression model above with Y and X1, X2, X3 as predictors. The full regression model is

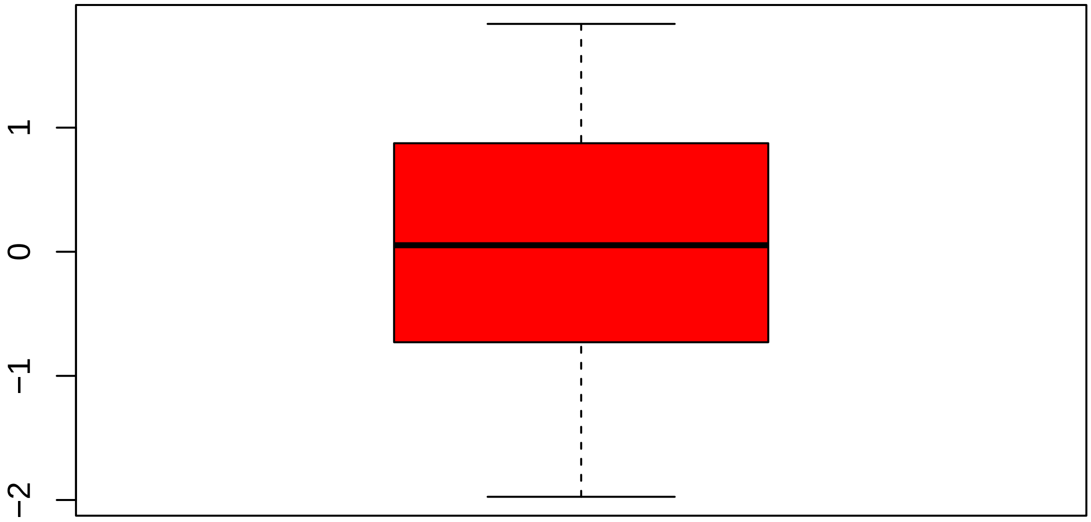
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$$

where we have defined

$$\hat{\beta}_2$$

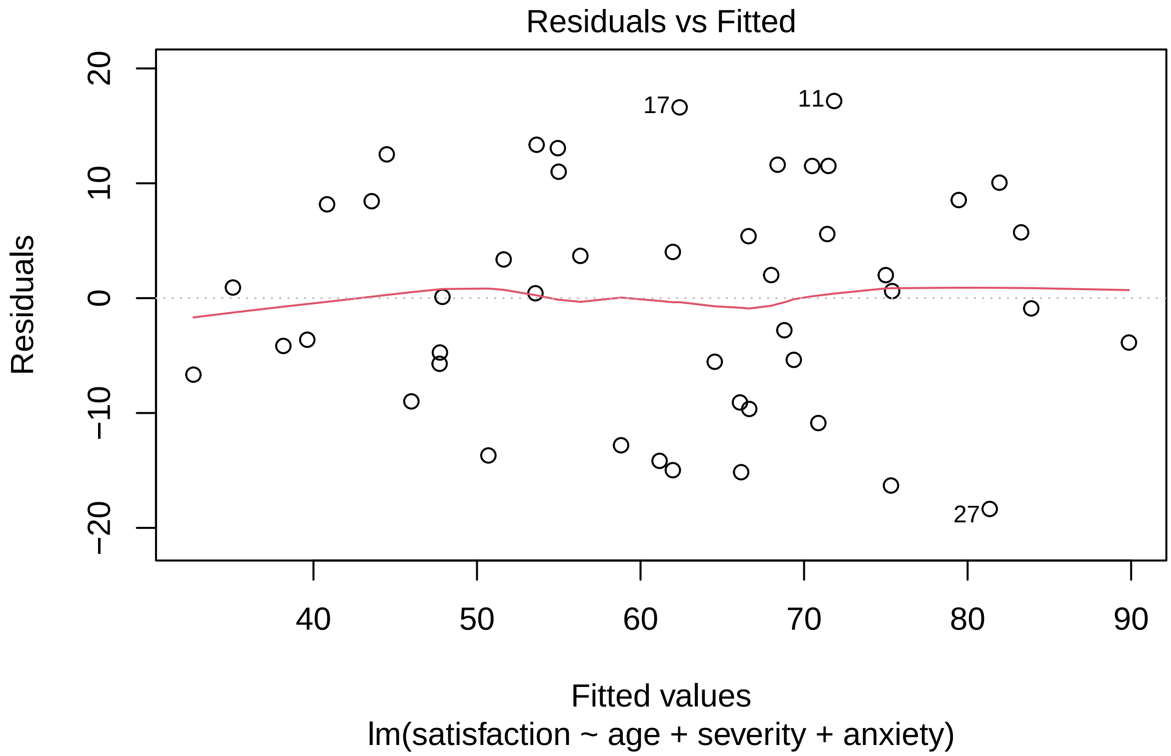
to be the coefficient of the severity predictor. Also, x1,x2,x3 are age, severity, and anxiety predictors respectively. For each additional unit increase in the severity index, there is an expected 0.4420 unit DECREASE in the satisfaction of a hospital patient.

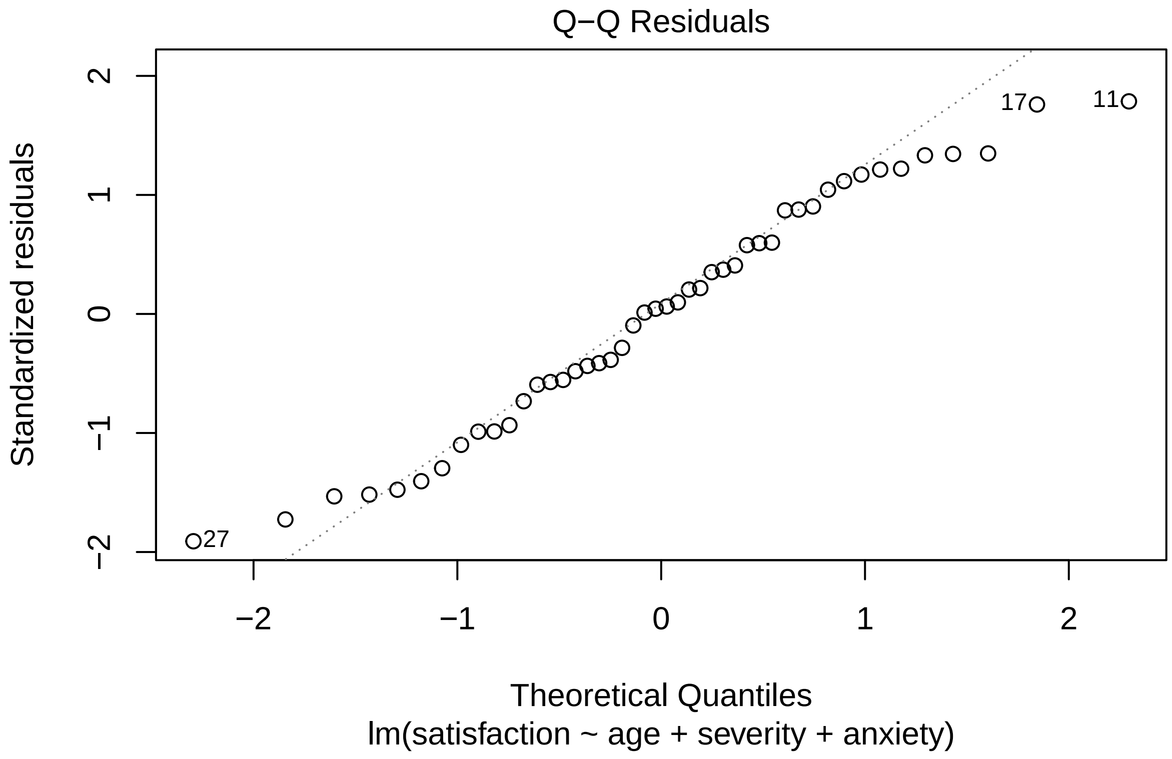
```
library (MASS)
boxplot(studres(lm_hosp),col="red")
```

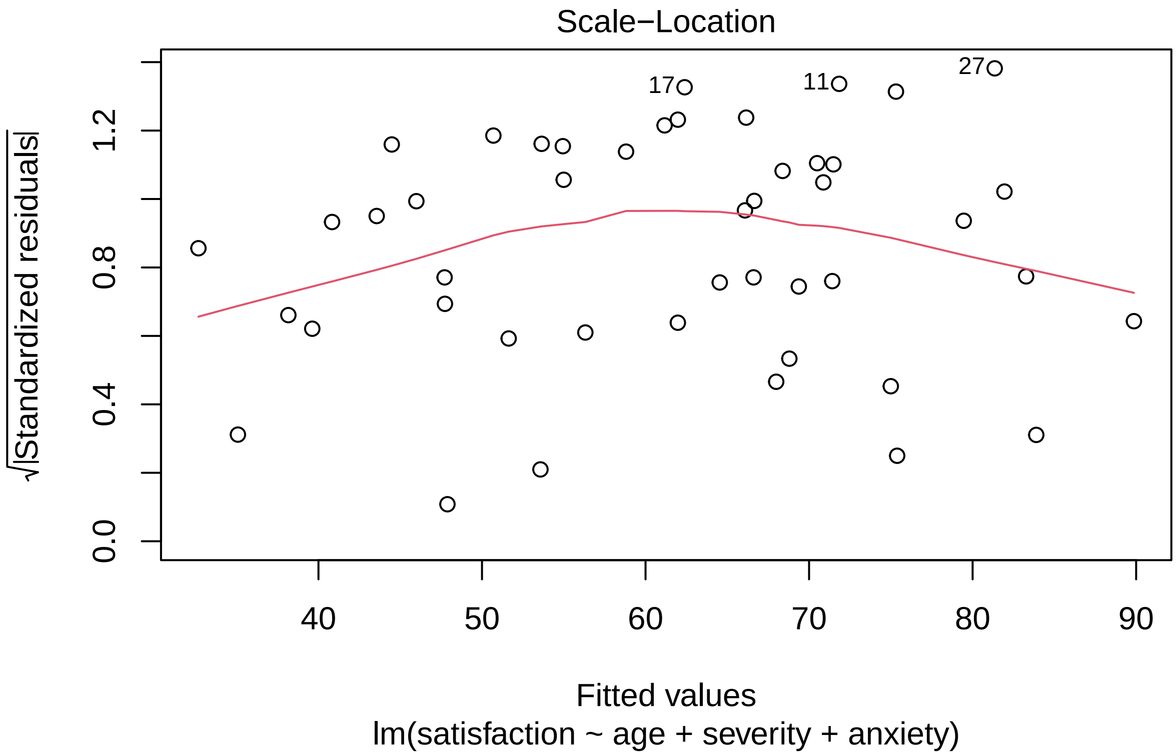


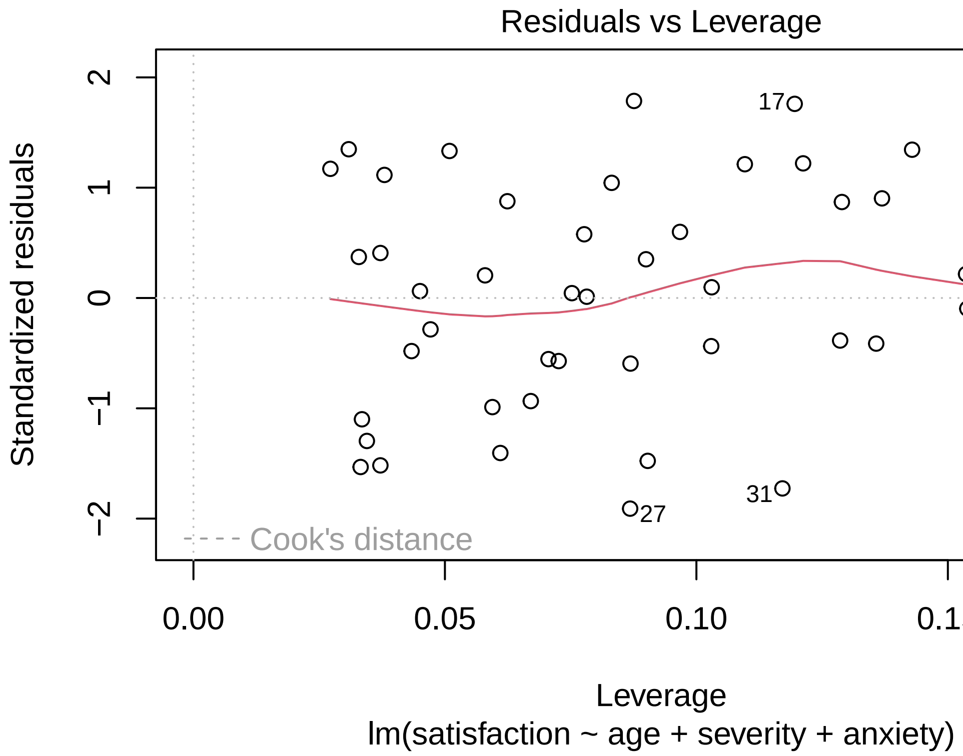
d). I utilize a boxplot of studentized residuals for the regression model from part C. There does not appear to be any outliers since all data points fall within whiskers of the box plot.

```
##plot of residuals vs fitted values
plot (lm_hosp)
```

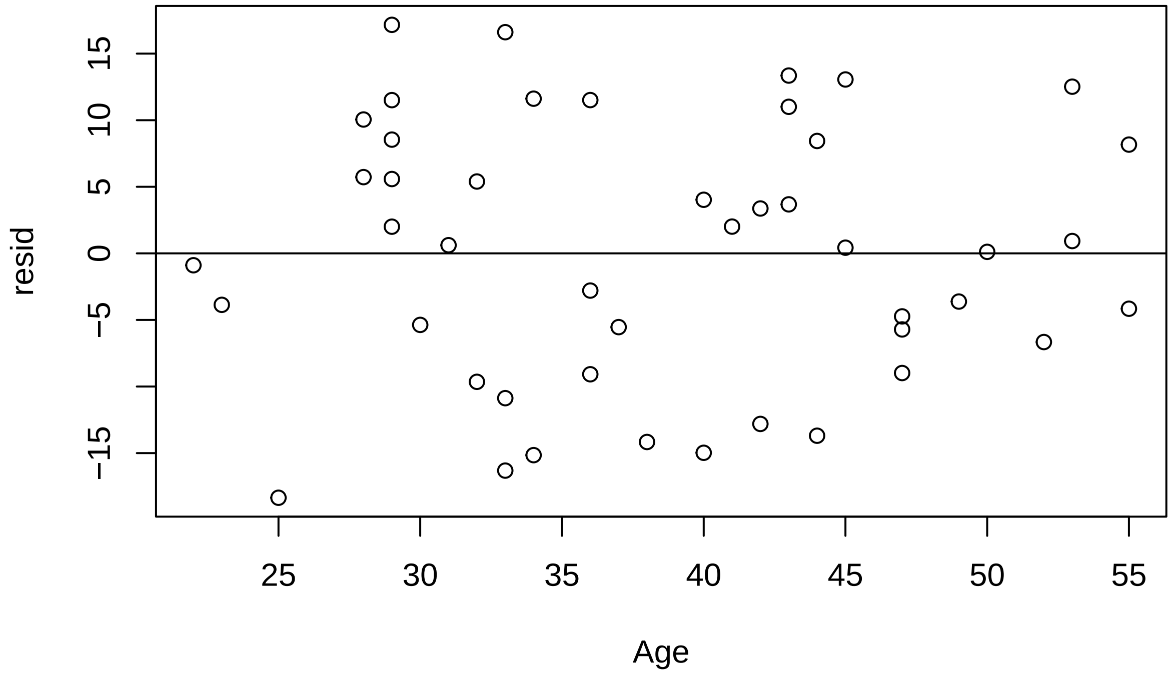

Residuals vs Fitted
Residuals
20
10
0
−10
−20
40
50
60
70
80
90
Fitted values
lm(satisfaction ~ age + severity + anxiety)
17
11
27


Q−Q Residuals
Standardized residuals
−2
−1
0
1
2
27
17
11
−2
−1
0
1
2
Theoretical Quantiles
lm(satisfaction ~ age + severity + anxiety)

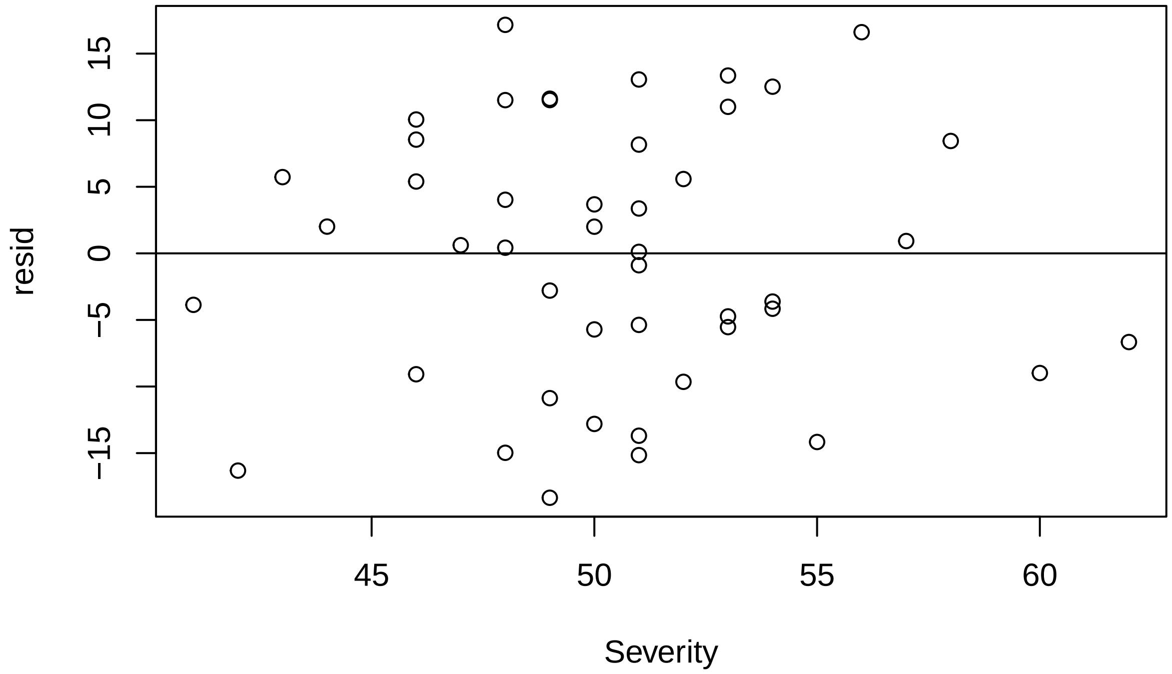

Scale−Location
√|Standardized residuals|
0.0
0.4
0.8
1.2
40
50
60
70
80
90
17
11
27
Fitted values
lm(satisfaction ~ age + severity + anxiety)



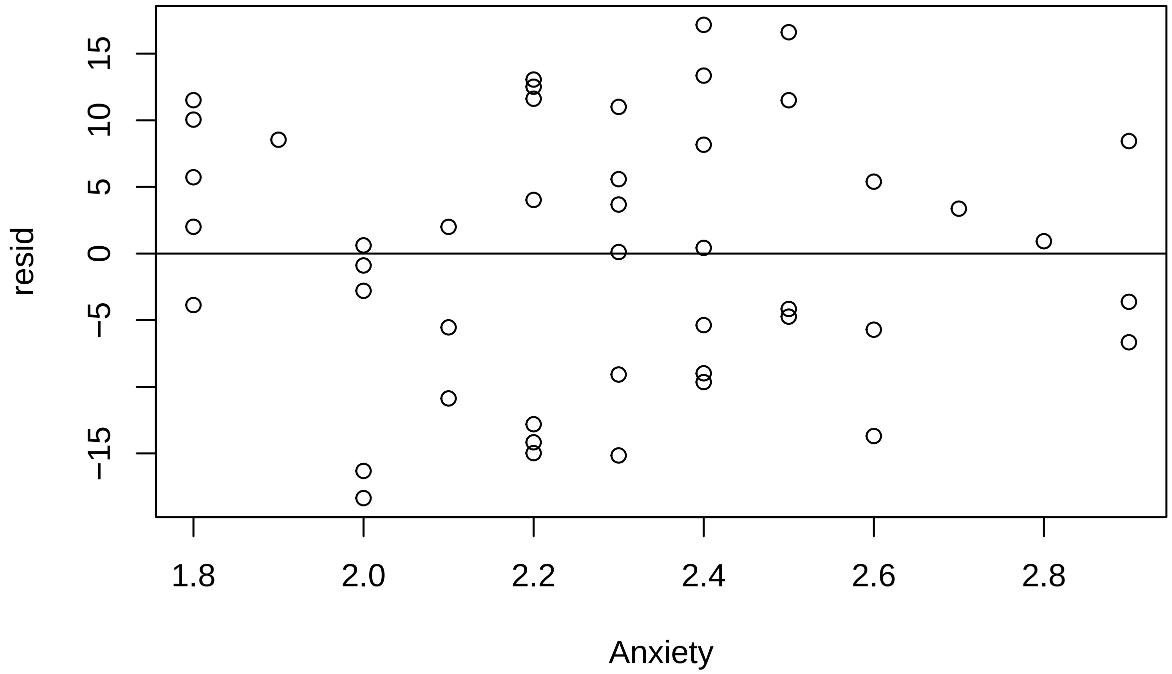
```
resid<-resid(lm_hosp)
Age<-hospital$age
Severity<-hospital$severity
Anxiety<-hospital$anxiety

plot(Age,resid)
abline (h=0)
```



```
plot(Severity,resid)
abline (h=0)
```



```r
plot(Anxiety,resid)
abline (h=0)
```



e) There doesn't seem to be a violation of homoskedacity for Residuals vs fitted values. There seems to be constant variance throughout entire range of fitted values. The individual residual vs Age and residual vs Severity plots seems to have no violation of homoskedacity as there seems to be a random scatter and constant variance. There may be a violation of homoskedacity for residuals vs anxiety as there are large number of residuals centered around the residuals=0 line, but also many observations with large residuals, but not in a very random/ scattered pattern.

```
qf(0.9,3,42)
```

```
## [1] 2.219059
```

```
pf(30.05, 3, 42, lower.tail = FALSE)
```

```
## [1] 1.543485e-10
```

f. Our

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$$

and

$$H_A :$$

Is at least one slope coefficient is not zero.

At significance level= 0.1 if

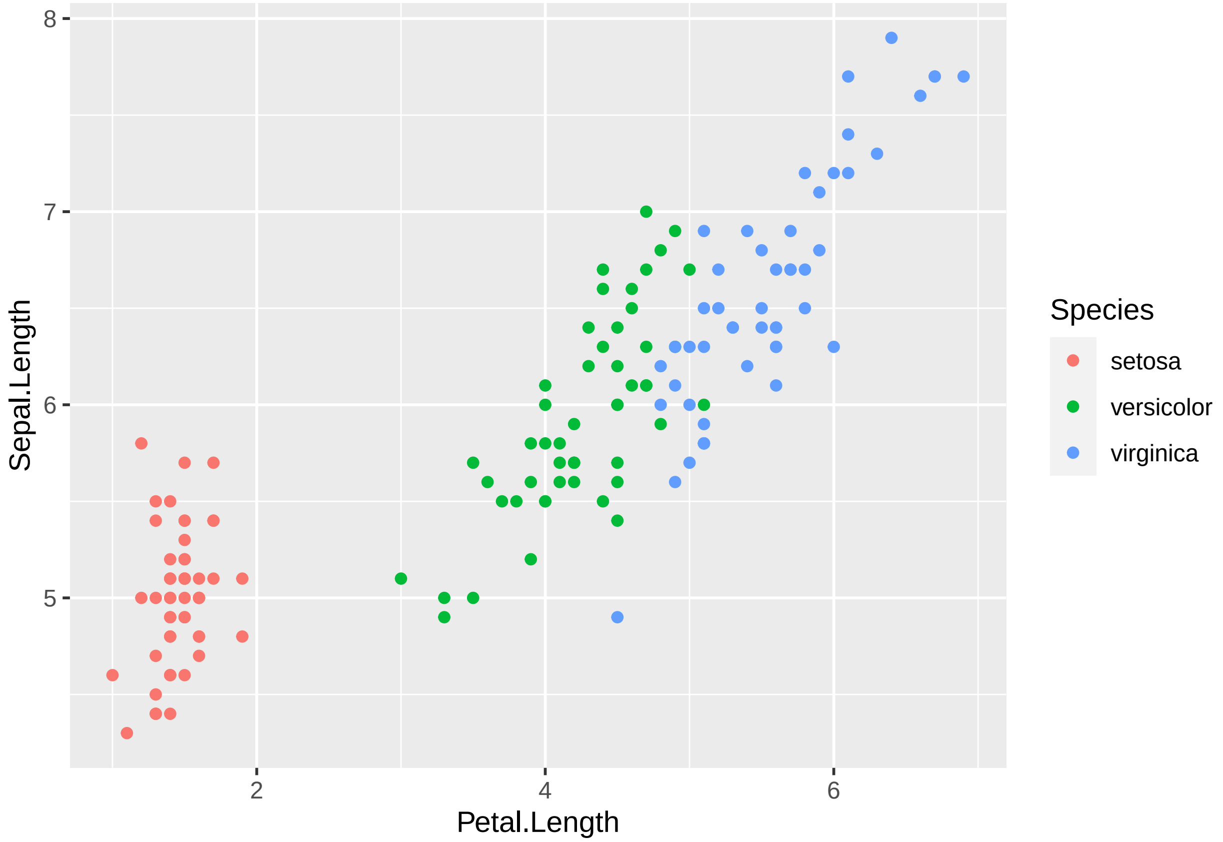
$$F^{\star} > F(0.9, 3, 42) = 2.22$$

then reject null hypothesis.From R output summary, we have F-stat of 30.05 > 2.22 so reject null hypothesis.P-value of F-Stat is 1.54*10^-10 which is the same p-value from summary output. There is sufficient evidence to conclude that at least one predictor variable is significant in the model such that the regression model is overall significant.

g. Coefficient of multiple regression is the absolute value of square root R^2 which is 0.826. Here, the correlation coefficient represents correlation between response variable and linear combination of predictors.

Problem 3

```
iris.df<- data(iris)
##head (iris)

ggplot(data = iris) +
  aes(x = Petal.Length, y = Sepal.Length) +
  geom_point(aes(color = Species))
```



a) As we can see the various species groups have different intercepts and slopes.If we were to plot a line of best fit separately for each group we'd get a positive regression coefficient for all, but setosa would have much larger coefficient value then virginica or versicolor. At the same time, whereas the intercept of the setosa line of best fit would be positive, it seems the intercept of the versicolor and virginica line would be negative.

```
sub.df<-subset(iris, Species=="setosa"| Species== "virginica")
##head(sub.df)
model_1<- lm(Sepal.Length~Petal.Length, data= sub.df)

summary(model_1)
```

```
##
## Call:
## lm(formula = Sepal.Length ~ Petal.Length, data = sub.df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.30238 -0.25500 -0.01849  0.27795  0.94480
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.36531    0.07917   55.14   <2e-16 ***
## Petal.Length  0.40824    0.01941   21.04   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.4045 on 98 degrees of freedom
## Multiple R-squared:  0.8187, Adjusted R-squared:  0.8169
## F-statistic: 442.6 on 1 and 98 DF,  p-value: < 2.2e-16
```

b) Here we have subset the data to remove observations of the species versicolor. Parameter estimates are as follows:

$$\beta_0 : 4.365$$

This is the intercept coefficient and can be interpreted as expected value of sepal length when petal length is zero.

$$\beta_1 : 0.40824$$

For every unit increase in petal length we have 0.4082 unit increase in sepal length.

```
model_2<- lm(Sepal.Length~Petal.Length+Species, data= sub.df)

summary(model_2)
```

```
##
## Call:
## lm(formula = Sepal.Length ~ Petal.Length + Species, data = sub.df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.71579 -0.24624  0.00451  0.18083  1.04418
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       3.60996    0.13023  27.719  < 2e-16 ***
## Petal.Length      0.95489    0.08294  11.513  < 2e-16 ***
## Speciesvirginica -2.32348    0.34581  -6.719 1.26e-09 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.3359 on 97 degrees of freedom
## Multiple R-squared:  0.8763, Adjusted R-squared:  0.8737
## F-statistic: 343.5 on 2 and 97 DF,  p-value: < 2.2e-16
```

C) Parameter estimates are as follows:

$$\beta_0 : 3.61$$

This is the intercept coefficient and can be interpreted as expected value of sepal length when all predictors are zero.

$$\beta_1 : 0.955$$

For every unit increase in petal length we have 0.955 unit increase in sepal length.

$$\beta_2 : -2.323$$

Since this is a dummy variable, this reflects difference in expected sepal length when species is Virgnica (compared to reference group setosa).

```r
model_3<- lm(Sepal.Length~ Petal.Length+Species + Petal.Length:Species, data= sub.df)

summary(model_3)
```

```
##
## Call:
## lm(formula = Sepal.Length ~ Petal.Length + Species + Petal.Length:Species,
##     data = sub.df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.73409 -0.23601 -0.03132  0.18695  0.93608
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      4.2132     0.4037  10.437  < 2e-16 ***
## Petal.Length                     0.5423     0.2742   1.978   0.0508 .
## Speciesvirginica                -3.1535     0.6283  -5.019 2.38e-06 ***
## Petal.Length:Speciesvirginica    0.4534     0.2875   1.577   0.1180
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.3334 on 96 degrees of freedom
## Multiple R-squared:  0.8794, Adjusted R-squared:  0.8756
## F-statistic: 233.4 on 3 and 96 DF,  p-value: < 2.2e-16
```

d)

$$\beta_0 : 4.21$$

This is the intercept coefficient and can be interpreted as expected value of sepal length when all predictors are zero.

$$\beta_1 : 0.54$$

For every unit increase in petal length we have 0.54 unit increase in sepal length.

$$\beta_2 : -3.15$$

Since this is a dummy variable, this reflects difference in expected sepal length when species is Virgnica (compared to reference group setosa).

$$\beta_3 : 0.453$$

This is an interaction term. When Species = 0 (or reference group in this case setosa) our regression equation becomes

$$Sepal\hat{l}ength = 4.213 + 0.54 * Petallength$$

since both species and interaction term becomes zero. If Species= 1 then our regression equation becomes

$$Sepal\hat{l}ength = 4.213 + 0.54 * Petallength - 3.153 + 0.453 * Petallength$$

.

e). Coefficient of determination for model 1,2,3 are 0.819, 0.8763, 0.8794 respectively. Coefficient of determination is expected to increase with each additional predictor added to the model. That does not mean model is necessarily getting better. The jump from model 1 to 2 by adding species as a predictor did significantly improve the coefficient of determination for the model.However, the increase from model 2 to 3 was minimal and by good modeling practice, we would prefer to use the simpler model (Model 2), if both models yield essentially the same coef of determination.