# p8130 HW4 Regression

*Eleanor Zhang*

*11/11/2018*

## Problem 2 Heart disease

We are interested in fi there is an association between **total cost** in dollars diagnosed with heart disease and the **number of ER visits**. Other factors will be adjusted later on.

### a) short description of data and look at the data

```
heart_disease <- read_csv("./data/HeartDisease.csv")
```

```
## Parsed with column specification:
## cols(
##   id = col_integer(),
##   totalcost = col_double(),
##   age = col_integer(),
##   gender = col_integer(),
##   interventions = col_integer(),
##   drugs = col_integer(),
##   ERvisits = col_integer(),
##   complications = col_integer(),
##   comorbidities = col_integer(),
##   duration = col_integer()
## )
```

```
head(heart_disease)
```

```
## # A tibble: 6 x 10
##      id totalcost   age gender interventions drugs ERvisits complications
##   <int>     <dbl> <int>  <int>         <int> <int>    <int>         <int>
## 1     1      179.    63      0             2     1        4             0
## 2     2      319     59      0             2     0        6             0
## 3     3     9311.    62      0            17     0        2             0
## 4     4      281.    60      1             9     0        7             0
## 5     5    18727.    55      0             5     2        7             0
## 6     6      453.    66      0             1     0        3             0
## # ... with 2 more variables: comorbidities <int>, duration <int>
```

In this dataset, there are 788 observations of patients with 10 variables:

- **id**: patient id
- **totalcost**: total cost ($) of patients who are diagnosed with heart disease
- **age**: age of patients
- **interventions**: number of interventions (integers)
- **drugs**: ? number of drugs.
- **ERvisits**: number of ER visits
- **complications**: number of complications
- **comorbidities**: number of co-presence of other diseases in additional to heart disease
- **duration**: duration of heart disease (in days)

Based our investigation interest, the main outcome is **total cost** of patients with heart disease and the main predictor is **ERvisits** (number of ER visits). Other important covariates also need to be considered because they could potential have differential effect on the association relationship between out main predictor and main outcome, including age, interventions, drugs used, complications, and duration of disease. We will first take a look at these variables:

i) First we took a look at the main outcome and main predictor

number summaries for variables:

```
variable_set1 <- dplyr::select(heart_disease, totalcost, ERvisits, everything(), -c(id, gender, complica
variable_set2 <- dplyr::select(heart_disease, gender, complications)
knitr::kable(summary(variable_set1))
```

| totalcost | ERvisits | age | interventions | drugs | comorbidities | duration |
|---|---|---|---|---|---|---|
| Min. : 0.0 | Min. : 0.000 | Min. :24.00 | Min. : 0.000 | Min. :0.0000 | Min. : 0.000 | Min. : 0.00 |
| 1st Qu.: 161.1 | 1st Qu.: 2.000 | 1st Qu.:55.00 | 1st Qu.: 1.000 | 1st Qu.:0.0000 | 1st Qu.: 0.000 | 1st Qu.: 41.75 |
| Median : 507.2 | Median : 3.000 | Median :60.00 | Median : 3.000 | Median :0.0000 | Median : 1.000 | Median :165.50 |
| Mean : 2800.0 | Mean : 3.425 | Mean :58.72 | Mean : 4.707 | Mean :0.4467 | Mean : 3.767 | Mean :164.03 |
| 3rd Qu.: 1905.5 | 3rd Qu.: 5.000 | 3rd Qu.:64.00 | 3rd Qu.: 6.000 | 3rd Qu.:0.0000 | 3rd Qu.: 5.000 | 3rd Qu.:281.00 |
| Max. :52664.9 | Max. :20.000 | Max. :70.00 | Max. :47.000 | Max. :9.0000 | Max. :60.000 | Max. :372.00 |

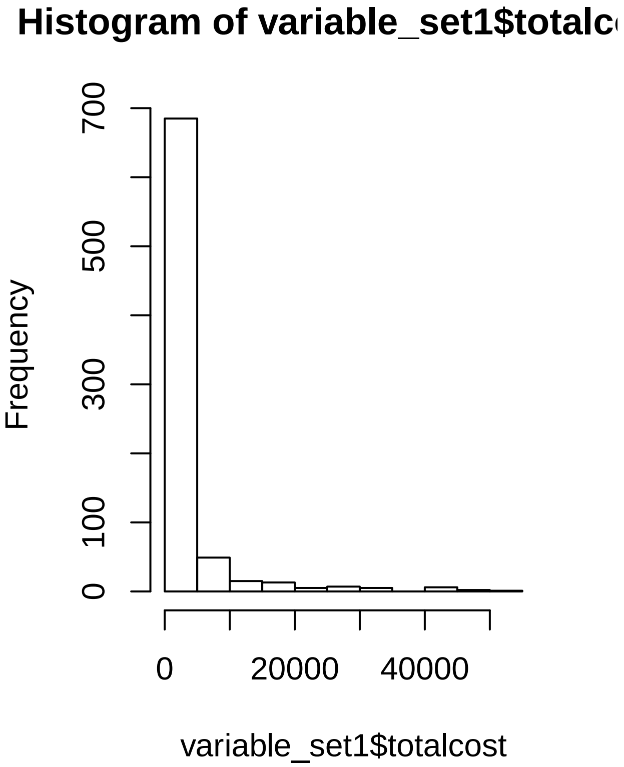
```
table(variable_set2)
```

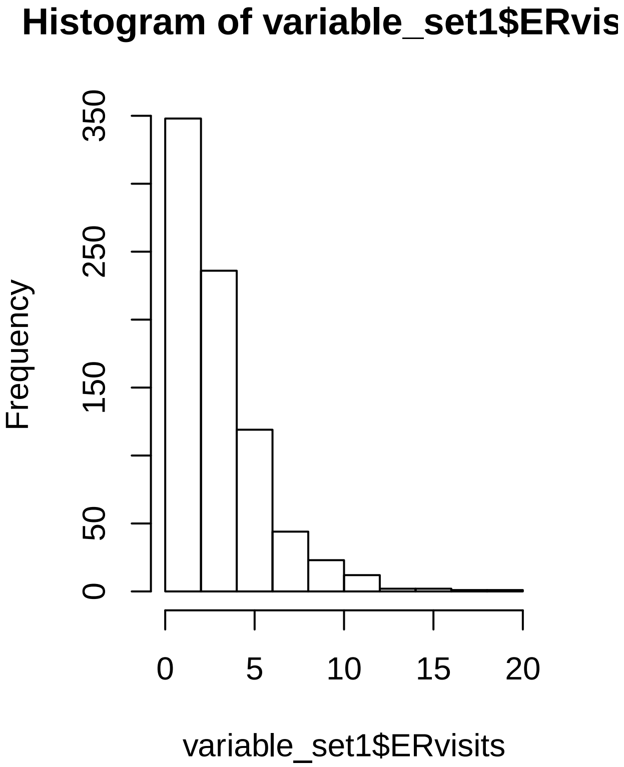
```
##       complications
## gender   0   1   3
##      0 576  32   0
##      1 169  10   1
```

```
#margin.table(table(variable_set2))
#prop.table(table(variable_set2))
```

Visualize the distribution of these variables

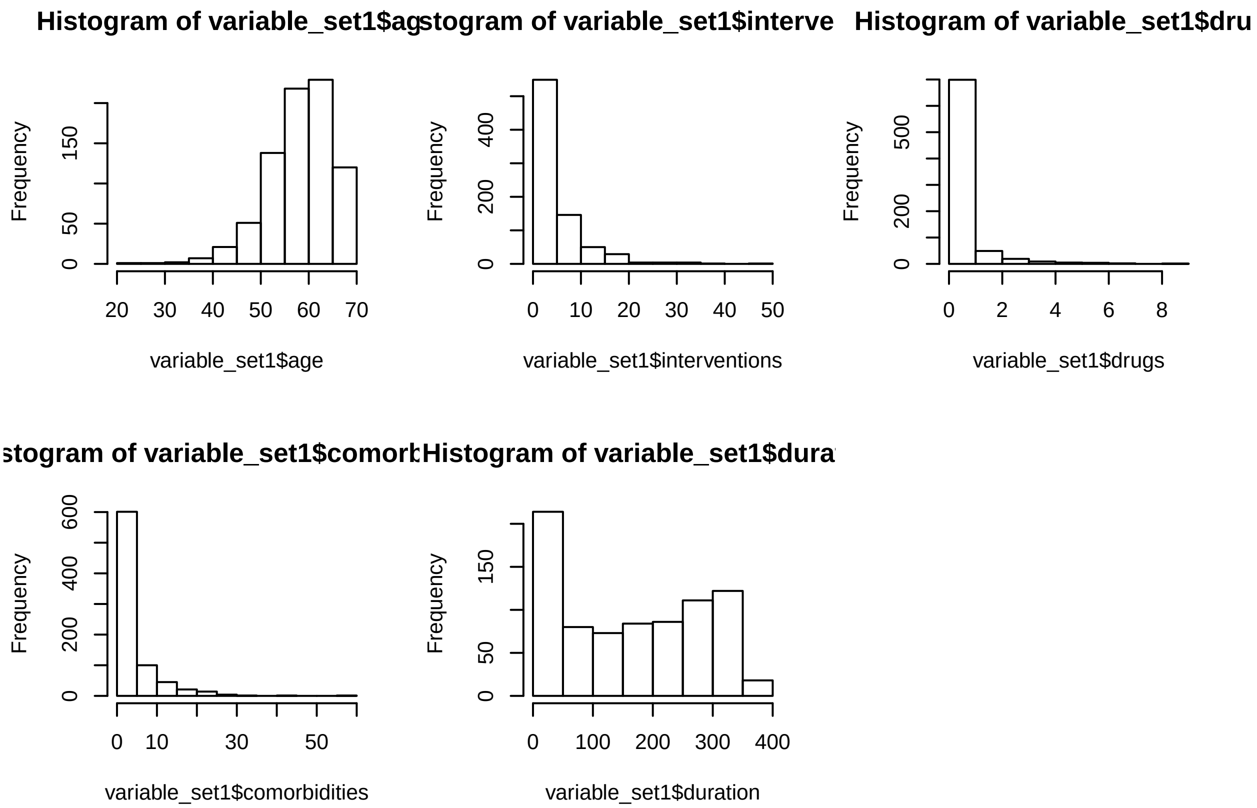
```
par(mfrow = c(1,2))
hist(variable_set1$totalcost)
hist(variable_set1$ERvisits)
```





Comment: Since total cost and ER visits are both heavily right skewed on the histograms, we better use median and IQR in the summay table to describe them. Especially for total cost, there are many extreme values at the right tail end which needed to be investigated further in the following analysis.

```r
par(mfrow = c(2,3))
hist(variable_set1$age)
hist(variable_set1$interventions)
hist(variable_set1$drugs)
hist(variable_set1$comorbidities)
hist(variable_set1$duration)
```

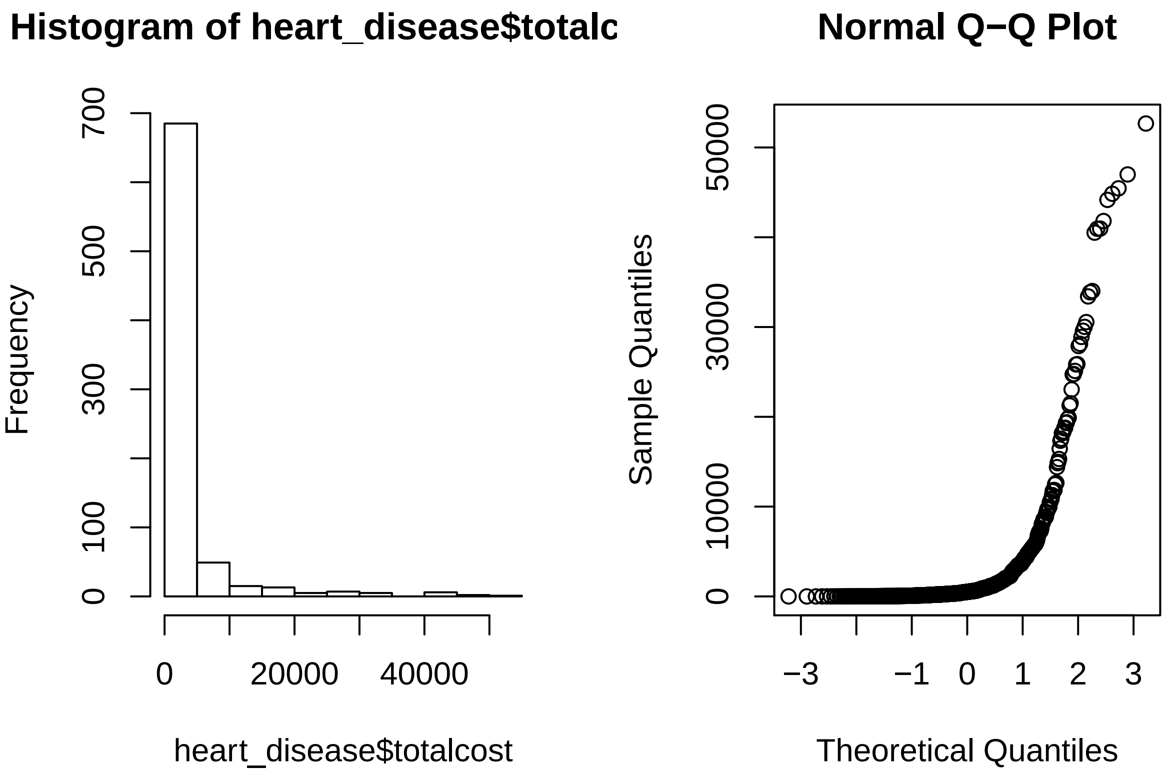


Comment: age is slightly left skewed which means elder people have been overly sampled. The median of intervention is about 5 with large IQR of 5. drugs?. Commordities have median of 3.7 with large IQR 5. Duration of heart disease is roughly uniformly distributed from 50 to 350 days with median 165 days and IQR 240 days. Therefore, these co-variables are not normally distributed in the sample, so we need to adjust for this in later analysis.

### b) investigate the shape of distribution for total cost

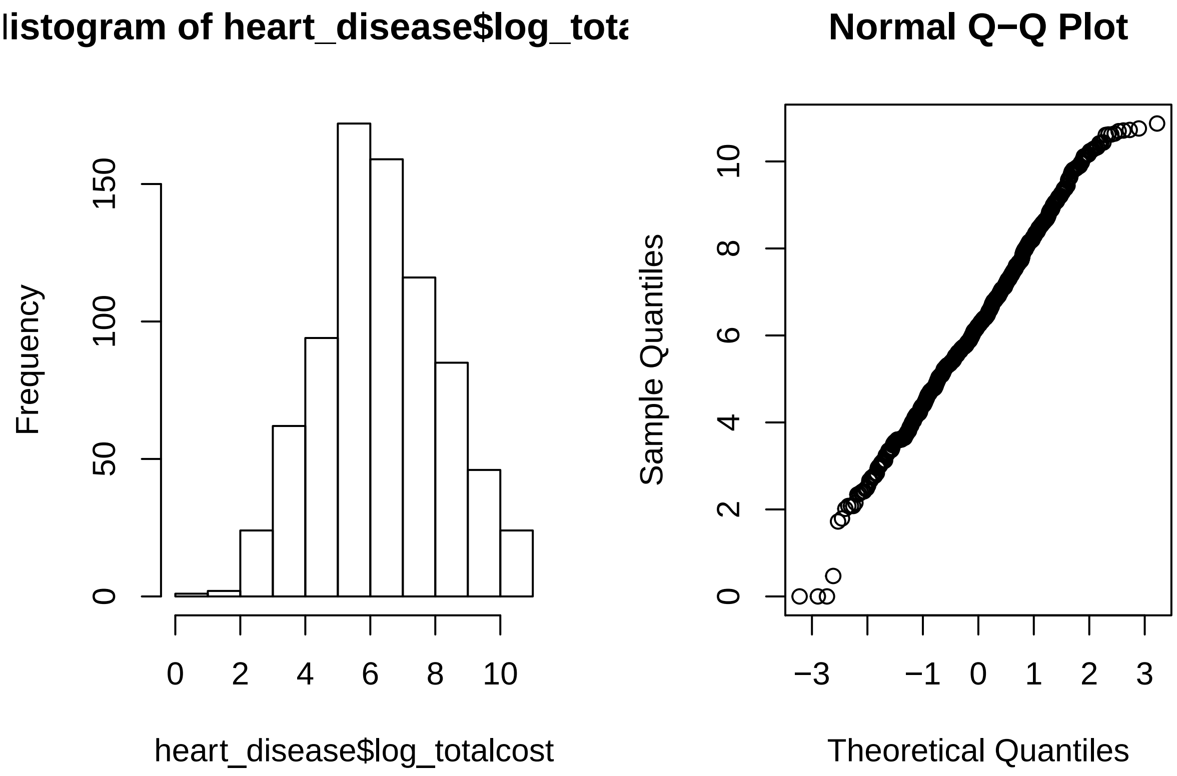
raw data of total cost

```
par(mfrow = c(1,2))
hist(heart_disease$totalcost)
qqnorm(heart_disease$totalcost)
```



Then we try log transformation on **totalcost** to see if this will transform the distribution.

```
heart_disease <- mutate(heart_disease, log_totalcost = log(totalcost))
par(mfrow = c(1,2))
hist(heart_disease$log_totalcost)
heart_disease$log_totalcost[is.infinite(heart_disease$log_totalcost)] = 0
qqnorm(heart_disease$log_totalcost)
```



comment: After log transformation, we saw a pretty good bell shape of the ditribution. So we will use this transformed data as needed.

### c) dichotomize complications

```
heart_disease <- heart_disease %>%
  mutate(comp_bin = ifelse(complications == 0, 0, 1))
```

### d) fit linear model SLR

From part (b), we saw the transformed data look better in normal shape, we will use the transformed data to fit SLR. So we fit a simple linear regression model between outcome **log_totalcost** and predictor **ERvisits**. Let $Y_i$ = response(total cost), $X_i$ = predictor (ERvisits).

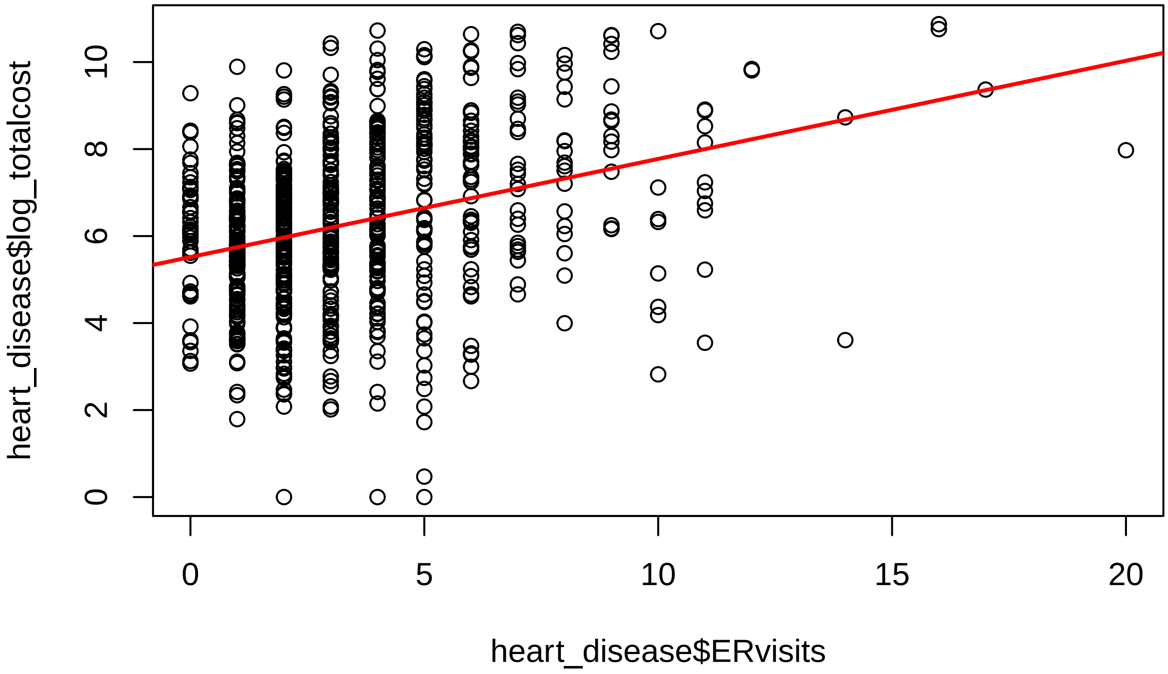
Then our model is $logY_i = \beta_0 + \beta_1 X_i + \epsilon_i$. Here it is reasonable to assume the error is normally distributed because the log transformation improve normality. then assume $\epsilon_i \sim N(0, \sigma^2)$

```
SLR <- lm(log_totalcost ~ ERvisits, data = heart_disease)
summary(SLR)
```

```
##
## Call:
## lm(formula = log_totalcost ~ ERvisits, data = heart_disease)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.6454 -1.1195  0.0371  1.2872  4.3046
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  5.51701    0.10585  52.119   <2e-16 ***
## ERvisits     0.22569    0.02449   9.214   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.812 on 786 degrees of freedom
## Multiple R-squared:  0.09749,    Adjusted R-squared:  0.09634 
## F-statistic: 84.91 on 1 and 786 DF,  p-value: < 2.2e-16
```

```r
plot(heart_disease$ERvisits, heart_disease$log_totalcost)
abline(SLR, col = "red", lwd = 2)
```



The result of regression tells that the fitted model is :

$$\hat{logY_i} = 5.517 + 0.23X_i$$

Interpretation: The fitted model indicates that for every unit increase in ER visits, the total cost in dollars on logarithm scale will increase by 0.23. when the ER visit is zero, the total cost in dollar on logarithm scale will be 5.517. The p value for two estimators $\beta_0$ and $\beta_1$ are well below 0.001. So we are very confident that there is a strong association between total cost and ER visits, and our simple regression model describes their relationship.

### e) fit MLR with comp_bin and ERvisits

i) test if **comp_bin** is an effect modifier of the relationship between **totalcost** and **ERvisits**

Let $Y_i$ = response(total cost), $X_{i1}$ = predictor (ERvisits), $X_{i2}$ = {when comp_bin equals 1, otherwise equals 0}

The full model is : $logY_i = \beta_0 + \beta_1 X_i + \beta_2 X_{i2} + \epsilon_i$

Now add in a potential modifier: $logY_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \epsilon_i$

So our hypothesis statement is: $H_0 : \beta_3 = 0$ vs. $H_a : \beta_3 \neq 0$

```
MLR_comp <- lm(log_totalcost ~ ERvisits + comp_bin, data = heart_disease)
MLR_comp_inter <- lm(log_totalcost ~ ERvisits + comp_bin + ERvisits*comp_bin, data = heart_disease)
summary(MLR_comp) %>% tidy
```

```
## # A tibble: 3 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)    5.50     0.104      53.1  3.58e-262
## 2 ERvisits       0.203    0.0242      8.39 2.29e- 16
## 3 comp_bin       1.71     0.281       6.09 1.72e-  9
```

```
summary(MLR_comp_inter) %>% tidy
```

```
## # A tibble: 4 x 5
##   term              estimate std.error statistic   p.value
##   <chr>                <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)         5.48      0.106      51.8  2.55e-255
## 2 ERvisits            0.210     0.0251      8.36 2.77e- 16
## 3 comp_bin            2.20      0.559       3.94 8.90e-  5
## 4 ERvisits:comp_bin  -0.0978    0.0970     -1.01 3.14e-  1
```

```
anova(MLR_comp, MLR_comp_inter)
```

```
## Analysis of Variance Table
##
## Model 1: log_totalcost ~ ERvisits + comp_bin
## Model 2: log_totalcost ~ ERvisits + comp_bin + ERvisits * comp_bin
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    785 2464.7
## 2    784 2461.5  1    3.1915 1.0165 0.3137
```

Based on the regression result, p value for the interaction coefficient $\beta_3$ is 0.314, which is quite large. So at 0.95 significance level, we do not have evidence to reject the null. Therefore we is no interaction or modifier effect of complications in the relationship between total cost and ER visits.

To visualize interaction model:

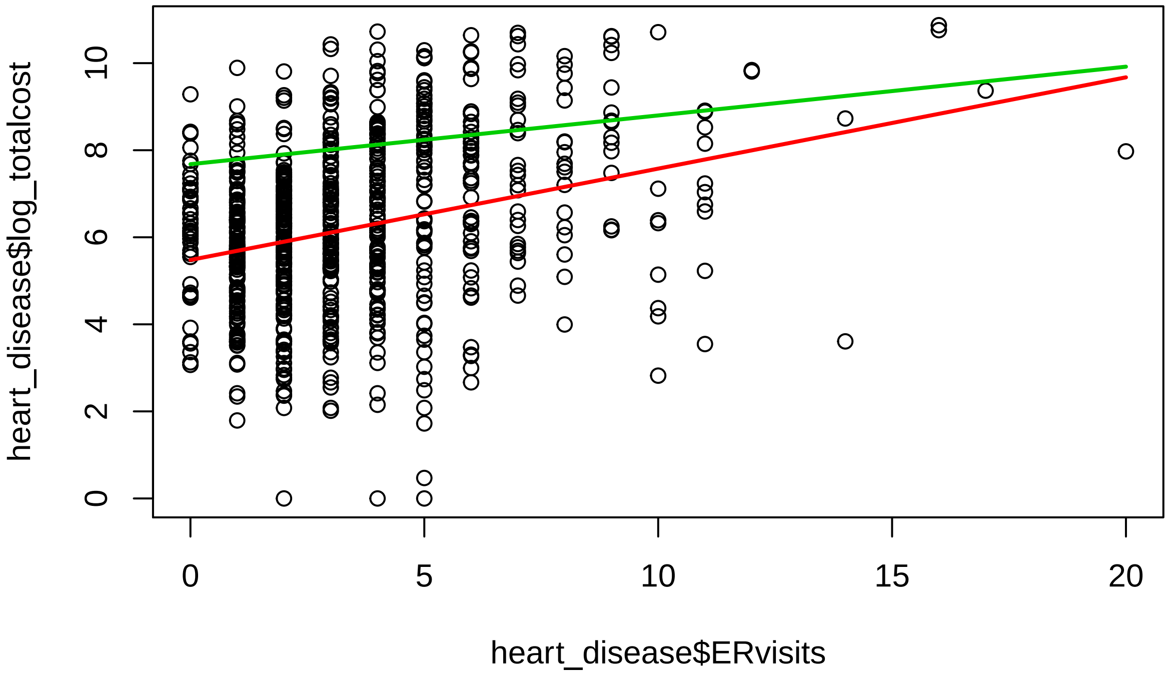
```
range(heart_disease$ERvisits)
```

```
## [1]  0 20
```

```
ER <- seq(0,20,0.5)
beta <- MLR_comp_inter$coefficients

# comp_bin = 0
yhat1 <- beta[1] + beta[2]*ER
# comp_bin greater than 0
yhat2 <- beta[1] + beta[3] + (beta[2] + beta[4])*ER

plot(heart_disease$ERvisits, heart_disease$log_totalcost)
lines(ER, yhat1, col = 2, lwd = 2) # comp_bin = 0
lines(ER, yhat2, col = 3, lwd = 2) # comp_bin greater than 0
```



ii) test if **comp_bin** is a confounder of relationship between total cost and ERvisits

Model 1 without comp_bin: $logY_i = \beta_0 + \beta_1 X_i + \epsilon_i$

Model 2 with comp_bin: $logY_i = \beta_0 + \beta_1 X_i + \beta_2 X_{i2} + \epsilon_i$

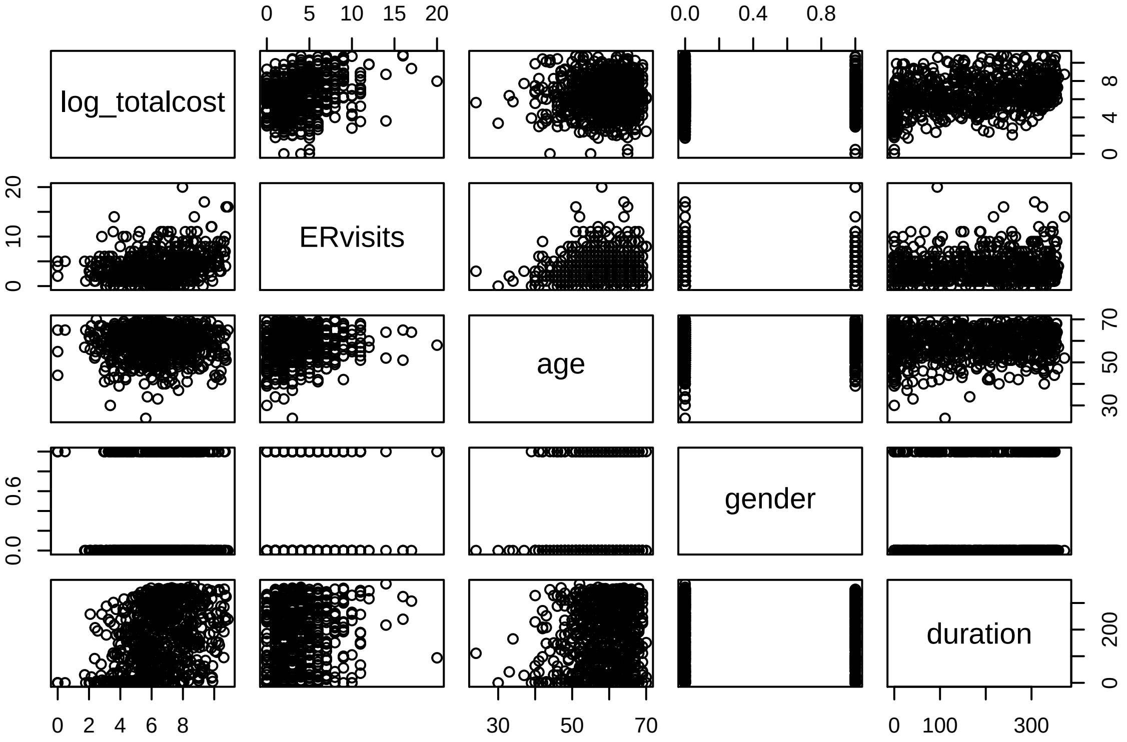
```
SLR <- lm(log_totalcost ~ ERvisits, data = heart_disease)
MLR_comp <- lm(log_totalcost ~ ERvisits + comp_bin, data = heart_disease)
anova(SLR, MLR_comp)
```

```
## Analysis of Variance Table
##
## Model 1: log_totalcost ~ ERvisits
## Model 2: log_totalcost ~ ERvisits + comp_bin
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    786 2581.3
## 2    785 2464.7  1    116.61 37.139 1.723e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

iii) decide if comp_bin should be included along with ERvisits

### f) examine additional covariates

```
heart_disease %>% dplyr::select(log_totalcost, ERvisits, age, gender, duration) %>% pairs()
```



## Problem 3

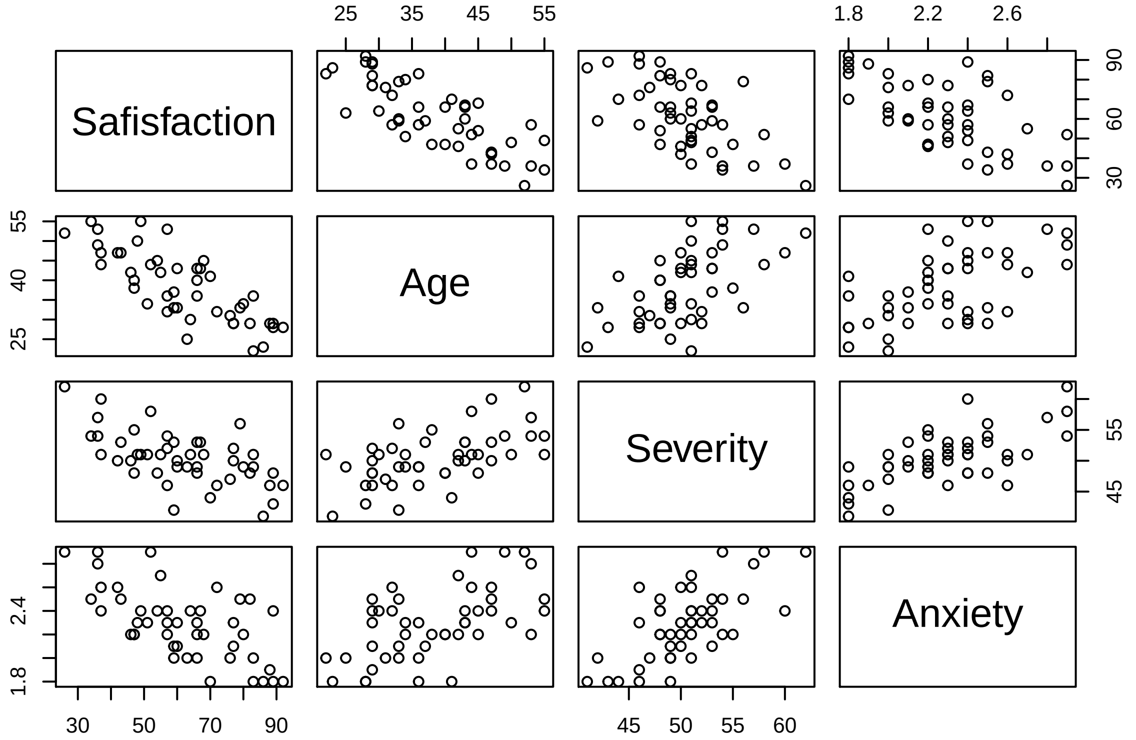
The investigators wants to test the relationship between patient's satisfaction (Y) and age, severity of illness, and anxiety level. The dataset contains 46 patients observations

### a) correlation matrix

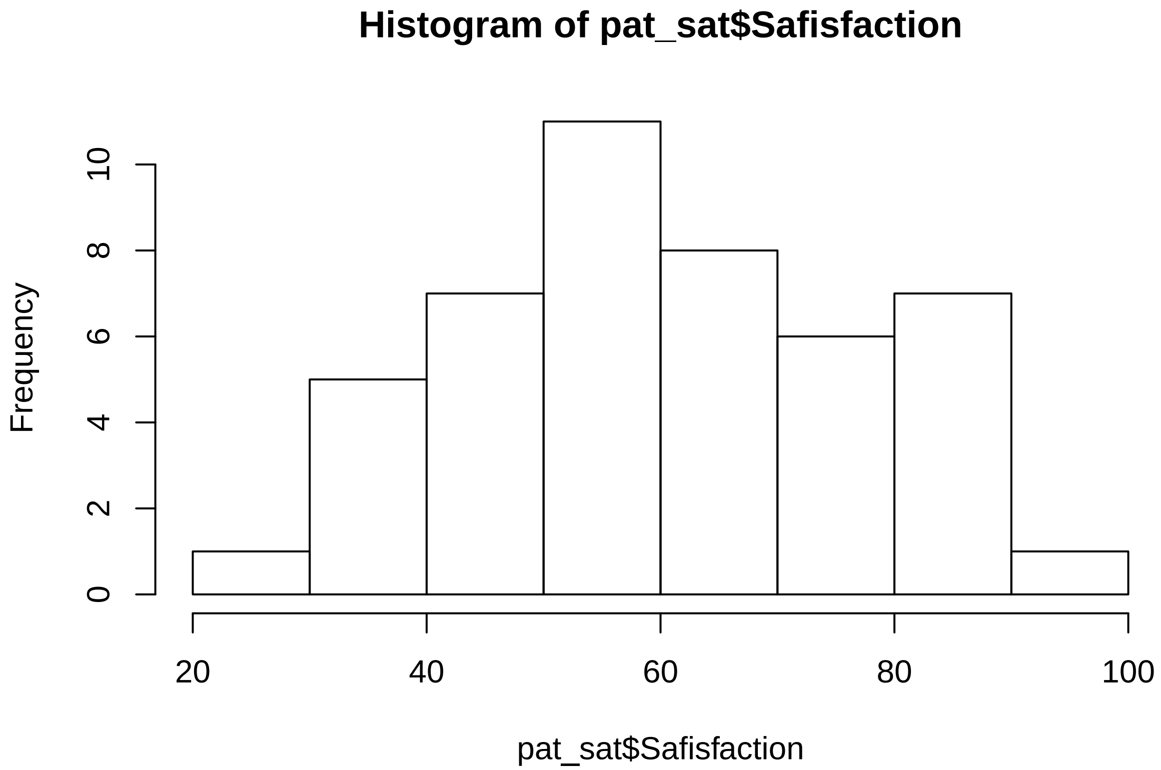
```
pat_sat <- readxl::read_excel("./data/PatSatisfaction.xlsx")
head(pat_sat)
```

```
## # A tibble: 6 x 4
##   Safisfaction   Age Severity Anxiety
##          <dbl> <dbl>    <dbl>   <dbl>
## 1           48    50       51     2.3
## 2           57    36       46     2.3
## 3           66    40       48     2.2
## 4           70    41       44     1.8
## 5           89    28       43     1.8
## 6           36    49       54     2.9
```

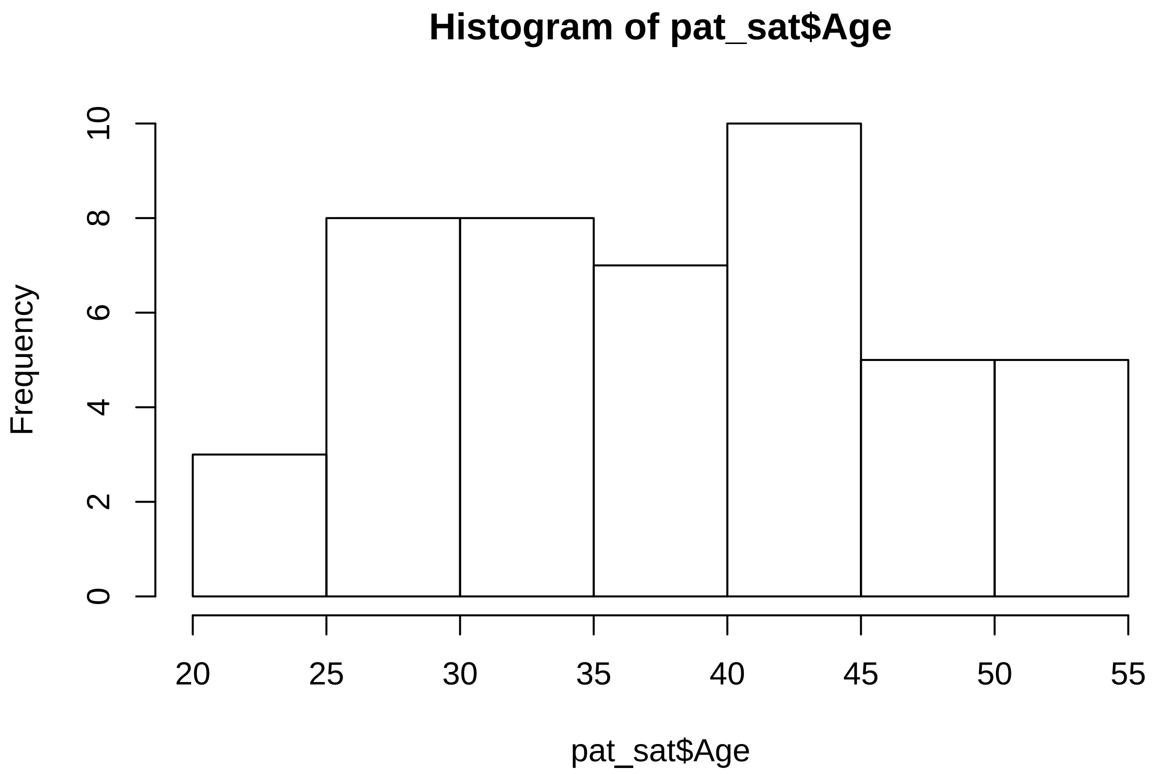
```
pairs(pat_sat)
```



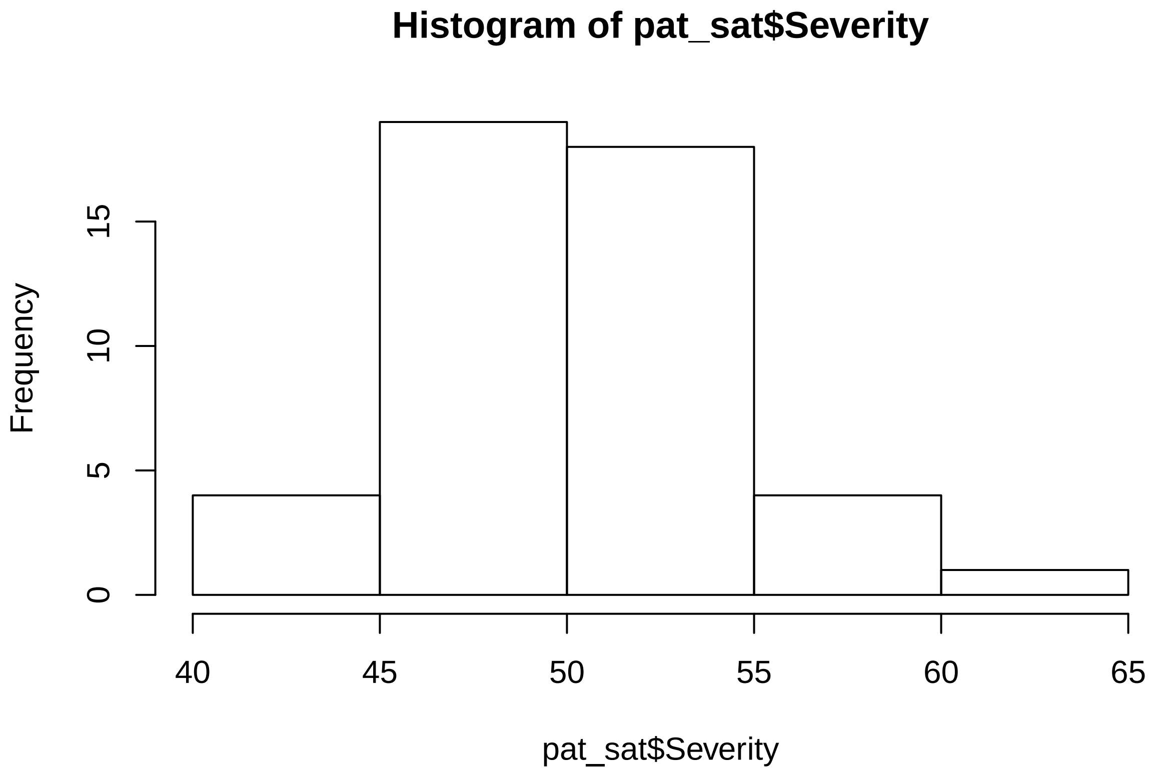
```
hist(pat_sat$Safisfaction)
```



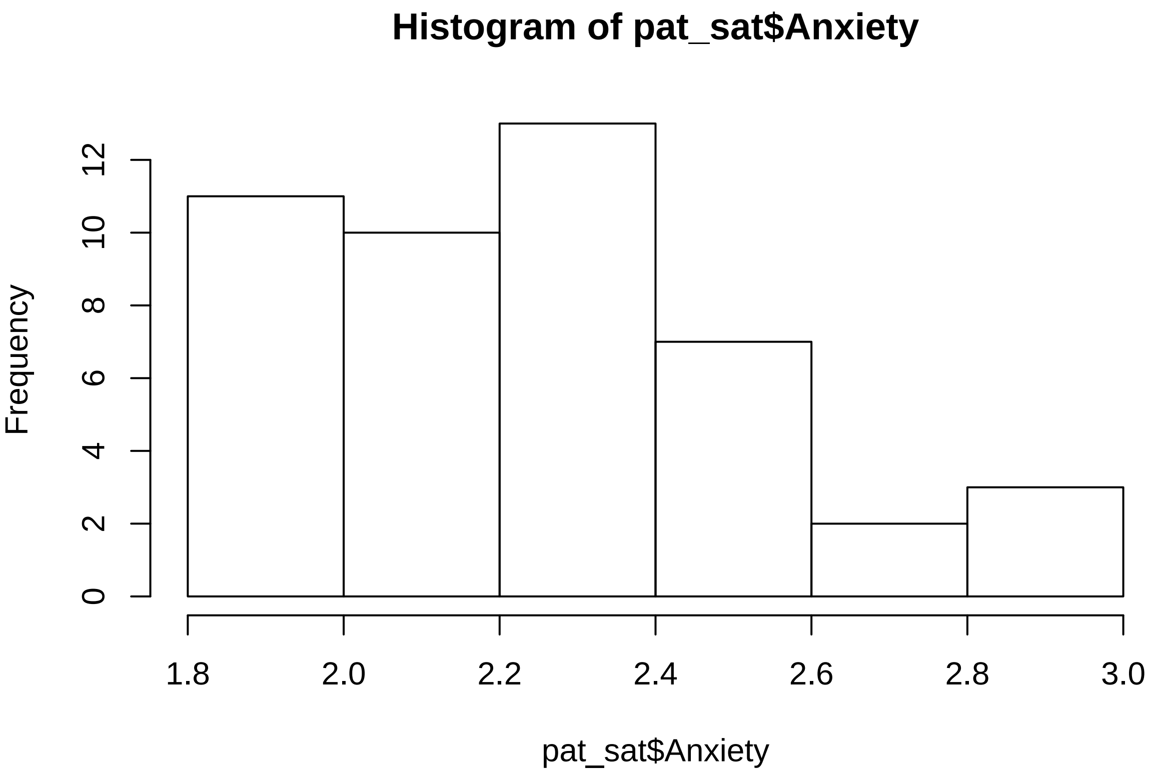
```
hist(pat_sat$Age)
```



```
hist(pat_sat$Severity)
```



```
hist(pat_sat$Anxiety)
```



```
cor(pat_sat) %>% knitr::kable()
```

| | Safisfaction | Age | Severity | Anxiety |
|---|---|---|---|---|
| Safisfaction | 1.0000000 | -0.7867555 | -0.6029417 | -0.6445910 |
| Age | -0.7867555 | 1.0000000 | 0.5679505 | 0.5696775 |
| Severity | -0.6029417 | 0.5679505 | 1.0000000 | 0.6705287 |
| Anxiety | -0.6445910 | 0.5696775 | 0.6705287 | 1.0000000 |

Comment: the correlation matrix shows that age, severity of illness and anxiety level are consistently negatively correlated with satisfaction score. Age seems to have the strongest correlation with satisfaction score while the other variables also have significant coefficient of correlations.

### b) fit a MLR and test whether there is a regression relation

In this MLR model, we will use the satisfaction as response while all other three variables as predictors. Let $Y_i$ = satisfaction (outcome), $X_{i1}$ = age, $X_{i2}$ = severity of illness, $X_{i3}$ = anxiety level

Full Model: $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_{i2} + \beta_2 X_{i2} + \epsilon_i$

```
names(pat_sat)
```

```
## [1] "Safisfaction" "Age"          "Severity"     "Anxiety"
```

```
MLR_all <- lm(Safisfaction ~ Age + Severity + Anxiety, data = pat_sat)
summary(MLR_all)
```

```
##
## Call:
## lm(formula = Safisfaction ~ Age + Severity + Anxiety, data = pat_sat)
```

```
##
## 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
```

```r
anova(MLR_all)
```

```
## Analysis of Variance Table
##
## Response: Safisfaction
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Age        1 8275.4  8275.4 81.8026 2.059e-11 ***
## Severity   1  480.9   480.9  4.7539   0.03489 *
## Anxiety    1  364.2   364.2  3.5997   0.06468 .
## Residuals 42 4248.8   101.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

First We need to do an overall F test for the three predictors:

```r
summary(MLR_all)
```

```
##
## Call:
## lm(formula = Safisfaction ~ Age + Severity + Anxiety, data = pat_sat)
##
## 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
```

State the hypothesis:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0 H_a : \text{at least one of the coefficient is nonzero}$$

Test Statistic: $F_{test} = \frac{MSR}{MSE} = 30.05 \sim F(4, 42)$

Decision Rule: at $\alpha = 0.05$, we will reject the null if $F_{test} > F(0.95, 4, 42) = 2.59$. Here we have $F_{test} = 30.05 > 2.59$, so we should reject the null and conclude that there is at least one linear association among these predictors with the outcome satisfaction level.

Then we can test these predictors one by one:

1. Test for $\beta_1$: the coefficent of age

State hypothesis: $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$

Test statisitc: we can use the generalized F test for SLR which is equivalent to t test for single regression coefficient. $F_{test} = \frac{SSR_L/1}{SSE_L/(46-2)} = 81.8 \sim F(1, 44)$

Decision rule: at $\alpha = 0.05$, we will reject the null if $F_{test} > F(1-\alpha, 2, 44) = 4.06$. Here we have $F_{test} = 81.8 > 4.06 $, with p value well below 0.001, so we will reject the null and conclude there is linear association between satisfaction score and age.

2. Test for $\beta_2$: the coefficent of Severity

State hypothesis: $H_0 : \beta_2 = 0$ vs. $H_a : \beta_2 \neq 0$

Test statisitc: $F_{test} = \frac{SSR_{X2|X1}/1}{SSE_{X1}/44} = 4.75 \sim F(1, 44)$

Rejection rule: we should reject the null if $F_{test} > 4.06$ at $\alpha = 0.05$. Here we have $F_{test} = 4.75 > 4.06$, so we should reject the null at $\alpha = 0.05$ and conclude there is linear association between Severity of illness and satisfaction level.

3. Test for $\beta_3$: the coefficent of Anxiety level

State hypothesis: $H_0 : \beta_3 = 0$ vs. $H_a : \beta_3 \neq 0$

Test statistic: $F_{test} = \frac{SSR_{X3|X1,X2}/1}{SSE_{X1,X2}/43} = 3.6 \sim F(1, 43)$

Rejection rule: at $\alpha = 0.05$, we should reject null if $F_{test} > F(0.95, 1, 43) = 4.07$. However, we obtained $F_{test} = 3.6 < 4.07$, so we do not have evidence to reject the null. Therefore we should not include anxiety level as one of the explaintary variable since it does not reduce SSE significantly in a model with exisiting variables age and Severity of illness.