biostatistical methods homework 4

```r
library(tidyverse)
library(knitr)
library(patchwork)
library(readxl)
```

Problem1

(a)

$$
\begin{aligned}
b_1 &= \frac{n\sum X_iY_i - \sum X_i \sum Y_i}{n\sum X_i^2 - (\sum X_i)^2} \\
&= \frac{\sum X_iY_i - n\bar{Y}\bar{X}}{\sum X_i^2 - n\bar{X}^2}
\end{aligned}
$$

$$
b_0 = \bar{Y} - b_1\bar{X}
$$

$$
\begin{aligned}
\sum X_iY_i - n\bar{Y}\bar{X} &= \sum X_iY_i - \bar{X}\sum Y_i \\
&= \sum (X_i - \bar{X})Y_i
\end{aligned}
$$

$$
\begin{aligned}
E\left\{\sum (X_i - \bar{X})Y_i\right\} &= \sum (X_i - \bar{X})E(Y_i) \\
&= \sum (X_i - \bar{X})(\beta_0 + \beta_1 X_i) \\
&= \beta_0 \sum X_i - n\bar{X}\beta_0 + \beta_1 \sum X_i^2 - n\bar{X}^2\beta_1 \\
&= \beta_1(\sum X_i^2 - n\bar{X}^2)
\end{aligned}
$$

$$
\begin{aligned}
E(b_1) &= \frac{E\left\{\sum (X_i - \bar{X})Y_i\right\}}{\sum X_i^2 - n\bar{X}^2} \\
&= \frac{\beta_1(\sum X_i^2 - n\bar{X}^2)}{\sum X_i^2 - n\bar{X}^2} \\
&= \beta_1
\end{aligned}
$$

$$
\begin{aligned}
E(b_0) &= E(\bar{Y} - b_1\bar{X}) \\
&= \frac{1}{n}\sum E(Y_i) - E(b_1)\bar{X} \\
&= \frac{1}{n}\sum [\beta_0 + \beta_1 X_i] - \beta_1\bar{X} \\
&= \frac{1}{n}\left[n\beta_0 + n\beta_1\bar{X}\right] - \beta_1\bar{X} \\
&= \beta_0
\end{aligned}
$$

(b)

$$\begin{aligned} Y_i &= \hat{\beta}_1 X_i + \hat{\beta}_0 \\ &= \hat{\beta}_1 X_i + \bar{Y} - \hat{\beta}_1 \bar{X} \end{aligned}$$

$$\begin{aligned} X_i &= \bar{X} \\ Y_i &= \hat{\beta}_1 \bar{X} + \bar{Y} - \hat{\beta}_1 \bar{X} \\ &= \bar{Y} \end{aligned}$$

So the Least Square line equation always goes through the point $(\bar{X}, \bar{Y})$

(c)

$$\begin{aligned} log_e L &= -\frac{n}{2} log_e 2\pi - \frac{n}{2} log_e \sigma^2 - \frac{1}{2\sigma^2} \sum (Y_i - \beta_0 - \beta_1 X_i)^2 \\ \frac{\partial (log_e L)}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum (Y_i - \beta_0 - \beta_1 X_i)^2 \end{aligned}$$

$$\begin{aligned} \hat{\sigma}^2 &= \frac{\sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2}{n} \\ &= \frac{\sum (Y_i - \hat{Y}_i)^2}{n} \end{aligned}$$

Find its expected value

$$\begin{aligned} E(\hat{\sigma}^2) &= E\left(\frac{SSE}{n}\right) \\ &= E\left(\frac{SSE}{n-2} \times \frac{n-2}{n}\right) \\ &= \frac{n-2}{n} \times E\left(\frac{SSE}{n-2}\right) \\ &= \frac{n-2}{n} \sigma^2 \end{aligned}$$

Comment on the unbiasness property

As the result shown above, $\hat{\sigma}^2$ is a biasd estimator of σ^2 as the unbiased estimator of σ^2 is MSE:

$$s^2 = MSE = \frac{SSE}{n-2} = \frac{\sum (Y_i - \hat{Y}_i)^2}{n-2} = \frac{\sum e_i^2}{n-2}$$

$$E\{MSE\} = \sigma^2$$

Problem 2

For this problem, you will be using data 'HeartDisease.csv'.

```
heart_data = read_csv("./data/HeartDisease.csv")
```

The investigator is mainly interested if there is an association between 'total cost' (in dollars) of patients diagnosed with heart disease and the 'number of emergency room (ER) visits'.

Further, the model will need to be adjusted for other factors, including 'age', 'gender', 'number of complications' that arose during treatment, and 'duration of treatment condition'.

a)

Provide a short description of the data set: what is the main outcome, main predictor and other important covariates.

This dataset include 10 variables and 788 observations. The main outcome is `totalcost` which represents the total cost (in dollars) of heart-diseased patients. The main predictor is the `ERvisits` which represents the number of emergency room (ER) visits. Other important covariates are `age`, `gender`, `complications` and `duration`.

Also, generate appropriate descriptive statistics for all variables of interest (continuous and categorical) – no test required.

```
mean_and_sd = function(x) {

  if (!is.numeric(x)) {
    stop("Argument x should be numeric")
  } else if (length(x) == 1) {
    stop("Cannot be computed for length 1 vectors")
  }

  mean_x = mean(x)
  sd_x = sd(x)

  tibble(
    mean = mean_x,
    sd = sd_x
  )
}
```

totalcost

```
mean_and_sd(heart_data$totalcost)
```

```
## # A tibble: 1 x 2
##    mean    sd
##   <dbl> <dbl>
## 1 2800. 6690.
```

The mean of the total cost is about 2800 with a standard deviation of 6690.26.

ERvisits

```
summary(heart_data$ERvisits)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   2.000   3.000   3.425   5.000  20.000
```

The minimum number of emergency room (ER) visits is 0 and the maximum is 20. The median is 3 with 1st Qu. of 2 and 3rd Qu. of 5.

age

```
mean_and_sd(heart_data$age)
```

```
## # A tibble: 1 x 2
##    mean    sd
##   <dbl> <dbl>
## 1  58.7  6.75
```

The distribution of age is centered at about 59 with a standard deviation of 6.75.

gender

```
(summary(as.factor(heart_data$gender)))
```

```
##   0   1
## 608 180
```

As 0 represents female and 1 represents male, there are 608 female and 180 male in the dataset.

complications

```
(summary(as.factor(heart_data$complications)))
```

```
##   0   1   3
## 745  42   1
```

As we observed from the dataset, there number of complications existing in this dataset is simply 0, 1 and 3. Using summary function, we can conclude that there are 745 patients have zero complicatoins and 42 patients have one complicatoins, and there is only 1 patient has 3 complicaton.

duration

```
mean_and_sd(heart_data$duration)
```

```
## # A tibble: 1 x 2
##    mean    sd
##   <dbl> <dbl>
## 1  164.  121.
```

The average duration of treatment condition is 164 with a standard deviation of 121.

b)

```r
totalcost_non = heart_data %>%
  ggplot(aes(x = totalcost)) +
  geom_density() +
  labs(
       x = 'Total Cost',
       y = 'Density'
       )
```

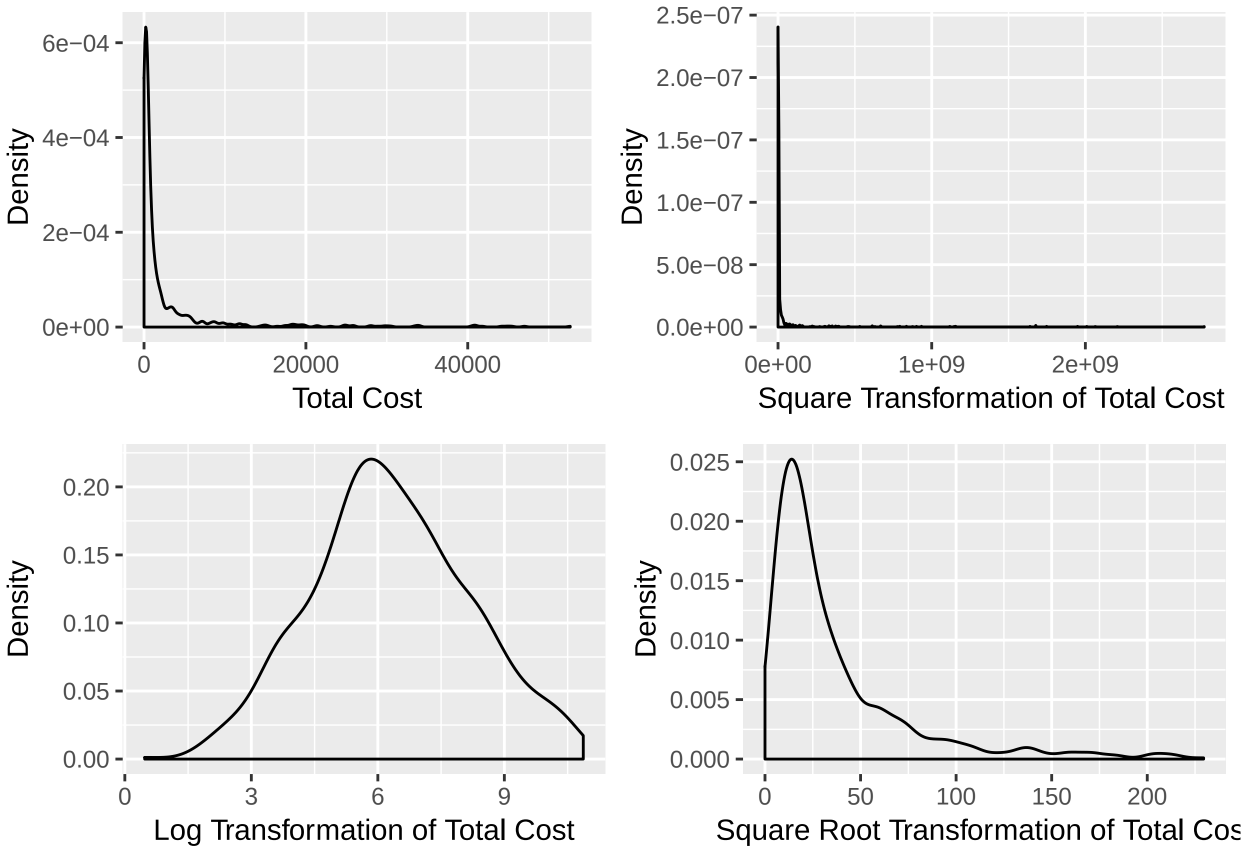
```r
totalcost_sq = heart_data %>%
  ggplot(aes(x = (totalcost)^2)) +
  geom_density() +
  labs(
       x = 'Square Transformation of Total Cost',
       y = 'Density'
       )
```

```r
totalcost_log = heart_data %>%
  ggplot(aes(x = log(totalcost))) +
  geom_density() +
  labs(
       x = 'Log Transformation of Total Cost',
       y = 'Density'
       )
```

```r
totalcost_sqrt = heart_data %>%
  ggplot(aes(x = sqrt(totalcost))) +
  geom_density() +
  labs(
      x = 'Square Root Transformation of Total Cost',
      y = 'Density'
      )
```

```r
(totalcost_non + totalcost_sq) / (totalcost_log + totalcost_sqrt)
```

```
## Warning: Removed 3 rows containing non-finite values (stat_density).
```



The shape of the distribution for `totalcost` is right skewed. After trying different transformation we find that the log transformation makes the plot approximate to normal distribution.

c)

Create a new variable called 'comp_bin' by dichotomizing 'complications': 0 if no complications, and 1 otherwise.

```
heart_data = heart_data %>%
  mutate(comp_bin = ifelse(complications == 0, 0, 1)) %>%
  mutate(comp_bin = as.character(comp_bin))
```

d)

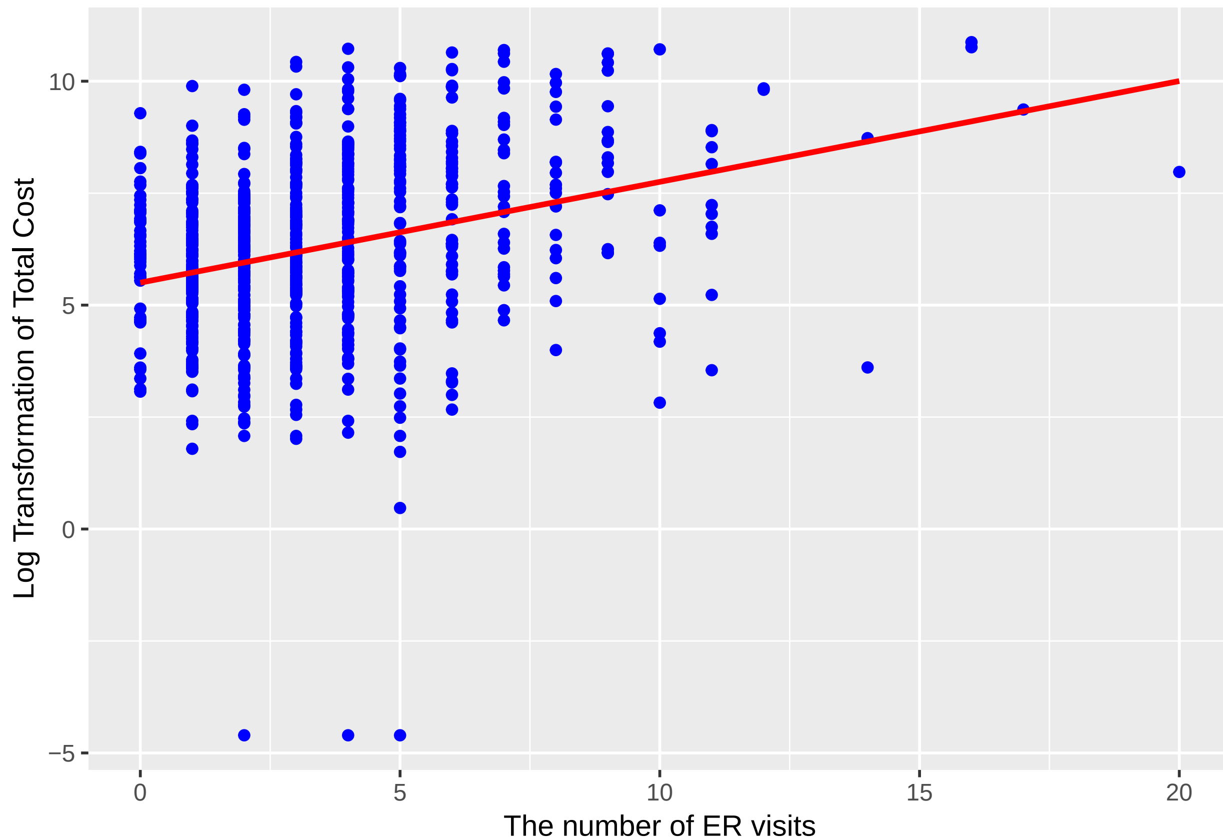
Based on our decision in part b), fit a simple linear regression (SLR) between the original or transformed 'total cost' and predictor 'ERvisits'. This includes a scatterplot and results of the regression, with appropriate comments on significance and interpretation of the slope.

```
heart_data_trans = heart_data

heart_data_trans$totalcost[heart_data_trans$totalcost==0]=0.01

heart_data_trans = heart_data_trans %>%
  mutate(totalcost = log(totalcost))
```

```r
heart_data_trans %>%
  ggplot(aes(x = ERvisits, y = totalcost)) +
  geom_point(color = 'blue') +
  geom_smooth(method = "lm", color = 'red', se = FALSE) +
  labs(
      x = 'The number of ER visits',
      y = 'Log Transformation of Total Cost'
      )
```




```r
fit_SLR = lm(totalcost ~ ERvisits, data = heart_data_trans)
summary(fit_SLR)
```

```
##
## Call:
## lm(formula = totalcost ~ ERvisits, data = heart_data_trans)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.2321  -1.1013   0.0529   1.3055   4.3224
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.5016     0.1106  49.725   <2e-16 ***
## ERvisits      0.2251     0.0256   8.792   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 1.894 on 786 degrees of freedom
## Multiple R-squared:  0.08954,    Adjusted R-squared:  0.08838
## F-statistic:  77.3 on 1 and 786 DF,  p-value: < 2.2e-16
```

$$
\begin{aligned}
log\left(\frac{Y_2}{Y_1}\right) &= \beta_1 = 0.22672 \\
\frac{Y_2}{Y_1} &= exp^{0.2251} = 1.25 \\
Y_2 &= 1.25Y_1
\end{aligned}
$$

The plot above shows the scatterplot and results of the regression. Using `summary` function, we can see that the estimate slope is 0.2251 with a p-value <2e-16, which strongly indicates that the slope is not equal to 0 and there is significant relationship with `ERvisits` and `totalcost`. The estimate of slope means that when the number of ER visits increase 1, total cost will increase 25%. However, the adjusted R-squated is 0.08838 which is really close to 0 indicating the SLR is not the best model.

e)

Fit a multiple linear regression (MLR) with 'comp_bin' and 'ERvisits' as predictors.

i)

Test if 'comp_bin' is an effect modifier of the relationship between 'total cost' and 'ERvisits'. Comment.

```r
fit_MLR_interaction = lm(totalcost ~ comp_bin * ERvisits, data = heart_data_trans)
summary(fit_MLR_interaction)
```

```
##
## Call:
## lm(formula = totalcost ~ comp_bin * ERvisits, data = heart_data_trans)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.1126  -1.0605   0.0257   1.2181   4.4258
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        5.46319    0.11072  49.343  < 2e-16 ***
## comp_bin1          2.21549    0.58466   3.789 0.000163 ***
## ERvisits           0.20884    0.02626   7.954 6.32e-15 ***
## comp_bin1:ERvisits -0.09686    0.10154  -0.954 0.340430
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.855 on 784 degrees of freedom
## Multiple R-squared:  0.1291, Adjusted R-squared:  0.1258
## F-statistic: 38.73 on 3 and 784 DF,  p-value: < 2.2e-16
```

The definition of modifier is when the magnitude of association differs at different levels of another variable (in this case `comp_bin`), it suggests that effect modification is present. From the result shown above, `comp_bin` is not a modifier according to the p-value of comp_bin1:ERvisits is way larger than 0.05.

ii)

Test if 'comp_bin' is a confounder of the relationship between 'total cost' and 'ERvisits'. Comment.

```r
lm(totalcost ~ ERvisits, data = heart_data_trans) %>%
  summary()
```

```
##
## Call:
## lm(formula = totalcost ~ ERvisits, data = heart_data_trans)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.2321  -1.1013   0.0529   1.3055   4.3224
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.5016     0.1106  49.725   <2e-16 ***
## ERvisits      0.2251     0.0256   8.792   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.894 on 786 degrees of freedom
## Multiple R-squared:  0.08954,    Adjusted R-squared:  0.08838
## F-statistic:  77.3 on 1 and 786 DF,  p-value: < 2.2e-16
```

```r
#coefficient estimate:
#ERvisits: 0.2251


lm(totalcost ~ comp_bin + ERvisits, data = heart_data_trans) %>%
  summary()
```

```
##
## Call:
## lm(formula = totalcost ~ comp_bin + ERvisits, data = heart_data_trans)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.1017  -1.0561   0.0165   1.2104   4.4301
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.48475    0.10838  50.607  < 2e-16 ***
## comp_bin1    1.73361    0.29432   5.890 5.72e-09 ***
## ERvisits     0.20236    0.02536   7.979 5.23e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.855 on 785 degrees of freedom
## Multiple R-squared:  0.1281, Adjusted R-squared:  0.1259
## F-statistic: 57.65 on 2 and 785 DF,  p-value: < 2.2e-16
```

```r
#coefficient estimate:
#ERvisits: 0.20236
```

To calculate the percentage change in the parameter estimate, we use the following formula:

$$\frac{|\beta_{crude} - \beta_{adjusted}|}{|\beta_{crude}|} = \frac{|0.2251 - 0.20236|}{|0.2251|} = 0.1010218$$

Here we use 10% rule of thumb. 0.1010218 is greater than 10%, so we consider `comp_bin` as a confounder.

iii)

Decide if 'comp_bin' should be included along with 'ERvisits'. Why or why not?

Hypothese:

$$Model\ 1:\ Y_i = \beta_0\ +\ \beta_1 X_{i\ ERvisits}\ +\ \varepsilon_i$$
$$Model\ 2:\ Y_i = \beta_0\ +\ \beta_1 X_{i\ ERvisits}\ +\ \beta_2 X_{i\ comp\_bin}\ +\ \varepsilon_i$$

$$H_0: \beta_2 = 0$$
$$H_1: \beta_2 \neq 0$$

Decision rule:

$$F^* = \frac{(SSE_S - SSE_L)/(df_L - df_S)}{\frac{SSE_L}{df_L}} \sim F_{df_L - df_S,\ df_L}$$
$$df_S = n - p_S - 1,\ df_L = n - p_L - 1$$

$$F^* = 34.694$$

$$Pr(F > F^*) < 0.05,\ reject\ H_0$$
$$Pr(F > F^*) \geq 0.05,\ fail\ to\ reject\ H_0$$

$$Pr(F > F^*) = 5.721e - 09 < 0.05,\ reject\ H_0$$

```
fit_without_comp = lm(totalcost ~ ERvisits, data = heart_data_trans)
fit_with_comp = lm(totalcost ~ ERvisits + comp_bin, data = heart_data_trans)

anova(fit_without_comp, fit_with_comp)
```

```
## Analysis of Variance Table
##
## Model 1: totalcost ~ ERvisits
## Model 2: totalcost ~ ERvisits + comp_bin
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    786 2820.0
## 2    785 2700.6  1    119.36 34.694 5.721e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

According to the result of partial F-test, the larger model including `comp_bin` is prefered. Besides, we already proof that `comp_bin` is a confounder so it should be included in the model along with ERvisits.

f)

Use your choice of model in part e) and add additional covariates (age, gender, and duration of treatment).

i)

Fit a MLR, show the regression results and comment.

Regression model in e):

```r
lm(totalcost ~ comp_bin + ERvisits, data = heart_data_trans) %>%
  summary()
```

```
##
## Call:
## lm(formula = totalcost ~ comp_bin + ERvisits, data = heart_data_trans)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.1017  -1.0561   0.0165   1.2104   4.4301
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.48475    0.10838  50.607  < 2e-16 ***
## comp_bin1    1.73361    0.29432   5.890 5.72e-09 ***
## ERvisits     0.20236    0.02536   7.979 5.23e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.855 on 785 degrees of freedom
## Multiple R-squared:  0.1281, Adjusted R-squared:  0.1259
## F-statistic: 57.65 on 2 and 785 DF,  p-value: < 2.2e-16
```

```r
lm(totalcost ~ comp_bin + ERvisits + age + gender + duration, data = heart_data_trans) %>%
  summary()
```

```
##
## Call:
## lm(formula = totalcost ~ comp_bin + ERvisits + age + gender +
##     duration, data = heart_data_trans)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.9436 -1.0080 -0.0886  0.9771  4.3492
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.8469672  0.5387065  10.854  < 2e-16 ***
## comp_bin1    1.5258252  0.2728204   5.593 3.09e-08 ***
## ERvisits     0.1736943  0.0238252   7.290 7.58e-13 ***
## age         -0.0198581  0.0091556  -2.169   0.0304 *
## gender      -0.2848042  0.1463906  -1.946   0.0521 .
## duration     0.0059649  0.0005159  11.561  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 1.714 on 782 degrees of freedom
## Multiple R-squared:  0.2583, Adjusted R-squared:  0.2536
## F-statistic: 54.47 on 5 and 782 DF,  p-value: < 2.2e-16
```

Comment:

The result shows that there is a relationship between Y and the set of covariates.

ii)

Compare the SLR and MLR models. Which model would you use to address the investigator's objective and why?

Hypothese:

$$\begin{aligned}
&Model\ 1:\ Y_i = \beta_0\ +\ \beta_1 X_{i\ ERvisits}\ +\ \varepsilon_i \\
&Model\ 2:\ Y_i = \beta_0\ +\ \beta_1 X_{i\ ERvisits}\ +\ \beta_2 X_{i\ comp_bin}\ +\ \beta_3 X_{i\ gender}\ +\ \beta_4 X_{i\ age}\ +\ \varepsilon_i
\end{aligned}$$

$$\begin{aligned}
&H_0: \beta_2 = \beta_3 = \beta_4 = 0 \\
&H_1: at\ least\ one\ \beta\ not\ equal\ to\ zero
\end{aligned}$$

Decision rule:

$$F^* = \frac{(SSE_S - SSE_L)/(df_L - df_S)}{\frac{SSE}{df_L}} \sim F_{df_L - df_S,\ df_L}$$
$$df_S = n - p_S - 1,\ df_L = n - p_L - 1$$

$$F^* = 44.49$$

$$\begin{aligned}
&Pr(F > F^*) < 0.05,\ reject\ H_0 \\
&Pr(F > F^*) \geq 0.05,\ fail\ to\ reject\ H_0
\end{aligned}$$

$$Pr(F > F^*) = 2.2e - 16 < 0.05,\ reject\ H_0$$

```r
fit_SLR = lm(totalcost ~ ERvisits, data = heart_data_trans)
fit_MLR = lm(totalcost ~ ERvisits + comp_bin + age + gender + duration, data = heart_data_trans)
anova(fit_SLR, fit_MLR)
```

```
## Analysis of Variance Table
##
## Model 1: totalcost ~ ERvisits
## Model 2: totalcost ~ ERvisits + comp_bin + age + gender + duration
##   Res.Df    RSS Df Sum of Sq     F    Pr(>F)
## 1    786 2820.0
## 2    782 2297.2  4    522.78 44.49 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusion:

Given the result of anova test, it's obviously that the MLR is prefered. So, I would choose MLR to address the investigator's objective.

Problem 3

A hospital administrator wishes to test the relationship between 'patient's satisfaction' (Y) and 'age', 'severity of illness', and 'anxiety level' (data 'PatSatisfaction.xlsx'). The administrator randomly selected 46 patients, collected the data, and asked for your help with the analysis.

```
sat_data = read_excel("./data/PatSatisfaction.xlsx")

colnames(sat_data)[1] <- "satisfaction"

sat_data = sat_data %>%
  janitor::clean_names()
```

a)

Create a correlation matrix and interpret your initial findings.

```
cor(sat_data, method = "pearson")
```

```
##              satisfaction        age   severity    anxiety
## satisfaction    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
```

The result is a table containing the correlation coefficients between each variable and the others. We can observe that age, severity and axiety have negative relationship with satisfaction. Among those three variables, age has the strongest negative relationship with satisfaction.

b)

Fit a multiple regression model and test whether there is a regression relation. State the hypotheses, decision rule and conclusion.

To build a multiple regression model, we add age, severity and anxiety as predictors:

$$Y_i = \beta_0 \ + \ \beta_1 X_{i\ age} \ + \ \beta_2 X_{i\ anxiety} \ + \beta_3 X_{i\ severity} + \varepsilon_i$$

Hypotheses:

$$\begin{aligned} H_0 &: \ \beta_1 = \beta_2 = \beta_3 = 0 \\ H_1 &: \ at\ least\ one\ \beta\ is\ not\ zero \end{aligned}$$

Decision rule:

$$\begin{aligned} & Test\ statistic: \\ & F^* = \frac{MSR}{MSE} > F(1-\alpha; p, n-p-1),\ reject\ H_0. \\ \\ & The\ null\ model\ contains\ only\ the\ intercept: \\ & F^* = \frac{MSR}{MSE} \leqslant F(1-\alpha; p, n-p-1),\ fail\ to\ reject\ H_0 \end{aligned}$$

$$F^* = 30.05$$

$$\begin{aligned} Pr(F > F^*) < 0.05,\ reject\ H_0 \\ Pr(F > F^*) \geq 0.05,\ fail\ to\ reject\ H_0 \end{aligned}$$

$$Pr(F > F^*) = 1.542e - 10 < 0.05,\ reject\ H_0$$

```
sat_fit = lm(satisfaction ~ age + severity + anxiety, data = sat_data)
summary(sat_fit)
```

```
##
## Call:
## lm(formula = satisfaction ~ age + severity + anxiety, data = sat_data)
##
## 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
```

Judging from the p-value, we reject the null, which means there is a relationship between predictors and outcome.

c)

Show the regression results for all estimated coefficients with 95% CIs. Interpret the coefficient and 95% CI associated with 'severity of illness'.

CI:

$$\begin{aligned} &E(Y_h) = \beta_0 + \beta_1 X_h \\ &\hat{Y_h} = \hat{\beta_0} + \hat{\beta_1} X_h \\ &\hat{\beta_0} + \hat{\beta_1} X_h \pm t_{n-2,1-\alpha/2} \times se(\hat{\beta_0} + \hat{\beta_1} X_h) \\ &se(\hat{\beta_0} + \hat{\beta_1} X_h) = \sqrt{MSE \left\{ \frac{1}{n} + [(X_h - \bar{X})^2 / \sum_{i=1}^{n} (X_i - \bar{X})^2] \right\}} \end{aligned}$$

```
sat_fit = lm(satisfaction ~ age + severity + anxiety, data = sat_data)
summary(sat_fit)
```

```
##
## Call:
## lm(formula = satisfaction ~ age + severity + anxiety, data = sat_data)
##
## 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
confint(sat_fit)
```

```
##                  2.5 %      97.5 %
## (Intercept) 121.911727 195.0707761
## age          -1.575093  -0.7081303
## severity     -1.434831   0.5508228
## anxiety     -27.797859   0.8575324
```

Interpretation:

The coefficient associated with `severity` means that while other predictors hold constant the satisfaction will decrease 0.44 for each additional unit of the severity of the illness.

The function of `confint` shows the CIs for each estimated coefficients. From the result we can conclude that we are 95% confidence that the mean satisfaction increases by somewhere between -1.434831 and 0.5508228 for each additional unit of the severity of the illness as other predictors hold constant.

d)

Obtain an interval estimate for a new patient's satisfaction when Age=35, Severity=42, Anxiety=2.1. Interpret the interval.

For a given value of x, the interval estimate of the dependent variable y is called the prediction interval.

PI:

$$\hat{\beta}_0 + \hat{\beta}_1 X_h \pm t_{n-2,1-\alpha/2} \times se(\hat{\beta}_0 + \hat{\beta}_1 X_h)$$
$$se(\hat{\beta}_0 + \hat{\beta}_1 X_h) = \sqrt{MSE\left\{\frac{1}{n} + [(X_h - \bar{X})^2 / \sum_{i=1}^{n}(X_i - \bar{X})^2] + 1\right\}}$$

```r
pi_data = data.frame(age = 35, severity = 42, anxiety = 2.1)
```

```r
predict(sat_fit, pi_data, interval="predict")
```

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

The 95% prediction interval of the satisfaction for the age is 35, severity is 42 and anxiety is 2.1 is between 50.06237 and 93.30426. The result means that the probability is 0.95 that this prediction interval will give a correct prediction for the satisfaction when age is 35, severity is 42 and anxiety is 2.1.

e)

Test whether 'anxiety level' can be dropped from the regression model, given the other two covariates are retained. State the hypotheses, decision rule and conclusion.

Hypothese:

$$\begin{aligned} Model\ 1: \ Y_i &= \beta_0 \ + \ \beta_1 X_{i\ age} \ + \ \beta_2 X_{i\ severity} \ + \ \varepsilon_i \\ Model\ 2: \ Y_i &= \beta_0 \ + \ \beta_1 X_{i\ age} \ + \ \beta_2 X_{i\ severity} \ + \ \beta_3 X_{i\ anxiety} \ + \ \varepsilon_i \end{aligned}$$

$$\begin{aligned} H_0 &: \beta_3 = 0 \\ H_1 &: \beta_3 \neq 0 \end{aligned}$$

Decision rule:

$$F^* = \frac{(SSE_S - SSE_L)/(df_L - df_S)}{\frac{SSE_L}{df_L}} \sim F_{df_L - df_S,\ df_L}$$

$$df_S = n - p_S - 1,\ df_L = n - p_L - 1$$

$$F^* = 3.5997$$

$$\begin{aligned} Pr(F > F^*) &< 0.05,\ reject\ H_0 \\ Pr(F > F^*) &\geq 0.05,\ fail\ to\ reject\ H_0 \end{aligned}$$

$$Pr(F > F^*) = 0.06468 > 0.05,\ fail\ to\ reject\ H_0$$

```r
anova(lm(satisfaction ~ age + severity, data = sat_data),
      lm(satisfaction ~ age + severity + anxiety, data = sat_data))
```

```
## Analysis of Variance Table
##
## Model 1: satisfaction ~ age + severity
## Model 2: satisfaction ~ age + severity + anxiety
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     43 4613.0
## 2     42 4248.8  1    364.16 3.5997 0.06468 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusion:

The result shows that the anxiety should NOT be included in the model due to the large p-value at 0.05 significance level and the desicion rule. So we tend to use the smaller model.