Homework 4

Pei Tian, pt2632

2023-12-10

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.

```
p1_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
)

alter_val = 120
n_star = sum(p1_data != alter_val)
C = sum(p1_data > alter_val)
# stats = (C - n_star / 2 + 0.5) / (sqrt(n_star / 4))
test_result = SIGN.test(p1_data, md = 120, alternative = "less", conf.level = 0.95)
```

Let Δ be the median of the blood sugar reading distribution of patients.

Hypothesis: $H_0 : \Delta = 120, H_1 = \Delta < 120$

Total number of non-zero difference: $n^{\star}$= 24

Number of positive difference: $C = 10$

Normal Approximation: $n^\star p(1-p) = n^\star/4 = 6 > 5$

Test Statistic: $stats = \frac{C - \frac{n^\star}{2} + \frac{1}{2}}{\sqrt{\frac{n^\star}{4}}} = 0.6066615$

p-value $= 0.2706281 > 0.05$

So we fail to reject the H_0, which means that median blood sugar readings equals to 120 in 0.05 significance level.

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.

```
test_result = wilcox.test(p1_data, mu = 120,
                          alternative = "less",
                          conf.level = 0.95,
                          correct = T)
```

Hypothesis: $H_0 : \Delta = 120, H_1 = \Delta < 120$

Let T_+ be the sum of the ranks for positive difference,

Statistic: (ties) $T = \frac{|T_+ - \frac{n^\star(n^\star+1)}{4}| - \frac{1}{2}}{\frac{n^\star(n^\star+1)(2n^\star+1)}{6} - \frac{\sum_{i=1}^{g}(t_i^3 - t_i)}{48}} =$ -1.0596327

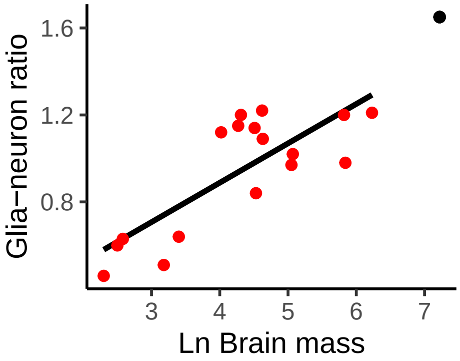
p-value $= 0.1446559 > 0.05$

So we fail to reject the H_0, which means that median blood sugar readings equals to 120 in 0.05 significance level.

Problem 2 (15 points)

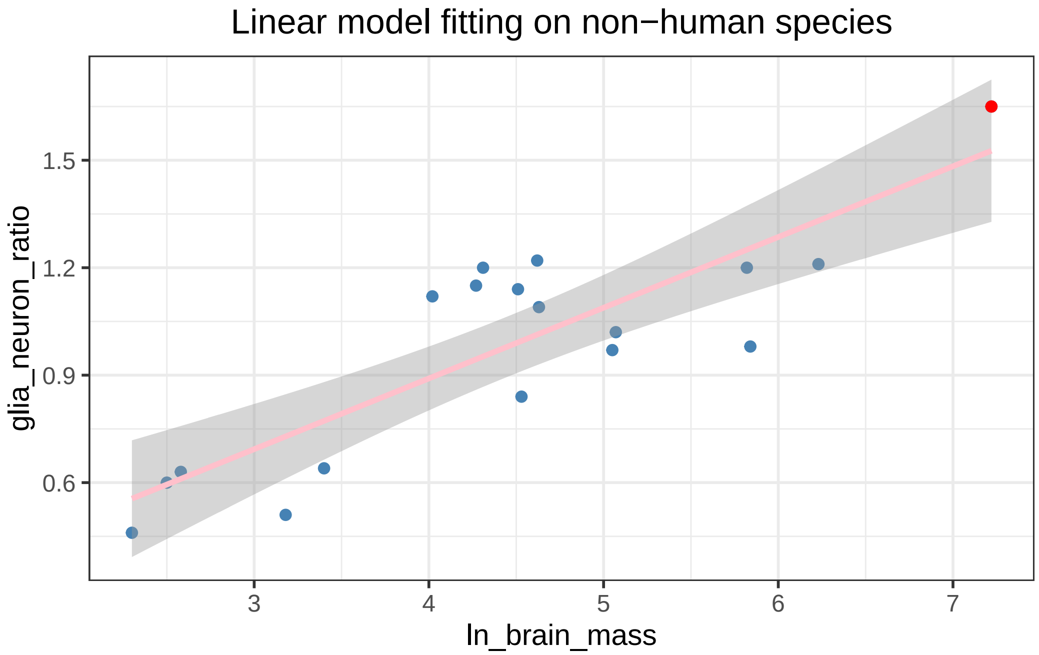
Human brains have a large frontal cortex with excessive metabolic demands compared with the brains of other primates. However, the human brain is also three or more times the size of the brains of other primates. Is it possible that the metabolic demands of the human frontal cortex are just an expected 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 non-human data using ln (brain mass) as a predictor. (Hint: Humans are “homo sapiens”.)



```
brain <- readxl::read_xlsx("./data/Brain.xlsx")

human = "Homo sapiens"
brain = brain |> janitor::clean_names()
non_human_fit = brain |>
  filter(species != human) |>
  lm(glia_neuron_ratio ~ ln_brain_mass, data = _)
modelr::add_predictions(brain, non_human_fit) |>
  ggplot(aes(x = ln_brain_mass, y = glia_neuron_ratio)) +
  geom_point(color = "steelblue") +
  geom_smooth(method = "lm", se = T,
              color = "pink", lwd = 1) +
  geom_point(data = filter(brain, species == human),
             aes(ln_brain_mass, glia_neuron_ratio), color = "red") +
  labs(
    title = "Linear model fitting on non-human species") +
  theme(plot.title = element_text(hjust = 0.5),
        legend.position = "none")
```



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

```
pred_input = brain |> filter(species == human)
pred_ratio = predict(non_human_fit, pred_input)
```

The predicted glia-neuron ratio for humans is 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?

```
mean_interval = predict(non_human_fit, pred_input,
                        interval = "conf", level = 0.95)
new_interval = predict(non_human_fit, pred_input,
                       interval = "pred", level = 0.95)
```

Interval for the predicted mean glia-neuron ratio at the given brain mass: (1.2295581, 1.7133578).

Interval for the prediction of a single new observation: (1.0360468, 1.9068691)

I think the later interval is more plausible, because the data of human is not included in the training data, which means it is new data for the fitted linear model.

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?

Given that the glia-neuron ratio of human equals to 1.65, which lies in the chosen confidence interval, so the human brain doesn't have an excessive glia-neuron ratio for its mass compared with other primates in 0.05 significance level.

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?

From the graph above, the human data point is located relatively far away from other data point, which indicates that it maybe an outlier in the dataset and the prediction result maybe unreliable.

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.

```
heart = read_csv("./data/HeartDisease.csv") |>
  janitor::clean_names()
```

Main outcome: total cost

Main predictor: number of emergency room (ER) visits

Other important covariates: age, gender, number of complications, duration of treatment condition

Descriptive statistics for all variables of interest (continuous and categorical):

```
# continuous
heart |>
  select(totalcost, e_rvisits, age, complications, duration) |>
  summary()
```

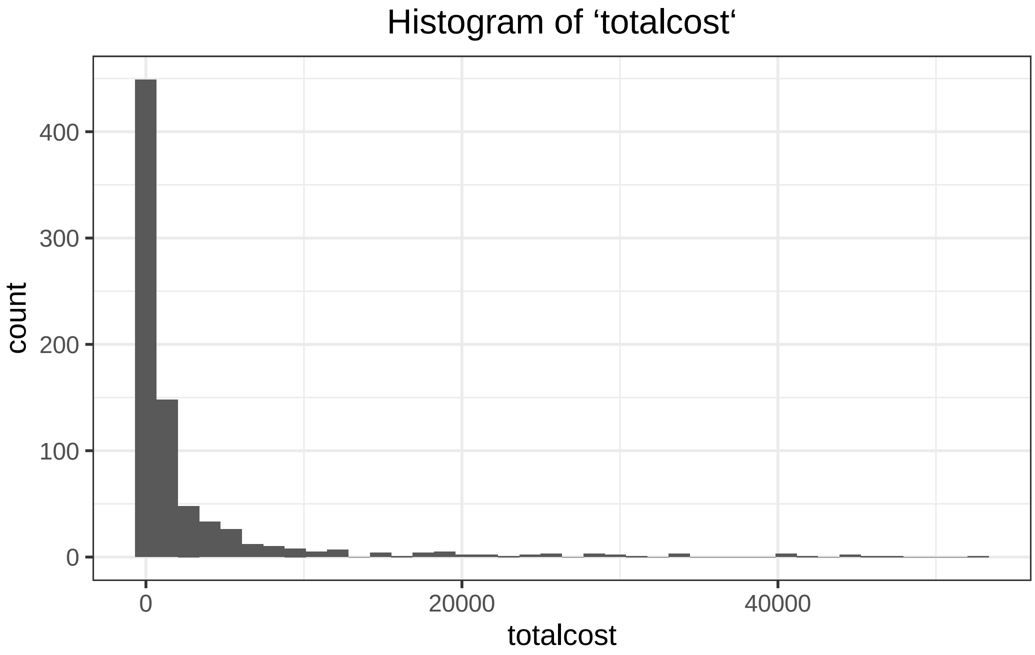
```
  totalcost          e_rvisits           age          complications    
 Min.   :    0.0   Min.   : 0.000   Min.   :24.00   Min.   :0.00000  
 1st Qu.:  161.1   1st Qu.: 2.000   1st Qu.:55.00   1st Qu.:0.00000  
 Median :  507.2   Median : 3.000   Median :60.00   Median :0.00000  
 Mean   : 2800.0   Mean   : 3.425   Mean   :58.72   Mean   :0.05711  
 3rd Qu.: 1905.5   3rd Qu.: 5.000   3rd Qu.:64.00   3rd Qu.:0.00000  
 Max.   :52664.9   Max.   :20.000   Max.   :70.00   Max.   :3.00000  
    duration     
 Min.   :  0.00  
 1st Qu.: 41.75  
 Median :165.50  
 Mean   :164.03  
 3rd Qu.:281.00  
 Max.   :372.00  
```

```
# categorical
heart |>
  group_by(gender) |>
  summarise(count = n()) |>
  knitr::kable()
```

gender	count
0	608
1	180

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

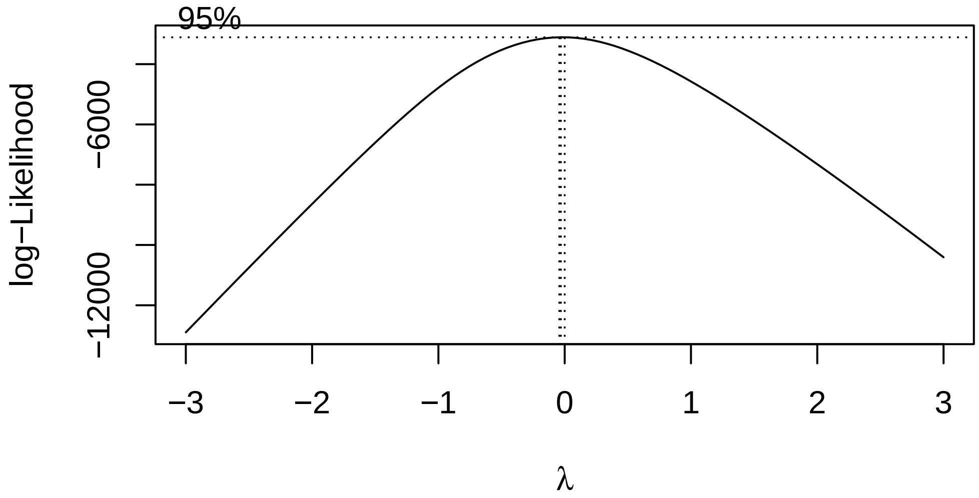
```
heart |>
  ggplot(aes(x = totalcost)) +
  geom_histogram(bins = 40) +
  labs(title = "Histogram of `totalcost`") +
  theme(plot.title = element_text(hjust = 0.5))
```



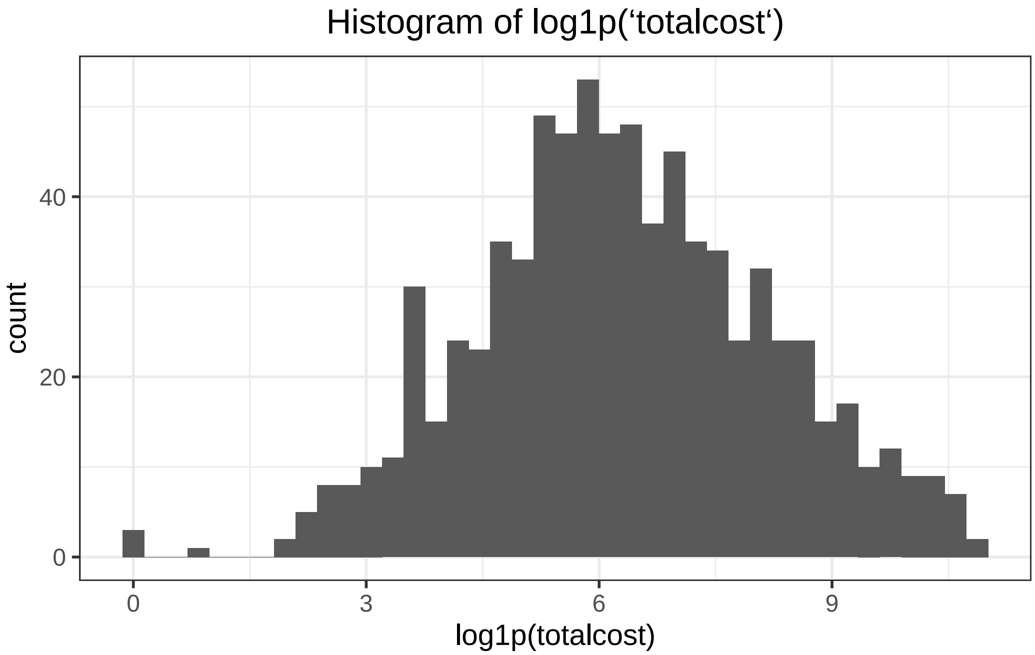
Distribution description: The distribution of `totalcost` is obviously right-skewed.

Use box-cox plot to determine the transformation power:

```
fit = heart |>
  filter(totalcost > 0) |>
  lm(totalcost ~ age, data = _)
MASS::boxcox(fit, lambda = seq(-3, 3, by = 0.25))
```



```
heart |>
  ggplot(aes(x = log1p(totalcost))) +
  geom_histogram(bins = 40) +
  labs(title = "Histogram of log1p(`totalcost`)") +
  theme(plot.title = element_text(hjust = 0.5))
```



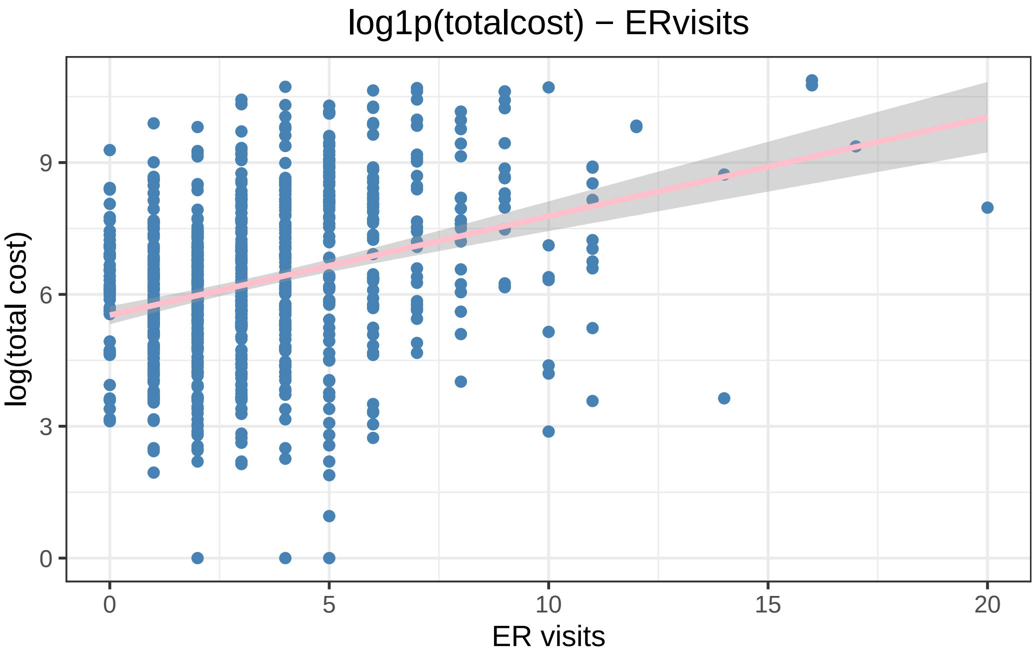
Given the result of `boxcox` plot, I tried to use logarithmic transformation on this variable. (Max likelihood achieved when $\lambda = 0$) After logarithmic transformation, the distribution of transformed `totalcost` is approximately symmetric and subject to normal distribution.

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

```
heart = heart |>
  mutate(comp_bin = factor(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.

```
heart |>
  mutate(log_totalcost = log1p(totalcost)) |>
  ggplot(aes(x = e_rvisits, y = log_totalcost)) +
  geom_point(color = "steelblue") +
  geom_smooth(method = "lm", se = T,
              color = "pink", lwd = 1) +
  labs(
    x = "ER visits", y = "log(total cost)",
    title = "log1p(totalcost) - ERvisits") +
  theme(plot.title = element_text(hjust = 0.5),
        legend.position = "none")
```



```
slr = heart |>
  mutate(log_totalcost = log1p(totalcost)) |>
  lm(log_totalcost ~ e_rvisits, data = _)
summary(slr)
```

```

Call:
lm(formula = log_totalcost ~ e_rvisits, data = mutate(heart,
    log_totalcost = log1p(totalcost)))

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 ***
e_rvisits    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
```

The linear regression model indicates a significant relationship between log-transformed `total cost` and the number of emergency room visits (`ERvisits`). The positive coefficient (0.22529) suggests that, on average, each additional ER visit is associated with 0.225 increase in total cost.

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

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

```
# Interaction effect
heart |>
  mutate(log_totalcost = log1p(totalcost)) |>
  lm(log_totalcost ~ e_rvisits * comp_bin, data = _) |>
  summary()
```

```
Call:
lm(formula = log_totalcost ~ e_rvisits * comp_bin, data = mutate(heart,
    log_totalcost = log1p(totalcost)))

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 ***
e_rvisits           0.20947    0.02490   8.412  < 2e-16 ***
comp_bin1           2.19096    0.55447   3.951 8.47e-05 ***
e_rvisits:comp_bin1 -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
```

From the result, the coefficient of combination term is not significant, indicating that `comp_bin` is not an effect modifier.

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

```r
heart |>
  mutate(log_totalcost = log1p(totalcost)) |>
  lm(log_totalcost ~ e_rvisits + comp_bin, data = _) |>
  summary()
```

```

Call:
lm(formula = log_totalcost ~ e_rvisits + comp_bin, data = mutate(heart, 
    log_totalcost = log1p(totalcost)))

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 ***
e_rvisits    0.20295    0.02405   8.437  < 2e-16 ***
comp_bin1    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
```

```r
heart |>
  mutate(log_totalcost = log1p(totalcost)) |>
  lm(log_totalcost ~ e_rvisits, data = _) |>
  summary()
```

```
Call:
lm(formula = log_totalcost ~ e_rvisits, data = mutate(heart, 
    log_totalcost = log1p(totalcost)))

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 ***
e_rvisits    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
```

From the result, the coefficient of `ERvisit` term decreases after adding `comp_bin` as predictor, indicating that `comp_bin` is a potential confounder.

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

Given that `comp_bin` is confounder, it should be included.

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.

```r
heart |>
  mutate(log_totalcost = log1p(totalcost)) |>
  lm(log_totalcost ~ e_rvisits + age + gender + duration + comp_bin, data = _) |>
  summary()
```

```
Call:
lm(formula = log_totalcost ~ e_rvisits + age + gender + duration + 
    comp_bin, data = mutate(heart, log_totalcost = log1p(totalcost)))
```

```
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 ***
e_rvisits    0.1745975  0.0225736   7.735 3.20e-14 ***
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 ***
comp_bin1    1.5044946  0.2584882   5.820 8.57e-09 ***
---
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
```

The model exhibits high significance, with key predictors `ERvisit`, `duration` and `comp_bin` significantly influencing the outcome variable. Predictor `age` still contribute meaningfully to the model with less impact, while `gender` is with non-significant effect to outcome.

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

I will choose to use MLR model. Because MLR excels in capturing complex relationships between multiple predictors and a response variable, while MLR is more precise given the model residual comparison.