p8130_hw4

2022-11-13

```
library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.
3.2 ──
## ✓ ggplot2 3.4.0      ✓ purrr   0.3.5
## ✓ tibble  3.1.8      ✓ dplyr   1.0.10
## ✓ tidyr   1.2.1      ✓ stringr 1.4.1
## ✓ readr   2.1.3      ✓ forcats 0.5.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflict
s() ──
## ✕ dplyr::filter() masks stats::filter()
## ✕ dplyr::lag()    masks stats::lag()

library(BSDA)

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

library(readxl)
library(arsenal)
library(knitr)
```

Problem 1

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

## Warning in wilcox.test.default(blood_data, mu = 120, alternative = "less")
:
## cannot compute exact p-value with ties

## Warning in wilcox.test.default(blood_data, mu = 120, alternative = "less")
:
## cannot compute exact p-value with zeroes

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

From the Sign test, the test statistic is 10, the p-value is 0.276, which is greater than 0.05. Therefore, we do not have significant evidence to reject the null hypothesis, there is no evidence that the blood sugar readings is less than 120.

From the Wilcoxon signed-rank test, the test statistic is 112.5, the p-value is 0.1447, which is greater than 0.05. Therefore, there is no significant evidence that the blood sugar level is less than 120.

Problem 2

a)

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

The relationship between glia-neuron ratio (denote as GR) and brain mass (denote as BM) is:

$$\widehat{GR} = 0.16370 + 0.18113 \times \ln(BM)$$

the glia-neuron ratio of Homo sapiens should be:

$$\widehat{GR} = 0.16370 + 0.18113 \times 7.22 = 1.471$$

c)

We find that the glia neuron ratio for human is 1.65, which is higher than other species. Therefore, the prediction interval interval for a single new observation is more appropriate since the value of glia neuron ratio for human can be considered as a new value. The predicted mean glia- neuron ratio at the given brain mass can only capture information of the given data.

d)

```
## Warning: `as.tibble()` was deprecated in tibble 2.0.0.
## ℹ Please use `as_tibble()` instead.
## ℹ The signature and semantics have changed, see `?as_tibble`.
```

fit	lwr	upr	category
1.471458	1.036047	1.906869	predict

```
##  glia_neuron_ratio
##  Min.   :0.46
##  1st Qu.:0.64
##  Median :1.02
##  Mean   :0.94
##  3rd Qu.:1.15
##  Max.   :1.22
```

the true value of human brain after log transformation falls in the prediction interval of non-human distribution, thus human brain do not have excessive glia-neuron ratio for its mass

e)

As seen from the plot, we can see that the glia neuron ration for human exceeds other specie's ratio. So the prediction of human from this model may not be appropriate enough.

Problem 3

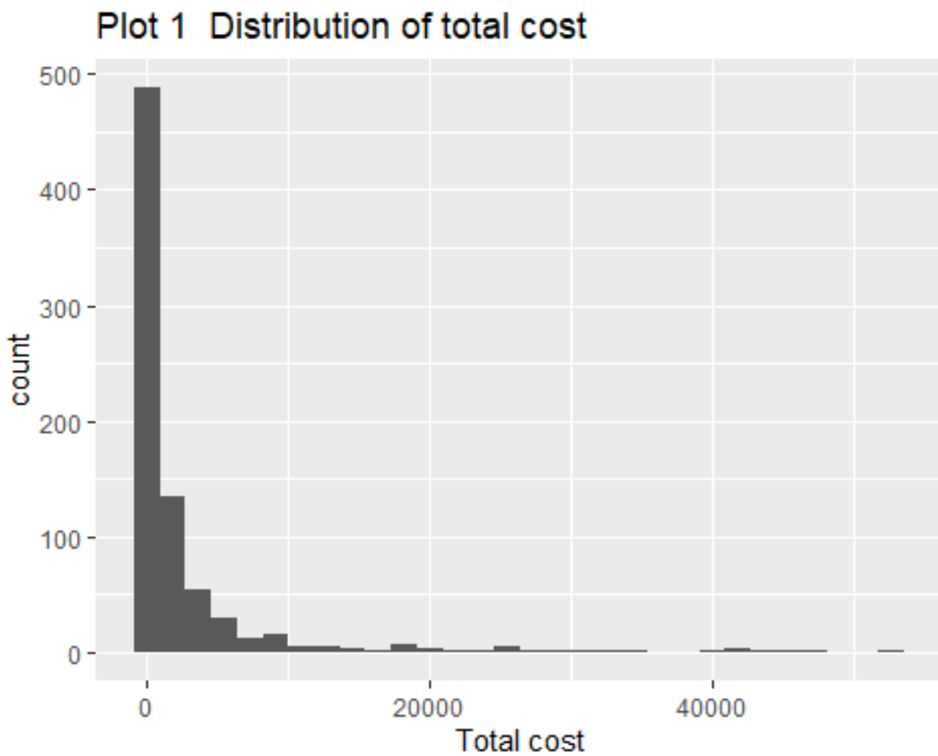
a)

```
##
##
```

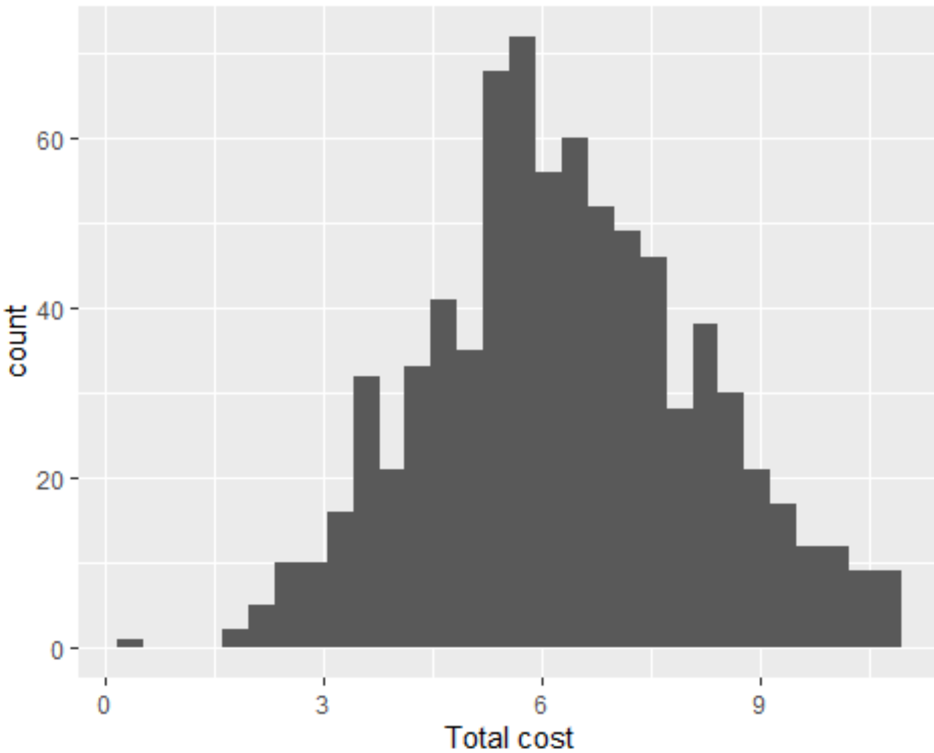
```
## |                   |       Overall (N=788)        |
## |:------------------|:----------------------------:|
## |totalcost          |                              |
## |-  Mean (SD)       |     2799.956 (6690.260)      |
## |-  Median (Q1, Q3) | 507.200 (161.125, 1905.450)  |
## |-  Range           |      0.000 - 52664.900       |
## |age                |                              |
## |-  Mean (SD)       |        58.718 (6.754)        |
## |-  Median (Q1, Q3) |   60.000 (55.000, 64.000)    |
## |-  Range           |       24.000 - 70.000        |
## |gender             |                              |
## |-  Mean (SD)       |        0.228 (0.420)         |
## |-  Median (Q1, Q3) |    0.000 (0.000, 0.000)      |
## |-  Range           |        0.000 - 1.000         |
## |interventions      |                              |
## |-  Mean (SD)       |        4.707 (5.595)         |
## |-  Median (Q1, Q3) |    3.000 (1.000, 6.000)      |
## |-  Range           |       0.000 - 47.000         |
## |drugs              |                              |
## |-  Mean (SD)       |        0.447 (1.064)         |
## |-  Median (Q1, Q3) |    0.000 (0.000, 0.000)      |
## |-  Range           |        0.000 - 9.000         |
## |e_rvisits          |                              |
## |-  Mean (SD)       |        3.425 (2.637)         |
## |-  Median (Q1, Q3) |    3.000 (2.000, 5.000)      |
## |-  Range           |       0.000 - 20.000         |
## |complications      |                              |
## |-  Mean (SD)       |        0.057 (0.248)         |
## |-  Median (Q1, Q3) |    0.000 (0.000, 0.000)      |
## |-  Range           |        0.000 - 3.000         |
## |comorbidities      |                              |
## |-  Mean (SD)       |        3.766 (5.951)         |
## |-  Median (Q1, Q3) |    1.000 (0.000, 5.000)      |
## |-  Range           |       0.000 - 60.000         |
## |duration           |                              |
## |-  Mean (SD)       |      164.030 (120.916)       |
## |-  Median (Q1, Q3) |  165.500 (41.750, 281.000)   |
## |-  Range           |       0.000 - 372.000        |
```

In this dataset, the main outcome is `total cost`. There are 788 rows and 10 variables. Other important covariate includes the age and gender, number of complications that happens during treatment, and duration of treatment condition.From the plot above, the possible important predictors are likely to be `complications`, `drugs` and `ERvisits` and `interventions`.

b)




```
## Warning: Removed 3 rows containing non-finite values (`stat_bin()`).
```



We find that after log transformation on totoal cost, the normality improved.

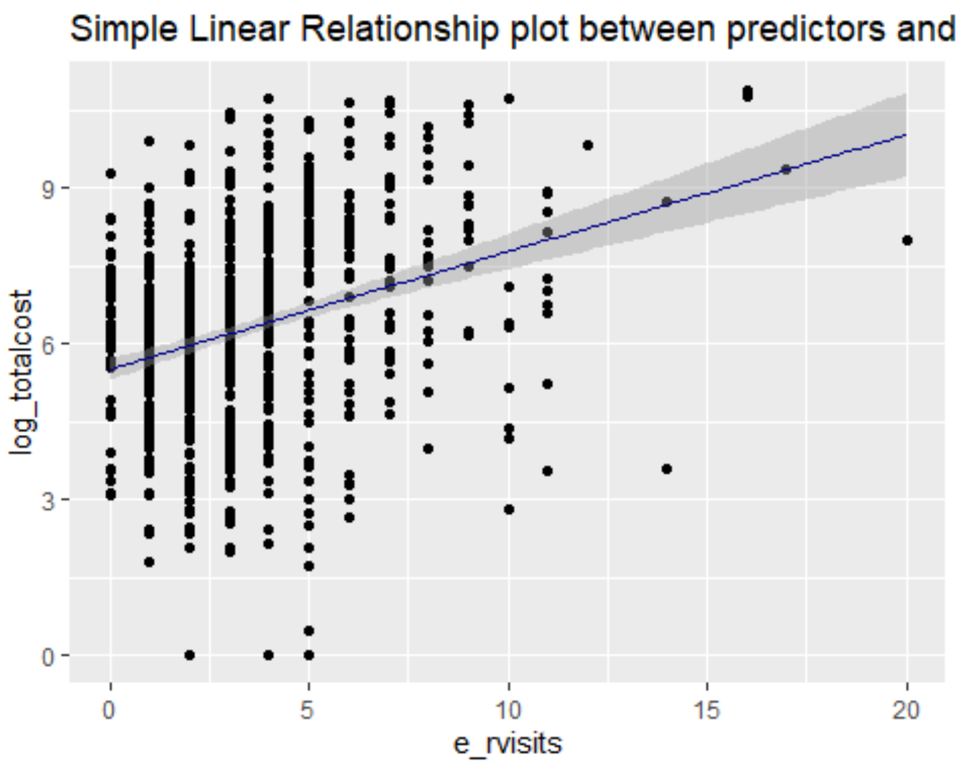
c) Add new variables

```
heart_new =
  heart_log %>%
  mutate(comp_bin =
           case_when(
             complications == 0 ~ "0",
             complications != 0 ~ "1"))
```

d)

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## ℹ Please use `linewidth` instead.

## $x
## [1] "e_rvisits"
##
## $y
## [1] "log(total cost)"
##
## $title
## [1] "Scatter plot of log(total cost) and e_rvisits"
##
## attr(,"class")
## [1] "labels"
```




```
##
## Call:
## lm(formula = log_totalcost ~ e_rvisits, data = heart_new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.6355 -1.1196  0.0371  1.2871  4.3045
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.51704    0.10585  52.123   <2e-16 ***
## e_rvisits    0.22569    0.02449   9.215   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.812 on 786 degrees of freedom
## Multiple R-squared:  0.09751,    Adjusted R-squared:  0.09636
## F-statistic: 84.92 on 1 and 786 DF,  p-value: < 2.2e-16
```

We can see that the p-value is extremely low, so we reject the null hypothesis that there isn't a linear relationship between total cost and number of emergency visits. The intercept represents the expected value of (total cost) after log transformation, in which case number of emergency visits equals to 0; The slope means that when one visit increases, the estimated value of (total cost) after log transformation will increase 0.22569 on average. Based on the regression results, the R^2 of this model is only 0.098, which is quite small, illustrating poor performance on predicting.

e)

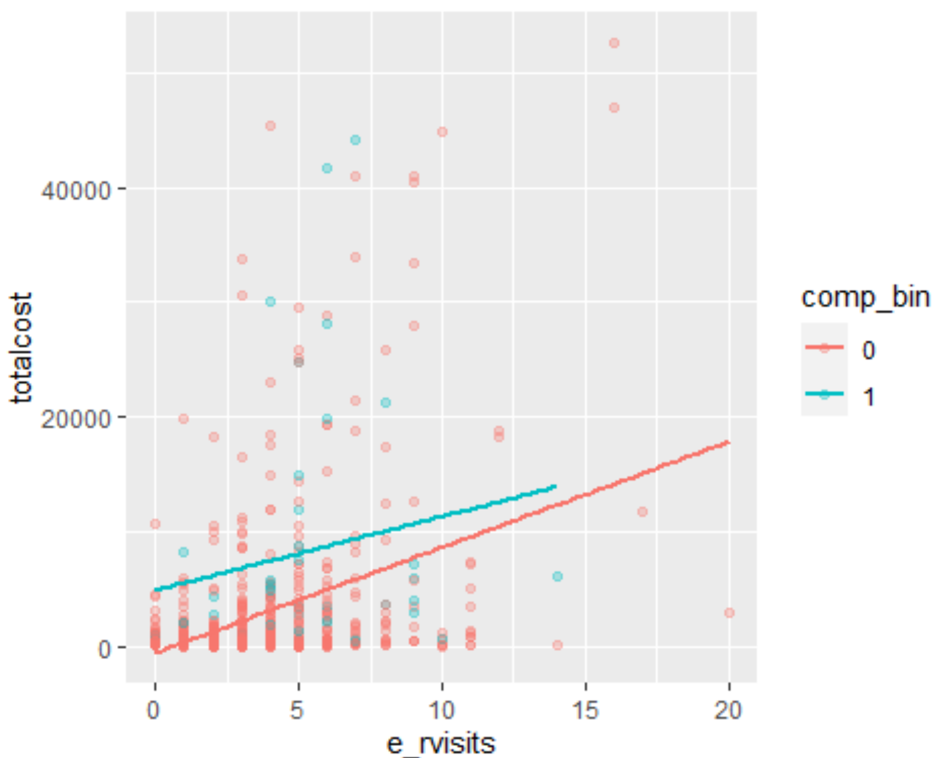
```
##
## Call:
## lm(formula = log_totalcost ~ e_rvisits + comp_bin, data = heart_new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.5066 -1.0745 -0.0009  1.1930  4.4109
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.50043    0.10353  53.129  < 2e-16 ***
## e_rvisits    0.20324    0.02423   8.389 2.27e-16 ***
## comp_bin1    1.71348    0.28115   6.094 1.72e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.772 on 785 degrees of freedom
## Multiple R-squared:  0.1383, Adjusted R-squared:  0.1361
## F-statistic: 62.98 on 2 and 785 DF,  p-value: < 2.2e-16
```

i)

```
##
## Call:
## lm(formula = log_totalcost ~ factor(comp_bin) + e_rvisits + factor(comp_bi
n) *
##     e_rvisits, data = heart_new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.5176 -1.0797  0.0104  1.2075  4.4065
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 5.47866    0.10576  51.805  < 2e-16 ***
## factor(comp_bin)1           2.20002    0.55846   3.939 8.89e-05 ***
## e_rvisits                   0.20978    0.02508   8.364 2.76e-16 ***
## factor(comp_bin)1:e_rvisits -0.09780    0.09699  -1.008    0.314
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.772 on 784 degrees of freedom
## Multiple R-squared:  0.1394, Adjusted R-squared:  0.1361
## F-statistic: 42.33 on 3 and 784 DF,  p-value: < 2.2e-16

##
## Call:
## lm(formula = log_totalcost ~ factor(comp_bin) * e_rvisits, data = heart_ne
w)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -6.5176 -1.0797  0.0104  1.2075  4.4065 
## 
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                  5.47866    0.10576  51.805  < 2e-16 ***
## factor(comp_bin)1            2.20002    0.55846   3.939 8.89e-05 ***
## e_rvisits                    0.20978    0.02508   8.364 2.76e-16 ***
## factor(comp_bin)1:e_rvisits -0.09780    0.09699  -1.008    0.314    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.772 on 784 degrees of freedom
## Multiple R-squared:  0.1394, Adjusted R-squared:  0.1361 
## F-statistic: 42.33 on 3 and 784 DF,  p-value: < 2.2e-16
```



From the plot we can see that the slope of e_rvisits change quite bit for different comp_bin, there might be an interaction between e_rvisits and comp_bin. From the above summary, the model with the term “comp_bin*e_rvisits”, we fail to reject the null hypothesis that the coefficient of comp_bin*e_rvisits is 0, therefore, the interaction effect is not significant. So the comp_bin is not a modifier.

ii)

When adding comp_bin into the model, the coefficient of e_rvisits decrease from 0.22569 to 0.20978, it decreases about 10% , so binary complication variable is a counfounder of association between number of emergency visits and total cost.

iii)

```
## Analysis of Variance Table
##
## Response: log_totalcost
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## e_rvisits   1  278.85 278.845  88.825 < 2.2e-16 ***
## comp_bin    1  116.60 116.601  37.143  1.72e-09 ***
## Residuals 785 2464.33   3.139
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Analysis of Variance Table
##
## Model 1: log_totalcost ~ e_rvisits
## Model 2: log_totalcost ~ e_rvisits + comp_bin
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1    786 2580.9
## 2    785 2464.3  1     116.6 37.143 1.72e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Total cost of comp_bin is significantly different. As a confounder, should be considered when finding the relationship between e_rvisits and total cost.

Small model:

$$\widehat{\text{total cost}} = \beta_0 + \beta_1 \cdot e\_rvisits$$

large model:

$$\widehat{\text{Totalcost}} = \beta_0 + \beta_1 \cdot e\_rvisits + \beta_2 \cdot comp\_bin$$

Hypothesis:

$H_0: \beta_2 = 0$

$H_a: \beta_2 \neq 0$

$$F_{test} = \frac{(SSE_L - SSE_S)/(df_L - df_S)}{\frac{SSE_L}{df_L}} = \frac{(2464.3-2580.9)/(-1)}{\frac{2580.9}{786}} = 37 \sim F_{1,785}$$

$F_{test} > F_{1,785}$

$\therefore$ We reject the null hypothesis, so at least one coefficient of age, gender and duration is not 1.
We should choose the large model

f)

```
##
## Call:
## lm(formula = log_totalcost ~ e_rvisits + age + gender + duration +
##     comp_bin, data = heart_new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.4442 -1.0367 -0.1109  0.9506  4.3478
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.9378825  0.5138860  11.555  < 2e-16 ***
## e_rvisits    0.1746131  0.0227275   7.683 4.66e-14 ***
## age         -0.0208988  0.0087337  -2.393    0.017 *
## gender      -0.2073611  0.1396457  -1.485    0.138
## duration     0.0057684  0.0004922  11.720  < 2e-16 ***
## comp_bin1    1.5102874  0.2602503   5.803 9.45e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.635 on 782 degrees of freedom
## Multiple R-squared:  0.269,  Adjusted R-squared:  0.2644
## F-statistic: 57.56 on 5 and 782 DF,  p-value: < 2.2e-16
```

```
## Analysis of Variance Table
##
## Response: log_totalcost
##            Df  Sum Sq Mean Sq  F value    Pr(>F)
## e_rvisits   1  278.85  278.85 104.3130 < 2.2e-16 ***
## age         1    3.46    3.46   1.2947    0.2555
## gender      1    5.02    5.02   1.8762    0.1712
## duration    1  392.02  392.02 146.6502 < 2.2e-16 ***
## comp_bin    1   90.02   90.02  33.6773  9.45e-09 ***
## Residuals 782 2090.41    2.67
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Analysis of Variance Table
##
## Model 1: log_totalcost ~ e_rvisits
## Model 2: log_totalcost ~ e_rvisits + age + gender + duration + comp_bin
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    786 2580.9
## 2    782 2090.4  4    490.52 45.875 < 2.2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Call:
## lm(formula = log_totalcost ~ e_rvisits, data = heart_new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.6355 -1.1196  0.0371  1.2871  4.3045
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.51704    0.10585  52.123   <2e-16 ***
## e_rvisits    0.22569    0.02449   9.215   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.812 on 786 degrees of freedom
## Multiple R-squared:  0.09751,    Adjusted R-squared:  0.09636
## F-statistic: 84.92 on 1 and 786 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = log_totalcost ~ e_rvisits + age + gender + duration +
##     comp_bin, data = heart_new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.4442 -1.0367 -0.1109  0.9506  4.3478
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.9378825  0.5138860  11.555  < 2e-16 ***
## e_rvisits    0.1746131  0.0227275   7.683 4.66e-14 ***
## age         -0.0208988  0.0087337  -2.393    0.017 *
## gender      -0.2073611  0.1396457  -1.485    0.138
## duration     0.0057684  0.0004922  11.720  < 2e-16 ***
## comp_bin1    1.5102874  0.2602503   5.803 9.45e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.635 on 782 degrees of freedom
## Multiple R-squared:  0.269,  Adjusted R-squared:  0.2644
## F-statistic: 57.56 on 5 and 782 DF,  p-value: < 2.2e-16
```

small model:

$$\widehat{totalcost} = \beta_0 + \beta_1 \cdot e\_rvisits$$

large model:

$$\widehat{totalcost} = \beta_0 + \beta_1 \cdot e\_rvisits + \beta_2 \cdot age + \beta_3 \cdot gender + \beta_4 \cdot duration + \beta_5 \cdot comp\_bin$$

Hypothesis:

$H_0: \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$

$H_a: \beta_2 \neq 0$ or $\beta_3 \neq 0$ or $\beta_4 \neq 0$ or $\beta_5 \neq 0$

$$F_{test} = \frac{(SSE_l - SSE_s)/(df_l - df_s)}{\frac{SSE_l}{df_l}}$$

$$= \frac{(7090.4 - 2580.9)/(1-4)}{\frac{2580.9}{786}}$$

$$= 46 \sim F_{4,782}$$

The Ftest is very large, thus we reject the null hypothesis, which means at least one coefficient of age, gender, and duration is not 1.

$\therefore$ We should choose the large model.

Additionally, the summary of the regression above, the adjusted R^2 is 0.26, which is greater than the small model (0.096).

This means, when adjusting other covariates, the model performs better than just considering the number of emergency room as the only predictor.

Therefore, we should choose the large model.