P8130: Homework 4

Alyssa Vanderbeek (amv2187)

16 November 2018

Problem 2

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.

The dataset contains information pertaining to patient demographics (age, gender), ER visits (frequency, duration), and healthcare cost for 788 (608 F, 180 M) patients with heart disease. The outcome of interest is the total cost per patient as a function of number of ER visits, primarily. The length of the ER visit and number of interventions are also likely to affect the total cost. Table 1 shows variable distributions for all patients; Table 2 examines distributions according to gender.

Table 1: Variable summaries across genders

	Min	1st Quartile	Median	Mean	3rd Quartile	Max
age	24	55.000	60.0	58.718	64.00	70.0
drugs	0	0.000	0.0	0.447	0.00	9.0
duration	0	41.750	165.5	164.030	281.00	372.0
ERvisits	0	2.000	3.0	3.425	5.00	20.0
totalcost	0	161.125	507.2	2799.956	1905.45	52664.9
comorbidities	0	0.000	1.0	3.766	5.00	60.0
complications	0	0.000	0.0	0.057	0.00	3.0
interventions	0	1.000	3.0	4.707	6.00	47.0

```
## Adding missing grouping variables: `gender`
```

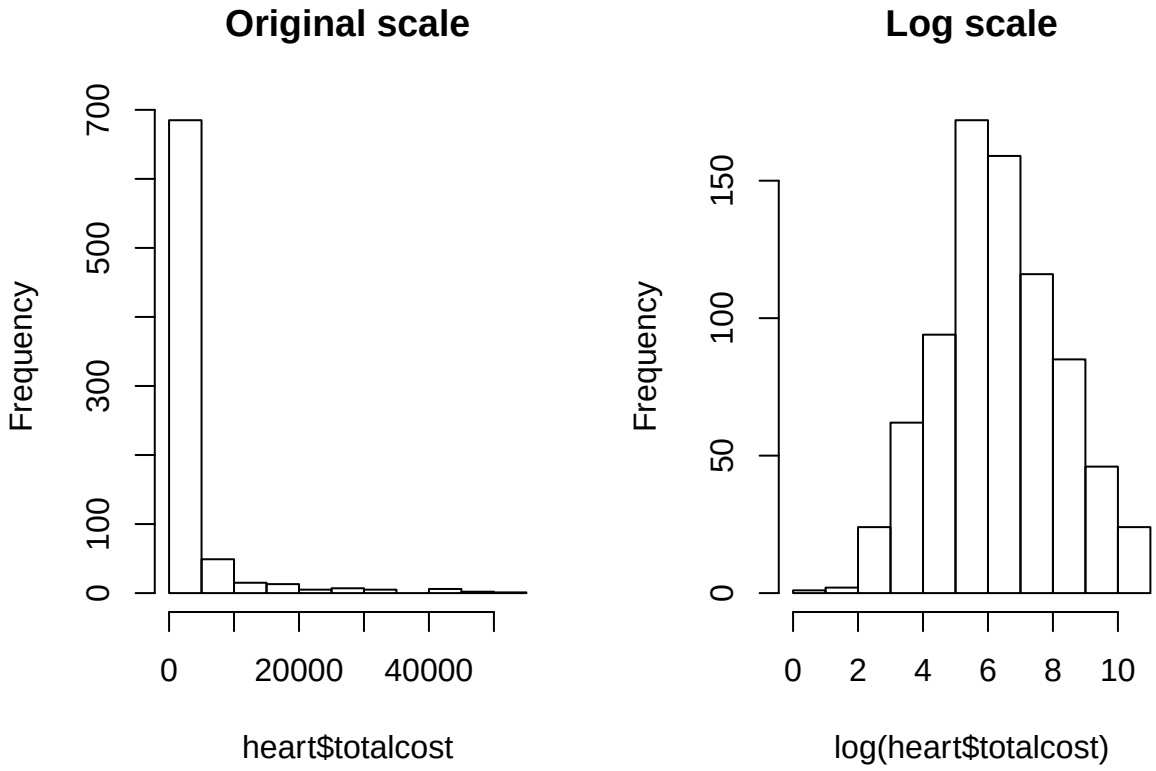
Table 2: Variable summaries by gender

	gender	1st Qu.	3rd Qu.	Max.	Mean	Median	Min.
age	F	55.00	64.00	70.00	58.78	60.00	24.00
age	M	54.00	64.00	70.00	58.49	60.00	39.00
drugs	F	0.0000	0.0000	9.0000	0.4293	0.0000	0.0000
drugs	M	0.0000	1.0000	7.0000	0.5056	0.0000	0.0000
duration	F	40.0	281.0	372.0	162.5	164.5	0.0
duration	M	59.25	283.50	351.00	169.26	168.00	0.00
ERvisits	F	2.000	4.000	17.000	3.266	3.000	0.000
ERvisits	M	2.000	5.000	20.000	3.961	3.000	0.000
totalcost	F	161.1	1975.0	52664.9	2867.3	502.0	5.6
totalcost	M	176.4	1881.3	44162.8	2572.3	546.2	0.0
comorbidities	F	0.0	5.0	60.0	3.9	2.0	0.0

	gender	1st Qu.	3rd Qu.	Max.	Mean	Median	Min.
comorbidities	M	0.000	4.000	28.000	3.317	1.000	0.000
complications	F	0.00000	0.00000	1.00000	0.05263	0.00000	0.00000
complications	M	0.00000	0.00000	3.00000	0.07222	0.00000	0.00000
interventions	F	1.000	6.000	47.000	4.607	3.000	0.000
interventions	M	1.000	7.000	34.000	5.044	3.000	0.000

(b) Investigate the shape of the distribution for variable 'total cost' and try different transformations, if needed.

Variable 'totalcost' is highly skewed; applying a log transformation normalizes the distribution. (See histograms below)




```
## null device
##           1
```

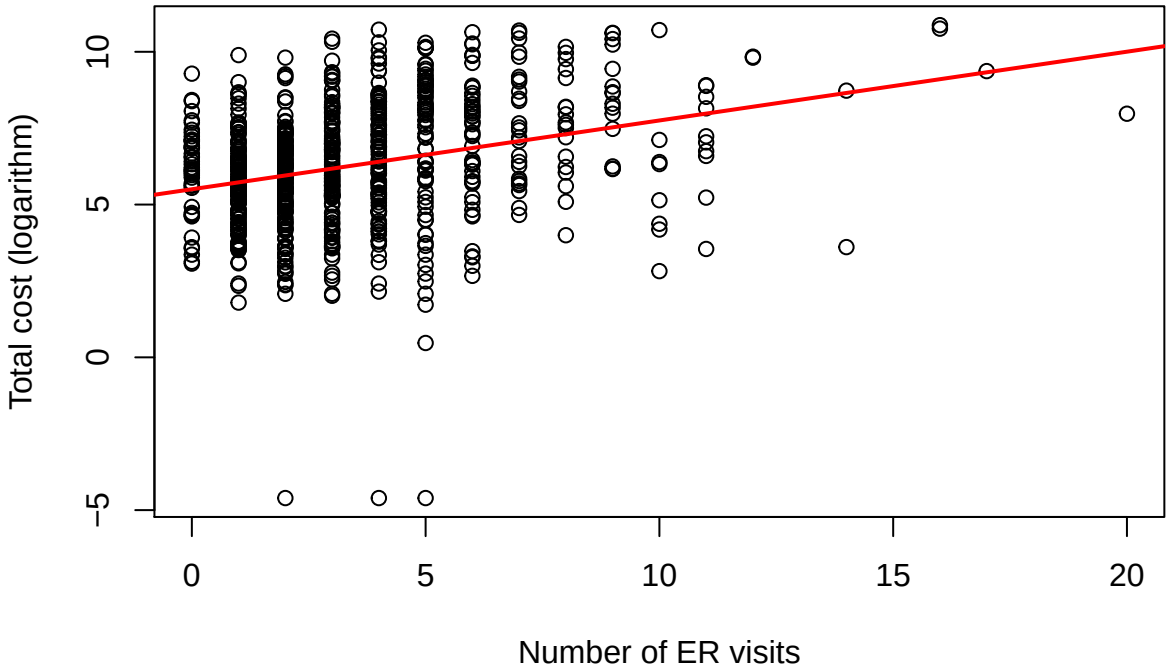
(c) Create a new variable called 'comp_bin' by dichotomizing 'complications': 0 if no complications, and 1 otherwise.

```
heart = heart %>% mutate(comp_bin = ifelse(complications == 0, 0, 1))
```

(d) Based on our decision in part b), fit a simple linear regression (SLR) between the original or transformed 'total cost' and predictor 'ERvisits'. This includes a scatterplot and results of the regression, with appropriate comments on significance and interpretation of the slope.

Using the log-transformed total cost variable (when total cost is $0, I added 0.01 before taking the log), I run an SLR using the number of ER vists as the predictor.

```
##
## Call:
## lm(formula = cost_transform ~ heart$ERvisits)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.2321  -1.1013   0.0529   1.3055   4.3224
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.5016     0.1106  49.725   <2e-16 ***
## heart$ERvisits   0.2251     0.0256   8.792   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.894 on 786 degrees of freedom
## Multiple R-squared:  0.08954,    Adjusted R-squared:  0.08838
## F-statistic:  77.3 on 1 and 786 DF,  p-value: < 2.2e-16
```



According to the fitted model, the number of ER visits a patient experiences is a significant predictor of their cost to the hospital ($p<0.001$). Specifically, for each additional ER visit, a patient's cost increases by about 25% on average. ($e^{0.2251} = 1.252$).

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

```
##
## Call:
## lm(formula = cost_transform ~ heart$ERvisits + heart$comp_bin)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.1017  -1.0561   0.0165   1.2104   4.4301
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     5.48475    0.10838  50.607  < 2e-16 ***
## heart$ERvisits  0.20236    0.02536   7.979 5.23e-15 ***
## heart$comp_bin  1.73361    0.29432   5.890 5.72e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.855 on 785 degrees of freedom
## Multiple R-squared:  0.1281, Adjusted R-squared:  0.1259
## F-statistic: 57.65 on 2 and 785 DF,  p-value: < 2.2e-16
```

(i) Test if 'comp_bin' is an effect modifier of the relationship between 'total cost' and 'ERvisits'. Comment.

(ii) Test if 'comp_bin' is a confounder of the relationship between 'total cost' and 'ERvisits'. Comment.

Note that the coefficient for ER visits changes from 0.2251 in the SLR to 0.202 in the MLR. Since this change is >10%, we suspect that the experience of complications is a confounder. We can also see this given that 'comp_bin' is correlated with ER visits (0.1520242) and the outcome, total cost (0.1919638).

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

(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.

```
##
## Call:
## lm(formula = cost_transform ~ heart$ERvisits + heart$comp_bin +
##     heart$age + heart$gender + heart$duration)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.9436 -1.0080 -0.0886  0.9771  4.3492
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     5.8469672  0.5387065  10.854  < 2e-16 ***
## heart$ERvisits  0.1736943  0.0238252   7.290 7.58e-13 ***
```

```
## heart$comp_bin  1.5258252  0.2728204   5.593 3.09e-08 ***
## heart$age      -0.0198581  0.0091556  -2.169   0.0304 *
## heart$gender   -0.2848042  0.1463906  -1.946   0.0521 .
## heart$duration  0.0059649  0.0005159  11.561  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.714 on 782 degrees of freedom
## Multiple R-squared:  0.2583, Adjusted R-squared:  0.2536
## F-statistic: 54.47 on 5 and 782 DF,  p-value: < 2.2e-16
```

According to the MLR model, age, presence of complications during treatment, and duration of treatment are significant predictors of total cost. Gender is marginal; there is not quite enough evidence at the 5% significance level to suggest that there is a difference in treatment cost between men and women (p=0.0521).

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

I would use the MLR to answer the investigator's question of whether there is an association between the number of ER visits and the total cost of treatment for heart disease patients. Although the conclusions drawn from both models are the same (there is an association), we get a better idea of the degree of this association by accounting for other demographic information in the MLR. Ultimately, after accounting for age, gender, complications, and the duration of treatment, the number of ER visits has a smaller effect than was shown in the SLR.

Problem 3

A hospital administrator wishes to test the relationship between 'patient's satisfaction' (Y) and 'age', 'severity of illness', and 'anxiety level' (data 'PatSatisfaction.xlsx'). The administrator randomly selected 46 patients, collected the data, and asked for your help with the analysis.

(a) Create a correlation matrix and interpret your initial findings.

```
##              Safisfaction        Age   Severity    Anxiety
## Safisfaction    1.0000000 -0.7867555 -0.6029417 -0.6445910
## Age            -0.7867555  1.0000000  0.5679505  0.5696775
## Severity       -0.6029417  0.5679505  1.0000000  0.6705287
## Anxiety        -0.6445910  0.5696775  0.6705287  1.0000000
```

As shown in the correlation matrix above, all parameters are pairwise correlated: as age increases, satisfaction decreases, severity of illness increases, and anxiety level increases; as severity of illness increases, satisfaction decreases and anxiety level increases; and as anxiety level increases, satisfaction decreases.

(b) Fit a multiple regression model and test whether there is a regression relation. State the hypotheses, decision rule and conclusion.

For any predictor in a regression model, we have the following hypotheses:

$$H_0 : \text{the variable is not predictive of the outcome} (\beta = 0)$$

$$H_1 : \text{the variable is predictive of the outcome} (\beta \neq 0)$$

We conclude that the variable is a significant predictor when the calculated t-statistic $t* \geq t_{1-\alpha,n-p-1} = t_{0.95,42}$

```
##
## Call:
## lm(formula = ptx$Safisfaction ~ ptx$Age + ptx$Severity + ptx$Anxiety)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.3524  -6.4230   0.5196   8.3715  17.1601
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  158.4913    18.1259   8.744 5.26e-11 ***
## ptx$Age       -1.1416     0.2148  -5.315 3.81e-06 ***
## ptx$Severity  -0.4420     0.4920  -0.898   0.3741
## ptx$Anxiety  -13.4702     7.0997  -1.897   0.0647 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.06 on 42 degrees of freedom
## Multiple R-squared:  0.6822, Adjusted R-squared:  0.6595
## F-statistic: 30.05 on 3 and 42 DF,  p-value: 1.542e-10
```

According to the MLR above, only age is a significant predictor of patient satisfaction at the 5% significance level.

(c) Show the regression results for all estimated coefficients with 95% CIs. Interpret the coefficient and 95% CI associated with 'severity of illness'.

(d) Obtain an interval estimate for a new patient's satisfaction when Age=35, Severity=42, Anxiety=2.1. Interpret the interval.

(e) Test whether 'anxiety level' can be dropped from the regression model, given the other two covariates are retained. State the hypotheses, decision rule and conclusion.

Testing for a regression relation (global F-test) has the following hypotheses:

$$H_0 : \text{no parameters predict patient satisfaction}$$

$$H_1 : \text{at least one parameter predictive of patient satisfaction}$$

The null hypothesis is rejected in favor of the alternative when the F statistics is $\geq F_{1-\alpha,p,n-p-1} = F_{0.95,3,42}$

```
## Analysis of Variance Table
##
## Response: ptx$Safisfaction
##              Df Sum Sq Mean Sq F value    Pr(>F)
## ptx$Age       1 8275.4  8275.4 81.8026 2.059e-11 ***
## ptx$Severity  1  480.9   480.9  4.7539   0.03489 *
## ptx$Anxiety   1  364.2   364.2  3.5997   0.06468 .
## Residuals    42 4248.8   101.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```