biostatistical methods homework 4

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

Problem1

(a)

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

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

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

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

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

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

(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)

$$
\begin{aligned}
log_e L &= -\frac{n}{2} log_e 2\pi - \frac{n}{2} log_e \sigma^2 - \frac{1}{2\sigma^2} \sum (Y_i - \beta_0 - \beta_1 X_i)^2 \\
\frac{\partial (log_e L)}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum (Y_i - \beta_0 - \beta_1 X_i)^2 \\
\hat{\sigma}^2 &= \frac{\sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2}{n} \\
&= \frac{\sum (Y_i - \hat{Y}_i)^2}{n}
\end{aligned}
$$

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’.

```r
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.

```r
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

```r
mean_and_sd(heart_data$totalcost) %>%
  knitr::kable()
```

mean	sd
2799.956	6690.26

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)

Investigate the shape of the distribution for variable 'total cost' and try different transformations, if needed.

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

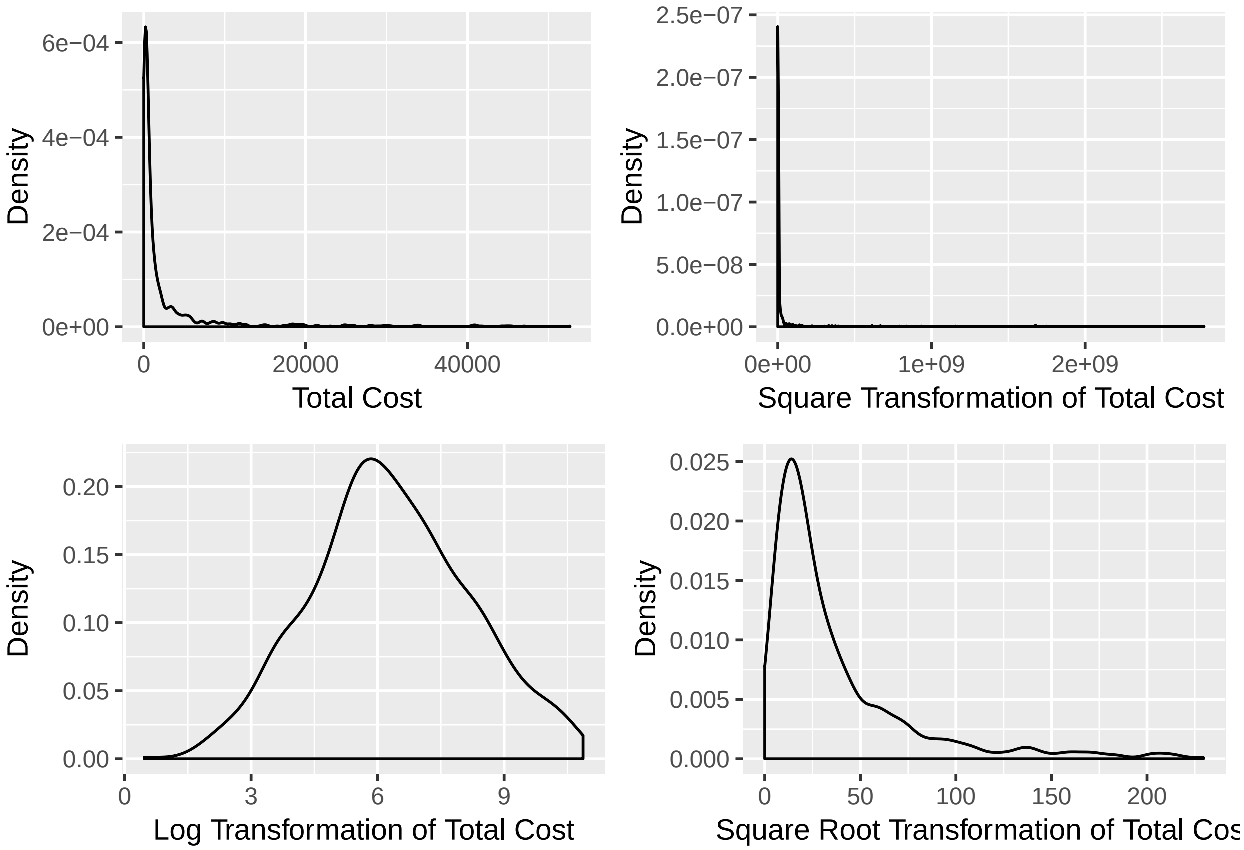
```
totalcost_sq = heart_data %>%
  ggplot(aes(x = (totalcost)^2)) +
  geom_density() +
  labs(
       x = 'Square Transformation of Total Cost',
       y = 'Density'
       )
```

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

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

```
(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))
```

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. (5p)

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

f) Test if 'comp_bin' is an effect modifier of the relationship between 'total cost' and 'ERvisits'. Comment. (2p)

ii)Test if 'comp_bin' is a confounder of the relationship between 'total cost' and 'ERvisits'. Comment. (2p)

iii) Decideif'comp_bin'shouldbeincludedalongwith'ERvisits.Whyorwhynot?(1p)

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

g) Fit a MLR, show the regression results and comment. (5p)

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

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.

a) Create a correlation matrix and interpret your initial findings. (2p)

b) Fit a multiple regression model and test whether there is a regression relation. State the hypotheses, decision rule and conclusion. (3p)

c) Show the regression results for all estimated coefficients with 95% CIs. Interpret the coefficient and 95% CI associated with 'severity of illness'. (5p)

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)