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

```r
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

```r
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

```r
(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

```r
(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

```r
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'
       )
```

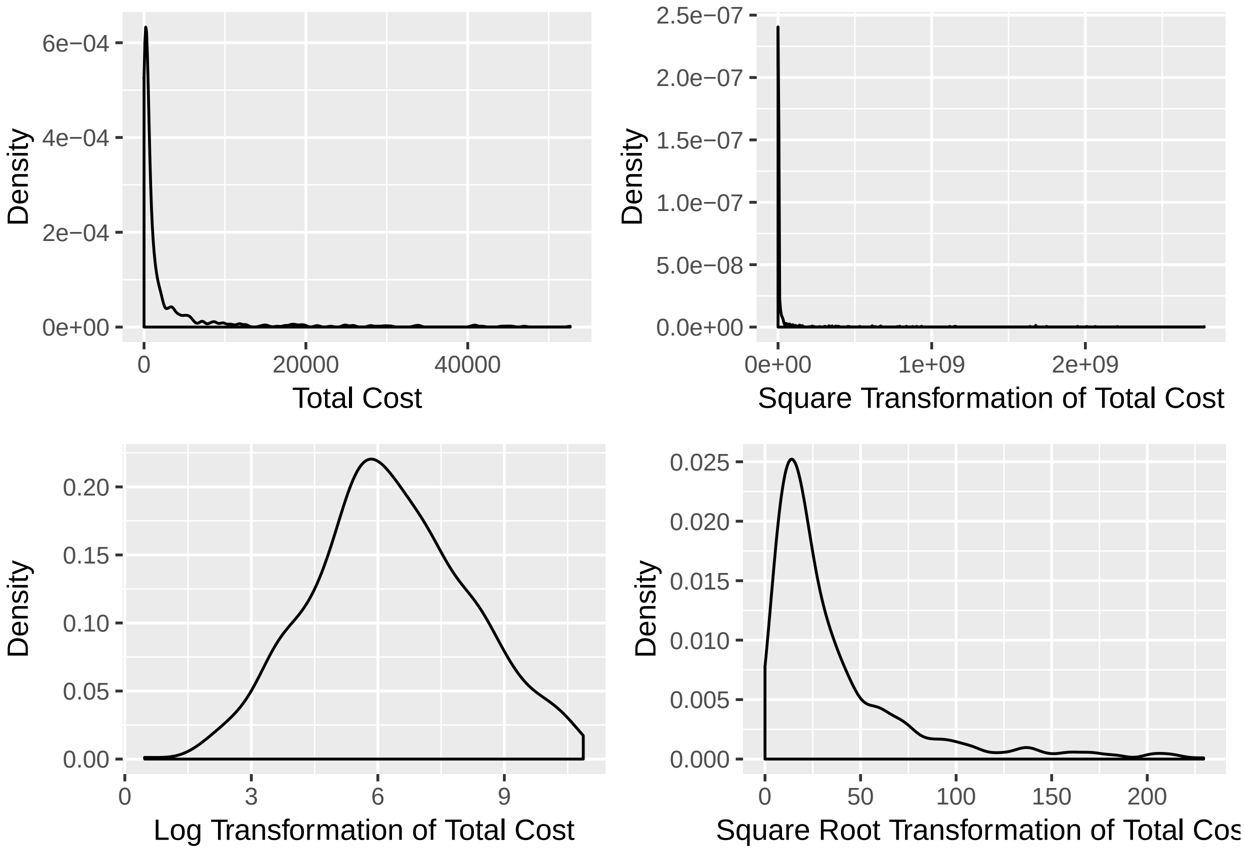
```r
totalcost_sq = heart_data %>%
  ggplot(aes(x = (totalcost)^2)) +
  geom_density() +
  labs(
       x = 'Square Transformation of Total Cost',
       y = 'Density'
       )
```

```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))
```

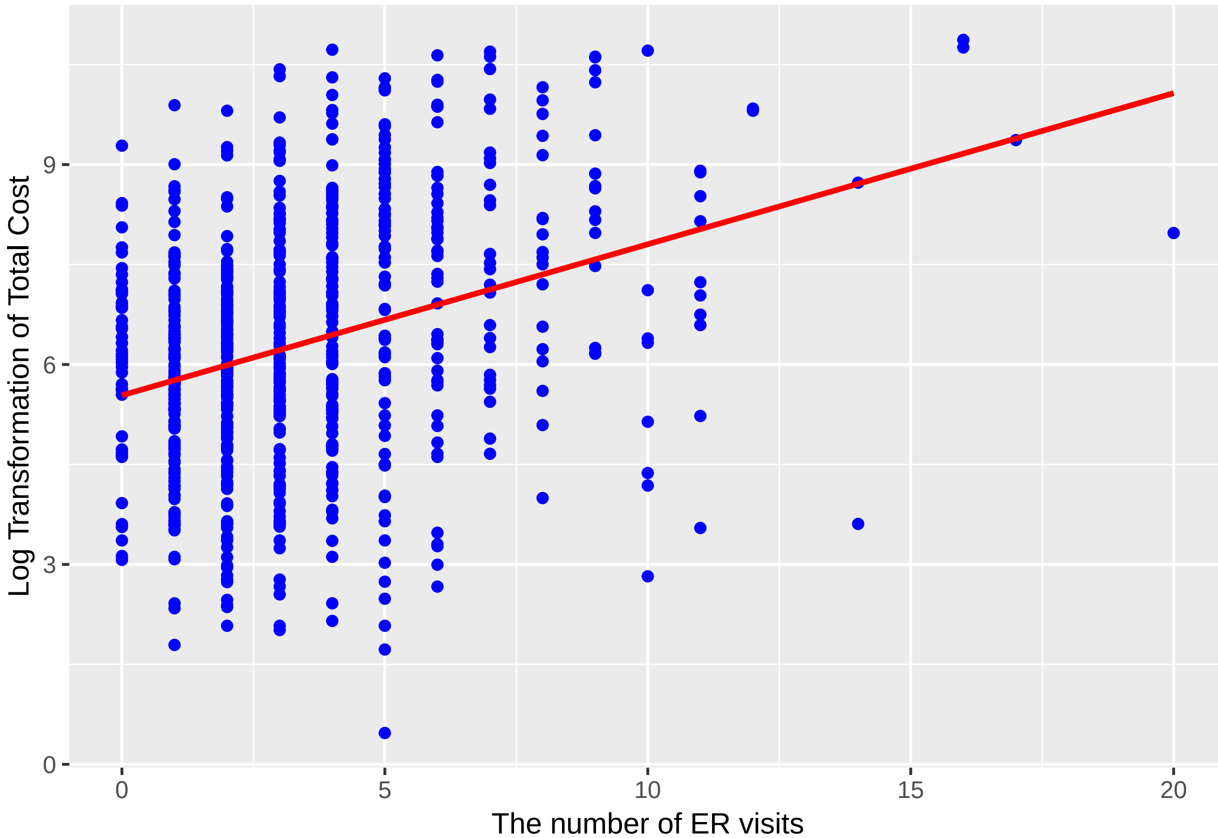
d)

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 %>%
  mutate(totalcost = log(totalcost)) %>%
  filter(totalcost != -Inf)
```

```
heart_data_trans %>%
  ggplot(aes(x = ERvisits, y = totalcost)) +
  geom_point(color = 'blue') +
  geom_smooth(method = "lm", color = 'red', se = FALSE) +
```

```r
  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
## -6.2013 -1.1265  0.0191  1.2668  4.2797
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.53771    0.10362   53.44   <2e-16 ***
## ERvisits     0.22672    0.02397    9.46   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.772 on 783 degrees of freedom
## Multiple R-squared:  0.1026, Adjusted R-squared:  0.1014
## F-statistic:  89.5 on 1 and 783 DF,  p-value: < 2.2e-16
```

The plot above shows the scatterplot and results of the regression. Using `summary` function, we can see that the estimate slope is 0.22672 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%.

$$log\left(\frac{Y_2}{Y_1}\right) = \beta_1 = 0.22672$$
$$\frac{Y_2}{Y_1} = e^{0.22672} = 1.25$$
$$Y_2 = 1.25Y_1$$

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.

```
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
## -6.0852 -1.0802 -0.0078  1.1898  4.3803
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        5.49899    0.10349  53.138  < 2e-16 ***
## comp_bin1          2.17969    0.54604   3.992 7.17e-05 ***
## ERvisits           0.21125    0.02453   8.610  < 2e-16 ***
## comp_bin1:ERvisits -0.09927    0.09483  -1.047    0.296
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.732 on 781 degrees of freedom
## Multiple R-squared:  0.1449, Adjusted R-squared:  0.1417
## F-statistic: 44.13 on 3 and 781 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 larger than 0.05.

ii)

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

```
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
## -6.0741 -1.0737 -0.0181  1.1810  4.3848
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.5211     0.1013  54.495  < 2e-16 ***
## comp_bin1     1.6859     0.2749   6.132 1.38e-09 ***
## ERvisits      0.2046     0.0237   8.633  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.732 on 782 degrees of freedom
## Multiple R-squared:  0.1437, Adjusted R-squared:  0.1416
## F-statistic: 65.64 on 2 and 782 DF,  p-value: < 2.2e-16
```

Using the summary function we can observe that after adding `comp_bin` as predictor the adjusted R-squared is increasing comparing with only using `ERvisits` as predictor. So `comp_bin` is a confounder.

iii)

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

`comp_bin` should be included along with `ERvisits` according to the test in ii). The p-value of comp_bin coefficient shows significance. Besides, judging from the adjusted R-squared, when including `comp_bin` the value increases comparing with only using `ERvisits` as predictor. So, `comp_bin` should be included 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. (5p)

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
## -6.0741 -1.0737 -0.0181  1.1810  4.3848
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.5211     0.1013  54.495  < 2e-16 ***
## comp_bin1     1.6859     0.2749   6.132 1.38e-09 ***
## ERvisits      0.2046     0.0237   8.633  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.732 on 782 degrees of freedom
## Multiple R-squared:  0.1437, Adjusted R-squared:  0.1416
## F-statistic: 65.64 on 2 and 782 DF,  p-value: < 2.2e-16
```

Add `age` as a covariate:

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

```
##
## Call:
## lm(formula = totalcost ~ comp_bin + ERvisits + age, data = heart_data_trans)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.0191 -1.0649  0.0098  1.2091  4.3112
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.062580   0.545373  11.116  < 2e-16 ***
## comp_bin1    1.672570   0.275253   6.076 1.92e-09 ***
## ERvisits     0.206229   0.023755   8.682  < 2e-16 ***
## age         -0.009302   0.009205  -1.010    0.313
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.732 on 781 degrees of freedom
## Multiple R-squared:  0.1449, Adjusted R-squared:  0.1416
## F-statistic:  44.1 on 3 and 781 DF,  p-value: < 2.2e-16
```

Judging from p-value and adjusted R-squared, `age` should NOT be included in the model.

Add `gender` as a covariate:

```r
heart_data_trans = heart_data_trans %>%
  mutate(gender = as.character(gender))

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

```
##
## Call:
## lm(formula = totalcost ~ comp_bin + ERvisits + gender, data = heart_data_trans)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.0127 -1.0579 -0.0132  1.2074  4.3654
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  5.53470    0.10431  53.060  < 2e-16 ***
## comp_bin1    1.68596    0.27506   6.129  1.4e-09 ***
## ERvisits     0.20604    0.02385   8.638  < 2e-16 ***
## gender1     -0.08222    0.14894  -0.552    0.581
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.733 on 781 degrees of freedom
## Multiple R-squared:  0.1441, Adjusted R-squared:  0.1408
## F-statistic: 43.82 on 3 and 781 DF,  p-value: < 2.2e-16
```

Judging from p-value and adjusted R-squared, `gender` should NOT be included in the model.

Add `duration` as a covariate:

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

```
##
## Call:
## lm(formula = totalcost ~ comp_bin + ERvisits + duration, data = heart_data_trans)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1450 -1.1008 -0.1479  0.9593  4.6166
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.7605975  0.1163153  40.928  < 2e-16 ***
## comp_bin1   1.5285357  0.2559535   5.972 3.56e-09 ***
## ERvisits    0.1708778  0.0222372   7.684 4.62e-14 ***
## duration    0.0053724  0.0004823  11.140  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.61 on 781 degrees of freedom
## Multiple R-squared:  0.2611, Adjusted R-squared:  0.2583
## F-statistic: 92.01 on 3 and 781 DF,  p-value: < 2.2e-16
```

Judging from p-value and adjusted R-squared, `duration` should be included in the model.

Test whether `duration` is an effect modifier:

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

```
##
## Call:
## lm(formula = totalcost ~ comp_bin * duration + ERvisits, data = heart_data_trans)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1300 -1.0881 -0.1405  0.9573  4.6307
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         4.7380358  0.1184976  39.984  < 2e-16 ***
```

```
## comp_bin1           1.9552348  0.4987652   3.920 9.63e-05 ***
## duration            0.0054804  0.0004943  11.087  < 2e-16 ***
## ERvisits            0.1723849  0.0222886   7.734 3.21e-14 ***
## comp_bin1:duration -0.0021385  0.0021454  -0.997    0.319
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.61 on 780 degrees of freedom
## Multiple R-squared:  0.2621, Adjusted R-squared:  0.2583
## F-statistic: 69.26 on 4 and 780 DF,  p-value: < 2.2e-16
```

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

```
##
## Call:
## lm(formula = totalcost ~ comp_bin + ERvisits * duration, data = heart_data_trans)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1592 -1.0697 -0.1308  0.9520  4.5454
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      4.9284866  0.1584830  31.098  < 2e-16 ***
## comp_bin1        1.4945266  0.2566500   5.823 8.43e-09 ***
## ERvisits         0.1157961  0.0417556   2.773  0.00568 **
## duration         0.0044308  0.0007729   5.732 1.42e-08 ***
## ERvisits:duration 0.0002911  0.0001868   1.558  0.11964
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.609 on 780 degrees of freedom
## Multiple R-squared:  0.2634, Adjusted R-squared:  0.2597
## F-statistic: 69.74 on 4 and 780 DF,  p-value: < 2.2e-16
```

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

```
##
## Call:
## lm(formula = totalcost ~ comp_bin * duration + ERvisits * duration,
##     data = heart_data_trans)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.2021 -1.0579 -0.1353  0.9294  4.5541
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         4.9218489  0.1585081  31.051  < 2e-16 ***
## comp_bin1           2.0389283  0.5004234   4.074 5.09e-05 ***
## duration            0.0044442  0.0007727   5.751 1.27e-08 ***
## ERvisits            0.1103900  0.0419569   2.631  0.00868 **
## comp_bin1:duration -0.0027511  0.0021713  -1.267  0.20551
```

```
## duration:ERvisits   0.0003299  0.0001893   1.743  0.08171 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.608 on 779 degrees of freedom
## Multiple R-squared:  0.2649, Adjusted R-squared:  0.2602
## F-statistic: 56.16 on 5 and 779 DF,  p-value: < 2.2e-16
```

Judging from the p-value and adjusted R-squared, `duration` is NOT an effective modifier.

Check:

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

```
## Analysis of Variance Table
##
## Response: totalcost
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## comp_bin    1  170.33  170.33  65.684 2.039e-15 ***
## ERvisits    1  223.68  223.68  86.257 < 2.2e-16 ***
## duration    1  321.78  321.78 124.089 < 2.2e-16 ***
## Residuals 781 2025.23    2.59
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In conclusion, the final MLR will include `comp_bin`, `duration` and `ERvisits`:

$$\hat{Y}_i = 1.53X_{i\ comp\_bin} + 0.17X_{i\ ERvisits} + 0.005X_{i\ duration} + 4.76$$

ii)

Compare the SLR and MLR models. Which model would you use to address the investigator's objective and why? (2p)

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

```
## Analysis of Variance Table
##
## Model 1: totalcost ~ ERvisits
## Model 2: totalcost ~ comp_bin + ERvisits + duration
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    783 2459.8
## 2    781 2025.2  2    434.62 83.803 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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 (15p)

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.

```r
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.

```r
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:

$$H_0: \ \beta_1 = \beta_2 = \beta_3 = 0$$
$$H_1: \ at\ least\ one\ \beta\ is\ not\ zero$$

Decision rule:

$$Test\ statistic:$$
$$F^* = \frac{MSR}{MSE} > F(1-\alpha; p, n-p-1),\ reject\ H_0.$$

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

```r
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
anova(sat_fit)
```

```
## Analysis of Variance Table
##
## Response: satisfaction
##           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
```

Judging from the p-value, we can see that at least one β is not zero, so we reject the null.

c)

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

d)

Obtain an interval estimate for a new patient's satisfaction when Age=35, Severity=42, Anxiety=2.1. Interpret the interval. (2p)

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. (3p)

The hypothese are as below:

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

$$H_0 : \beta_3 = 0$$
$$H_1 : \beta_3 \neq 0$$

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

```
## 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
```

The result shows that the anxiety should NOT be included in the model.