p8130_hw4_xj2249

xj2249

11/12/2019

Problme1

a)

The Least Squares line equation:

$$\hat{Y} = \hat{\beta_0} + \hat{\beta_1} X_i$$

Proof: According to method of Least Squares, $\hat{\beta_0} = \bar{Y} - \hat{\beta_1}\bar{X}$, therefore, when X = $\bar{X}$

$$\begin{aligned}\hat{Y} &= \hat{\beta_0} + \hat{\beta_1}\bar{X} \\ &= \bar{Y} - \hat{\beta_1}\bar{X} + \hat{\beta_1}\bar{X} \\ &= \bar{Y}\end{aligned}$$

Therefore, it always goes through the point $(\bar{X}, \bar{Y})$.

b)

First, there are some useful properties of LS:

$$\begin{aligned}&\sum e_i = 0 \\ &\sum X_i e_i = 0 \\ &\sum \hat{Y}_i e_i = 0\end{aligned}$$

Proof: To minimize Q, we have:

$$\begin{aligned}\frac{\partial Q}{\partial \beta_0} &= 2\sum_i (Y_i - \beta_0 - \beta_1 X_i) \cdot (-\beta_0)' \\ &= -2\sum_i (Y_i - \beta_0 - \beta_1 X_i) = 0 \\ \frac{\partial Q}{\partial \beta_1} &= -2\sum_i (Y_i - \beta_0 - \beta_1 X_i) \cdot X_i = 0\end{aligned}$$

Then these two equation imply that $\sum e_i = 0$ and $\sum X_i e_i = 0$

$$\begin{aligned}\sum_i \hat{Y}_i e_i &= \sum_i \left(\hat{\beta}_0 + \hat{\beta}_1 X_i\right) e_i \\ &= \hat{\beta}_0 \sum_i e_i + \hat{\beta}_1 \sum_i e_i X_i \\ &= 0\end{aligned}$$

Therefore,

$$\begin{aligned}
corr(e_i, \hat{Y}_i) &= corr(Y_i - \hat{Y}_i, \hat{Y}_i) \\
&= corr(Y_i, \hat{Y}_i) - corr(\hat{Y}_i, \hat{Y}_i) \\
&= \frac{1}{n}\sum (Y_i - \bar{Y})(\hat{Y}_i - \bar{\hat{Y}}) - \frac{1}{n}\sum (\hat{Y}_i - \bar{\hat{Y}})^2 \\
&= \frac{1}{n}\sum (Y_i - \bar{Y})(\hat{Y}_i - \bar{Y}) - \frac{1}{n}\sum (\hat{Y}_i - \bar{Y})^2 \\
&= \frac{1}{n}\sum (Y_i - \bar{Y})(\hat{Y}_i - \bar{Y}) - (\hat{Y}_i - \bar{Y})^2 \\
&= \frac{1}{n}\sum \hat{Y}_i(Y_i - \hat{Y}_i) \\
&= \frac{1}{n}\sum \hat{Y}_i e_i \\
&= 0
\end{aligned}$$

One of the possible explanations is that the variance of ϵ is not constant.

Problem2

a)

$$\begin{aligned}
Q &= (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\
&= \mathbf{Y}'\mathbf{Y} - \boldsymbol{\beta}'\mathbf{X}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \\
&= \mathbf{Y}'\mathbf{Y} - 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{Y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}
\end{aligned}$$

Therefore, let

$$\frac{\partial}{\partial \beta}(Q) = -2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = 0$$

then

$$\underset{2\times1}{\hat{\boldsymbol{\beta}}} = \underset{2\times2}{(\mathbf{X}'\mathbf{X})}^{-1}\underset{2\times1}{\mathbf{X}'\mathbf{Y}}$$

$$\begin{aligned}
E(\hat{\boldsymbol{\beta}}) &= E((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}) \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{Y}) \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'X\beta \\
&= \boldsymbol{I}\boldsymbol{\beta} \\
&= \boldsymbol{\beta}
\end{aligned}$$

b)

$$
\begin{aligned}
\mathrm{Var}(\tilde{\beta}) &= \mathrm{Var}\left(\left(X'X\right)^{-1}X'Y\right)\\
&= \left(X'X\right)^{-1}X'\sigma^2 IX\left(X'X\right)^{-1}\\
&= \sigma^2 I\left(X'X\right)^{-1}X'X\left(X'X\right)^{-1}\\
&= \sigma^2\left(X'X\right)^{-1}\\
&= \sigma^2\begin{bmatrix} n & \sum_{i=1}^{n} X_i \\ \sum_{i=1}^{n} X_i & \sum_{i=1}^{n} X_i^2 \end{bmatrix}^{-1}\\
&= \sigma^2\begin{bmatrix} \frac{1}{n}+\frac{\bar{X}^2}{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2} & \frac{-\bar{X}}{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2} \\ \sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2 & \frac{1}{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2} \end{bmatrix}
\end{aligned}
$$

Problem3

a) Regression model for the nonhuman data

```r
lm_non_human <-
        brain_df %>%
        filter(species != "Homo sapiens") %>%
        lm(glia_neuron_ratio ~ ln_brain_mass, data = .)
summary(lm_non_human)
```

```
##
## 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
```

Therefore, the regression equation is:

$$\hat{Y} = 0.164 + 0.181\ln(\text{brain mass})$$

b)

$$
\begin{aligned}
\hat{Y}_{human} &= 0.164 + 0.181 \ln(\text{brain mass}) \\
&= 0.164 + 0.181 \times 7.22 \\
&= 1.471
\end{aligned}
$$

Therefore, the predicted glia-neuron ratio for humans is 1.471

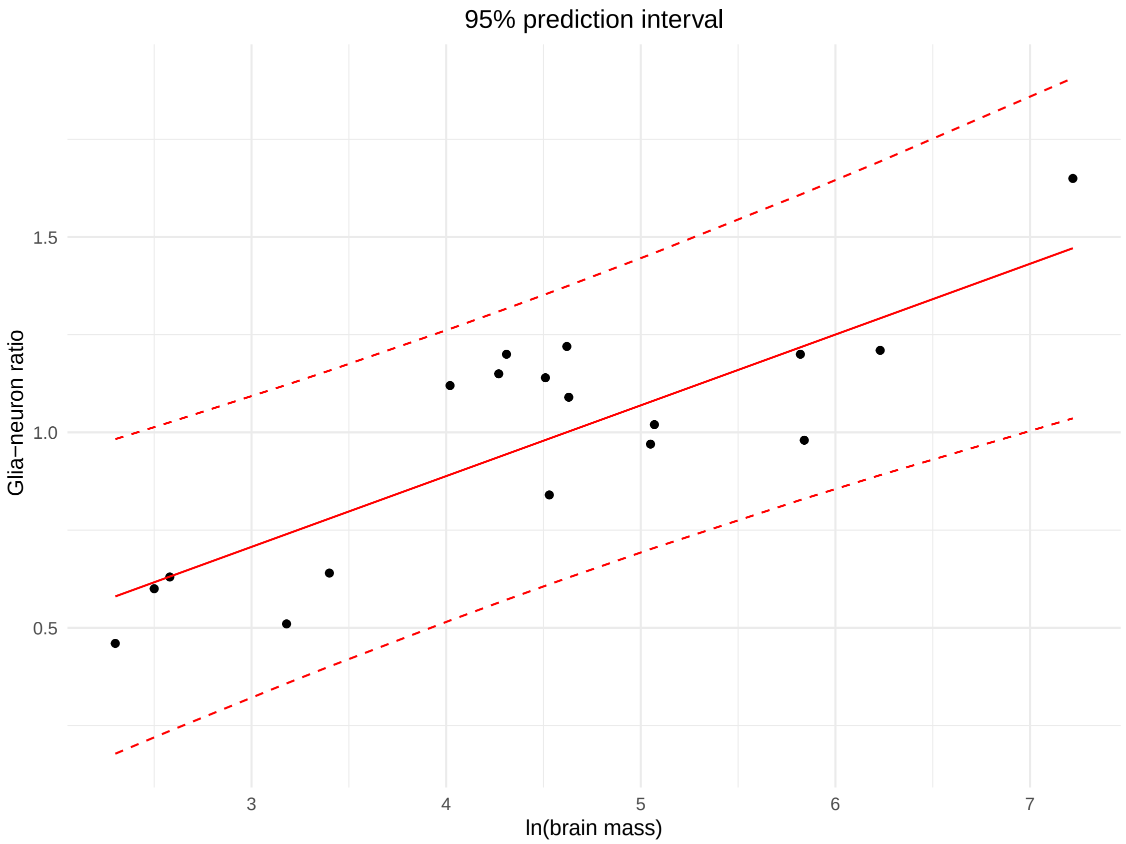
c)

The most-plausible range of values for the prediction is (2.30,6.23), which is the range of ln(brain mass) for non-human. An interval for the prediction of a single new observation is more relevant for our prediction of human glia-neuron ratio, because in the case, we want to predict the value of a new observation(human data).

d)

```r
pred_pi = predict.lm(lm_non_human,brain_df,interval = "prediction")

# prediction interval
brain_df <- cbind(brain_df,pred_pi)
brain_df %>%
        ggplot(aes(y = glia_neuron_ratio, x = ln_brain_mass)) +
        geom_point() +
        geom_line(aes(y = fit), color = "red" ) +
        geom_line( aes(y = lwr), color = "red", linetype = "dashed" ) +
        geom_line( aes(y = upr), color = "red", linetype = "dashed") +
        labs( x = "ln(brain mass)", y = " Glia-neuron ratio", title = "95% prediction interval")
```



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

$$\text{se}\left(\widehat{\beta_0} + \widehat{\beta_1} X_h\right) = \sqrt{MSE\left\{\frac{1}{n} + \left[\left(X_h - \bar{X}\right)^2 / \sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right] + 1\right\}}$$

Therefore, 95% prediction interval for huamn is (1.04,1.91). As we can see, the huamn data point falls within the prediction interval for its given ln(brain mass), therefore, human brain doesn't have an excessive glia-neuron ratio for its mass compared with other primates

e)

As we can see in the graph, the human data point falls outside the scope of non-human data points, from which we generate the regression line. Therefore, though it's not far away from the scope, we may still need to be cautious in using the regression line to predict.

Problem4

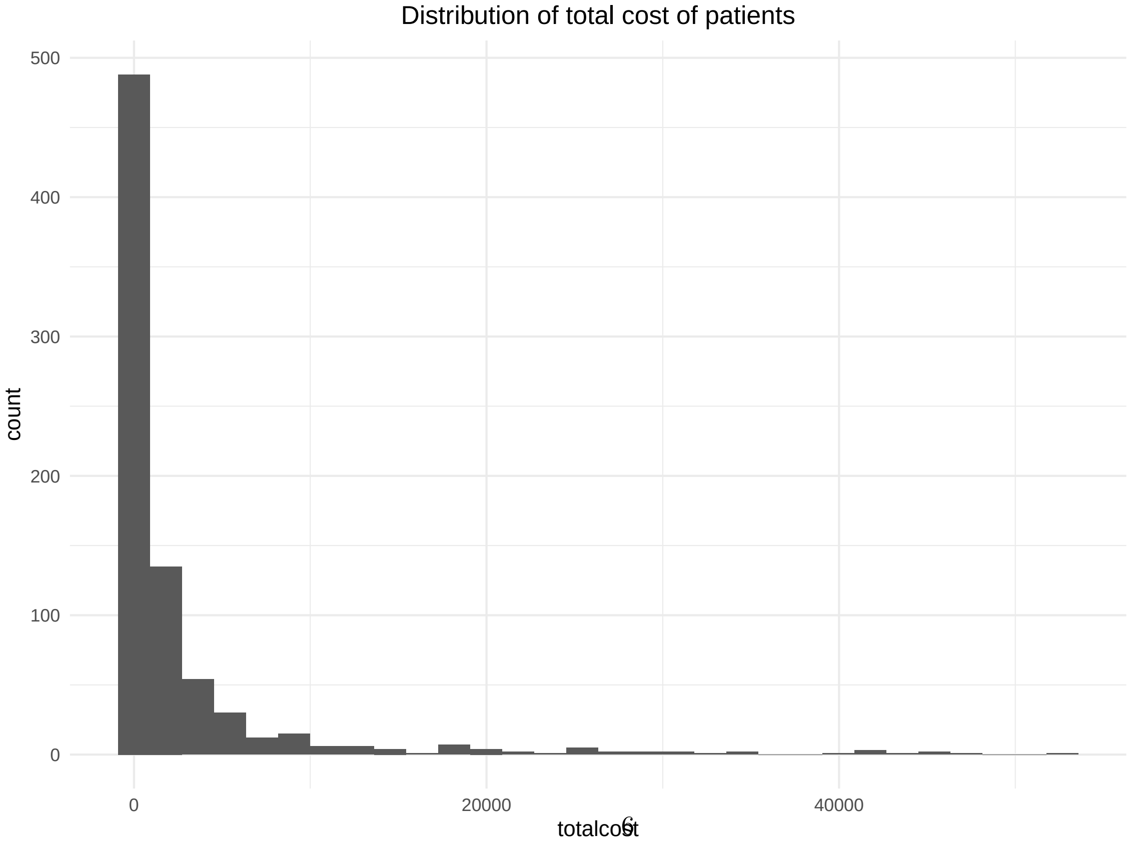
a)

The main outcome is "total cost" (in dollars) of patients diagnosed with heart disease, and the main predictor is the "number of emergency room (ER) visits". Other important covariates include age, gender, number or intervetion carried out, drugs prescribed, complications, comorbidities, and duraion of treatment condition.

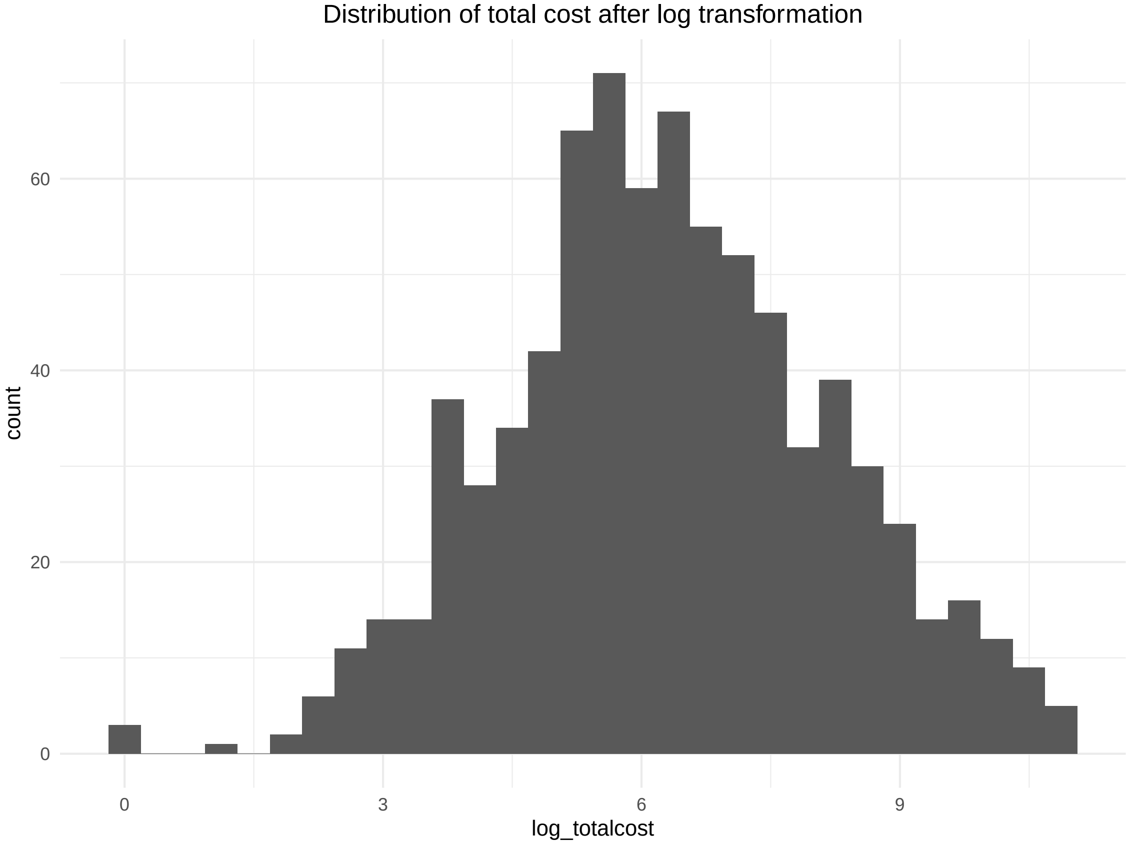
Table 1: Characcteristics of patients

	Overall (N=788)
totalcost	
- Mean (SD)	2799.96 (6690.26)
- Median (Q1, Q3)	507.20 (161.12, 1905.45)
- Min - Max	0.00 - 52664.90
e_rvisits	
- Mean (SD)	3.43 (2.64)
- Median (Q1, Q3)	3.00 (2.00, 5.00)
- Min - Max	0.00 - 20.00
age	
- Mean (SD)	58.72 (6.75)
- Median (Q1, Q3)	60.00 (55.00, 64.00)
- Min - Max	24.00 - 70.00
gender	
- female	608 (77.2%)
- male	180 (22.8%)
complications	
- Mean (SD)	0.06 (0.25)
- Median (Q1, Q3)	0.00 (0.00, 0.00)
- Min - Max	0.00 - 3.00
duration	
- Mean (SD)	164.03 (120.92)
- Median (Q1, Q3)	165.50 (41.75, 281.00)
- Min - Max	0.00 - 372.00

b)



As we can see in the first histgram, the distribution for “total cost” is highly right-skewed. Therefore, we may need to use *Logarithm transformation* to reduce right skewness.



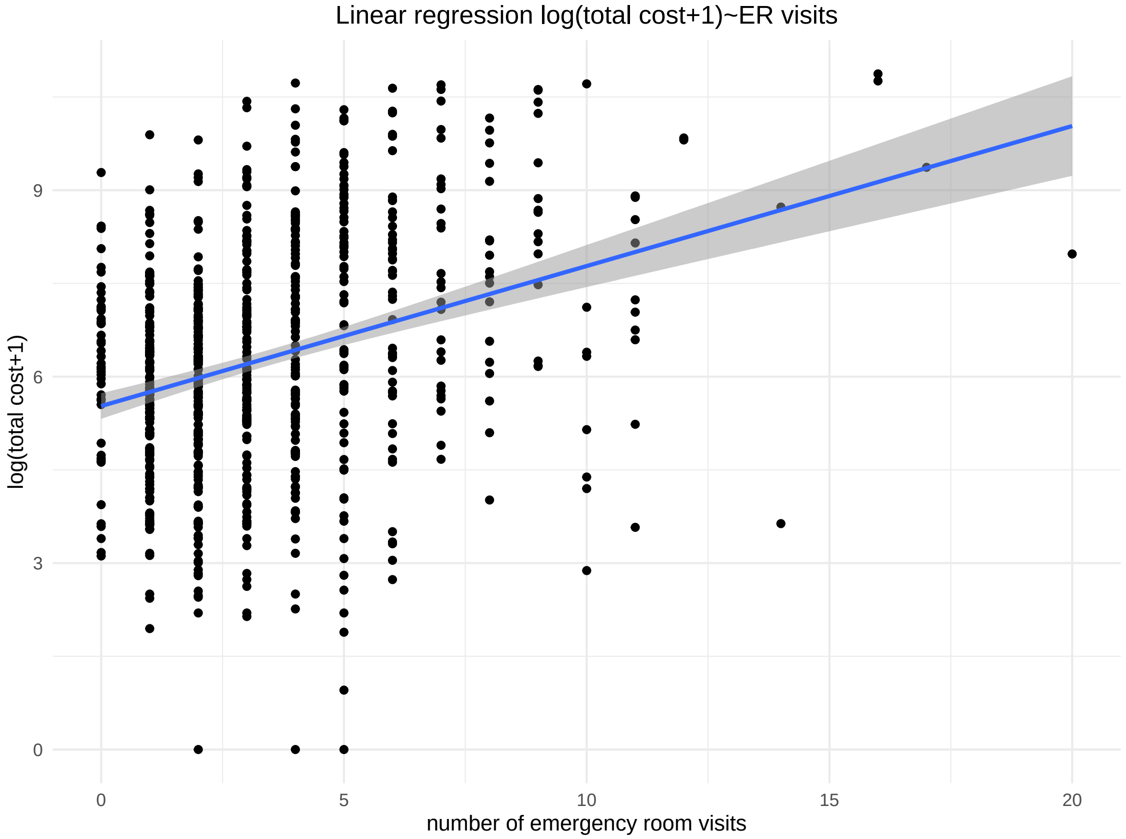
After log tranformation, the distribution for “total cost” is a bell-shaped curve, and the transformed total cost follows a normal distribution.

c) Creat “comp_bin” variable.

```
hr_df <-
        hr_df %>%
        mutate(comp_bin = case_when(complications == 0 ~ 0,TRUE ~ 1),
               comp_bin = factor(comp_bin))
```

d) SLR with transformed “total cost” and predictor “ERvisits”

```
##
## Call:
## lm(formula = log_totalcost ~ e_rvisits, data = .)
##
## 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
```



Therefore, the regression equation is:

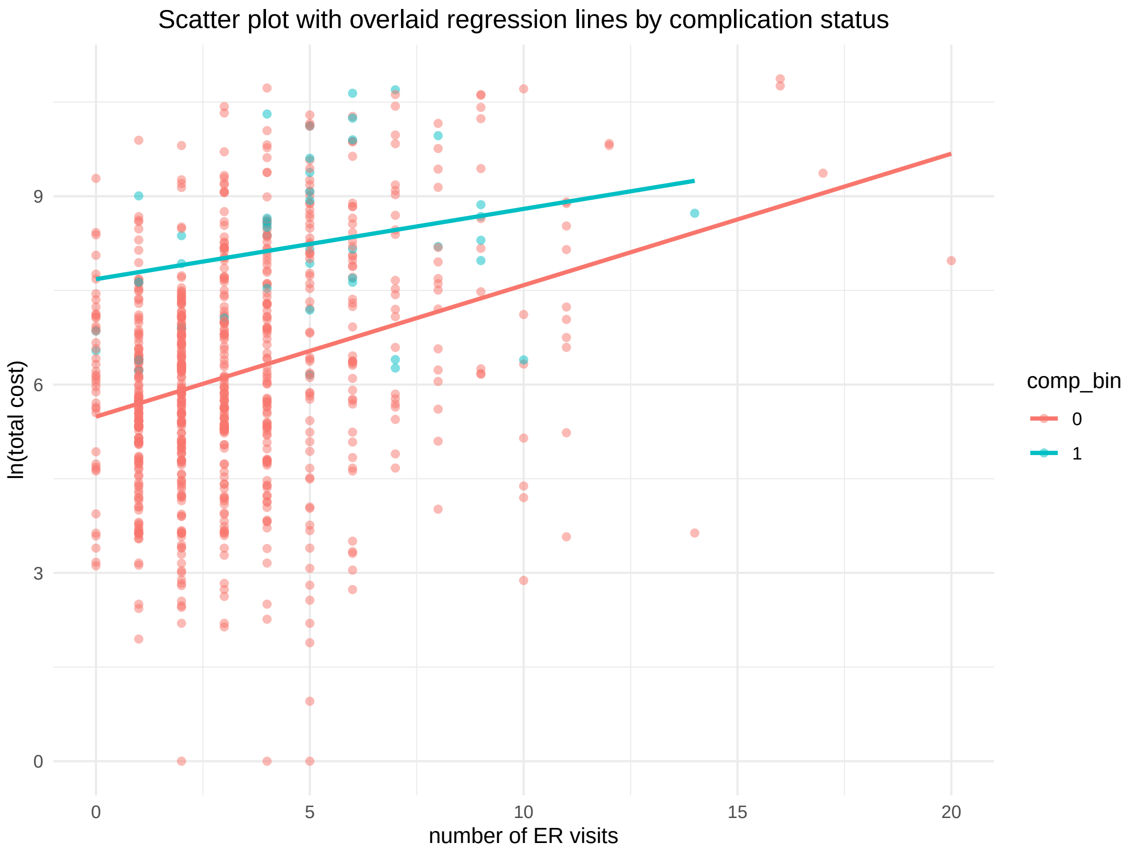
$$\hat{Y} = 5.527 + 0.225X$$

where $\hat{Y}$ =ln(total cost+1), X = number of ER visits

The estimated slope is 0.225 and it's P-value is it's significant, indicates that if the number of emergency room visits increase by 1 , the log(total cost+1) will increase by 0.225 on avarage. In other word, with the number of emergency room visits increase by 1, the total cost for a patient will increase by 25% on avarage.

e) MLR with "comp_bin" and "ERvisits" as predictors

i) Is "comp_bin" an effect modifier?



As the plot shows, there are non-parallel slopes for different category of "com_bin", therefore, we can add interaction effects to the model to test.

```
hr_lm2 <-
        hr_df %>%
        lm(log_totalcost ~ e_rvisits*comp_bin, data = .)
summary(hr_lm2)
```

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

As the regression model shows, the interaction term is not significant(P-value 0.311). Therefore, we can conclude that “comp_bin” is not an effect modifier of the relationship between “total cost” and “ERvisits”.

ii) Is “comp_bin” a confounder?

```
## 
## Call:
## lm(formula = log_totalcost ~ e_rvisits + comp_bin, data = .)
## 
## 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
```

Notice that ER visit remains statistically significantly associated with transformed total cost (p<0.001), but the magnitude of the association is lower after adjustment. (0.203 versus 0.225). The regression coefficent decreases by about 10%.

As rule of thumb, “comp_bin” meets the criteria for confounders. Thus, “comp_bin”is a confounder of the relationship between “total cost” and “ERvisits”.

iii) Should “comp_bin” be included along with “ERvisits”?

```r
anova(hr_lm1,hr_lm3)
```

```
## 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 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
```

“comp_bin” should be included along with “ERvisits” for two reasons.
First, as mentioned in ii), “comp_bin” is a potential confounder and thus need to be adjusted. Second, I compare these two models using partial ANOVA. Since the P-value < 0.05, at a significane level of 0.05, we reject H_0, and conclude that the model including “comp_bin” is “superior”.
Therefore, “comp_bin” should be included along with “ERvisits”.

f)

i) Fit a new MLR

```
hr_lm4 <-
    hr_df %>%
    lm(log_totalcost ~ e_rvisits + comp_bin + age + gender + duration, data = .)
summary(hr_lm4)
```

```
##
## Call:
## lm(formula = log_totalcost ~ e_rvisits + comp_bin + age + gender +
##     duration, data = .)
##
## 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 ***
## comp_bin1    1.5044946  0.2584882   5.820 8.57e-09 ***
## 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 ***
## ---
## 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
```

Therefore, the regression equation is:

$$\hat{Y} = 5.940 + 0.175X_1 + 1.504X_2 - 0.021X_3 - 0.207X_4 + 0.006X_5$$

where $\hat{Y}$ =ln(total cost+1);X1 = number of ER visits; X2 = having complication; X3 = age; X4 = being male; X5 = duration of treatment.

The regression results show that variables of the model, except for gender are all significant associated with of total cost of patients. Holding all other variables constant, if the number of emergency room visits increase by 1, the total cost for a patient will increase by 19% on avarage. Compared with the coefficient from SLR, the coefficient decrease.

ii) Compare the SLR and MLR models.

```
anova(hr_lm1,hr_lm4)
```

```
## Analysis of Variance Table
##
## Model 1: log_totalcost ~ e_rvisits
## Model 2: log_totalcost ~ e_rvisits + comp_bin + age + gender + duration
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)    
## 1    786 2544.8                                  
## 2    782 2062.2  4    482.62 45.753 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

I would choose the MLR models to address the investigator's objective. First, by comparing the two models using partial ANOVA, we can reject H_0 and conclude that the MLR model is suprior.

Second, $R^2_{adjusted}$ of the SLR and the MLR model is 0.097 and 0.265 respectively, which indicates that the MLR model can better explain the variation of the dependent variable.