Bios Methods Homework 4

Alison Elgass

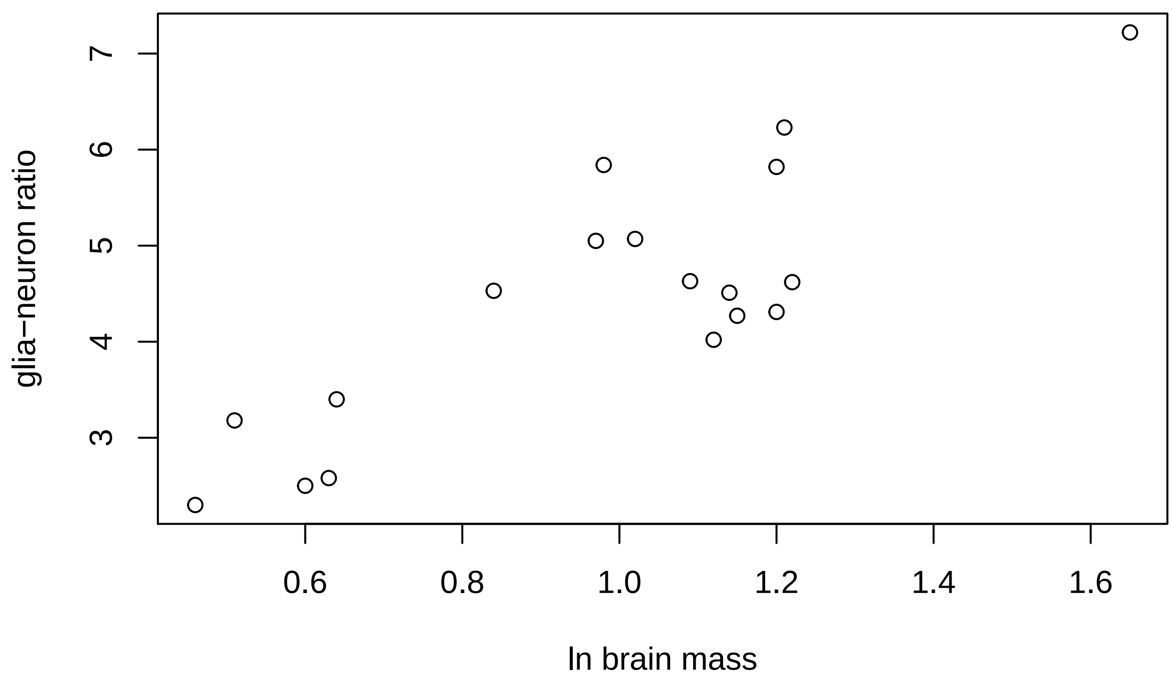
Problems 1 & 2

See hand-written scanned section to follow...

Problem 3

```
brain_data = read_excel(path = "./Brain.xlsx") %>%
  janitor::clean_names()

plot(brain_data$glia_neuron_ratio, brain_data$ln_brain_mass,
     xlab = "ln brain mass",
     ylab = "glia-neuron ratio")
```




```
nonhuman = brain_data %>% filter(species != "Homo sapiens")
```

Part A - Regression

```
brain_reg = lm(glia_neuron_ratio ~ ln_brain_mass, data = nonhuman)
summary(brain_reg)
```

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

The regression equation is $\hat{Y} = 0.1637 + 0.18113\mathbf{X_i}$

Part B - Prediction

For humans, ln(brain mass) = 7.22, so the regression equation would predict a glia neuron ratio of 0.1637 + 0.18113(7.22) = 1.471

Part C - Prediction Interval

The most meaningful estimate would be a prediction interval rather than a confidence interval at the given brain mass.

Part D - 95% Interval

95% PI = $\hat{\beta}_0 + \hat{\beta}_1 \mathrm{X_h}$ +- ($\mathrm{t}_{\text{n-2,1-}\alpha/2)(\text{se}(\hat{\beta}}0\sim + \hat{\beta}_1 \mathrm{X_h}$))

where se = $\sqrt{\ }$ MSE ((1/n) + ((x$_\mathrm{h}$ - $\overline{x}$)/ $\sum_{i=1}^{n}$ x$_\mathrm{i}$ - $\overline{x}$) + 1)

```r
xh = 7.22
predict.lm(brain_reg, data_frame(ln_brain_mass = xh),
        interval = "predict", level = 0.95)
```

```
## Warning: `data_frame()` is deprecated, use `tibble()`.
## This warning is displayed once per session.

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

We find that our 95% prediction interval for the human glia-neuron ratio is (1.036, 1.907). The actual measured ratio is 1.65, which is within this range, so we conclude that humans are not abnormal compared to other primates.

Since the human data point is somewhat of an outlier we should be cautious since the regression line might not accurately apply to a value of ln(brain mass) that is so high

Problem 4

```r
heart_data = read_csv(file = "./HeartDisease.csv")
```

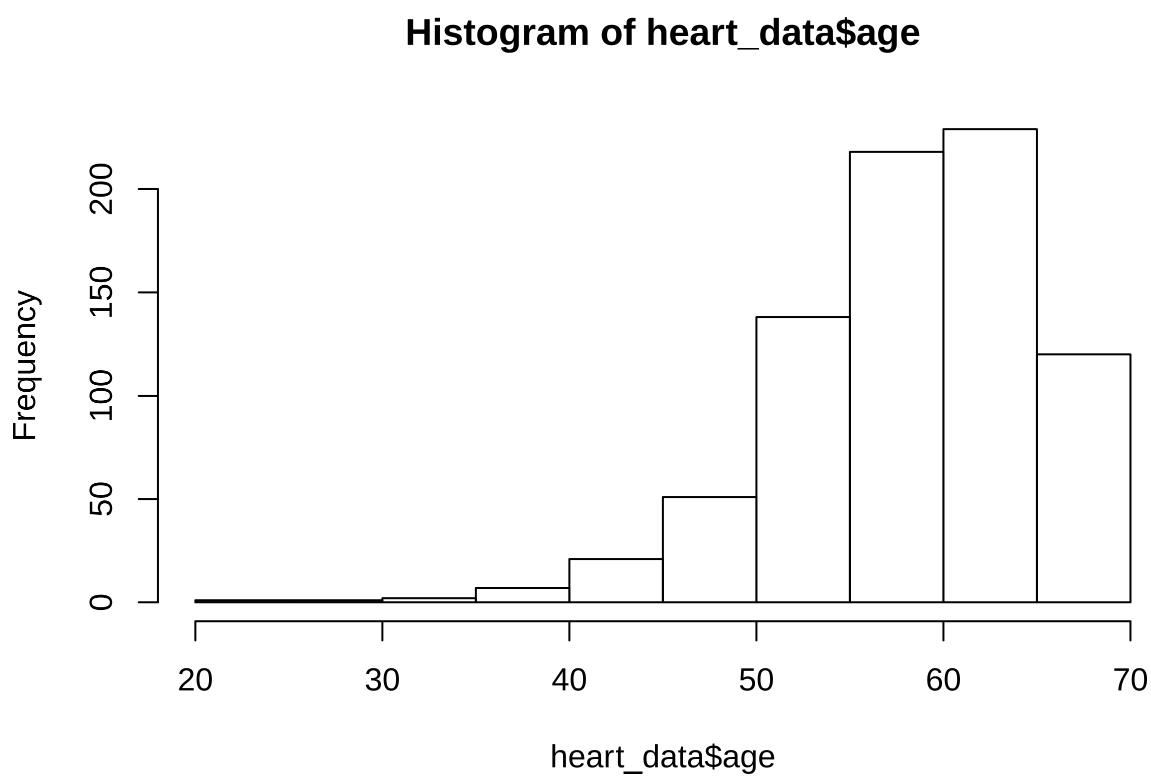
```
## Parsed with column specification:
## cols(
##   id = col_double(),
##   totalcost = col_double(),
##   age = col_double(),
##   gender = col_double(),
##   interventions = col_double(),
##   drugs = col_double(),
##   ERvisits = col_double(),
```

```
##   complications = col_double(),
##   comorbidities = col_double(),
##   duration = col_double()
## )
```

```r
heart_data %>%
  count(gender)
```

```
## # A tibble: 2 x 2
##   gender     n
##    <dbl> <int>
## 1      0   608
## 2      1   180
```

```r
hist(heart_data$age)
```




```r
summary(heart_data$ERvisits)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   2.000   3.000   3.425   5.000  20.000
```

```r
summary(heart_data$totalcost)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   161.1   507.2  2800.0  1905.5 52664.9
```

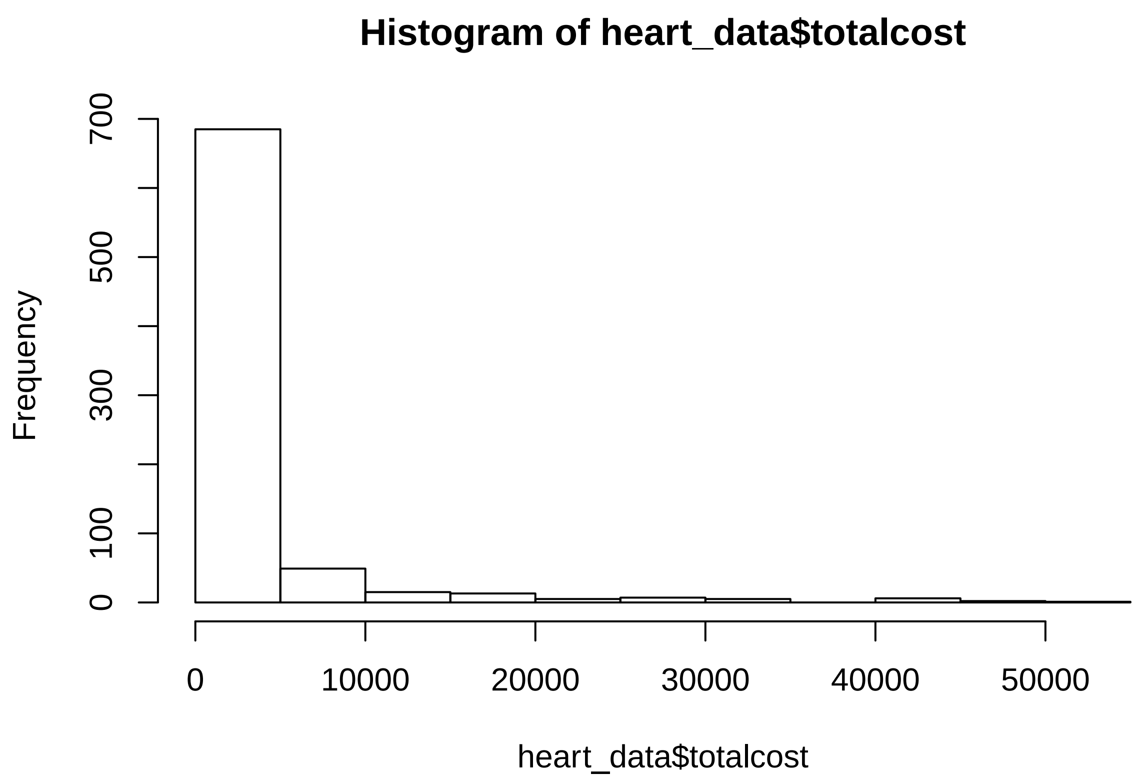
Part A - Description

The dataset contains data on 788 patients who made insurance claims for coronary heart disease. The main predictor is the number of ER visits, and the main outcome is total cost. Covariates include age, gender, number of complications, and condition duration.

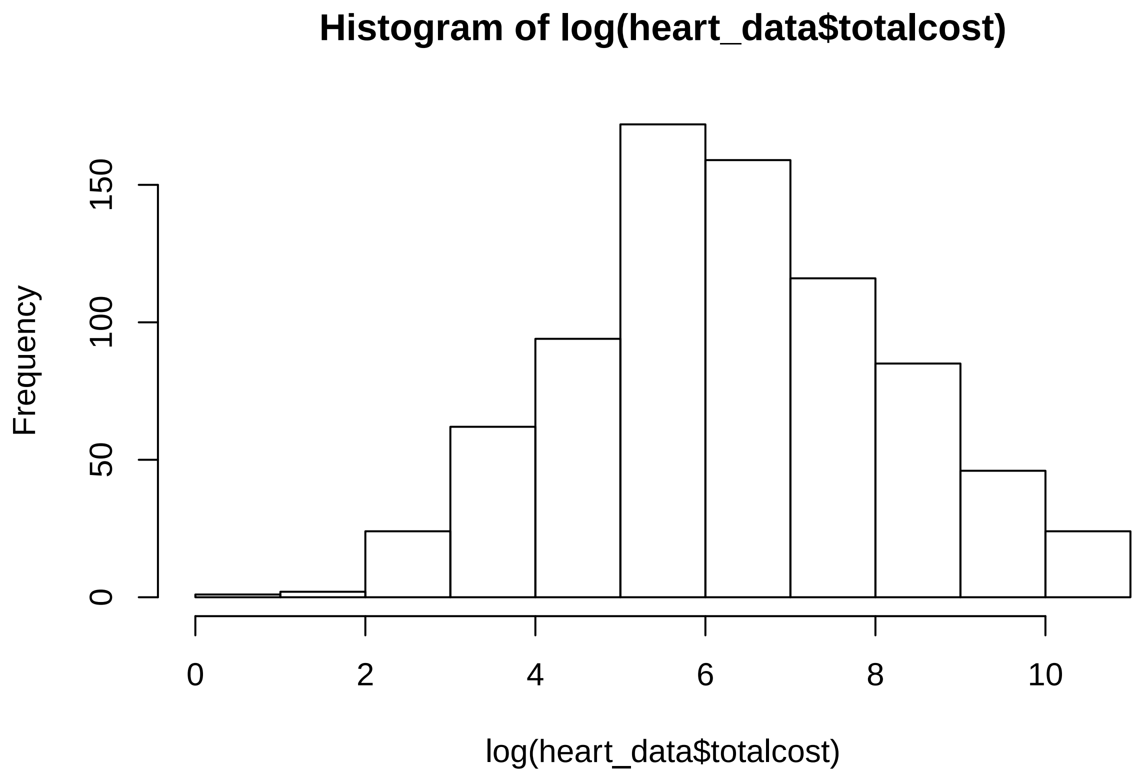
Part B - Total Cost

Altering total cost to ln(totalcost) makes a more normal distribution

```
hist(heart_data$totalcost)
```




```
hist(log(heart_data$totalcost))
```



Part C - Complications Factor

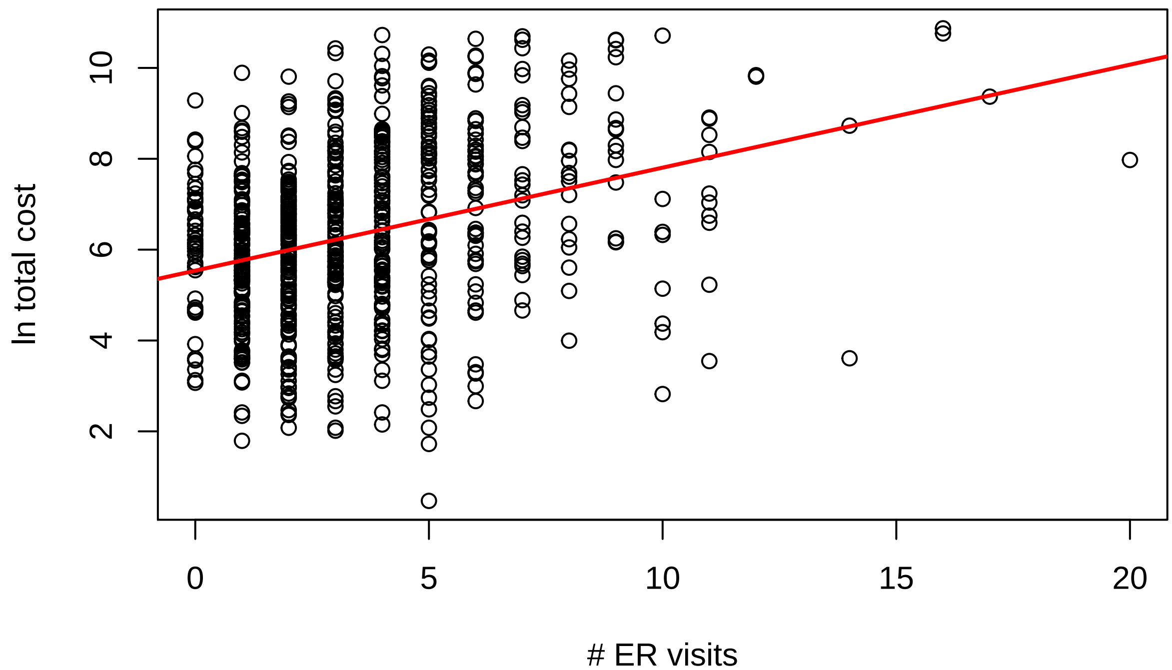
```
new_heart = heart_data %>%
  filter(totalcost != 0) %>%
  mutate(
    #one 3 in row 79
    compbin = factor(complications,
                     levels = c(0,1,3), labels = c(0,1,1)),
    lncost = log(totalcost)
  )

#str(new_heart$compbin)
#new_heart$compbin[79]
```

Part D - SLR

Note that in creating our `new_heart` dataset above, we also excluded 3 data points where cost was 0, since these will not work with a log transformation. This seems reasonable since these ER costs are likely missing or were not recorded by the insurance company, so they are not helpful for our regression analysis.

```
slr_heart = lm(lncost ~ ERvisits, data = new_heart)
plot(new_heart$ERvisits, new_heart$lncost,
     xlab = "# ER visits",
     ylab = "ln total cost")
abline(slr_heart, lwd = 2, col = 2)
```




```r
summary(slr_heart)
```

```
##
## Call:
## lm(formula = lncost ~ ERvisits, data = new_heart)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.2013 -1.1265  0.0191  1.2668  4.2797
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.53771    0.10362   53.44   <2e-16 ***
## ERvisits     0.22672    0.02397    9.46   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.772 on 783 degrees of freedom
## Multiple R-squared:  0.1026, Adjusted R-squared:  0.1014
## F-statistic:  89.5 on 1 and 783 DF,  p-value: < 2.2e-16
```

```r
b1_simple = summary(slr_heart)$coefficients[2,1] #0.2267
```

The regression equation is $\ln\hat{Y} = 5.537 + 0.226\mathbf{X_i}$
These results are highly significant, as a test of $\hat{\beta}_1 = 0$ gives the test statistic 9.46 which is $>$ $t_{785\text{-}2,1\text{-}0.95/2} =$ 1.9629983 so we conclude the slope is not 0.
Our estimated slope is 0.2267, meaning that for every 1 additional ER visit, we would predict that total cost

will increase by $100(e^{0.2267} - 1) = 25.4\%$.

Part E - MLR

```
mlr_heart = lm(lncost ~ ERvisits + compbin, data = new_heart)
summary(mlr_heart)
```

```
##
## Call:
## lm(formula = lncost ~ ERvisits + compbin, data = new_heart)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.0741 -1.0737 -0.0181  1.1810  4.3848
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.5211     0.1013  54.495  < 2e-16 ***
## ERvisits      0.2046     0.0237   8.633  < 2e-16 ***
## compbin1      1.6859     0.2749   6.132 1.38e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.732 on 782 degrees of freedom
## Multiple R-squared:  0.1437, Adjusted R-squared:  0.1416
## F-statistic: 65.64 on 2 and 782 DF,  p-value: < 2.2e-16
```

```
b1_mult = summary(mlr_heart)$coefficients[2,1] #0.2046
```

Using multiple linear regression, our regression equation is
$\ln\hat{Y} = 5.5211 + 0.2046\mathbf{X_{i1}} + 1.686\mathbf{X_{i2}}$
where X_1 = # ER visits, and X_2 = whether patient experienced complications (reference category = 0, no complications)

i. Testing for interaction To test for interaction, we run a MLR with an interaction term

```
mlr_interact = lm(lncost ~ ERvisits*compbin, data = new_heart)
summary(mlr_interact)
```

```
##
## Call:
## lm(formula = lncost ~ ERvisits * compbin, data = new_heart)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.0852 -1.0802 -0.0078  1.1898  4.3803
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        5.49899    0.10349  53.138  < 2e-16 ***
## ERvisits           0.21125    0.02453   8.610  < 2e-16 ***
## compbin1           2.17969    0.54604   3.992 7.17e-05 ***
## ERvisits:compbin1 -0.09927    0.09483  -1.047    0.296
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 1.732 on 781 degrees of freedom
## Multiple R-squared:  0.1449, Adjusted R-squared:  0.1417
## F-statistic: 44.13 on 3 and 781 DF,  p-value: < 2.2e-16
```

```r
-0.1 + c(-1,1)*(0.09483)*qt(0.975,781)
```

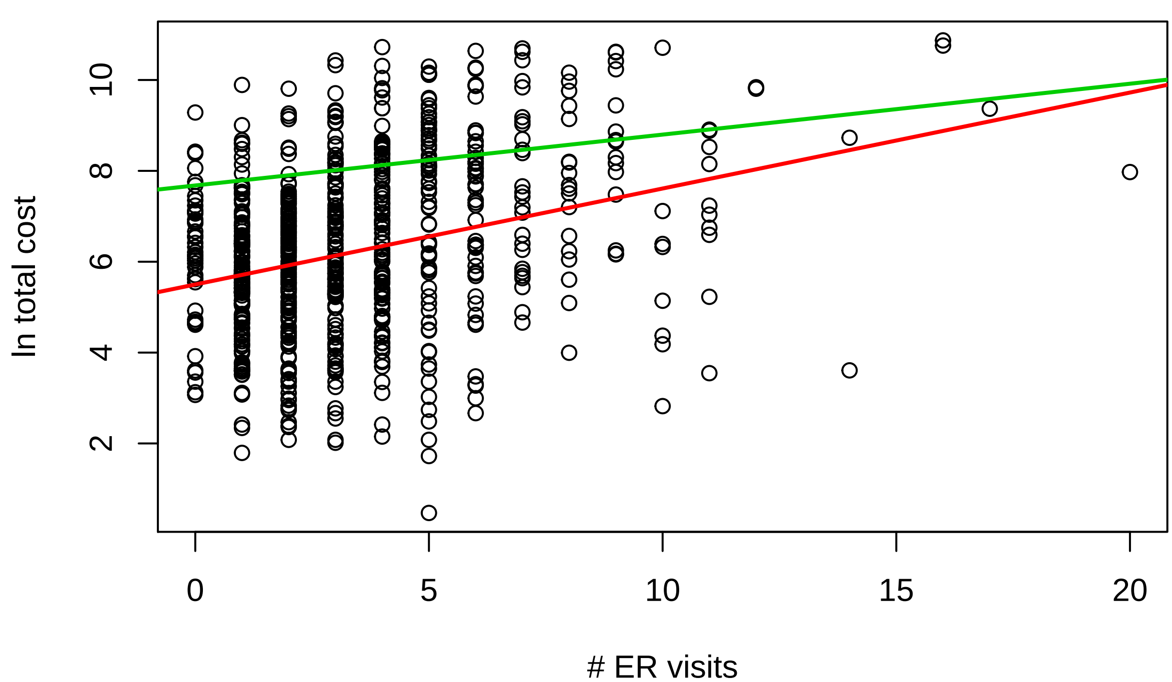
```
## [1] -0.28615187  0.08615187
```

This gives the model $\ln\hat{Y}$ = 5.499 + 0.211$\mathbf{X_{i1}}$ + 2.18$\mathbf{X_{i2}}$ - 0.1$\mathbf{X_{i1}}\mathbf{X_{i2}}$

The interaction term = -0.1, with a t-value that is insignifcant.
Also the confidence interval would be -0.1 +- (0.09483)($t_{0.975,781}$) = (-0.286, 0.862) which includes zero, so we conclude there is not a significant interaction between ER visits and complication status.

We can also look at the graph of ln(cost) vs. # ER visits, stratified by compbin = 0 or 1, and the respective slopes of the regression lines

```r
#make a plot of slr lines, stratified by compbin
plot(new_heart$ERvisits, new_heart$lncost,
     xlab = "# ER visits",
     ylab = "ln total cost")
heart0 = new_heart %>% filter(compbin == 0)
heart1 = new_heart %>% filter(compbin == 1)
lm(lncost ~ ERvisits, data = heart0) %>% abline(lwd = 2, col=2)
lm(lncost ~ ERvisits, data = heart1) %>% abline(lwd = 2, col = 3)
```



For the 742 observations without complications (compbin = 0, in red), $\hat{\beta}_1$ = 0.211
For the remaining 43 observations with complications (compbin = 1), $\hat{\beta}_1$ = 0.112

ii. Testing for confounding To test for confounding, we compare the SLR (ER visits as the only predictor of cost) and MLR (adding in/adjusting for complications)

```
(b1_mult - b1_simple)/b1_simple #decreases by 9.75%
```

```
## [1] -0.09755288
```

In the simple linear regression, we found $\hat{\beta}_1 = 0.2267$.
Adding in the compbin variable and doing a multiple linear regression, now $\hat{\beta}_1 = 0.2046$, a 9.8% decrease, so I don't believe it is a significant confounder.

iii. Compbin in the model

```
anova(slr_heart, mlr_heart)
```

```
## Analysis of Variance Table
##
## Model 1: lncost ~ ERvisits
## Model 2: lncost ~ ERvisits + compbin
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    783 2459.8
## 2    782 2347.0  1    112.84 37.598 1.379e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
partial_r2 = 112.84/2459.8  #= 0.0459
dfS = 785 - 1 - 1
dfL = 785 - 2 - 1
qf(0.95,dfS-dfL,dfL)
```

```
## [1] 3.853378
```

Testing the effect of the complications variable (X_2),
H_0: $\beta_2 = 0$
H_1: $\beta_2 \neq 0$
F = MSR/MSE = 37.598 > $F_{1,782,0.05} = 3.85$, so we reject the null hypothesis and conclude that β_2 is not equal to 0

We also calculate from the previous ANOVA tables the partial R^2 from the marginal contribution of compbin to be 112.84 / 2459.8 = 0.046.
In other words, about 4.6% of the variation in cost can be attributed to the complications factorm holding the ER visits variable fixed.

Because of this I would include compbin in our model, so
$\ln\hat{Y} = 5.5211 + 0.2046\mathbf{X_{i1}} + 1.686\mathbf{X_{i2}}$

Part F - Final MLR

```
mlr_big = lm(lncost ~ ERvisits + compbin + age + gender + duration,
               data = new_heart)
summary(mlr_big)
```

```
##
## Call:
## lm(formula = lncost ~ ERvisits + compbin + age + gender + duration,
##     data = new_heart)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0823 -1.0555 -0.1352  0.9533  4.3462
```

```
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.0449619  0.5063454  11.938  < 2e-16 ***
## ERvisits     0.1757486  0.0223189   7.874 1.15e-14 ***
## compbin1     1.4921110  0.2554883   5.840 7.65e-09 ***
## age         -0.0221376  0.0086023  -2.573   0.0103 *
## gender      -0.1176181  0.1379809  -0.852   0.3942
## duration     0.0055406  0.0004848  11.428  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.605 on 779 degrees of freedom
## Multiple R-squared:  0.268,  Adjusted R-squared:  0.2633
## F-statistic: 57.03 on 5 and 779 DF,  p-value: < 2.2e-16
```

$\ln\hat{Y} = 6.05 + 0.176\mathbf{X_{i1}} + 1.492\mathbf{X_{i2}}$ - $0.022\mathbf{X_{i3}}$ - $0.118\mathbf{X_{i4}} + 0.0055\mathbf{X_{i5}}$

where
X_{i1} = # ER visits
X_{i2} = complications (yes compared to no)
X_{i3} = age (years)
X_{i4} = gender (male compared to female)
X_{i5} = duration (days)

The adjusted R^2 is 0.2633, meaning 26.3% of the variation in cost can be attributed to these covariates as in this model.

Which model to use?

```
a1 = lm(lncost ~ ERvisits + compbin + age, data = new_heart)
a2 = lm(lncost ~ ERvisits + compbin + gender, data = new_heart)
a3 = lm(lncost ~ ERvisits + compbin + duration, data = new_heart)
tibble("Added Covariate" =
         c("simple","compbin","age","gender","duration"),
       "Adj R2" =
         c(summary(slr_heart)$adj.r.squared,
         summary(mlr_heart)$adj.r.squared,
         summary(a1)$adj.r.squared, #age
         summary(a2)$adj.r.squared, #gender
         summary(a3)$adj.r.squared) #duration
       ) %>% knitr::kable()
```

Added Covariate	Adj R2
simple	0.1014291
compbin	0.1415533
age	0.1415764
gender	0.1407894
duration	0.2582995

I would include the complications and duration covariates to adjust the relationship between cost vs. ER visits for these factors. These increase the adjusted R^2 value. Gender and age do not appear to have a significan affect. It's interesting to note that in the large model, age comes back with a significant p-value (0.0103), however because it does not affect the R^2 I would not include it.

```r
mlr_final = lm(lncost ~ ERvisits + compbin + duration, data = new_heart)
summary(mlr_final)
```

```
##
## Call:
## lm(formula = lncost ~ ERvisits + compbin + duration, data = new_heart)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1450 -1.1008 -0.1479  0.9593  4.6166
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.7605975  0.1163153  40.928  < 2e-16 ***
## ERvisits    0.1708778  0.0222372   7.684 4.62e-14 ***
## compbin1    1.5285357  0.2559535   5.972 3.56e-09 ***
## duration    0.0053724  0.0004823  11.140  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.61 on 781 degrees of freedom
## Multiple R-squared:  0.2611, Adjusted R-squared:  0.2583
## F-statistic: 92.01 on 3 and 781 DF,  p-value: < 2.2e-16
```

Our **final model** then is
$\ln\hat{Y} = 4.76 + 0.171\mathbf{X_{i1}} + 1.529\mathbf{X_{i2}} + 0.005\mathbf{X_{i3}}$

Using the formula % change in Y = 100(e^{β_1} - 1),
we conclude that an additional ER visit increases total cost by about 18%, having complications increases it by 361%, and an additional day of stay increases it by 0.54%.