HW4

2022-11-12

```
library(tidyverse)
library(BSDA)
library(readxl)
library(readr)

knitr::opts_chunk$set(
    echo = TRUE,
    warning = FALSE)
```

Problem1

```
blood_data = c(125, 123, 117, 123, 115, 112, 128, 118, 124, 111, 116, 109, 125, 120, 113, 123, 112, 118
blood_test_data = blood_data - 120
test1 = SIGN.test(blood_test_data, alternative = "less", conf.level = 0.95)
```

a)

According to the sign-test above, we can see that the test statistic is **10** with p-value **0.2706281**. Thus we fail to reject the null hypothesis under a 0.05 significant level and claim that the median blood sugar readings was 120 in the population from which the 25 patients were selected.

```
test2 = wilcox.test(blood_test_data, alternative = "less", conf.level = 0.95)
```

b)

According to Wilcoxon signed-rank test, we can see that the test statistic is **112.5** with p-value **0.1446559**. Thus we fail to reject the null hypothesis under a 0.05 significant level and claim that the median blood sugar readings was 120 in the population from which the 25 patients were selected.

Problem 2

```
brain_data =
  read_xlsx("./Brain.xlsx")[-c(1),] %>%
  janitor::clean_names()

brain_fit =
  lm(glia_neuron_ratio ~ ln_brain_mass, data = brain_data)

brain_fit %>%
  broom::tidy()
```

```
## # A tibble: 2 x 5
##   term          estimate std.error statistic  p.value
##   <chr>            <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)      0.164    0.160       1.02 0.322
```

```
## 2 ln_brain_mass    0.181    0.0360      5.03 0.000151
```

a)

The linear model for the nonhuman data using ln (brain mass) as the predictor is : $\mathbf{y = 0.181x + 0.164}$.

```
y_0 = 0.181*7.22 + 0.164
```

b)

The predicted glia-neuron ratio for humans is **1.47082**.

c)

I assume the most plausible range of values for the prediction is an interval for the prediction of a single new observation.

d)

```
new.data = data.frame(ln_brain_mass = 7.22)
predict(brain_fit, newdata = new.data, interval = "prediction", level = 0.95)
```

```
##        fit      lwr      upr
## 1 1.471458 1.036047 1.906869
```

The 95% prediction interval for human glia-neuron ratio is **(1.04, 1.91)**. Based on this result, we can conclude that human brain doesn't have an excessive glia-neuron ratio for its mass compared with other primates.

e)

Considering the position of the human data point relative to those data used to generate the regression line, we are not certain that the regression line could be used to predict the glia_neuron ratio of humans, as this point falls beyond the range of the variable used to fit the line.

Problem 3

```
heart_data =
  read.csv("./HeartDisease.csv") %>%
  mutate(
    gender = as.factor(gender),
    gender = recode(gender, "0" = "otherwise", "1" = "male")
         )
```

a)

The main predictor of this dataset is **total cost**, The main outcome is **number of emergency room visits**. Other important covariates: **age, gender, complications, duration**. Here are some descriptions of the important variables:

```
summary(heart_data$totalcost)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   161.1   507.2  2800.0  1905.5 52664.9
```

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

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

```
summary(heart_data$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   24.00   55.00   60.00   58.72   64.00   70.00
```

```
summary(heart_data$complications)
```

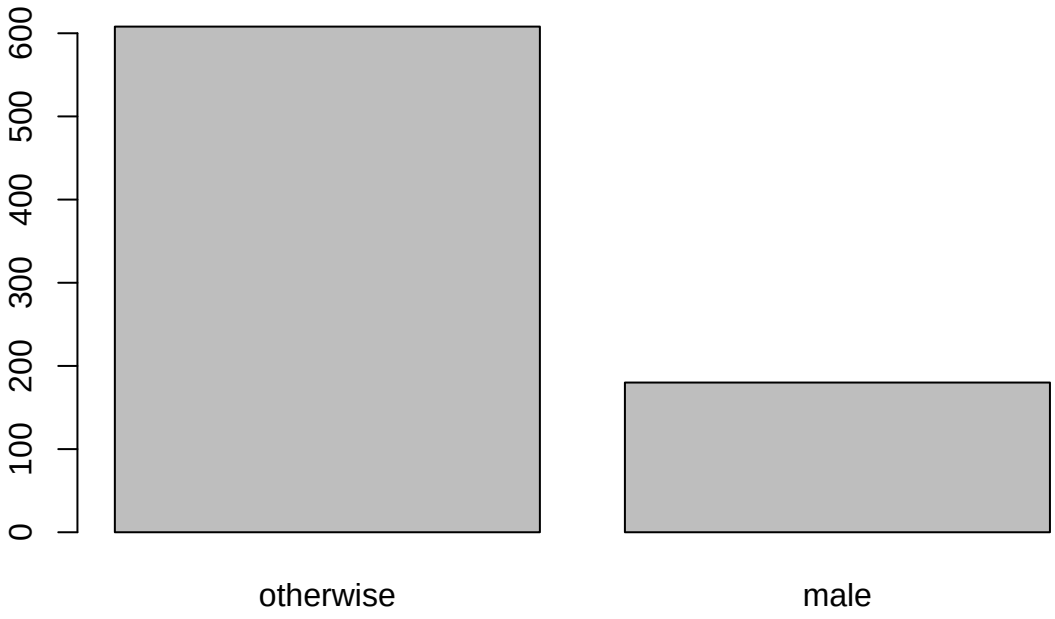
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00000 0.00000 0.00000 0.05711 0.00000 3.00000
```

```
summary(heart_data$duration)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   41.75  165.50  164.03  281.00  372.00
```

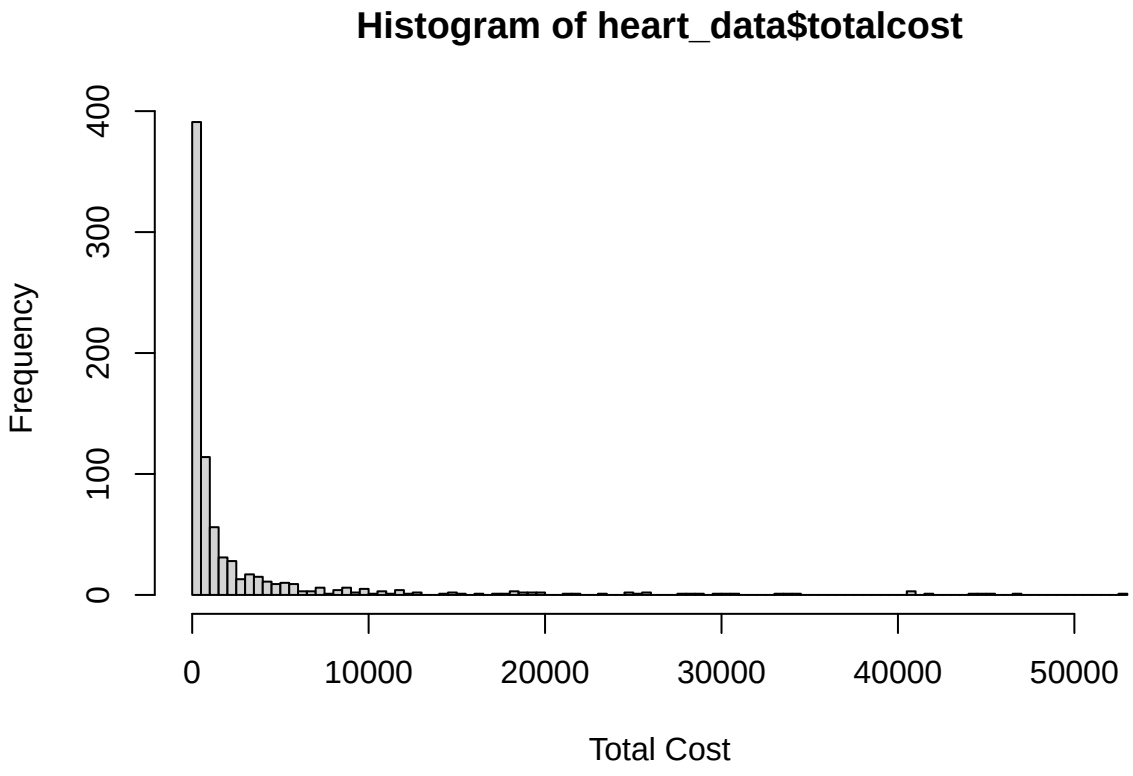
```
gender_sum =
  heart_data %>%
  group_by(gender) %>%
  summarise(count = n())

  barplot(height = gender_sum$count,
          names = gender_sum$gender)
```

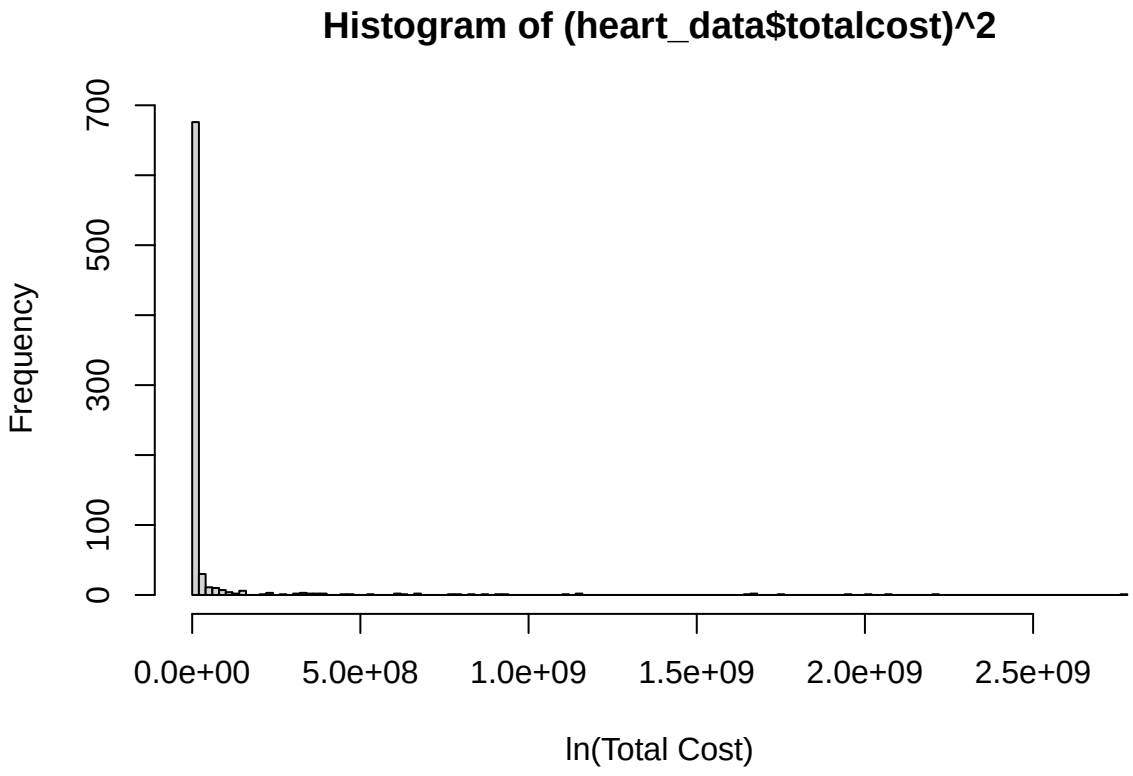


b)

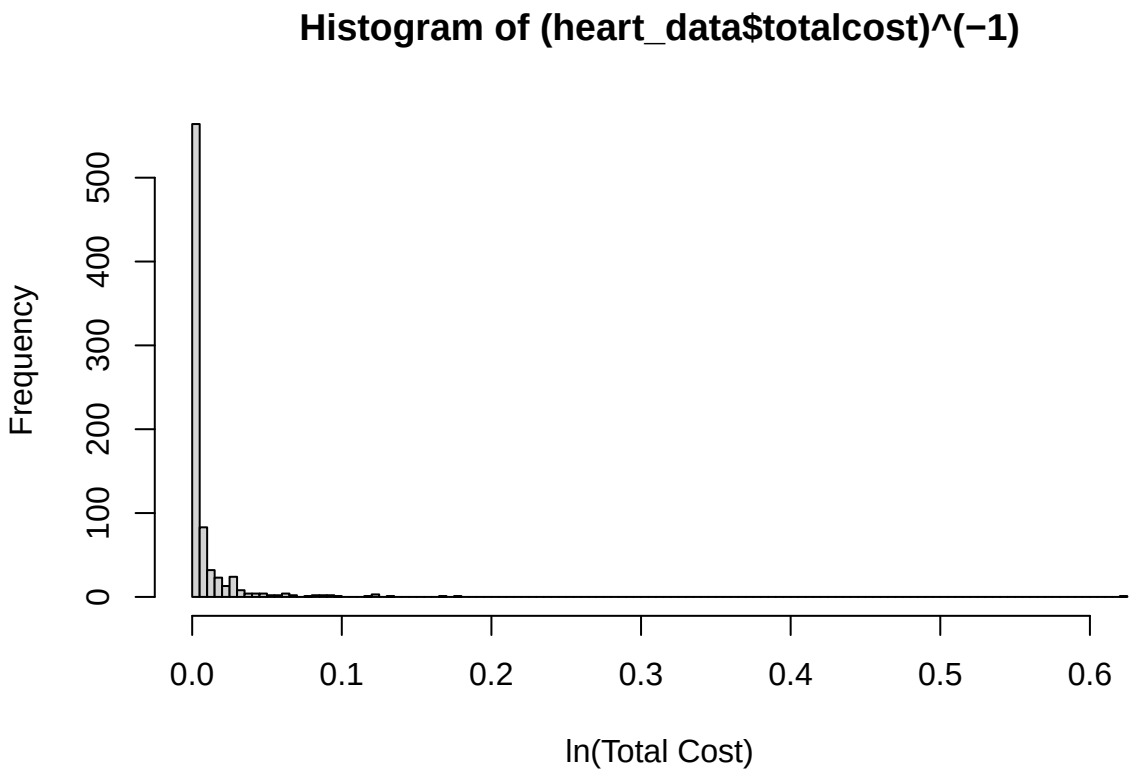
```
hist(heart_data$totalcost, xlab = "Total Cost", breaks = 100)
```



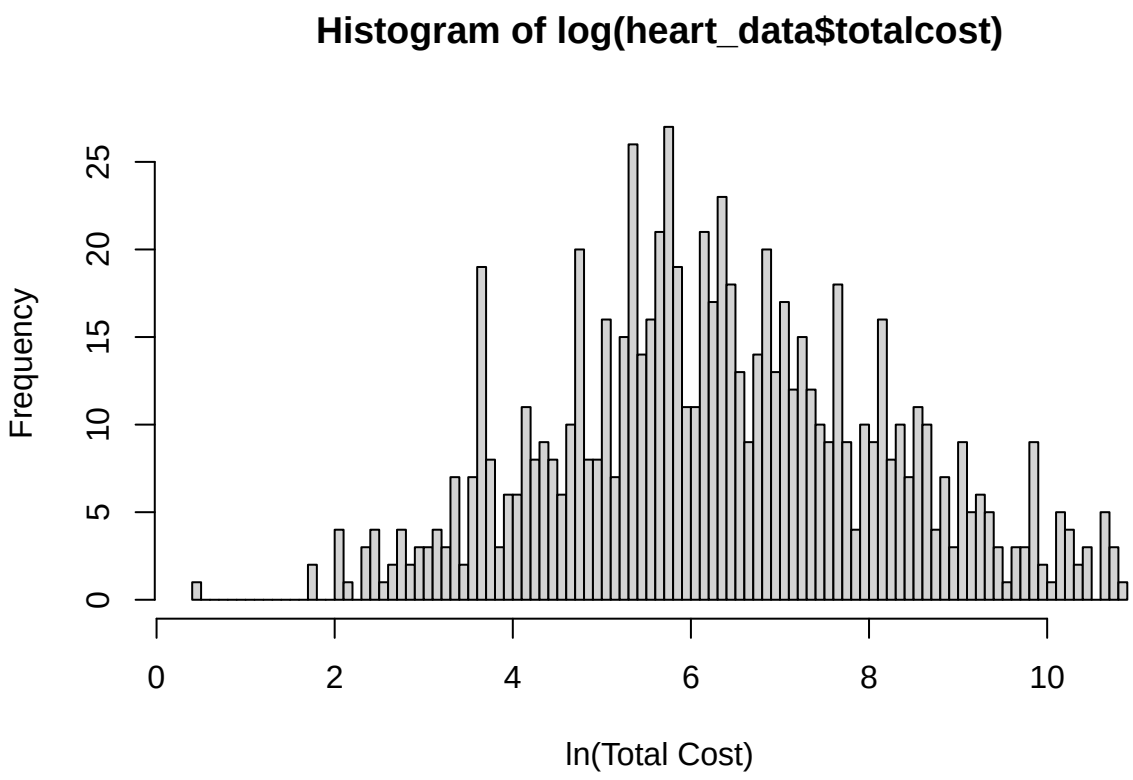

```
hist((heart_data$totalcost)^2, xlab = "ln(Total Cost)", breaks = 100)
```




```
hist((heart_data$totalcost)^(-1), xlab = "ln(Total Cost)", breaks = 100)
```




```
hist(log(heart_data$totalcost), xlab = "ln(Total Cost)", breaks = 100)
```



It seems that the plot best fits normality after ln-transformation.

```
heart_data =
  heart_data %>%
```

```
  filter(totalcost > 0) %>%
  mutate(ln_cost = log(totalcost))

shapiro.test(heart_data$ln_cost)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  heart_data$ln_cost
## W = 0.9952, p-value = 0.01488
```

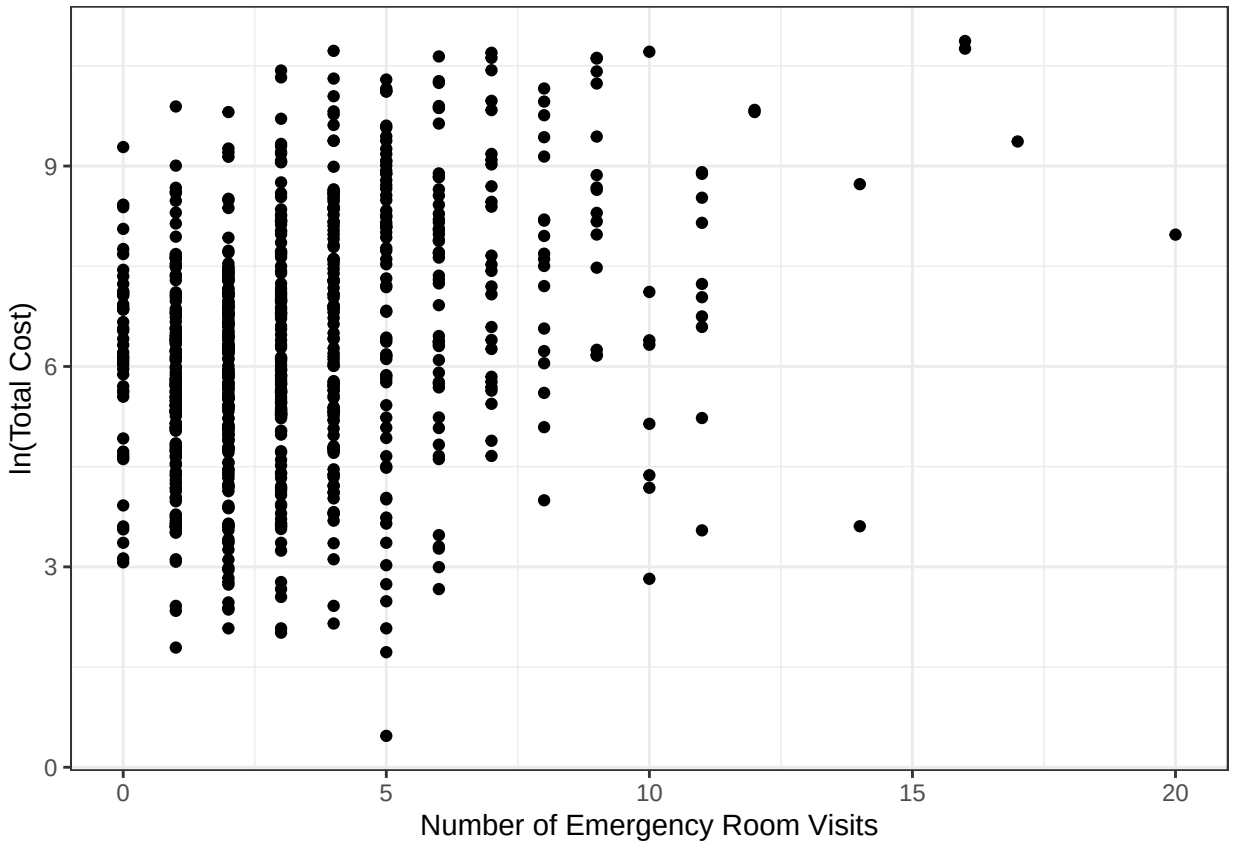
The Shapiro Test shows that total cost data doesn't follow normal distribution after ln-transformation.

c)

```
heart_data =
  heart_data %>%
  mutate(
    comp_bin =
      case_when(
        complications == 0 ~ "0",
        TRUE ~ "1"
      ))
```

d)

```
heart_data %>%
  ggplot()+
  geom_point(aes(x = ERvisits, y = ln_cost))+
  theme_bw()+
  labs(x = "Number of Emergency Room Visits",
       y = "ln(Total Cost)")
```




```
heart_fit =
  lm(ln_cost ~ ERvisits, data = heart_data)

summary(heart_fit)
```

```
##
## Call:
## lm(formula = ln_cost ~ ERvisits, data = heart_data)
##
## 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
```

```
t_cri = qt(p=.05/2, df=783, lower.tail=FALSE)
t_cri
```

```
## [1] 1.962998
```

The slope is 0.22672, at a 5% significance level, t > t783,0.975, we reject the null and conclude that there is a significant linear association between the number of Emergency room visits and ln(Total cost). Which also means that holding all other variable constantm, as the risk of ERvisits goes up by 1 percent point, the predicted ln(Total cost) will increase by approximately 0.22672 dollars.

e)

1)

```
fit_inter =
  lm(totalcost ~ ERvisits*comp_bin, data = heart_data)
summary(fit_inter)
```

```
##
## Call:
## lm(formula = totalcost ~ ERvisits * comp_bin, data = heart_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -14973  -2187   -973    247  42326
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -566.69     367.27  -1.543  0.12325
## ERvisits             922.13      87.07  10.590  < 2e-16 ***
## comp_bin1           5423.48    1937.91   2.799  0.00526 **
## ERvisits:comp_bin1  -277.03     336.56  -0.823  0.41069
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6148 on 781 degrees of freedom
## Multiple R-squared:  0.1614, Adjusted R-squared:  0.1582
## F-statistic:  50.1 on 3 and 781 DF,  p-value: < 2.2e-16
```

We can tell from the test above that comp_bin is not an effect modifier of the relationship between totalcost and ERvisit, as the p-value for the coefficient of ERvisits:comp_bin is not significant.

2)

```
fit_1 =
  lm(ln_cost ~ ERvisits, data = heart_data)
fit_2 =
  lm(ln_cost ~ ERvisits + comp_bin, data = heart_data)
fit_1$coefficients
```

```
## (Intercept)    ERvisits
##   5.5377096   0.2267218
```

```
fit_2$coefficients
```

```
## (Intercept)    ERvisits   comp_bin1
##   5.5210974   0.2046044   1.6858626
```

The coefficients of ERvisits in the regression model with or without comp_bin did not show much difference, indicating that comp_bin might no be considered a confounder of the relationship between totalcost and ERvisits.

3)

```
fit_2|>anova()
```

```
## Analysis of Variance Table
##
## Response: ln_cost
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## ERvisits    1  281.16 281.160  93.680 < 2.2e-16 ***
## comp_bin    1  112.84 112.842  37.598 1.379e-09 ***
## Residuals 782 2347.01   3.001
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the ANOVA test above we can tell that comp_bin should be included with ERvisits as the p-value for the coefficient of comp_bin is less than 0.05 after adding it as an additional variable to the regression model.

f)

1)

```
fit_more =
  lm(ln_cost ~ ERvisits + comp_bin + age + gender + duration, data = heart_data)
fit_more|>summary()
```

```
##
## Call:
## lm(formula = ln_cost ~ ERvisits + comp_bin + age + gender + duration,
##     data = heart_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0823 -1.0555 -0.1352  0.9533  4.3462
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.0449619  0.5063454  11.938  < 2e-16 ***
## ERvisits     0.1757486  0.0223189   7.874 1.15e-14 ***
## comp_bin1    1.4921110  0.2554883   5.840 7.65e-09 ***
## age         -0.0221376  0.0086023  -2.573   0.0103 *
## gendermale  -0.1176181  0.1379809  -0.852   0.3942
## duration     0.0055406  0.0004848  11.428  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.605 on 779 degrees of freedom
## Multiple R-squared:  0.268,  Adjusted R-squared:  0.2633
## F-statistic: 57.03 on 5 and 779 DF,  p-value: < 2.2e-16
```

```
fit_more|>anova()
```

```
## Analysis of Variance Table
##
## Response: ln_cost
##            Df  Sum Sq Mean Sq  F value    Pr(>F)
## ERvisits    1  281.16  281.16 109.1541 < 2.2e-16 ***
## comp_bin    1  112.84  112.84  43.8083 6.738e-11 ***
## age         1    3.06    3.06   1.1896    0.2757
```

```
## gender      1    0.99    0.99   0.3832    0.5361
## duration    1  336.40  336.40 130.6016 < 2.2e-16 ***
## Residuals 779 2006.55    2.58
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The fitted model is **ln(totalcost) = 6.0449619 + 0.1757486ERvisits + 1.4921110comp_bin + 0.0055406duration**. As the covariates **age and gender** didn't make any significant difference to the model under a 5% confidence level, they should not be included along with other variables.

2)

```
anova(fit_2, fit_more)
```

```
## Analysis of Variance Table
##
## Model 1: ln_cost ~ ERvisits + comp_bin
## Model 2: ln_cost ~ ERvisits + comp_bin + age + gender + duration
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    782 2347.0
## 2    779 2006.5  3    340.46 44.058 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

I would choose the MLR model, as the p-value of anova test is less than 0.05, we would reject the null hypotheses and conclude that the larger model is superior.