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") %>%
  mutate(gender = as.factor(gender),
         complications = as.factor(complications))
```

```
## 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> <fct>          <int> <int>    <int> <fct>
## 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
- **gender**: gender of patient
- **interventions**: number of interventions (integers)
- **drugs**: number of drugs the patients take.
- **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 potentially have effect on the association relationship between our main predictor and main outcome, including age, interventions, drugs used, complications, and duration of disease. We will first take a look at the availible variables:

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

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

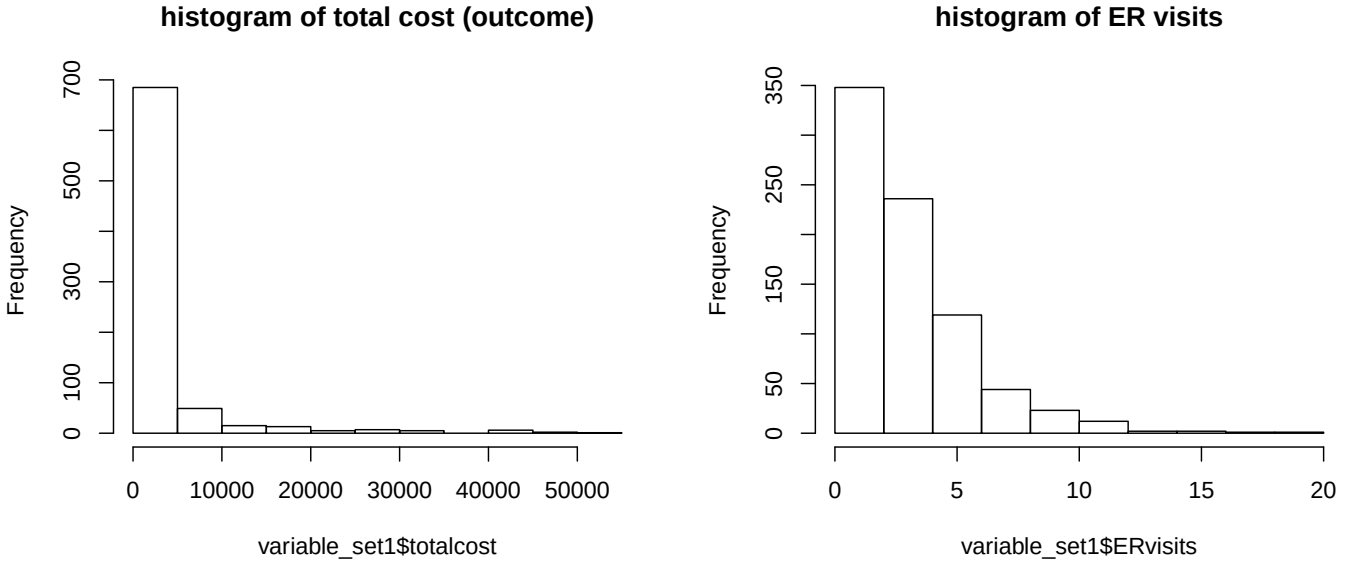
totalcost	ERvisits	age	interventions	drugs	comorbidities	duration
Min. : 0.0	Min. : 0.000	Min. :24.00	Min. : 0.000	Min. :0.0000	Min. : 0.000	Min. : 0.00
1st Qu.: 161.1	1st Qu.: 2.000	1st Qu.:55.00	1st Qu.: 1.000	1st Qu.:0.0000	1st Qu.: 0.000	1st Qu.: 41.75
Median : 507.2	Median : 3.000	Median :60.00	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

```
knitr::kable(summary(variable_set2))
```

gender	complications
0:608	0:745
1:180	1: 42
NA	3: 1

Visualize the distribution of these variables:

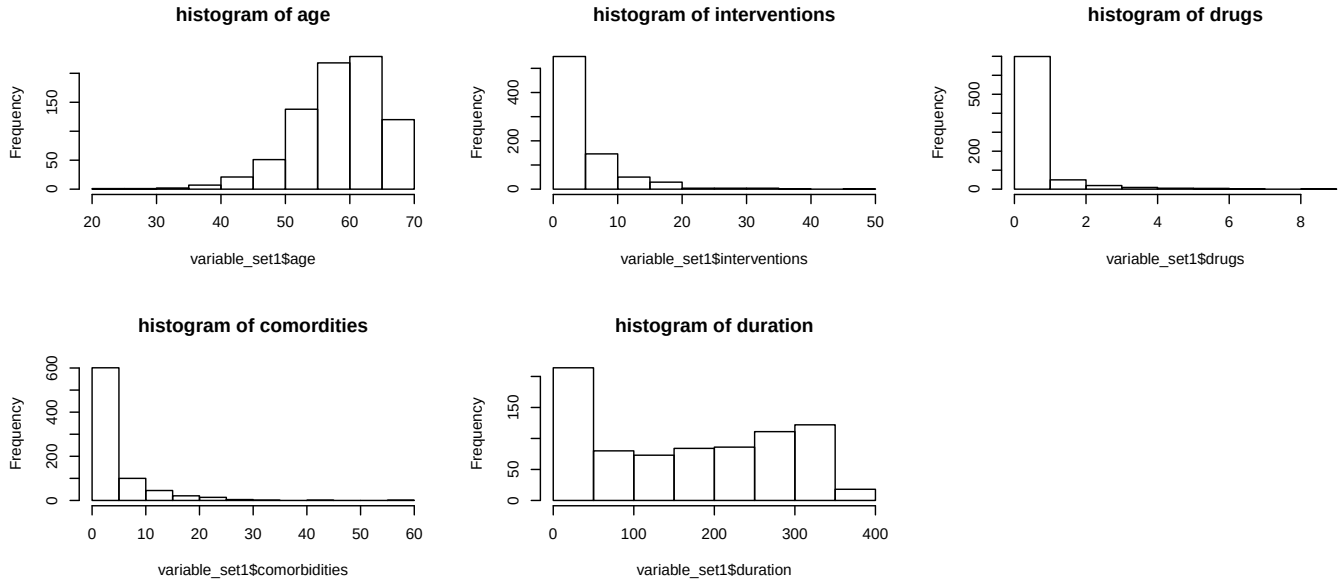
```
par(mfrow = c(1,2))
hist(variable_set1$totalcost, main = "histogram of total cost (outcome)")
hist(variable_set1$ERvisits, main = "histogram of ER visits")
```



Describe the main outcome and main predictor:
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. We categorized two other remaining variables

gender and complications as categorical variables. From the summary table, we saw

```
par(mfrow = c(2,3))
hist(variable_set1$age, main = "histogram of age")
hist(variable_set1$interventions, main = "histogram of interventions")
hist(variable_set1$drugs, main = "histogram of drugs")
hist(variable_set1$comorbidities, main = "histogram of comordities")
hist(variable_set1$duration, main = "histogram of duration")
```

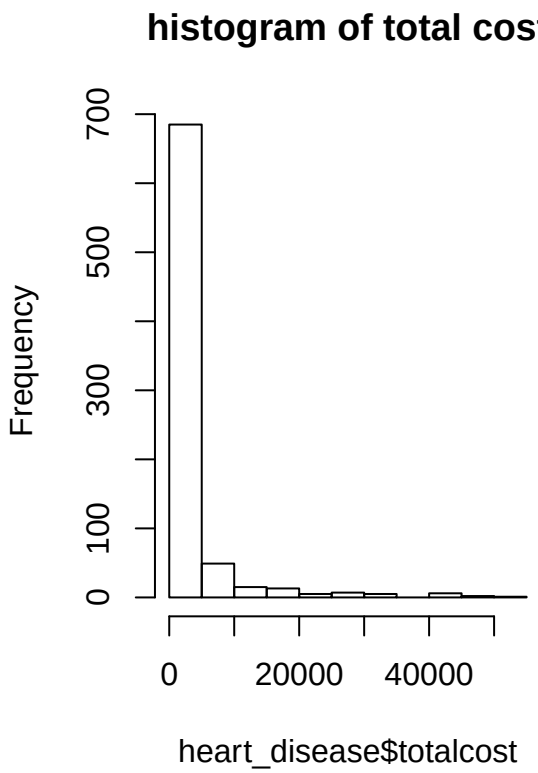


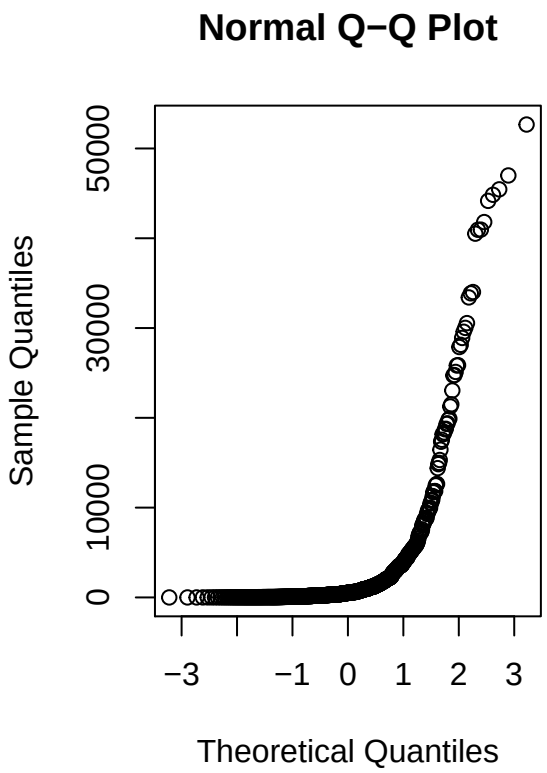
Describe other covariables: age is slightly left skewed which means elder people have been overly sampled. The median of intervention is about 5 with large IQR of 5. drugs?. Commordities have median of 3.7 with large IQR 5. Duration of heart disease is roughly uniformly distributed from 50 to 350 days with median 165 days and IQR 240 days. Therefore, these co-variables are not normally distributed in the sample. We categorized two other remaining variables gender and complications as categorical variables. From the summary table, we saw one sex is oversampled (608). The majority of patients do not have any complications.

b) investigate the shape of distribution for total cost

First we examined the distribution of raw data of total cost and check its normality:

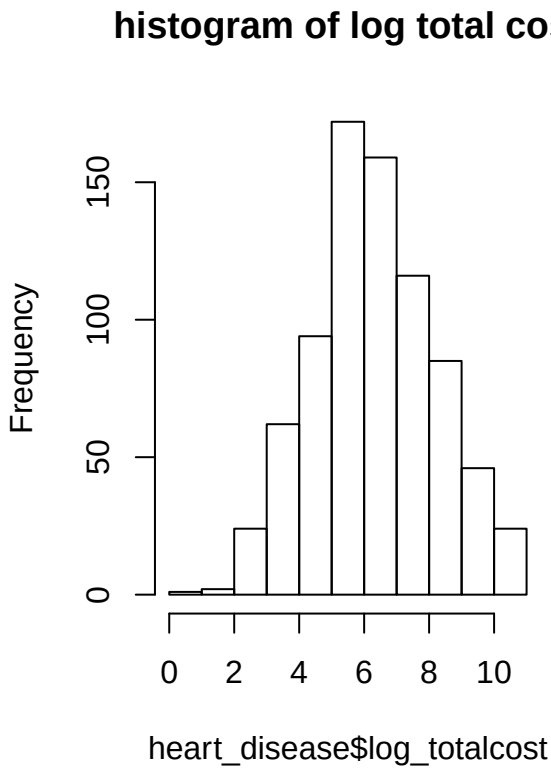
```
par(mfrow = c(1,2))
hist(heart_disease$totalcost, main = "histogram of total cost")
qqnorm(heart_disease$totalcost)
```

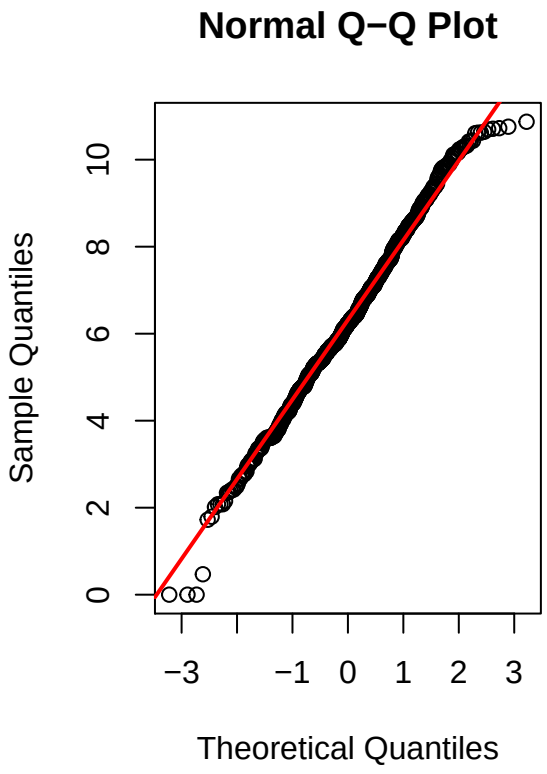




Then we try log transformation on **totalcost** to see if this will improve the normality.

```
heart_disease <- mutate(heart_disease, log_totalcost = log(totalcost))
par(mfrow = c(1,2))
hist(heart_disease$log_totalcost, main = "histogram of log total cost")
heart_disease$log_totalcost[is.infinite(heart_disease$log_totalcost)] = 0.001
qqnorm(heart_disease$log_totalcost)
qqline(heart_disease$log_totalcost, col = "red", lwd = 2)
```



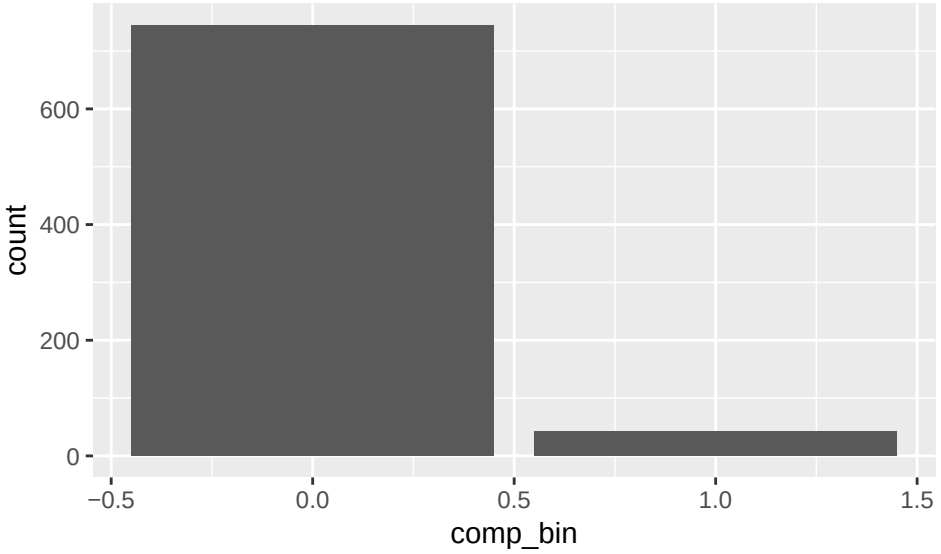


Comment: After log transformation, we saw a pretty good bell shaped ditribution. So we will use this transformed data in the linear model fitting and interpretation.

c) dichotomize complications

0 represents no complications; 1 represents having complications

```
heart_disease <- heart_disease %>%
  mutate(comp_bin = ifelse(complications == 0, 0, 1))
heart_disease %>% ggplot(aes(x = comp_bin)) + geom_bar()
```



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).

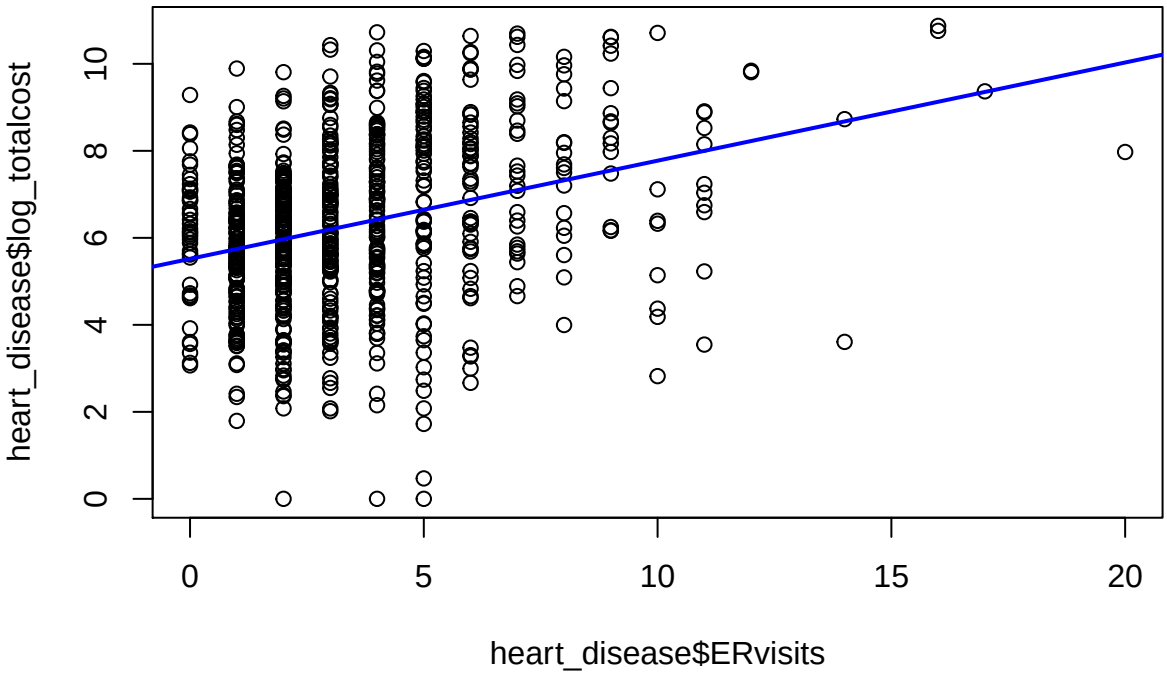
Then our model is $logY_i = \beta_0 + \beta_1 X_i + \epsilon_i$. Here we assume the error is normally distribued. then $\epsilon_i \sim N(0, \sigma^2)$

```
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
## -6.6444 -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   9.215   <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:  0.09635
```

```
## F-statistic: 84.91 on 1 and 786 DF,  p-value: < 2.2e-16
```

```
plot(heart_disease$ERvisits, heart_disease$log_totalcost)
abline(SLR, col = "blue", 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 expected total cost in dollars *on logarithm scale* will increase by 0.23. When the ER visit is 0, the expected total cost in dollar *on logarithm scale* will be 5.517. The p value for two estimators β_0 and β_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} = comp_bin(factor with two levels)

The full model is :

$logY_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$

Now add a potential modifier (interaction):

$logY_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \epsilon_i$

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_inter) %>% tidy %>% knitr::kable()
```

term	estimate	std.error	statistic	p.value
(Intercept)	5.4786282	0.1057633	51.800833	0.0000000
ERvisits	0.2097772	0.0250825	8.363489	0.0000000
comp_bin	2.2000533	0.5584979	3.939233	0.0000890

term	estimate	std.error	statistic	p.value
ERvisits:comp_bin	-0.0977939	0.0969959	-1.008228	0.3136561

```
anova(MLR_comp, MLR_comp_inter) %>% tidy
```

```
## Warning: Unknown or uninitialised column: 'term'.
```

```
## # A tibble: 2 x 6
##   res.df   rss    df sumsq statistic p.value
## *  <dbl> <dbl> <dbl> <dbl>     <dbl>   <dbl>
## 1    785 2465.    NA NA         NA     NA
## 2    784 2461.     1  3.19      1.02   0.314
```

Comment: Based on the regression summary and anova result, p value for the interaction coefficient β_3 is 0.314, which is quite large. Anova F test for comparing two models indicate adding the interaction term does not increase SSR by significant amount. 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.

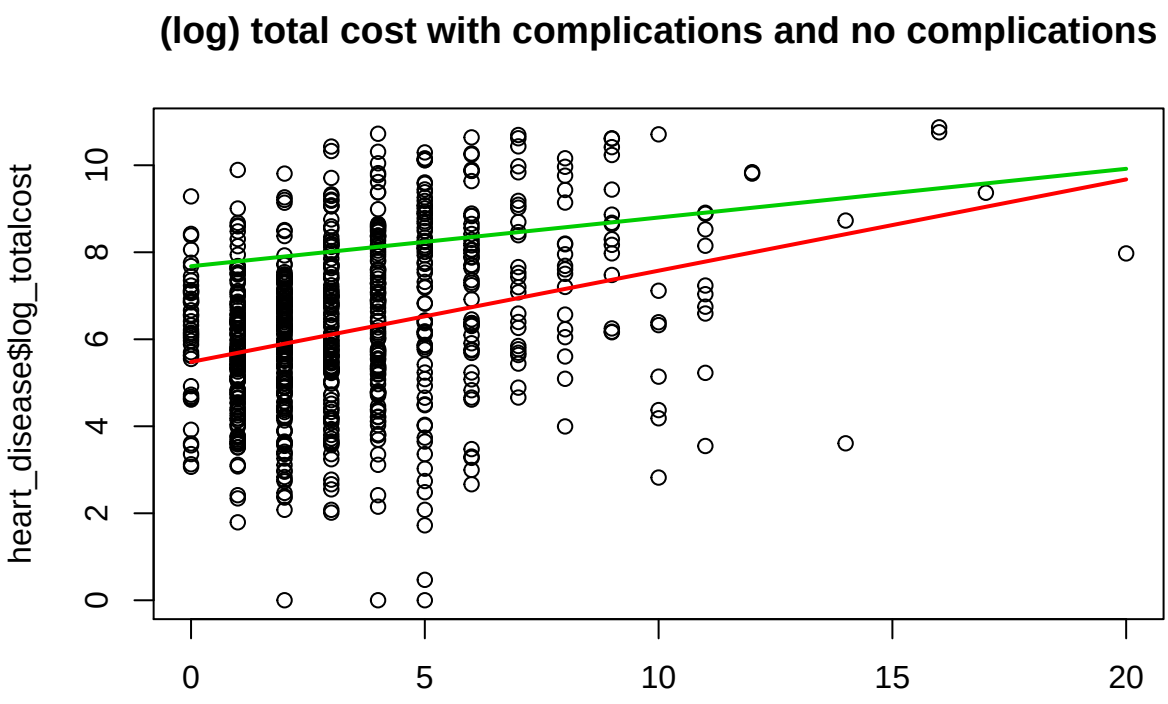
We can also visualize this interaction model:

```
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 = 1
yhat2 <- beta[1] + beta[3] + (beta[2] + beta[4])*ER

plot(heart_disease$ERvisits, heart_disease$log_totalcost,
     main = "(log) total cost with complications and no complications",
     xlab = "")
lines(ER, yhat1, col = 2, lwd = 2) # total cost of comp_bin = 0 with fixed ER
lines(ER, yhat2, col = 3, lwd = 2) # total cost of comp_bin greater than 0 with fixed ER
```



model-1.bb

Comment: although we expect two parallel lines on the plot if there is truly no interaction effect. However, the statistical test we conducted above indicate that although there is some interaction effect, the effect is not significant. As a conclusion, we will not consider this mediator effect.

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: $logY_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$

```
comp_ER <- cor(heart_disease$comp_bin, heart_disease$ERvisits)
comp_total <- cor(heart_disease$comp_bin, heart_disease$log_totalcost)
SLR <- lm(log_totalcost ~ ERvisits, data = heart_disease)
MLR_comp <- lm(log_totalcost ~ ERvisits + factor(comp_bin), data = heart_disease)
SLR %>% tidy %>% knitr::kable()
```

term	estimate	std.error	statistic	p.value
(Intercept)	5.5170148	0.1058538	52.119203	0
ERvisits	0.2256856	0.0244923	9.214538	0

```
MLR_comp %>% tidy %>% knitr::kable()
```

term	estimate	std.error	statistic	p.value
(Intercept)	5.5003974	0.1035370	53.124944	0
ERvisits	0.2032377	0.0242296	8.387989	0
factor(comp_bin)1	1.7135233	0.2811723	6.094210	0

```
anova(SLR, MLR_comp)
```

```
## Analysis of Variance Table
##
## Model 1: log_totalcost ~ ERvisits
## Model 2: log_totalcost ~ ERvisits + factor(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
```

```
(0.226 - 0.203)/ 0.226
```

```
## [1] 0.1017699
```

Comment: From the regression result, we saw the coefficient of **comp_bin** is quite significant with p value well below 0.001. The anova result for two model comparison shows that adding complication variable greatly reduce the overall SSTO while increasing SSR, with p value well below 0.001. So we should include complications in our linear model.

iii) decide if comp_bin should be included along with ERvisits

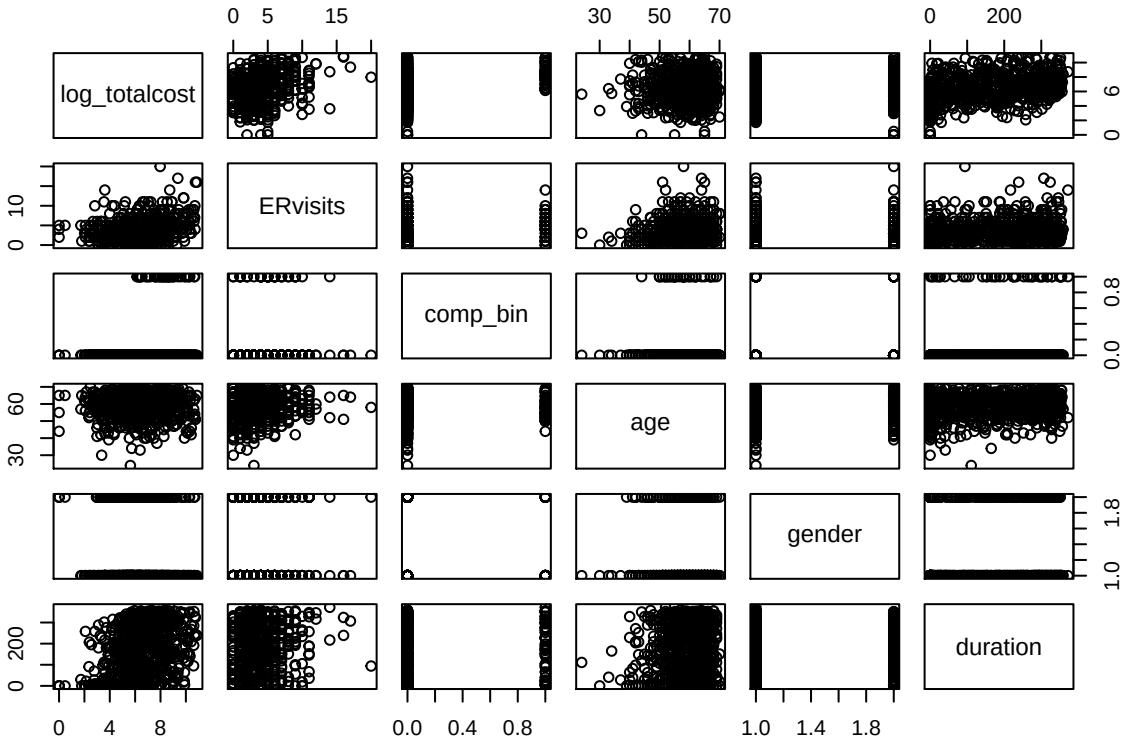
From above test, we should include **comp_bin** as a predictor in our additive linear model. Adding **com_bin** in the model increase SSR significantly. The coefficient of **comp_bin** is also significant in the linear model from above discussion.

f) examine additional covariates

(i) fit a MLR

We start with screen for any colinearility of variables among **ERvisits**, **comp_bin**, **age**, **gender**, and **duration**

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



Then we fit a MLR with all variables of interests:

```
fit_all <- lm(log_totalcost ~ ERvisits + comp_bin + age + factor(gender) + duration,
              data = heart_disease)
summary(fit_all)
```

```
##
## Call:
## lm(formula = log_totalcost ~ ERvisits + comp_bin + age + factor(gender) +
##     duration, data = heart_disease)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.4529 -1.0367 -0.1108  0.9507  4.3478
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       5.9377052  0.5139206  11.554  < 2e-16 ***
## ERvisits          0.1746113  0.0227290   7.682 4.68e-14 ***
## comp_bin          1.5103177  0.2602679   5.803 9.46e-09 ***
## age              -0.0208968  0.0087343  -2.392    0.017 *
## factor(gender)1 -0.2075121  0.1396551  -1.486    0.138
## duration          0.0057688  0.0004922  11.720  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.635 on 782 degrees of freedom
## Multiple R-squared:  0.269,  Adjusted R-squared:  0.2643
## F-statistic: 57.56 on 5 and 782 DF,  p-value: < 2.2e-16
```

```
vif(fit_all)
```

```
##        ERvisits        comp_bin             age factor(gender)1
##        1.057863        1.030044        1.024444        1.013165
##        duration
##        1.042700
```

Comment: From the VIF test, there is no significant colinearity between predictors. Based on the t test statistics for each regression coefficients along with their p values, we observe linear relationship between gender and total cost is weak. Then we will perform ANOVA test and make decision.

(ii) compare SLR and MLR

Here we construct several nested MLR to determine if we want to include the predicor in the model or not.

```
anova(fit_all)
```

```
## Analysis of Variance Table
##
## Response: log_totalcost
##                  Df  Sum Sq Mean Sq  F value    Pr(>F)
## ERvisits          1  278.84  278.84 104.2978 < 2.2e-16 ***
## comp_bin          1  116.61  116.61  43.6154  7.38e-11 ***
## age               1    1.83    1.83   0.6828    0.4089
## factor(gender)    1    4.91    4.91   1.8368    0.1757
## duration          1  367.24  367.24 137.3612 < 2.2e-16 ***
## Residuals       782 2090.69    2.67
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comment: From F test for each nested model indicate that we should definitely include **duration** in our model and better exclude age and gender since they do not much additional information in the exisiting model. Therefore from all above analysis, we come up with a fitted model of main outcome with its main predictor along with other covariates is the follow:

```
MLR_final <- lm(log_totalcost ~ ERvisits + comp_bin + duration,
              data = heart_disease)
summary(MLR_final)
```

```
##
## Call:
## lm(formula = log_totalcost ~ ERvisits + comp_bin + duration,
##     data = heart_disease)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5497 -1.0961 -0.1131  0.9654  4.6552
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.7094831  0.1181700  39.853  < 2e-16 ***
## ERvisits    0.1682334  0.0226476   7.428 2.88e-13 ***
## comp_bin    1.5451974  0.2608238   5.924 4.69e-09 ***
## duration    0.0056087  0.0004897  11.453  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.641 on 784 degrees of freedom
## Multiple R-squared:  0.2618, Adjusted R-squared:  0.2589
## F-statistic: 92.67 on 3 and 784 DF,  p-value: < 2.2e-16
```

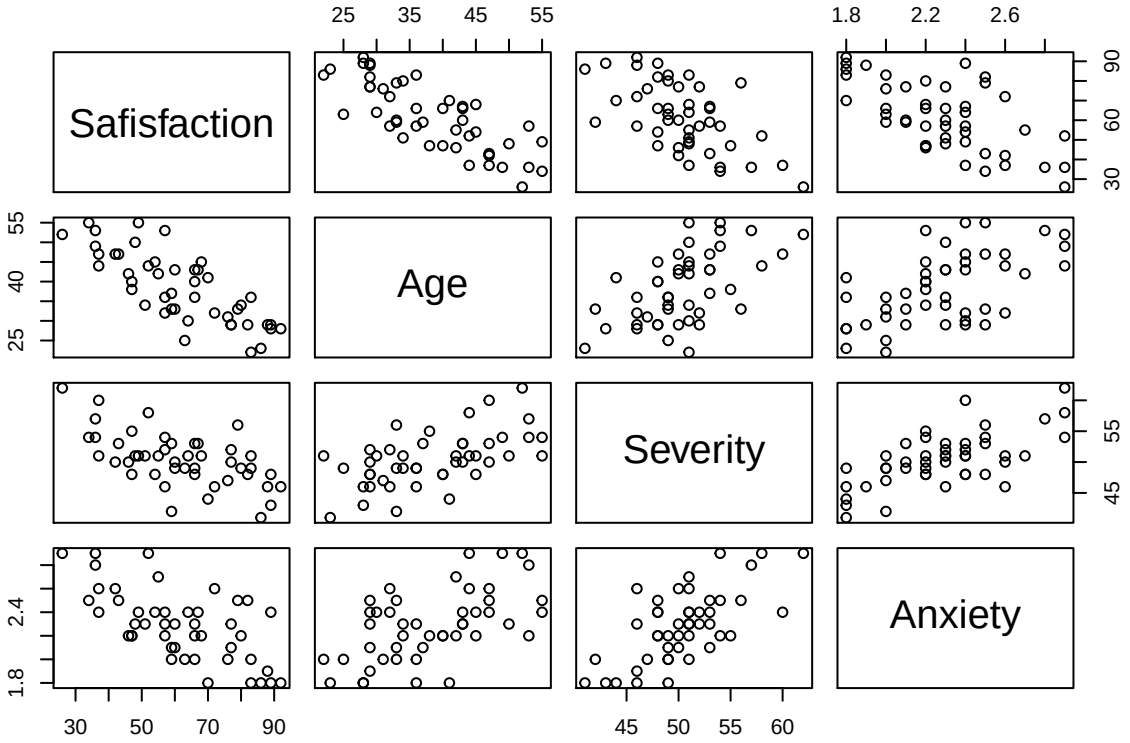
total cost (log) = 4.7 + 0.17ERvisit + 1.55comp_bin + 0.0056duration

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")
pairs(pat_sat)
```




```
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. However, covariates are positively correlated with each other significantly as well. So we should keep this in mind.

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: $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_2 X_{i2} + \epsilon_i$

```
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
```

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

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: $F_{test} = \frac{MSR}{MSE} = \frac{9120.5/3}{4248.8/42} = 30.05 \sim F(3, 42)$

Decision Rule: at $\alpha = 0.05$, we will reject the null if $F_{test} > F(0.95, 3, 42) = 2.83$. Here we have $F_{test} = 30.05 > 2.83$, so we should reject the null and conclude that there is at least one linear association among these predictors with the outcome satisfaction level.

c) compute 95% CI for estimated coefficients

create a table with estimator and 95% Confidence Interval:

```
summary(MLR_all) %>%
  tidy %>%
  mutate(lower_bound = estimate - qt(0.975, 42) * std.error,
         upper_bound = estimate + qt(0.975, 42) * std.error) %>%
  dplyr::select(term, estimate, std.error, lower_bound, upper_bound)
```

```
## # A tibble: 4 x 5
##   term        estimate std.error lower_bound upper_bound
##   <chr>          <dbl>     <dbl>       <dbl>       <dbl>
## 1 (Intercept)  158.       18.1        122.       195.
## 2 Age           -1.14      0.215       -1.58      -0.708
## 3 Severity      -0.442     0.492       -1.43       0.551
## 4 Anxiety      -13.5       7.10       -27.8        0.858
```

Interpret severity of illness:

While holding age and anxiety level constant, the expected **decrease** of satisfaction score with an unit increase in severity of illness is 0.442. We are 95% confident that the true mean change of satisfaction score with one unit increase in severity of illness is between -1.43 to 0.551.

d) Obtain interval estimate for a new patient

```
new_data <- tibble(Age = 35,
                   Severity = 42,
                   Anxiety = 2.1)
predict.lm(MLR_all, new_data, interval="prediction", conf.level = 0.95)
```

```
##        fit      lwr      upr
## 1 71.68332 50.06237 93.30426
```

Comment: The point estimator for this new patient's satisfaction score is 71.7. We are 95% confident that the predicted satisfaction score for this new patient will be between 50.1 to 93.3

e) test whether anxiety level can be dropped from the MLR

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

Test statistic: $F_{test} = \frac{SSR_{X3|X1,X2}/1}{SSE_{X1,X2}/43} = \frac{364.16}{4613/43} = 3.6 \sim F(1, 43)$

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 SSTO significantly in a model with exisiting variables age and Severity of illness.

Perform test in R:

```
MLR_Age_Sev <- lm(Safisfaction ~ Age + Severity, data = pat_sat)
MLR_all <- lm(Safisfaction ~ Age + Severity + Anxiety, data = pat_sat)
anova(MLR_Age_Sev) %>% tidy
```

```
## # A tibble: 3 x 6
##   term         df sumsq meansq statistic   p.value
##   <chr>     <int> <dbl>  <dbl>     <dbl>     <dbl>
## 1 Age           1 8275.  8275.     77.1   3.80e-11
## 2 Severity      1  481.   481.      4.48  4.01e- 2
## 3 Residuals    43 4613.   107.     NA    NA
```

```
anova(MLR_all) %>% tidy
```

```
## # A tibble: 4 x 6
##   term         df sumsq meansq statistic   p.value
##   <chr>     <int> <dbl>  <dbl>     <dbl>     <dbl>
## 1 Age           1 8275.  8275.     81.8   2.06e-11
## 2 Severity      1  481.   481.      4.75  3.49e- 2
## 3 Anxiety       1  364.   364.      3.60  6.47e- 2
## 4 Residuals    42 4249.   101.     NA    NA
```

```
anova(MLR_Age_Sev, MLR_all) %>% tidy
```

```
## Warning: Unknown or uninitialised column: 'term'.
```

```
## # A tibble: 2 x 6
##   res.df   rss    df sumsq statistic p.value
## *  <dbl> <dbl> <dbl> <dbl>     <dbl>   <dbl>
## 1     43 4613.    NA   NA      NA     NA
```

```
## 2     42 4249.     1  364.      3.60  0.0647
```