P8130hw4

Ze Li

2023-11-25

Problem 1

(a)

H0: The median blood sugar readings was equal to 120 in the population from which the 25 patients were selected

Ha:The median blood sugar readings was less than 120 in the population from which the 25 patients were selected.

```
data1 = c(125,123, 117, 123, 115, 112, 128, 118, 124, 111, 116, 109, 125, 120, 113, 123, 112, 118, 121,
data1
```

```
##  [1] 125 123 117 123 115 112 128 118 124 111 116 109 125 120 113 123 112 118 121
## [20] 118 122 115 105 118 131
```

```
medianvalue = 120
diff = data1-medianvalue
num_neg = sum(diff<0)
result1.1 <- binom.test(num_neg, length(diff), p = 0.5, alternative = "less")
result1.1
```

```
##
##  Exact binomial test
##
## data:  num_neg and length(diff)
## number of successes = 14, number of trials = 25, p-value = 0.7878
## alternative hypothesis: true probability of success is less than 0.5
## 95 percent confidence interval:
##  0.0000000 0.7301469
## sample estimates:
## probability of success
##                   0.56
```

The test statistics 0.56. Since the p-value is 0.7878, which is greater than 0.05, so we fail to reject the null hypothesis, indicating we have no evidence that the median blood sugar levels are less than 120.

(b)

```r
result1.2 <- wilcox.test(data1, mu = medianvalue, alternative = "less")
```

```
## Warning in wilcox.test.default(data1, mu = medianvalue, alternative = "less"):
## cannot compute exact p-value with ties
```

```
## Warning in wilcox.test.default(data1, mu = medianvalue, alternative = "less"):
## cannot compute exact p-value with zeroes
```

```r
result1.2
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  data1
## V = 112.5, p-value = 0.1447
## alternative hypothesis: true location is less than 120
```

The test statistics is 112.5. Since the p-value is 0.1447, which is greater than 0.05, so we fail to reject the null hypothesis, indicating we have no evidence that the median blood sugar levels are less than 120.

Problem 2

(a)

```r
data2 = read_excel("Brain.xlsx") |>
  janitor::clean_names()
data2_nohomo = data2 |>
  filter(species != "Homo sapiens")
model=lm(glia_neuron_ratio ~ ln_brain_mass,data2_nohomo)
summary(model)
```

```
##
## Call:
## lm(formula = glia_neuron_ratio ~ ln_brain_mass, data = data2_nohomo)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.24150 -0.12030 -0.01787  0.15940  0.25563
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.16370    0.15987   1.024 0.322093
## ln_brain_mass  0.18113    0.03604   5.026 0.000151 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1699 on 15 degrees of freedom
## Multiple R-squared:  0.6274, Adjusted R-squared:  0.6025
## F-statistic: 25.26 on 1 and 15 DF,  p-value: 0.0001507
```

(b)

```
predicted = model$coefficients[1]+7.22*model$coefficients[2]
predicted
```

```
## (Intercept)
##    1.471458
```

(c)

The interval for the prediction of a single new observation is more relevant for your prediction of human glia-neuron ratio than an interval for the predicted mean glia-neuron ratio at the given brain mass.

(d)

```
# Method1
prediction_interval = predict(model, newdata = data.frame(ln_brain_mass= 7.22), interval = "prediction"
prediction_interval
```

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

```
# Method2
tcrit = qt(df=15,0.975)
se=sqrt(0.1699)
lowerbound = predicted - tcrit*se
upperbound = predicted + tcrit*se
c(lowerbound,upperbound)
```

```
## (Intercept) (Intercept)
##   0.5928973   2.3500186
```

The 95% prediction interval for human glia-neuron ratio is (1.04, 1.91). So 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 can see that the point falls beyond the range of the variable used to fit the line, so we are not certain that the regression line could be used to predict the glia_neuron ratio of humans.

Problem 3

(a)

The main outcome is total cost (in dollars) of patients diagnosed with heart disease. The main predictor is number of emergency room (ER) visits. And other important covariates are age, gender, number of complications that arose during treatment, and duration of treatment condition.

```r
data3 = read_csv("HeartDisease.csv") |>
  janitor::clean_names() |>
  mutate(
    gender = as.factor(gender),
    gender = recode(gender, "0" = "female","1" = "male")
  )
```

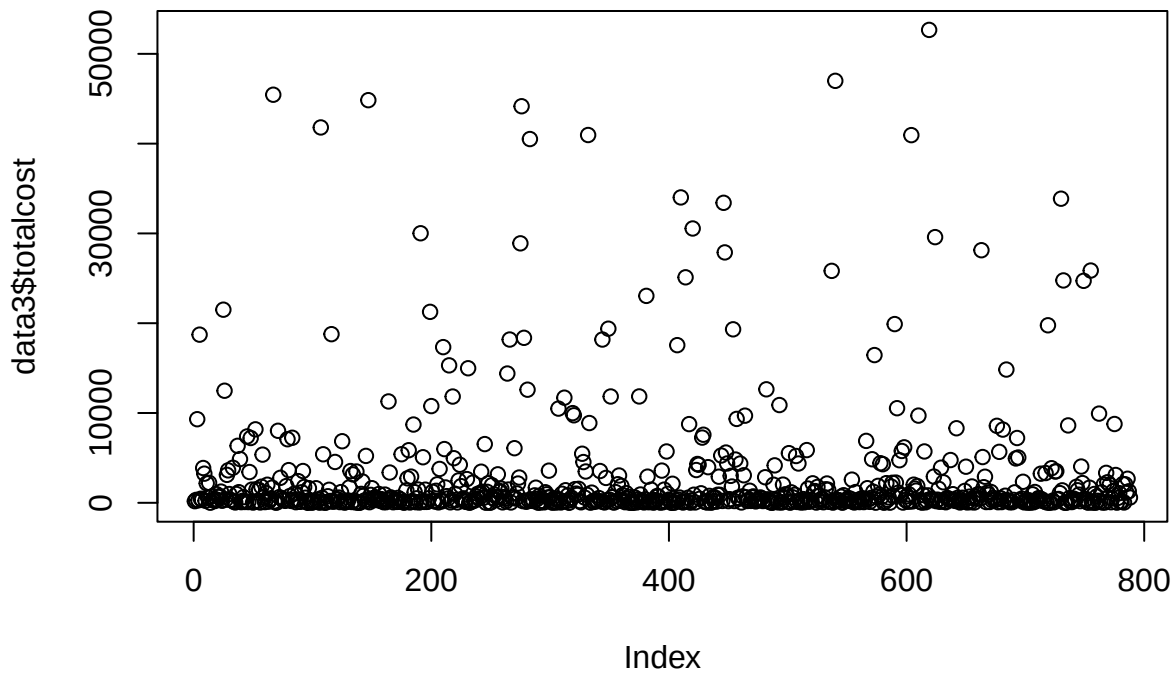
```
## Rows: 788 Columns: 10
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (10): id, totalcost, age, gender, interventions, drugs, ERvisits, compli...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
summary(data3)
```

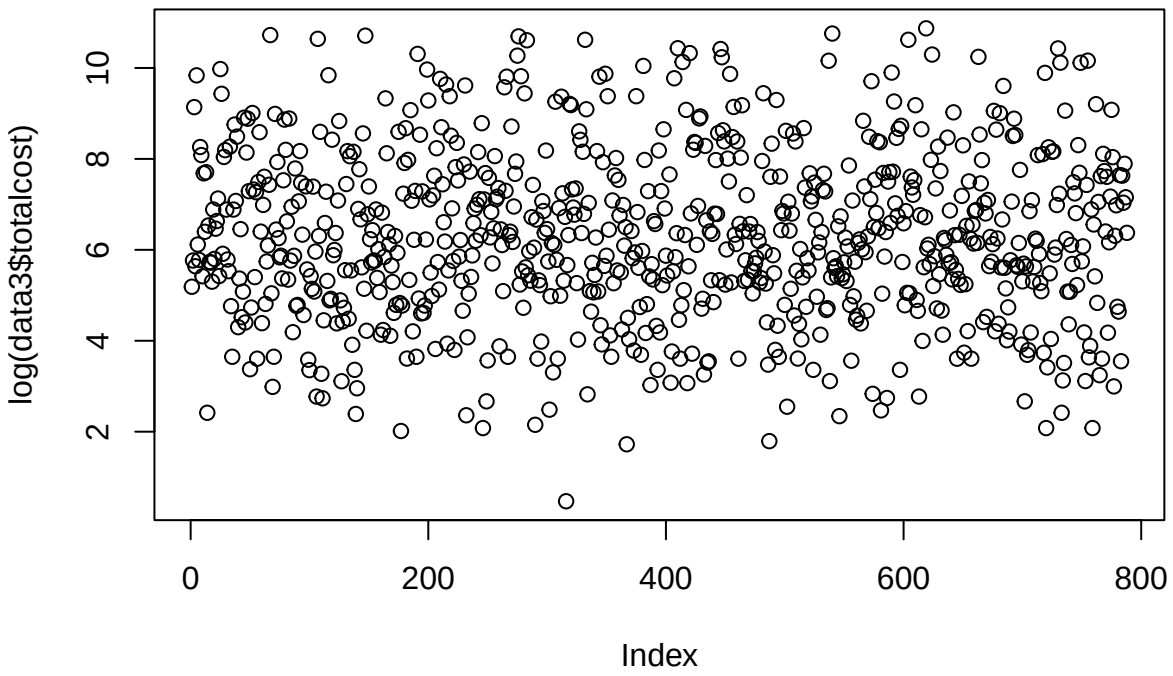
```
##        id          totalcost           age          gender
##  Min.   :  1.0   Min.   :    0.0   Min.   :24.00   female:608
##  1st Qu.:197.8   1st Qu.:  161.1   1st Qu.:55.00   male  :180
##  Median :394.5   Median :  507.2   Median :60.00
##  Mean   :394.5   Mean   : 2800.0   Mean   :58.72
##  3rd Qu.:591.2   3rd Qu.: 1905.5   3rd Qu.:64.00
##  Max.   :788.0   Max.   :52664.9   Max.   :70.00
##  interventions        drugs          e_rvisits      complications
##  Min.   : 0.000   Min.   :0.0000   Min.   : 0.000   Min.   :0.00000
##  1st Qu.: 1.000   1st Qu.:0.0000   1st Qu.: 2.000   1st Qu.:0.00000
##  Median : 3.000   Median :0.0000   Median : 3.000   Median :0.00000
##  Mean   : 4.707   Mean   :0.4467   Mean   : 3.425   Mean   :0.05711
##  3rd Qu.: 6.000   3rd Qu.:0.0000   3rd Qu.: 5.000   3rd Qu.:0.00000
##  Max.   :47.000   Max.   :9.0000   Max.   :20.000   Max.   :3.00000
##  comorbidities       duration
##  Min.   : 0.000   Min.   :  0.00
##  1st Qu.: 0.000   1st Qu.: 41.75
##  Median : 1.000   Median :165.50
##  Mean   : 3.767   Mean   :164.03
##  3rd Qu.: 5.000   3rd Qu.:281.00
##  Max.   :60.000   Max.   :372.00
```

(b)

```r
plot(data3$totalcost)
```

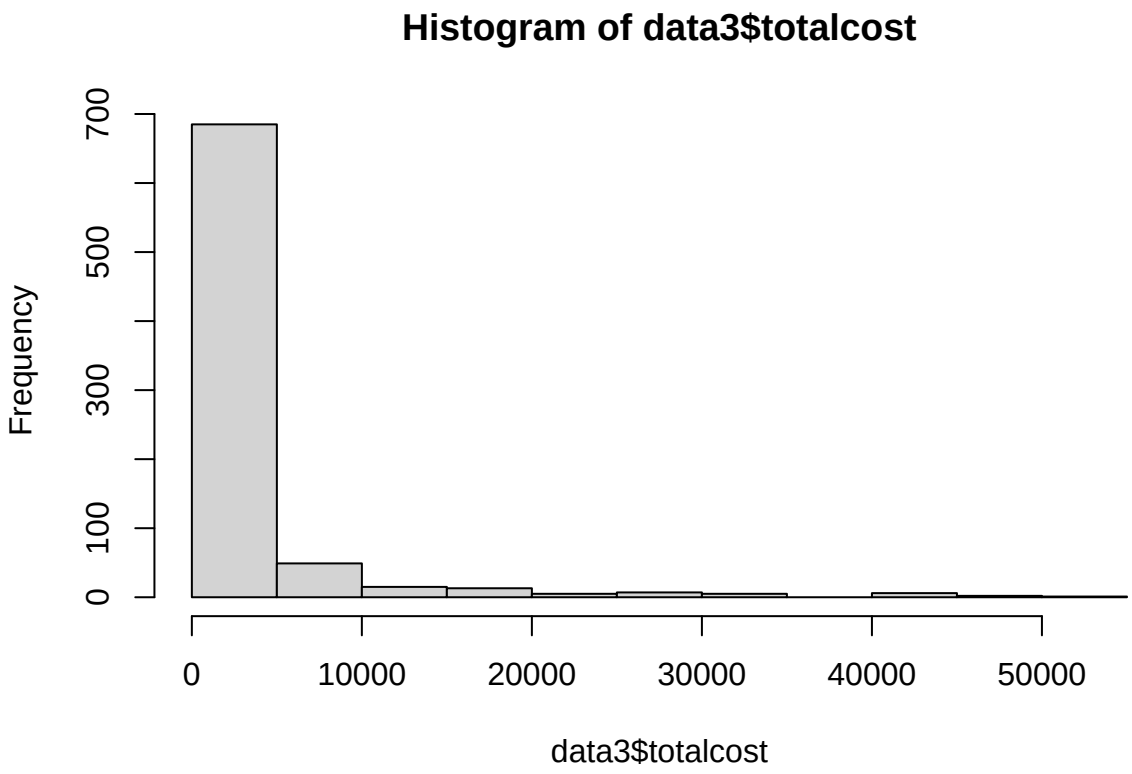



```r
plot(log(data3$totalcost))
```

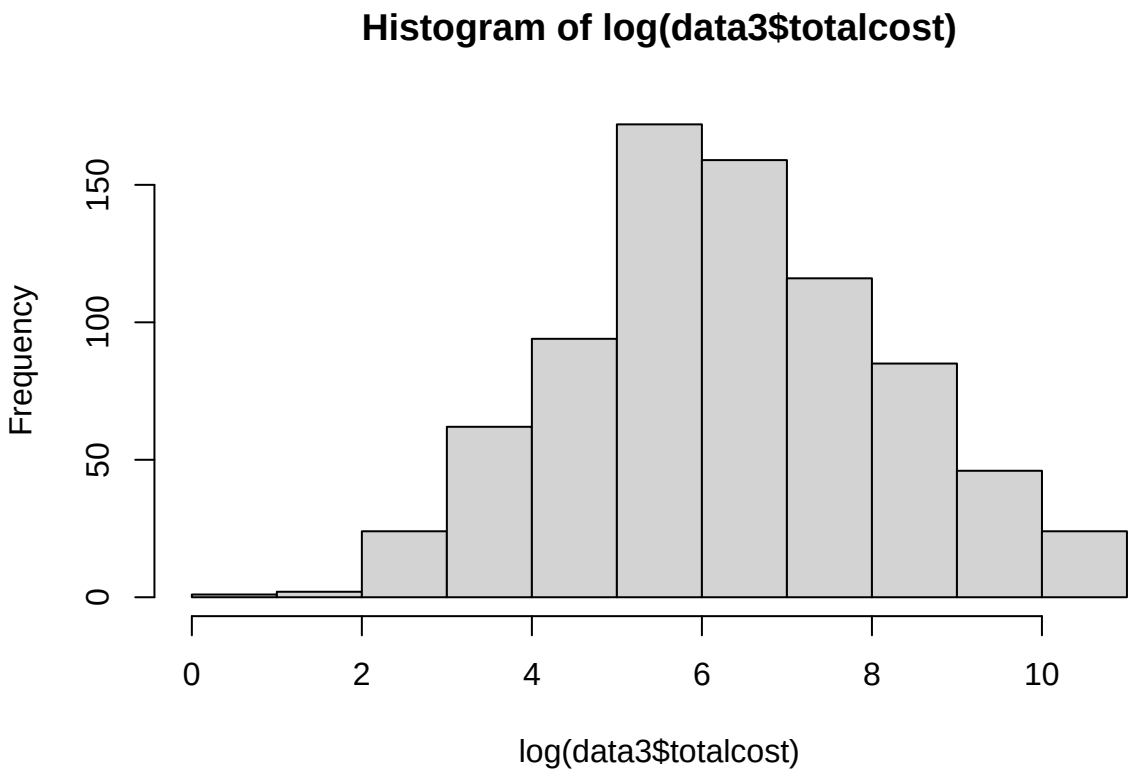


It seems that after log transformation, the plot is approximately to normality since the points are randomly distributed.

```
hist(data3$totalcost)
```




```
hist(log(data3$totalcost))
```



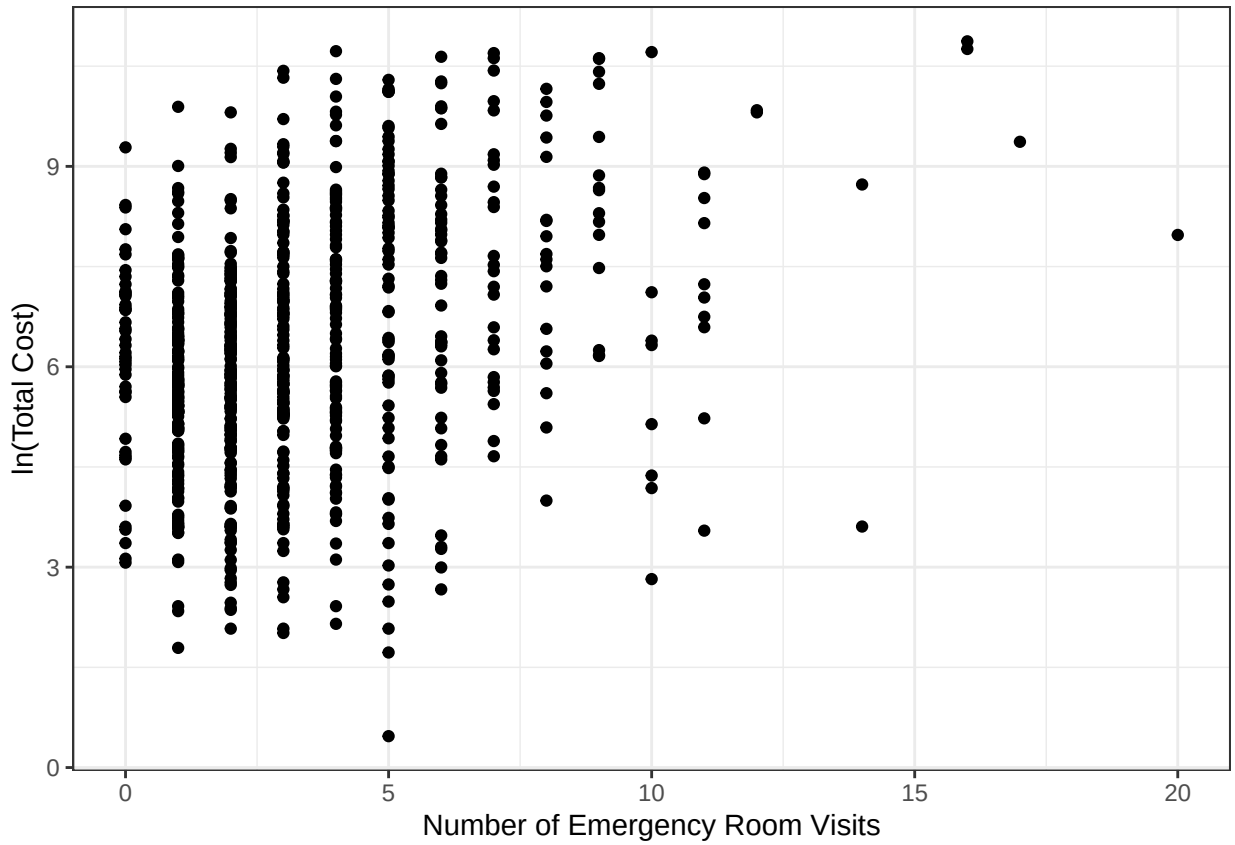
We can also see from the histogram plot that log transformation helps improve normality.

(c)

```
heart_data =
  data3 |>
  mutate(
    comp_bin =
      case_when(
        complications == 0 ~ "0",
        TRUE ~ "1"
      )) |>
  filter(totalcost > 0) |>
  mutate(ln_cost = log(totalcost))
```

(d)

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




```
model3 = lm(ln_cost~e_rvisits,heart_data)
summary(model3)
```

```
##
## Call:
## lm(formula = ln_cost ~ e_rvisits, 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 ***
## e_rvisits    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_crit = qt(p=.05/2, df=783, lower.tail=FALSE)
t_crit
```

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

The slope is 0.22672, at a 5% significance level, t > t_783,0.975, we reject the null and conclude that there is a significant linear association between the number of Emergency room visits and ln(Total cost).

It means that holding all other variable constant, as the risk of ERvisits goes up by 1 percent point, the predicted ln(Total cost) will increase by approximately 0.22672 dollars.

(e)

(i) Test if comp_bin is an effect modifier of the relationship between totalcost and ERvisits. Comment.

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

```
##
## Call:
## lm(formula = totalcost ~ e_rvisits * 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
## e_rvisits             922.13      87.07  10.590  < 2e-16 ***
## comp_bin1            5423.48    1937.91   2.799  0.00526 **
## e_rvisits: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
```

The comp_bin is not an effect modifier of the relationship between totalcost and ERvisit, since the p-value for the coefficient of e_rvisits*comp_bin is not significant.

(ii) Test if comp_bin is a confounder of the relationship between totalcost and ERvisits. Comment.

```
fit1 = lm(ln_cost ~ e_rvisits, data = heart_data)
fit2 = lm(ln_cost ~ e_rvisits + comp_bin, data = heart_data)
fit1$coefficients
```

```
## (Intercept)   e_rvisits
##   5.5377096   0.2267218
```

```
fit2$coefficients
```

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

The coefficients of e_rvisits in the regression model with or without comp_bin did not show much difference, showing that comp_bin might not be considered a confounder of the relationship between totalcost and ERvisits.

(iii) Decide if comp_bin should be included along with ERvisits. Why or why not?

```
fit2 |>
  anova()
```

```
## Analysis of Variance Table
##
## Response: ln_cost
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## e_rvisits   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, the comp_bin should be included with ERvisits as the p-value for the coefficient of comp_bin is less than 0.05 in this model.

(f)

(i) Fit a MLR, show the regression results and comment.

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

```
##
## Call:
## lm(formula = ln_cost ~ e_rvisits + 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 ***
## e_rvisits    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
```

```r
fit_mlr |>
  anova()
```

```
## Analysis of Variance Table
##
## Response: ln_cost
##            Df  Sum Sq Mean Sq  F value    Pr(>F)
## e_rvisits   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 + 0.1757486 ERvisits + 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.

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

```r
anova(fit2,fit_mlr)
```

```
## Analysis of Variance Table
##
## Model 1: ln_cost ~ e_rvisits + comp_bin
## Model 2: ln_cost ~ e_rvisits + 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
```

Since the p-value of anova test is less than 0.05, we reject the null hypotheses and conclude that the larger model is superior. As a result, we will choose MLR models.