az2852_hw4

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(BSDA)
```

```
## Loading required package: lattice
##
## Attaching package: 'BSDA'
##
## The following object is masked from 'package:datasets':
##
##     Orange
```

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

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.

H0: The median blood sugar level is equal to 120. H1: The median blood sugar level is less than 120.

```
data = 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)

SIGN.test(data, md = 120, alternative = "less")
```

```
##
##  One-sample Sign-Test
##
## data:  data
## 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
```

The number of negative differences s = 10, p-value = 0.2706, larger than 0.05, so we fail to reject the null hypothesis.We don't have enough evidence that median population blood sugar readings is less than 120.

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.

H0: The median blood sugar level is equal to 120. H1: The median blood sugar level is less than 120.

```
wilcox.test(data, mu = 120, alternative = "less", exact = FALSE)
```

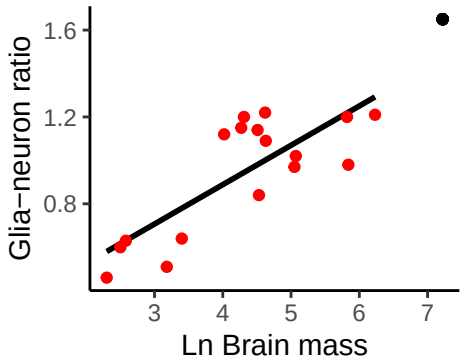
```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  data
## V = 112.5, p-value = 0.1447
## alternative hypothesis: true location is less than 120
```

The sum of negative ranks V = 112.5, p-value = 0.1447, larger than 0.05, so we fail to reject the null hypothesis.We don't have enough evidence that median population blood sugar readings is less than 120.

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.

```
## Warning in geom_point(aes(x = brain$`Ln Brain mass`[1], y = brain$`Glia-neuron ratio`[1])): All aestl
## i Please consider using `annotate()` or provide this layer with data containing
##   a single row.
```



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

```
# Filter out humans
nonhuman_data = subset(brain, Species != "Homo sapiens")

model = lm(`Glia-neuron ratio` ~ `Ln Brain mass`, data = nonhuman_data)

summary(model)
```

```
##
## Call:
## lm(formula = `Glia-neuron ratio` ~ `Ln Brain mass`, data = nonhuman_data)
##
## 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) Using the nonhuman primate relationship, what is the predicted glia-neuron ratio for humans, given their brain mass?

```
human_lnBrainMass = brain %>%
  filter(Species == "Homo sapiens") %>%
  pull(`Ln Brain mass`)

predict(model, newdata = tibble(`Ln Brain mass` = human_lnBrainMass))
```

```
##        1
## 1.471458
```

The predicted glia-neuron ratio for humans is 1.471.

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?

The prediction interval is more relevant because it accounts for the uncertainty in predicting a single data point.

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? construct a 95% PI:

```
predict(model, newdata = tibble(`Ln Brain mass` = human_lnBrainMass), interval = "prediction", level = (
```

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

The 95% PI for the average human glia-neuron ratio is [1.036047, 1.906869].

The observed value is 1.65, inside the CI. We do not have enough evidence to suggest that the human brain has 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 (see graph above), what additional caution is warranted?

Considering that the human data point lies far above the regression line, it might be an outlier that can't be appropriately estimated with our regression model. The deviation of the human data point suggests that additional factors, beyond brain mass alone, may need to be considered when predicting human brain glia-neuron ratio.

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.

```
#load the data
heart_df = read_csv("HeartDisease.csv")
```

```
## 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.
```

The main outcome is `totalcost` (total cost of claims by the subscriber in dollars), the main predictor is `ERvisits` (number of ER visits).

Other important covariates include `age`, `gender`(1 for male, 0 for female), `interventions`(total number of interventions or procedures), `drugs`(Number of drugs prescribed), `complications`(Number of complications during treatment), `comorbidities`(Number of other diseases the patient had), and `duration`(Duration of treatment in days).

Descriptive statistics:

```
summary(heart_df)
```

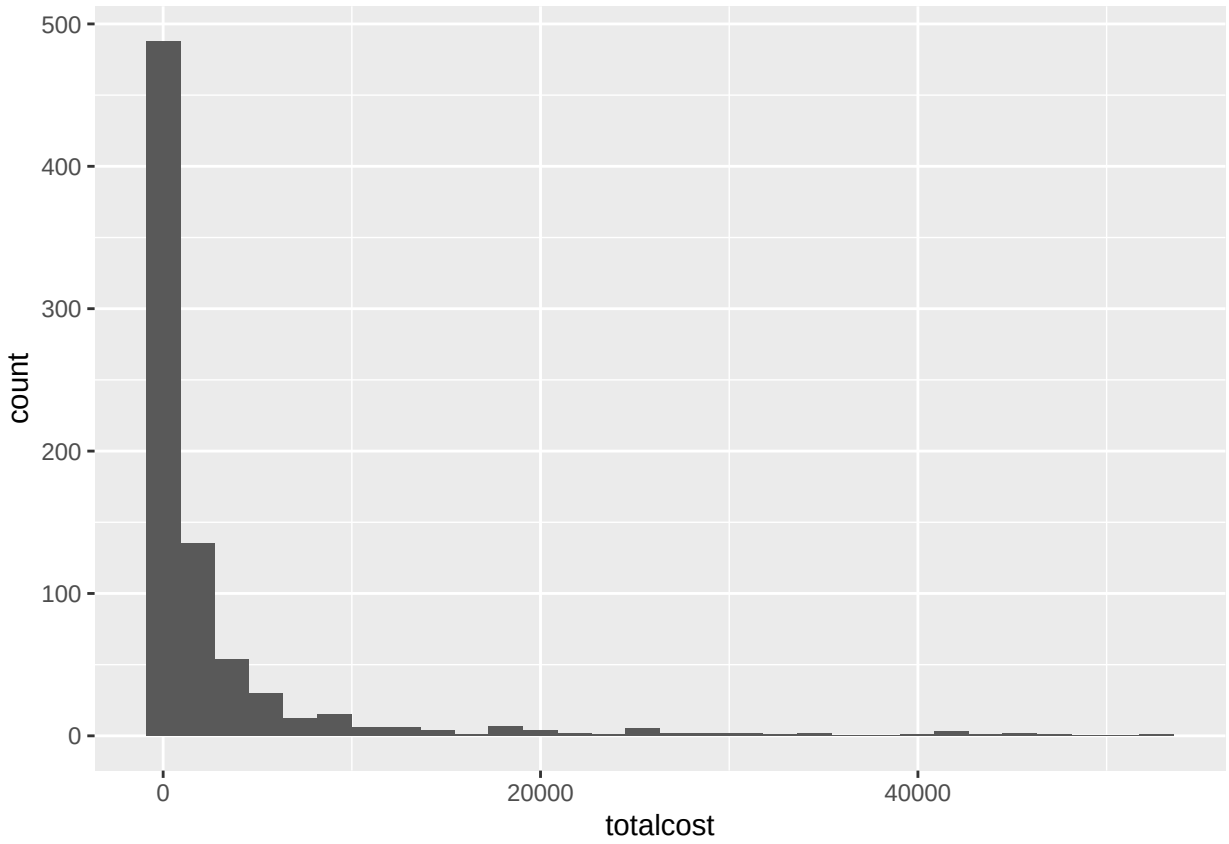
```
##        id          totalcost            age            gender
##  Min.   :  1.0   Min.   :    0.0   Min.   :24.00   Min.   :0.0000
##  1st Qu.:197.8   1st Qu.:  161.1   1st Qu.:55.00   1st Qu.:0.0000
##  Median :394.5   Median :  507.2   Median :60.00   Median :0.0000
##  Mean   :394.5   Mean   : 2800.0   Mean   :58.72   Mean   :0.2284
##  3rd Qu.:591.2   3rd Qu.: 1905.5   3rd Qu.:64.00   3rd Qu.:0.0000
##  Max.   :788.0   Max.   :52664.9   Max.   :70.00   Max.   :1.0000
##  interventions        drugs           ERvisits      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) Investigate the shape of the distribution for variable `totalcost` and try different transformations, if needed.

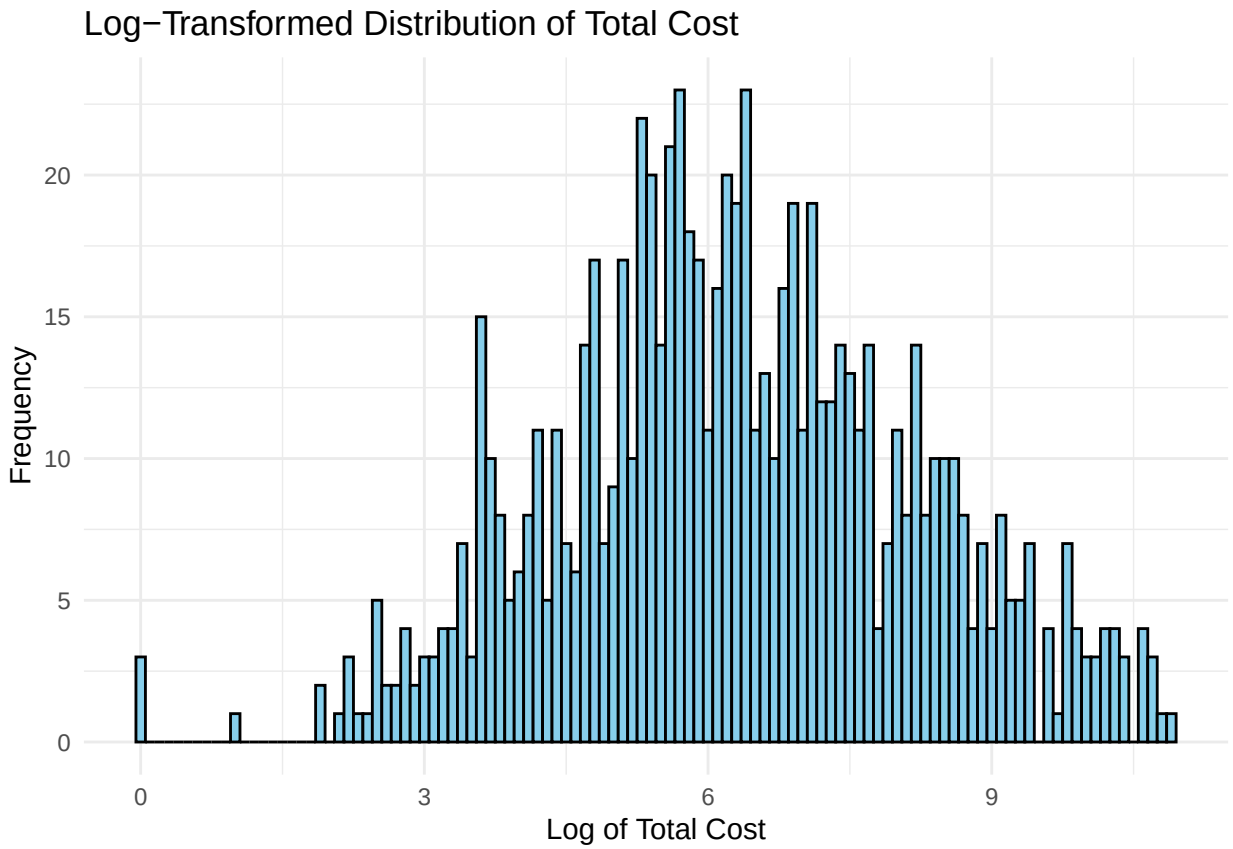
The distribution is severely right skewed. We will use a log transformation.

```
ggplot(heart_df, aes(x = totalcost))+
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
# Apply log transformation to 'totalcost'
heart_df$totalcost_log = log(heart_df$totalcost + 1)  # +1 to handle zero values

# Visualize the transformed distribution
ggplot(heart_df, aes(x = totalcost_log)) +
  geom_histogram(binwidth = 0.1, fill = "skyblue", color = "black") +
  theme_minimal() +
  labs(title = "Log-Transformed Distribution of Total Cost", x = "Log of Total Cost", y = "Frequency")
```




```
# Calculate basic statistics for the log-transformed variable
summary(heart_df$totalcost_log)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   5.088   6.231   6.298   7.553  10.872
```

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

```
heart_df = heart_df %>%
  mutate(comp_bin = ifelse(heart_df$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.

The p values is <2e-16, indicating that the relationship between number of ER visits and total cost is strongly statistically significant.

The slope is 0.22529, suggesting a positive linear relationship between totalcost and ERvists. For each additional ER visit, the log of total cost increases by 0.22529.

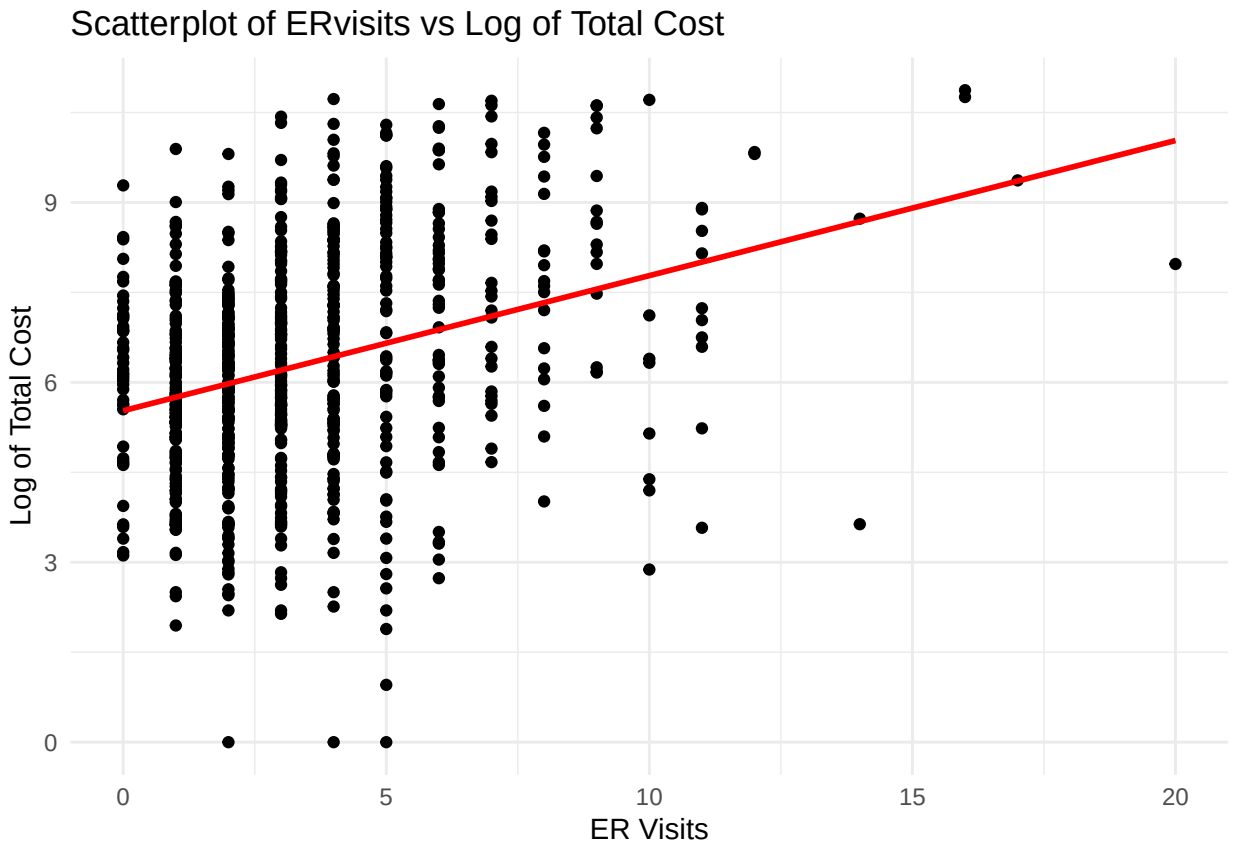
```
# Fit a simple linear regression model with log-transformed totalcost and ERvisits
model_simple = lm(totalcost_log ~ ERvisits, data = heart_df)

# Summary of the regression model
summary(model_simple)
```

```
##
## Call:
```

```
## lm(formula = totalcost_log ~ ERvisits, data = heart_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.6532 -1.1230  0.0309  1.2797  4.2964
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.52674    0.10510  52.584   <2e-16 ***
## ERvisits     0.22529    0.02432   9.264   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.799 on 786 degrees of freedom
## Multiple R-squared:  0.09844,    Adjusted R-squared:  0.09729
## F-statistic: 85.82 on 1 and 786 DF,  p-value: < 2.2e-16
```

```r
# Scatterplot with regression line for log-transformed totalcost
ggplot(heart_df, aes(x = ERvisits, y = totalcost_log)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  theme_minimal() +
  labs(title = "Scatterplot of ERvisits vs Log of Total Cost", x = "ER Visits", y = "Log of Total Cost")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



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

```r
# Fit the MLR model for log transformed total cost with 'ERvisits', and 'comp_bin' as predictors
model_mlr = lm(totalcost_log ~ ERvisits + comp_bin, data = heart_df)

# Summary of the model
summary(model_mlr)
```

```
##
## Call:
## lm(formula = totalcost_log ~ ERvisits + comp_bin, data = heart_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.5249 -1.0769 -0.0074  1.1847  4.4024
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.51020    0.10279  53.606  < 2e-16 ***
## ERvisits     0.20295    0.02405   8.437  < 2e-16 ***
## comp_bin     1.70573    0.27915   6.111 1.56e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.759 on 785 degrees of freedom
## Multiple R-squared:  0.1394, Adjusted R-squared:  0.1372
## F-statistic: 63.57 on 2 and 785 DF,  p-value: < 2.2e-16
```

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

The p-value for the interaction term is 0.311, larger than 0.05. We conclude that comp_bin does not modify the relationship between ERvisits and totalcost_log.

```r
# Fit the model with interaction term between 'ERvisits' and 'comp_bin'
model_interaction = lm(totalcost_log ~ ERvisits * comp_bin, data = heart_df)

# Summary of the interaction model
summary(model_interaction)
```

```
##
## Call:
## lm(formula = totalcost_log ~ ERvisits * comp_bin, data = heart_df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.536 -1.083  0.004  1.200  4.398
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        5.48849    0.10500  52.271  < 2e-16 ***
## ERvisits           0.20947    0.02490   8.412  < 2e-16 ***
## comp_bin           2.19096    0.55447   3.951 8.47e-05 ***
## ERvisits:comp_bin -0.09753    0.09630  -1.013    0.311
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.759 on 784 degrees of freedom
```

```
## Multiple R-squared:  0.1405, Adjusted R-squared:  0.1372
## F-statistic: 42.72 on 3 and 784 DF,  p-value: < 2.2e-16
```

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

```
# Extract coefficients for ERvisits from both models & calculate the difference
coef_simple = summary(model_simple)$coefficients["ERvisits", "Estimate"]
coef_mlr = summary(model_mlr)$coefficients["ERvisits", "Estimate"]

diff_coef = coef_mlr - coef_simple
diff_coef
```

```
## [1] -0.02234588
```

The difference between the coefficients is -0.0223459. The difference is quite small, suggesting that comp_bin is not a confounder.

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

We don't need to include `comp-bin` when trying to understand the relationship between total cost and ERvisits as `comp_bin` is neither an effect modifier nor a confounder.

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.

```
model_cov = lm(totalcost_log ~ ERvisits + comp_bin + age + gender + duration, data = heart_df)
summary(model_cov)
```

```
##
## Call:
## lm(formula = totalcost_log ~ ERvisits + comp_bin + age + gender +
##     duration, data = heart_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.4711 -1.0340 -0.1158  0.9493  4.3372
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.9404610  0.5104064  11.639  < 2e-16 ***
## ERvisits     0.1745975  0.0225736   7.735 3.20e-14 ***
## comp_bin     1.5044946  0.2584882   5.820 8.57e-09 ***
## age         -0.0206475  0.0086746  -2.380   0.0175 *
## gender      -0.2067662  0.1387002  -1.491   0.1364
## duration     0.0057150  0.0004888  11.691  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.624 on 782 degrees of freedom
## Multiple R-squared:  0.2694, Adjusted R-squared:  0.2647
## F-statistic: 57.68 on 5 and 782 DF,  p-value: < 2.2e-16
```

In this MLR, significant predictors are ERvisits, comp_bin, age, and duration of treatment, while gender does not appear to be a significant predictor. The coefficients suggest:

- `ERvisits`: The coefficient (0.175) is significant ($p < 0.001$), indicating that ER visits have a positive association with totalcost_log.

- `Comp_bin`: The coefficient (1.504) is significant ($p < 0.001$), suggesting that total cost increases when there are complications.
- `Age`: The coefficient (-0.0206) is significant ($p = 0.0175$), showing a small negative effect on log_totalcost.
- `Gender`: Not significant ($p = 0.1364$).
- `Duration`: The coefficient (0.005) is significant ($p < 0.001$), indicating a small positive effect.

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

The MLR model has an R-squared of 0.269, much higher than the SLR model's 0.098. This indicates a better fit and more explained variance.
The MLR model is more preferable because it accounts for potential confounders in the covariates.