P8130-HW4

Yuqi Miao

11/13/2019

Problem 3

a) Fit a regression model for the non-human data

```
##
## 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
```

As shown in the summary table, the coefficient of brain mass after log transformation is significantly different from 0, with adjusted R^2 equal to 60.25%.

b) Predict human brain mass using the model

```
## # A tibble: 2 x 5
##   term          estimate std.error statistic  p.value
##   <chr>            <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)      0.164    0.160       1.02 0.322
## 2 ln_brain_mass    0.181    0.0360      5.03 0.000151
```

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

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

Thus using this relationship, the glia-neuron ratio of Homo sapiens should be:

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

c) CI or PI?

In this model, Homo sapiens species provides a `new value` for the model which beyond the original range of glia-neuron ratio, which means that in this way, the prediction interval is more plausible because the expected prediction can only capture the information of the given data, which is narrower than prediction interval.

d) Construct PI

Prediction interval of glia_neuron ratio of Homo sapiens.

$$se(\widehat{\beta_0} + \widehat{\beta_1} X_h) = \sqrt{MSE\left\{\frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2} + 1\right\}} = \sqrt{0.02885 \times (\frac{1}{17} + \frac{(7.22 - \overline{X})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}) + 1}$$

fit	lwr	upr	category
1.471458	1.036047	1.906869	predict

summary of the non-human species data.

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

As shown above, although the mean value of other species smaller than for Homo sapiens, the true value of human brain after log transformation (which is 1.65) falls in the prediction interval of non-human distribution, thus human brains do not have excessive glia-neuron ratio for its mass.

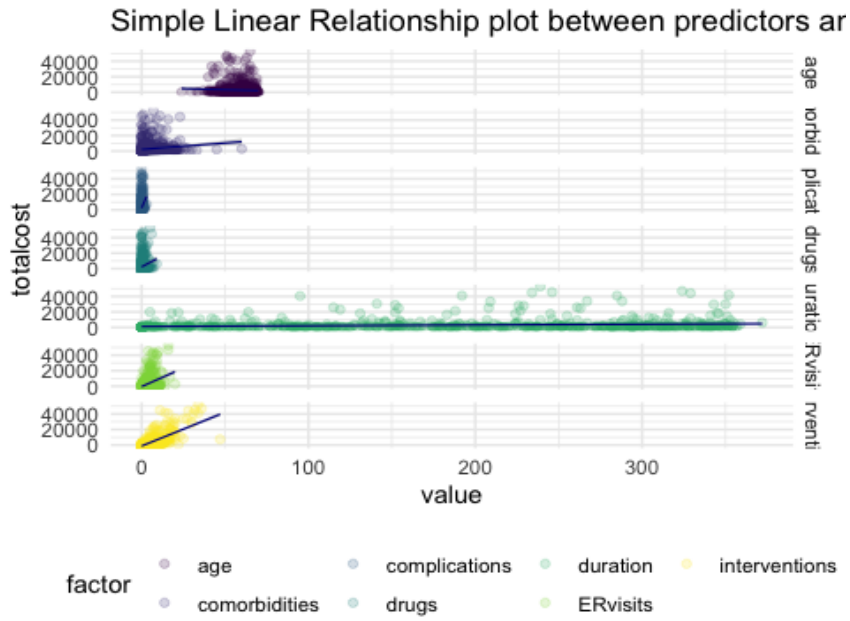
e) other notions.

1. As shown in the plot, the $ln(BM)$ for Homo sapiens exceeds the range of non_human species which fitted the model. In this case, the prediction of human beings from this model may not defensible enough.

Problem 4

a) Description

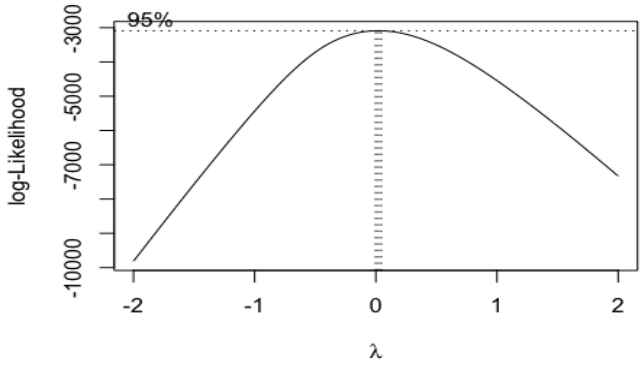
	female (N=608)	male (N=180)
totalcost		
- Mean (SD)	2867.3 (6867.2)	2572.3 (6067.1)
- Median (Q1, Q3)	502.0 (161.1, 1975.1)	546.2 (176.4, 1881.4)
- Min - Max	5.6 - 52664.9	0.0 - 44162.8
age		
- Mean (SD)	58.8 (6.7)	58.5 (6.9)
- Median (Q1, Q3)	60.0 (55.0, 64.0)	60.0 (54.0, 64.0)
- Min - Max	24.0 - 70.0	39.0 - 70.0
interventions		
- Mean (SD)	4.6 (5.6)	5.0 (5.6)
- Median (Q1, Q3)	3.0 (1.0, 6.0)	3.0 (1.0, 7.0)
- Min - Max	0.0 - 47.0	0.0 - 34.0
drugs		
- Mean (SD)	0.4 (1.0)	0.5 (1.1)
- Median (Q1, Q3)	0.0 (0.0, 0.0)	0.0 (0.0, 1.0)
- Min - Max	0.0 - 9.0	0.0 - 7.0
ERvisits		
- Mean (SD)	3.3 (2.5)	4.0 (3.0)
- Median (Q1, Q3)	3.0 (2.0, 4.0)	3.0 (2.0, 5.0)
- Min - Max	0.0 - 17.0	0.0 - 20.0
complications		
- Mean (SD)	0.1 (0.2)	0.1 (0.3)
- Median (Q1, Q3)	0.0 (0.0, 0.0)	0.0 (0.0, 0.0)
- Min - Max	0.0 - 1.0	0.0 - 3.0
comorbidities		
- Mean (SD)	3.9 (6.1)	3.3 (5.4)
- Median (Q1, Q3)	2.0 (0.0, 5.0)	1.0 (0.0, 4.0)
- Min - Max	0.0 - 60.0	0.0 - 28.0
duration		
- Mean (SD)	162.5 (121.3)	169.3 (119.7)
- Median (Q1, Q3)	164.5 (40.0, 281.0)	168.0 (59.2, 283.5)
- Min - Max	0.0 - 372.0	0.0 - 351.0

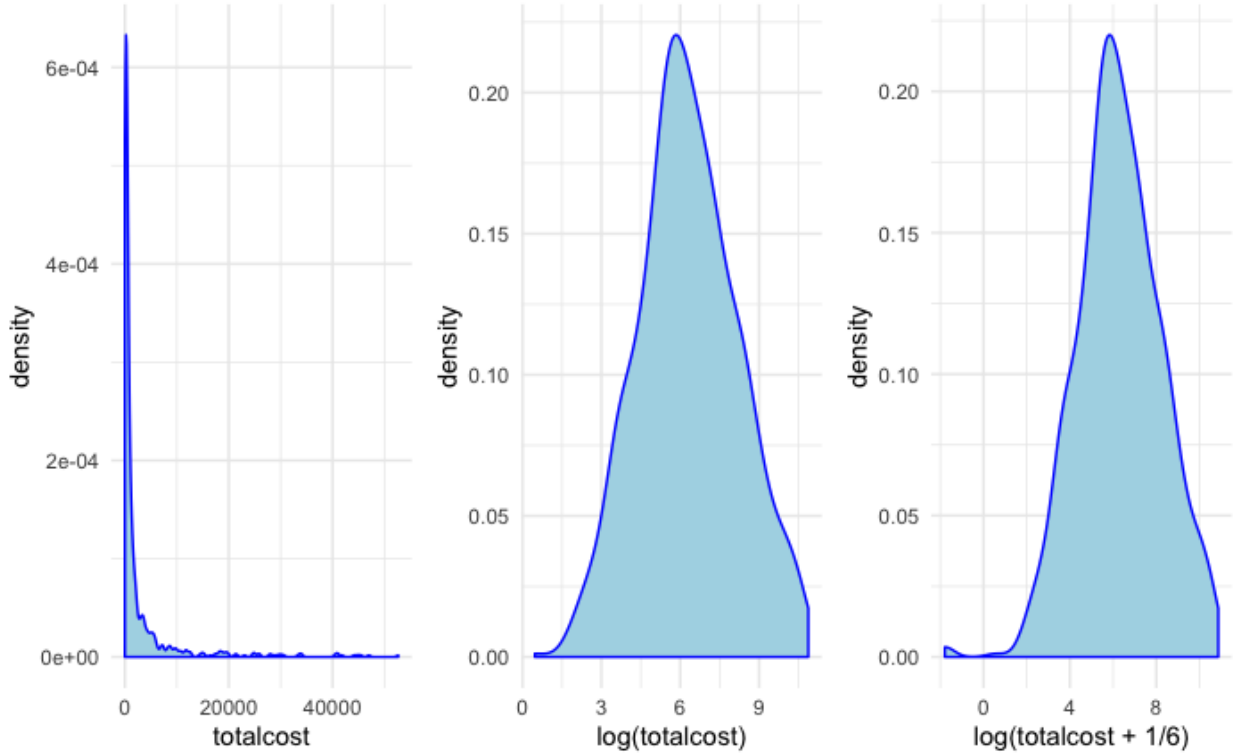


In this dataset, the main outcome is `total cost`, which is the total cost (in dollars) of patients diagnosed with heart disease, the main predictor is `ERvisits`, which is the number of emergency room visits for every observation. Other important covariate including the age and gender of the subscriber(indicated by `age`,`gender`), number of complications that arouse during treatment(indicated by `complications`) and duration of treatment condition(indicated by `duration`). According to the plot above, The possible important predictors are likely to be `complications`, `drugs` and `ERvisits` and `interventions`.

distribution of total cost

```
## # A tibble: 3 x 10
##      id totalcost   age gender interventions drugs ERvisits complications
##   <dbl>     <dbl> <dbl> <fct>          <dbl> <dbl>    <dbl>         <dbl>
## 1    96         0    55 male               0     0        5             0
## 2   370         0    65 male               1     1        4             0
## 3   402         0    44 male               1     0        2             0
## # … with 2 more variables: comorbidities <dbl>, duration <dbl>
```



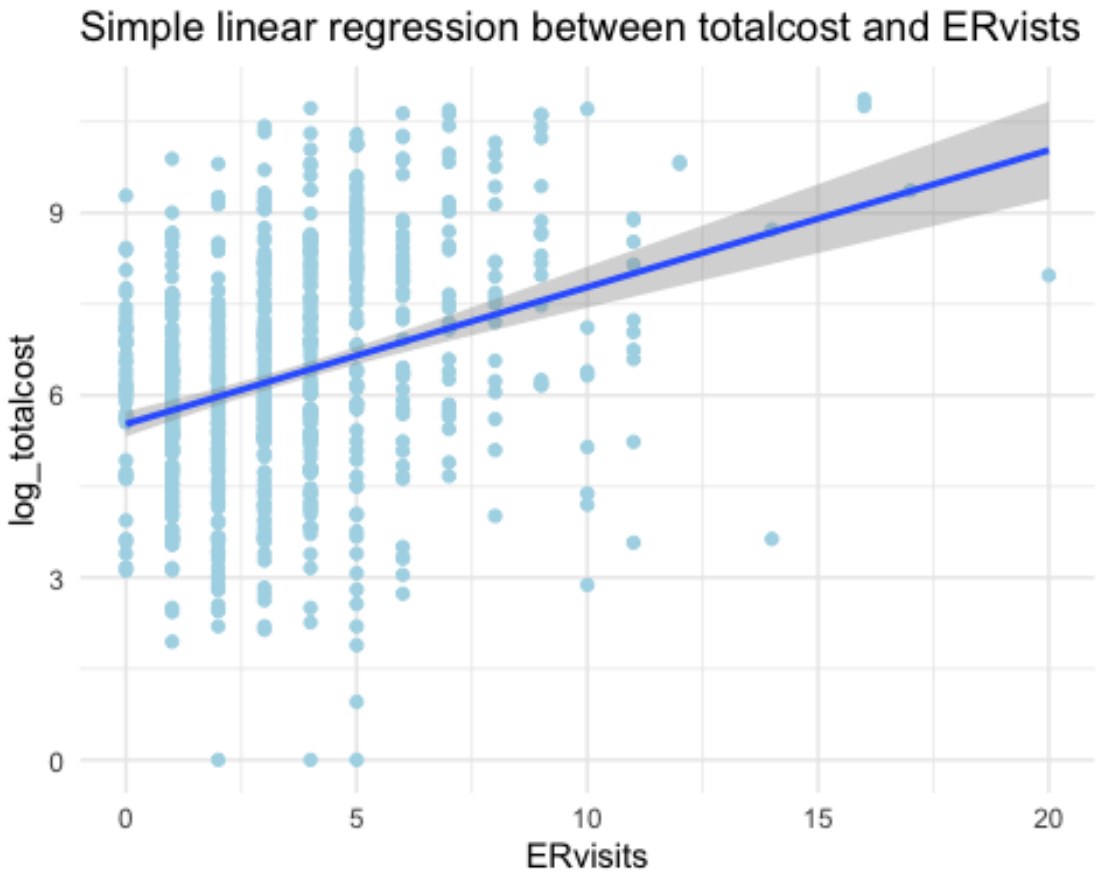


As shown above, because there are 3 0s in the data, when we perform log trasformation, these data would go to inf. So we add 1 to the totalcost variable to make it meaningful. Based on the results of Box-cox transformations, taking log transformation after adding 1 to the total cost makes the distribution approach normality.As shown above, after adding 1 (adjust for the 0 cases) and then take log transformation to the totalcost, the density plot plot shows a great symmetry.

c) add a binary variable.

d)

```
##
## Call:
## lm(formula = log_totalcost ~ ERvisits, data = hd)
##
## 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
```



Fitted model:

$$\widehat{Totalcost} = 5.52674 + 0.22529 \times ERvisits$$

significance test:

* Hypothesis:

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

* Test statistics:

$$t_{test} = \frac{\beta_1}{se(\beta_1)} = \frac{0.22529}{0.02432} = 9.26 \overset{H_0}{\sim} t_{786}$$

- results $t_{test} > t_{786,0.975} = 1.96$, reject the null hypothesis, which means that the coefficient of number of emergency room visits is significantly different different from 0.

Interpretation:

With extremely low p-value, 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 + 1) after log transformation at the baseline, in which case number of emergency visits equals to 0; The slope means that when one visit increases, the estimated value of (total cost + 1) after log transformation will increase 0.22529 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 ~ ERvisits + comp_bin, data = hd)
##
## 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_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
```

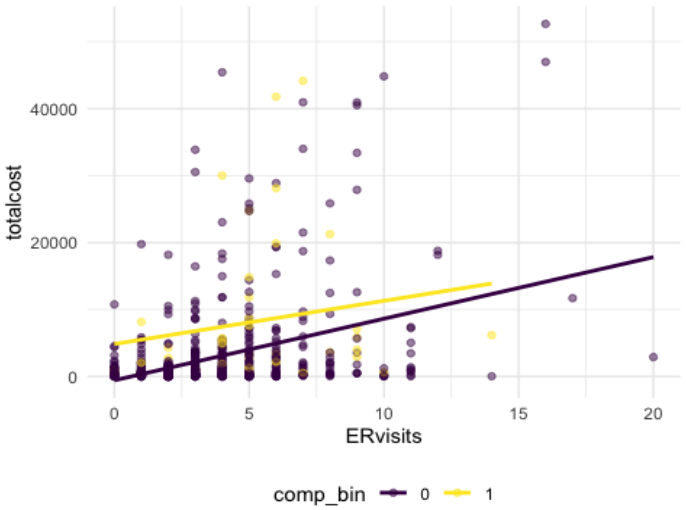
```
## Multiple R-squared:  0.1394, Adjusted R-squared:  0.1372
## F-statistic: 63.57 on 2 and 785 DF,  p-value: < 2.2e-16
```

counfounder?

When add comp_bin into the model, the coefficient of ERvisits decrease from 0.22529 to 0.20295, the decrease rate is approximately 10%, so binary complication variable is a counfounder of association between number of emergency visits and total cost.

effect modifier test

```
##
## Call:
## lm(formula = log_totalcost ~ factor(comp_bin) + ERvisits + factor(comp_bin
) *
##     ERvisits, data = hd)
##
## 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 ***
## factor(comp_bin)1             2.19096    0.55447   3.951 8.47e-05 ***
## ERvisits                      0.20947    0.02490   8.412  < 2e-16 ***
## factor(comp_bin)1:ERvisits -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
```



Showing in the plot, the slope of ERvisits is change slightly in differnt categories of comp_bin, which meand there might be interactions between comp_bin and ERvistis. When itegrating interaction term(comp_bin*ERvisits) in to the model,we fail to reject the null hypothesis that the coeffeicient of comp_bin*ERvisits term is 0, so the interaction effect is not significant. In this way, the binary complication variable is a confounder but not an modifier to the association between number of emergency visits and total cost.

include binary complication variable?

global test for binary complication variable:

```
## Analysis of Variance Table
##
## Response: log_totalcost
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## ERvisits    1  277.87 277.870  89.792 < 2.2e-16 ***
## comp_bin    1  115.55 115.549  37.339 1.563e-09 ***
## Residuals 785 2429.27   3.095
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As seen in global anova test, total cost of different categories of binary complication variable is significantly differenct, and it is also a confounder that should be considered when finding the relationship between number of emergency visits and total cost.

F-test for the parameter

```
## Analysis of Variance Table
##
## Model 1: log_totalcost ~ ERvisits
## Model 2: log_totalcost ~ ERvisits + comp_bin
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    786 2544.8
## 2    785 2429.3  1    115.55 37.339 1.563e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Small model:

$$\widehat{Totalcost} = \beta_0 + \beta_1 \times ERvisits$$

large model

$$\widehat{Totalcost} = \beta_0 + \beta_1 \times ERvisits + \beta_2 \times comp\_bin$$

- Hypothesis:

$$H_0: \beta_2 = 0$$

$$H_1: \beta_2 \neq 0$$

* Test statistics:

$$F_{test} = \frac{(SSE_l - SSE_s)/(df_l - df_s)}{\frac{SSE_l}{df_l}} = \frac{(2429.3 - 2544.8)/(785 - 786)}{\frac{2544.82}{786}} = 37.339 \overset{H_0}{\sim} F_{1,786}$$

- results $F_{test} > F_{1,786,0.975} = 5.04$, reject the null hypothesis, which means at least one coeeficient of age, gender and duration isn't equal to 1, thus we choose the large model.

Above all, we should take the compication bianary variable into the model.

f)

Small model:

$$\widehat{Totalcost} = \beta_0 + \beta_1 \times ERvisits$$

large model

$$\widehat{Totalcost} = \beta_0 + \beta_1 \times ERvisits + \beta_2 \times age + \beta_3 \times gender + \beta_4 \times duration + \beta_5 \times comp\_bin$$

```
##
## Call:
## lm(formula = log_totalcost ~ ERvisits + age + gender + duration +
##     comp_bin, data = hd)
##
## 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 ***
## age         -0.0206475  0.0086746  -2.380   0.0175 *
## gendermale  -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

## Analysis of Variance Table
##
## Response: log_totalcost
##            Df  Sum Sq Mean Sq  F value    Pr(>F)
## ERvisits    1  277.87  277.87 105.3703 < 2.2e-16 ***
## age         1    3.36    3.36   1.2758    0.2590
```

```
## gender      1    4.99    4.99   1.8932    0.1692
## duration    1  384.93  384.93 145.9677 < 2.2e-16 ***
## comp_bin    1   89.34   89.34  33.8766 8.566e-09 ***
## Residuals 782 2062.20    2.64
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As shown in the summary table of regression, the adjusted R^2 is 0.2647, and the fitted model is

$$\widehat{Totalcost} = 5.94 + 0.17 \times ERvisits - 0.02 \times age - 0.21 \times gender + 0.01 \times duration + 1.50 \times comp\_bin$$

Shown in the global anova table of this linear model, the age and gender are not showing significantly different partial variance, but ERvisits, duration, and comp_bin show significant partial variance.

- Hypothesis:

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

$$H_1: \beta_2 \neq 0 \ or \ \beta_3 \neq 0 \ or \ \beta_4 \neq 0 \ or \ \beta_5 \neq 0$$

* Test statistics:

$$F_{test} = \frac{(SSE_l - SSE_s)/(df_l - df_s)}{\frac{SSE_l}{df_l}} = \frac{(2062.20 - 2544.82)/(782 - 786)}{\frac{2544.82}{786}} = 45.753 \overset{H_0}{\sim} F_{4,786}$$

- results
  $F_{test} > F_{4,786,0.975} = 2.802079$, reject the null hypothesis, which means at least one coeeficient of age, gender and duration isn't equal to 1, thus we choose the large model.

Regression summary table for large model:

```
## 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 ***
## age         -0.0206475  0.0086746  -2.380   0.0175 *
## gendermale  -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
```

According to the regression results as above, the adjusted R^2 for large model is 0.26, which is bigger than the small model. In this way, by adjusting other covariates, the model performs better than just considering number of emergency room visits as predictor. As shown above, we choose the large model.