Homework 4 SOLUTIONS

P8130 Fall 2022

Due: November 13, 2022 at midnight Eastern

P8130 Guidelines for Submitting Homework

- Your homework must be submitted through Courseworks. No email submissions!
- Only one PDF file should be submitted, including all derivations, graphs, output, and interpretations. When handwriting is allowed (this will be specified), scan the derivations and merge ALL PDF files (http: //www.pdfmerge.com/).
- You are encouraged to use R for calculations, but you must show all mathematical formulas and derivations. Please include the important parts of your R code in the PDF file but also submit your full, commented code as a separate R/RMD file.
- To best follow these guidelines, we suggest using Word (built in equation editor), R Markdown, Latex, or embedding a screenshot or scanned picture to compile your work.

DO NOT FORGET: You are encouraged to collaborate on homeworks, explain things to each other, and test each other's knowledge. But Do NOT hand out answers to someone who has not done any work. Everyone ought to have ideas about the possible answers or at least some thoughts about how to probe the problem further. Write your own solutions!

Problem 1 (10 points)

A new device has been developed which allows patients to evaluate their blood sugar levels. The most widely device currently on the market yields widely variable results. The new device is evaluated by 25 patients having nearly the same distribution of blood sugar levels yielding the following data:

125 123 117 123 115 112 128 118 124 111 116 109 125 120 113 123 112 118 121 118 122 115 105 118 131

```
# import data
blood_sugar_levels <- c(125, 123, 117, 123, 115, 112, 128, 118, 124,
                        111, 116, 109, 125, 120, 113, 123, 112, 118,
                        121, 118, 122, 115, 105, 118, 131)
```

a) Is there significant ($\alpha = 0.05$) evidence that median blood sugar readings was less than 120 in the population from which the 25 patients were selected? Use the sign test and report the test statistic and p-value. (5 points)

$H_0 : median = 120$ or $H_0 : median \geq 120$

$H_1 : median < 120$

```
# find the number of observations equal to 120
sum(blood_sugar_levels == 120)
```

```
[1] 1
```

Since there is one observation that is equal to 120, $n^* = 24$. Let C be the number of positive observations that are greater than 120 –

```
# find the number of observations greater than 120
sum(blood_sugar_levels > 120)
```

```
[1] 10
```

Therefore, $C = 10$.

We will reject the null if $C \geq \left(\frac{n^*}{2} + \frac{1}{2} + z_{1-\alpha/2}\sqrt{\frac{n^*}{4}}\right) = \left(\frac{24}{2} + \frac{1}{2} + 1.96\sqrt{\frac{24}{4}}\right) = 17.3$ or $C \leq \left(\frac{n^*}{2} - \frac{1}{2} - z_{1-\alpha/2}\sqrt{\frac{n^*}{4}}\right) = \left(\frac{24}{2} - \frac{1}{2} - 1.96\sqrt{\frac{24}{4}}\right) = 6.7$. In other words, we fail to reject if C is between 6.7 and 17.3 Since $C = 10$ and 10 is between 6.7 and 17.3, we fail to reject H_0 and

conclude that there is not significant evidence that the median blood sugar reading is less than 120 in the population from which these 25 patients were selected from.

Alternative solution:

```
library(BSDA)
BSDA::SIGN.test(
  blood_sugar_levels,
  md = 120,
  alternative = "less"
)
```

```
    One-sample Sign-Test

data:  blood_sugar_levels
s = 10, p-value = 0.2706
alternative hypothesis: true median is less than 120
95 percent confidence interval:
     -Inf 122.1203
sample estimates:
median of x
        118

Achieved and Interpolated Confidence Intervals:

                  Conf.Level L.E.pt   U.E.pt
Lower Achieved CI     0.9461   -Inf 122.0000
Interpolated CI       0.9500   -Inf 122.1203
Upper Achieved CI     0.9784   -Inf 123.0000
```

From this ouput, we see that the p-value is 0.2706, which is greater than $\alpha = 0.05$, therefore, we fail to reject the null hypothesis and conclude that there is not significant evidence that the median blood sugar reading is less than 120 in the population from which these 25 pateints were selected from.

b) Is there significant ($\alpha = 0.05$) evidence that median blood sugar readings was less than 120 in the population from which the 25 patients were selected? Use the Wilcoxon signed-rank test and report the test statistic and p-value. (5 points)

$H_0 : median = 120$ or $H_0 : median \geq 120$

$H_1 : median < 120$

Using R:

```r
wilcox.test(blood_sugar_levels, alternative = "less", mu = 120)
```

```
Warning in wilcox.test.default(blood_sugar_levels, alternative = "less", :
cannot compute exact p-value with ties
```

```
Warning in wilcox.test.default(blood_sugar_levels, alternative = "less", :
cannot compute exact p-value with zeroes
```

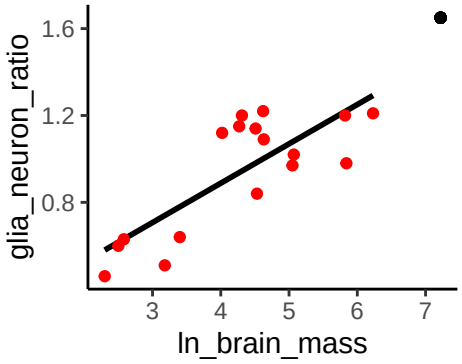
```
    Wilcoxon signed rank test with continuity correction

data:  blood_sugar_levels
V = 112.5, p-value = 0.1447
alternative hypothesis: true location is less than 120
```

From the output above, we see the p-value is 0.1447, which is greater than our α of 0.05. Therefore, we fail to reject the null hypothesis and conclude that there is not significant evidence that the median blood sugar reading is less than 120 in the population from which these 25 pateints were selected from.

Problem 2 (15 points)

Human brains have a large frontal cortex with excessive metabolic demands compared with the brains of other primates. However, the human brain is also three or more times the size of the brains of other primates. Is it possible that the metabolic demands of the human frontal cortex are just an expected consequence of greater brain size? A data file containing the measurements of glia-neuron ratio (an indirect measure of the metabolic requirements of brain neurons) and the log-transformed brain mass in nonhuman primates was provided to you along with the following graph.



a) Fit a regression model for the nonhuman data using $\ln$(brain mass) as a predictor. (Hint: Humans are "homo sapiens".) (4 points)

```r
# first remove "homo sapiens" which are humans
brain_no_human <-
  brain %>%
  filter(species != "Homo sapiens")

# fit the linear model
fit <- lm(glia_neuron_ratio ~ ln_brain_mass, data = brain_no_human)
summary(fit)
```

```
Call:
lm(formula = glia_neuron_ratio ~ ln_brain_mass, data = brain_no_human)

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
```

The fitted model can be written as: $\widehat{\text{glia-neuron ratio}} = 0.164 + 0.181 \cdot \ln(\text{brain mass})$

b) Using the nonhuman primate relationship, what is the predicted glia-neuron ratio for humans, given their brain mass? (4 points)

To find the predicted value for glia-neuron ratio for humans, we can use the fitted model:

$\widehat{\text{glia-neuron ratio}} = 0.164 + 0.181 \cdot \ln(1373.3) = 1.47$

Alternatively, we can use R:

```r
predict(fit, newdata = brain[1,])
```

```
       1
1.471458
```

c) Determine the most plausible range of values for the prediction. Which is more relevant for your prediction of human glia-neuron ratio: an interval for the predicted mean glia-neuron ratio at the given brain mass, or an interval for the prediction of a single new observation? (1 point)

We should use a prediction interval because we are interested in getting a range for a new, individual data point (not the predicted mean of glia-neuron ratio).

d) Construct the 95% interval chosen in part (c). On the basis of your result, does the human brain have an excessive glia-neuron ratio for its mass compared with other primates? (3 points)

The $100 \cdot (1-\alpha)\%$ prediction interval for a *single, new value* in SLR is given by:

$\hat{\beta}_0 + \hat{\beta}_1 X_h \pm t_{n-1,1-\alpha/2} \cdot se\left(\hat{\beta}_0 + \hat{\beta}_1 X_h\right)$

where $se\left(\hat{\beta}_0 + \hat{\beta}_1 X_h\right) = \sqrt{MSE \cdot \left\{\frac{1}{n} + \left[\frac{(X_h - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right] + 1\right\}}$

In R, we can compute the 95% prediction interval simply using the following code:

```
predict(fit, newdata = brain[1,], interval = "predict", level = 0.95)
```

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

This prediction interval tells us that the predicted glia-neuron ratio for a new value of brain mass (from humans) is between 1.04 and 1.91. Since the upper limit is as high as 1.91 (which is much higher than the other ratios corresponding to primates), the human brain has an excessive glia-neuron ratio for its mass compared with other primates. In addition, the lower bound (1.04) is at the high end of glia-neuron ratio for other primates, further supporting the idea that humans might have an excessive glia-neuron ratio compared to primates.

e) Considering the position of the human data point relative to those data used to generate the regression line (see graph above), what additional caution is warranted? (3 points)

Since humans are primates, we are not concerned about our data coming from different populations, necessarily. The main problem here is that the log-transformed brain mass for humans is beyond the range of log-transformed brain mass for the other primates. Hence, we run into the danger of extrapolation. This means that the linear model we fit may not be applicable for an observation outside the range of data we used to fit our model.

Problem 3 (25 points)

For this problem, you will be using 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. Also, generate appropriate descriptive statistics for all variables of interest (continuous and categorical) – no test required. (2 points)

This data set consists of information about patients who are subscribed to a particular health insurance company who submitted claims relating to coronary heart disease. There are 788 patients in the data set. In total, there are 10 variables, but we are interested in 6 of them: total cost (totalcost), number of ER visits (ERvisits), age (age), gender (gender), number

of complications (complications), and number of days of duration of the treatment condition (duration).

Total cost, age, number of ER visits, number of complications, and duration are quantitative variables, though all are recorded as discrete variables in this data set except for total cost.

Gender is categorical, but coded with numbers: 1 if male, 0 otherwise.

Our outcome of interest is total cost.

Our main predictor of interest is number of ER visits.

The covariates we want to adjust for are: age, gender, number of complications, and duration.

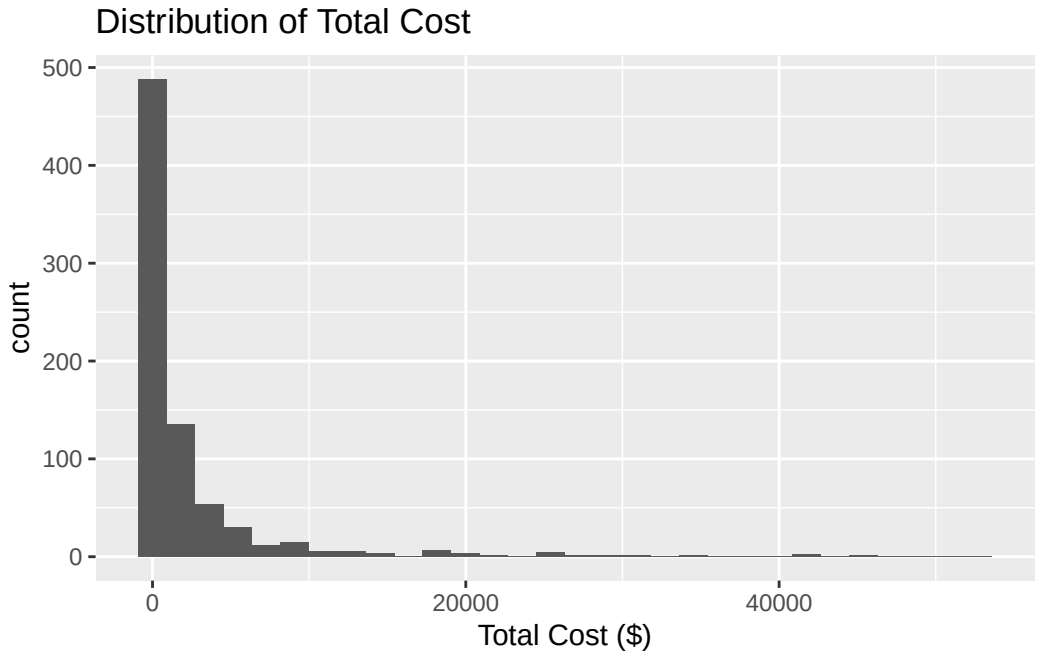
```
library(gtsummary)

heart_data %>%
  select(totalcost, ERvisits, age, gender, complications, duration) %>%
  gtsummary::tbl_summary() %>%
  gtsummary::bold_labels() %>%
  gtsummary::italicize_levels()
```

Characteristic	**N = 788**
totalcost	507 (161, 1,905)
ERvisits	3.00 (2.00, 5.00)
age	60 (55, 64)
gender	180 (23%)
complications	
0	745 (95%)
1	42 (5.3%)
3	1 (0.1%)
duration	166 (42, 281)

b) Investigate the shape of the distribution for variable `totalcost` and try different transformations, if needed. (3 points)

```
heart_data %>%
  ggplot(aes(x = totalcost)) +
  geom_histogram() +
  labs(x = "Total Cost ($)",
       title = "Distribution of Total Cost")
```

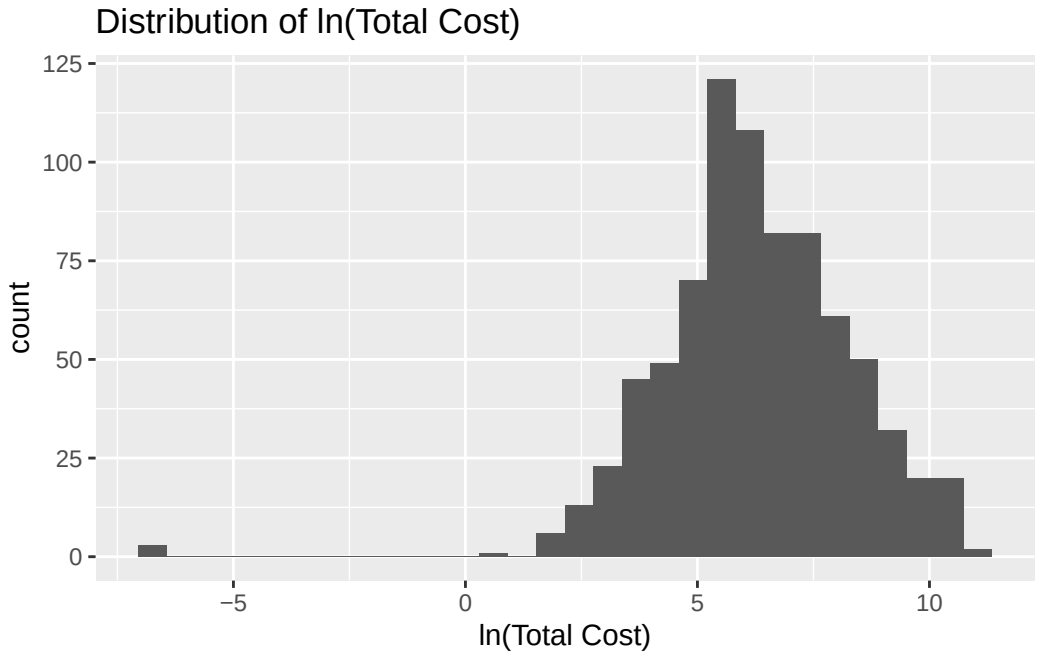


This distribution is extremely right skewed. Before we try to fit any model, we should transform this variable before using it in a linear model.

A common transformation for highly skewed data is the natural log. Alternatively, you could apply a Box Cox transformation to find the ideal power transformation. (Since we have some 0 observations, we will add a small number (0.001) to each value before we take the natural log.)

```
heart_data <-
  heart_data %>%
  mutate(ln_totalcost = log(totalcost + 0.001))
#the default base in the log function is e, hence the natural log

## check the transformation with another histogram:
heart_data %>%
  ggplot(aes(x = ln_totalcost)) +
  geom_histogram() +
  labs(x = "ln(Total Cost)",
       title = "Distribution of ln(Total Cost)")
```



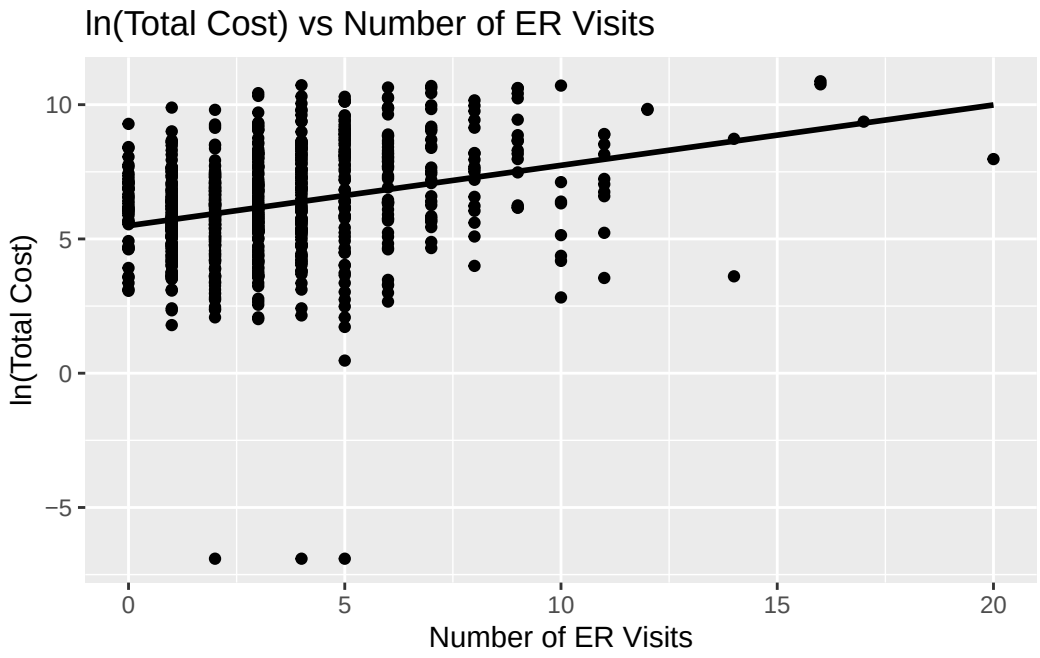
Using the natural log makes this data look much more "normal"/symmetric. This transformation caused a few outliers (the observations with a total cost = 0), but the rest of the data looks good after this transformation.

c) Create a new variable called `comp_bin` by dichotomizing `complications`: 0 if no complications, and 1 otherwise. (1 point)

```
heart_data <-
  heart_data %>%
  mutate(comp_bin = if_else(complications == 0, 0, 1))
```

d) Based on your decision in part (b), fit a simple linear regression (SLR) between the original or transformed `totalcost` and predictor `ERvisits`. This includes a scatterplot and results of the regression, with appropriate comments on significance and interpretation of the slope. (3 points)

```
heart_data %>%
  ggplot(aes(x = ERvisits, y = ln_totalcost)) +
  geom_point() +
  labs(x = "Number of ER Visits",
       y = "ln(Total Cost)",
       title = "ln(Total Cost) vs Number of ER Visits") +
  geom_smooth(method = "lm", se = FALSE, color = "black")
```



Based on the scatterplot above and the fitted line, it seems like a linear relationship between number of ER visits and the natural log of total cost is plausible, so we will continue with our regression analysis.

```
fit <- lm(ln_totalcost ~ ERvisits, data = heart_data)
summary(fit)
```

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

Residuals:
     Min       1Q   Median       3Q      Max
-13.5255  -1.0922   0.0608   1.3147   4.3314

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.49385    0.11387  48.248   <2e-16 ***
ERvisits     0.22477    0.02635   8.531   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 1.949 on 786 degrees of freedom
Multiple R-squared:  0.08475,   Adjusted R-squared:  0.08359
F-statistic: 72.79 on 1 and 786 DF,  p-value: < 2.2e-16
```

The resulting formula to predict the natural log of total costs is:

$\widehat{\ln(\text{totalcost})} = 5.49 + 0.22 \cdot X_{\text{ERvisits}}$

Our slope is 0.22 – this means that for every additional ER visit, the predicted natural log of total cost will increase by 0.22 on average.

The slope estimate is significant (has an extremely small p-value), meaning the number of ER visits is significantly associated with the natural log of the total costs. (Statistically speaking, this p-value tells us that this estimate is significantly different from 0.)

e) Fit a multiple linear regression (MLR) with `comp_bin` and `ERvisits` as predictors. (1 point)

```r
fit2 <- lm(ln_totalcost ~ ERvisits + comp_bin,
           data = heart_data)
summary(fit2)
```

```
Call:
lm(formula = ln_totalcost ~ ERvisits + comp_bin, data = heart_data)

Residuals:
     Min       1Q   Median       3Q      Max
-13.3943  -1.0451   0.0252   1.2191   4.4397

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.47694    0.11165  49.054  < 2e-16 ***
ERvisits     0.20193    0.02613   7.728 3.33e-14 ***
comp_bin     1.74365    0.30321   5.751 1.27e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.911 on 785 degrees of freedom
Multiple R-squared:  0.1218,    Adjusted R-squared:  0.1195
F-statistic: 54.41 on 2 and 785 DF,  p-value: < 2.2e-16
```

$\widehat{\ln(\text{totalcost})} = 5.48 + 0.20 \cdot X_{\text{ERvisits}} + 1.74 \cdot X_{\text{comp-bin}}$

i) Test if `comp_bin` is an effect modifier of the relationship between `totalcost` and `ERvisits`. Comment. (3 points)

We want to see if `comp_bin` and `ERvisits` have a significant interaction.

```
fit3 <- lm(ln_totalcost ~ ERvisits * comp_bin,
           data = heart_data)
summary(fit3)
```

```
Call:
lm(formula = ln_totalcost ~ ERvisits * comp_bin, data = heart_data)

Residuals:
     Min       1Q   Median       3Q      Max
-13.4051  -1.0559   0.0325   1.2269   4.4353

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        5.45549    0.11406  47.828  < 2e-16 ***
ERvisits           0.20837    0.02705   7.703 4.01e-14 ***
comp_bin           2.22319    0.60233   3.691 0.000239 ***
ERvisits:comp_bin -0.09639    0.10461  -0.921 0.357101
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.911 on 784 degrees of freedom
Multiple R-squared:  0.1227,    Adjusted R-squared:  0.1193
F-statistic: 36.55 on 3 and 784 DF,  p-value: < 2.2e-16
```

Based on this output, there is not a significant interaction between `ERvisits` and `comp_bin`.

Specifically, the p-value associated with the `ERvisits` and `comp_bin` interaction term is testing: $H_0 : \beta_{\text{ERvisits*comp-bin}} = 0$ vs $H_1 : \beta_{\text{ERvisits*comp-bin}} \neq 0$. Since this p-value is non-significant when compared to $\alpha = 0.05$, we do not have evidence that this coefficient should be something other than 0. This implies that `comp_bin` is not an effect modifier of the relationship between `ERvisits` and `ln_totalcost`. This suggests that the effect of the number of ER visits on the natural log of the total cost is not different based on whether the patient had complications or did not have complications.

Based on these results, we can safely remove the interaction term from our model.

ii) Test if `comp_bin` is a confounder of the relationship between `totalcost` and `ERvisits`. Comment. (3 points)

To see if `comp_bin` is a confounder, we will add it to our model as another covariate and observe how β_{ERvisits} changes.

Recall: in the model with number of ER visits as the only predictor, $\beta_{\text{ERvisits}} = 0.22$.

```
fit4 <- lm(ln_totalcost ~ ERvisits + comp_bin,
           data = heart_data)
summary(fit4)
```

```
Call:
lm(formula = ln_totalcost ~ ERvisits + comp_bin, data = heart_data)

Residuals:
     Min       1Q   Median       3Q      Max
-13.3943  -1.0451   0.0252   1.2191   4.4397

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.47694    0.11165  49.054  < 2e-16 ***
ERvisits     0.20193    0.02613   7.728 3.33e-14 ***
comp_bin     1.74365    0.30321   5.751 1.27e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.911 on 785 degrees of freedom
Multiple R-squared:  0.1218,    Adjusted R-squared:  0.1195
F-statistic: 54.41 on 2 and 785 DF,  p-value: < 2.2e-16
```

After adding `comp_bin` to our model, β_{ERvisits} changes to 0.20.

While the estimate of the β for ER visits changes by about 10% ($\frac{0.22-0.20}{0.22} = 0.091$), the overall estimate does not change much. If you follow the rule of thumb that says "if the effect changes by about 10% then it's a confounder", then we might consider `comp_bin` to be a counfounder of the relationship between `ERvisits` and `ln_totalcost`. However, in this instance, it's not a strong conclusion.

iii) Decide if `comp_bin` should be included along with `ERvisits`. Why or why not? (3 points)

To determine if we should include `comp_bin` in the model or not, we can do a partial F-test.

Recall, `fit` is the model that has `ERvisits` as the only predictor and `fit4` is the model that includes both `ERvisits` and `comp_bin` as predictors.

Using $\alpha = 0.05$, we can test the hypotheses

$H_0 : \beta_{\text{comp-bin}} = 0$ vs $H_1 : \beta_{\text{comp-bin}} \neq 0$

```
anova(fit, fit4)
```

```
Analysis of Variance Table

Model 1: ln_totalcost ~ ERvisits
Model 2: ln_totalcost ~ ERvisits + comp_bin
  Res.Df    RSS Df Sum of Sq     F    Pr(>F)
1    786 2986.8
2    785 2866.1  1    120.74 33.07 1.273e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We end up with an $F_{stat} = 33.07$ and a very small p-value, which is less than our $\alpha = 0.05$. Therefore, we would reject the null hypothesis and conclude that `comp_bin` should be included along with `ERvisits`. (In other words, we have evidence that the β associated with `comp_bin` is non-zero, and therefore important.)

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. (3 points)

```
fit5 <- lm(ln_totalcost ~ ERvisits + comp_bin +
             age + gender + duration,
           data = heart_data)
summary(fit5)
```

```
Call:
lm(formula = ln_totalcost ~ ERvisits + comp_bin + age + gender +
    duration, data = heart_data)

Residuals:
     Min       1Q   Median       3Q      Max
-12.1885  -0.9962  -0.0838   1.0099   4.3499
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.8016094  0.5559875  10.435  < 2e-16 ***
ERvisits     0.1732359  0.0245895   7.045 4.07e-12 ***
comp_bin     1.5335712  0.2815721   5.446 6.89e-08 ***
age         -0.0193387  0.0094493  -2.047   0.0410 *
gender      -0.3234404  0.1510866  -2.141   0.0326 *
duration     0.0060628  0.0005325  11.386  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.769 on 782 degrees of freedom
Multiple R-squared:  0.2502,    Adjusted R-squared:  0.2454
F-statistic: 52.18 on 5 and 782 DF,  p-value: < 2.2e-16
```

$$\widehat{\ln(\text{totalcost})} = 5.8 + 0.17 \cdot X_{\text{ERvisits}} + 1.53 \cdot X_{\text{comp-bin}} - 0.02 \cdot X_{\text{age}} - 0.32 \cdot X_{\text{gender}} + 0.006 \cdot X_{\text{duration}}$$

Using an $\alpha = 0.05$, all of our predictors are significant in this model.

Our model can give us the following information:

- For every additional ER visit, we predict that the natural log of total cost will increase by 0.17 on average, after adjusting for presence of complications, age, gender, and duration of treatment condition.
- The predicted natural log of total cost will increase by about 1.53 on average for patients who have complications vs those who do not have complications, after adjusting for number of ER visits, age, gender, and duration of treatment condition.
- For every one year increase in age, we predict that the natural log of total cost will decrease by 0.02 on average, after adjusting for the number of ER visits, presence of complications, gender, and duration of treatment condition.
- We expect the natural log of total cost to be 0.32 less than females on average, after adjusting for number of ER visits, presence of complications, age, and duration of treatment condition.
- For every additional day of treatment condition duration, we expect the natural log of total cost to increase by about 0.006 on average, after adjusting for number of ER visits, presence of complications, age, and gender.

ii) Compare the SLR and MLR models. Which model would you use to address the investigator's objective and why? (3 points)

We can use a partial F-test at the $\alpha = 0.05$ significance level to compare the SLR model (only `ERvisits`) with the full model including all additional covariates of interest.

$H_0 : \beta_{\text{comp-bin}} = \beta_{\text{age}} = \beta_{\text{gender}} = \beta_{\text{duration}} = 0$ vs H_1 : at least one is not equal to 0

```
anova(fit, fit5)
```

```
Analysis of Variance Table

Model 1: ln_totalcost ~ ERvisits
Model 2: ln_totalcost ~ ERvisits + comp_bin + age + gender + duration
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    786 2986.8
2    782 2447.0  4    539.86 43.132 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Our $F_{stat} = 43.1$ with 4 degrees of freedom and the p-value is extremely small. Since our p-value is less than α, we will reject the null hypothesis and conclude that the larger model including `comp_bin`, `age`, `gender`, and `duration` as additional predictors includes at least one significant predictor and should be included in the final model.

Additionally, the larger model has an adjusted R^2 of 0.2452, which is higher than the adjusted R^2 in the SLR model (0.0836). This means that the larger model can explain about 25% of the variation in the natural log of total cost. In the model that only uses number of ER visits as a predictor, only 8% of the total variance in the natural log of total costs can be explained by ER visits.