Module 3 Quarto Project Myrie

Joseph Myrie

2024-06-10

Introduction

This cohort data set has 5 variables including: smoking status, gender, age, cardiac status, and then cost. One could imagine this to be a cohort of patients that was assess for cardiac morbidity and then followed to see what their yearly healthcare expenditures were. We could make the assumption that costs will be higher for those with a positive smoking and cardiac status, also for those who are male, and for those who are older. Unclear if these are yearly costs. Also unclear if this is what patients owed or if this overall healthcare costs for the individual patients.

Overall this dataset is pretty simple with only 5 variables, but there does not seem to be alot of missing data which is great. There is also are large patient sample which is also great about 5000 patients.

```
library(readr)                                                                    ①
cohort_1_ <- read_csv("C:/Users/myriej01/Downloads/cohort (1).csv")
```

① Imported cohort CSV

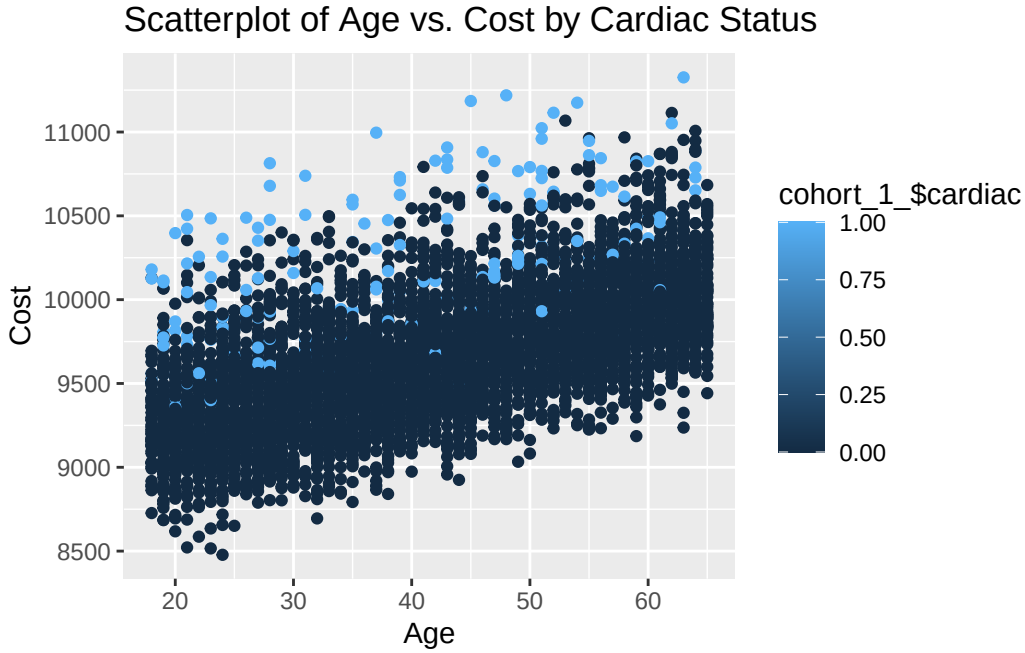
```
Rows: 5000 Columns: 5
-- Column specification --------------------------------------------------------
Delimiter: ","
dbl (5): smoke, female, age, cardiac, cost

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Methods

First, we tried to see what relationships there were in the dataset by producing some scatter plots of the variables. First, we assessed the relationship of Age vs Cost by Cardiac status.

```
library(ggplot2)                                                                    ①
ggplot(cohort_1_, aes(x = cohort_1_$age, y = cohort_1_$cost, color = cohort_1_$cardiac)) +
  geom_point() +
  labs(title = "Scatterplot of Age vs. Cost by Cardiac Status", x = "Age", y = "Cost")
```

① Created a scatterplot of Age vs Cost by Cardiac Status

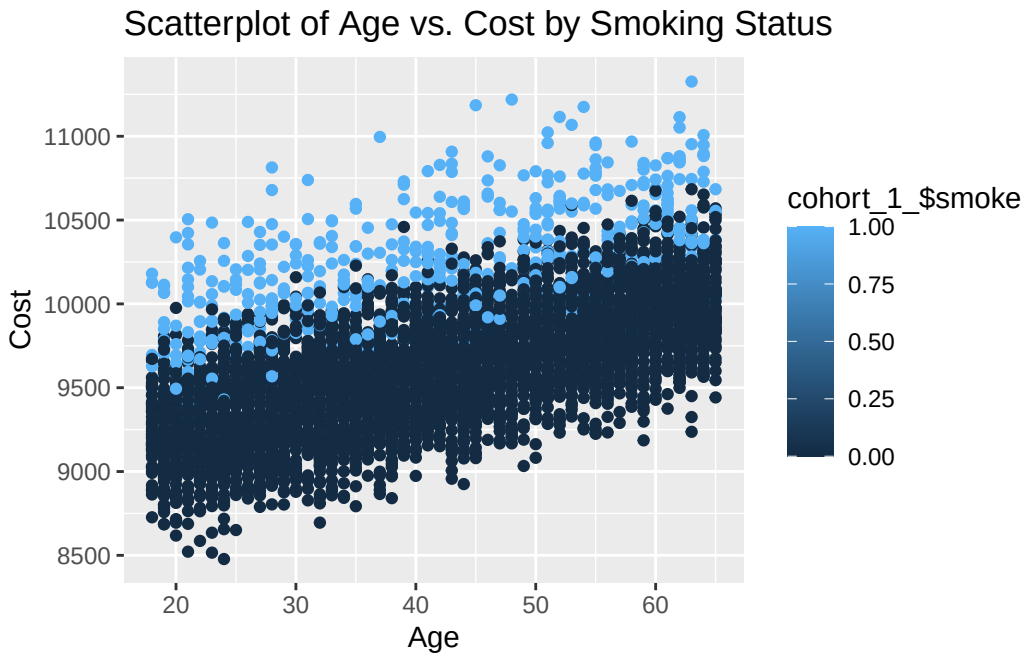


<Age vs Cost by cardiac status>

We next made a scatterplot of Age vs Cost by Smoking status.

```
ggplot(cohort_1_, aes(x = cohort_1_$age, y = cohort_1_$cost, color = cohort_1_$smoke)) +
  geom_point() +
  labs(title = "Scatterplot of Age vs. Cost by Smoking Status", x = "Age", y = "Cost") ①
```

① Scatter plot of Age vs Cost by Smoking Status

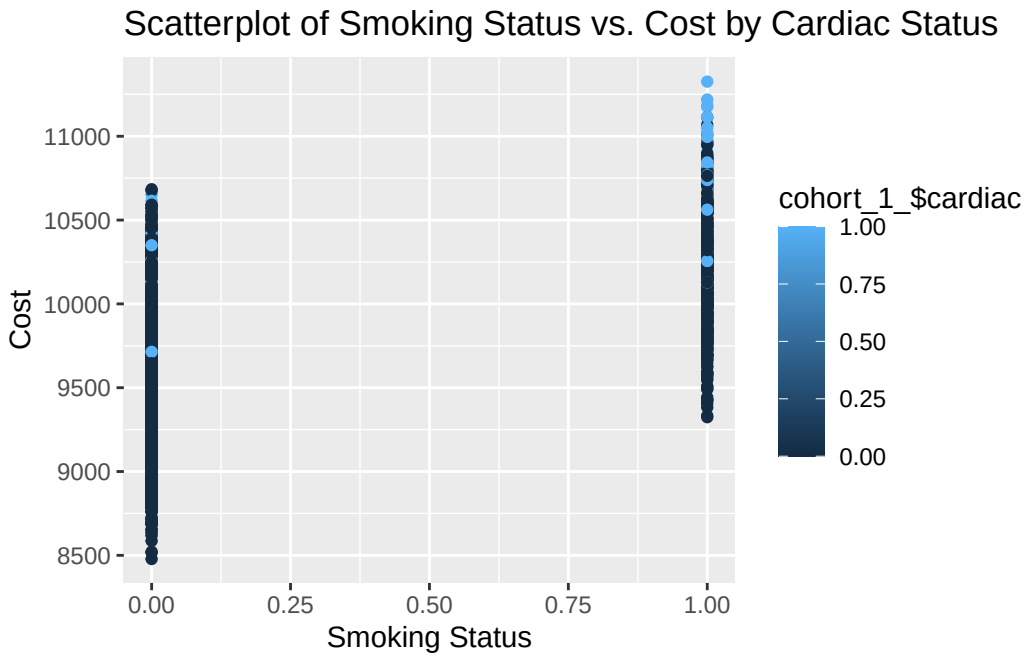


<Smoking Status vs Cost by Cardiac status>

We next looked at the relationship between smoking status and cost by cardiac status.

```
ggplot(cohort_1_, aes(x = cohort_1_$smoke, y = cohort_1_$cost, color = cohort_1_$cardiac)) +
  geom_point() +
  labs(title = "Scatterplot of Smoking Status vs. Cost by Cardiac Status", x = "Smoking Statu
```

① Scatterplot of Smokign Status vs Cost by Cardiac Status



Results

First we have a table of the summary statics below:

```
library(gtsummary)
tbl_summary(cohort_1_, statistic = list(
    all_continuous() ~ "{mean} ({sd}) [{min}, {max}]"))                    ①
```

① Used gtsummary to create a summary table for our 5 variables.

```
Table printed with `knitr::kable()`, not {gt}. Learn why at
https://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html
To suppress this message, include `message = FALSE` in code chunk header.
```

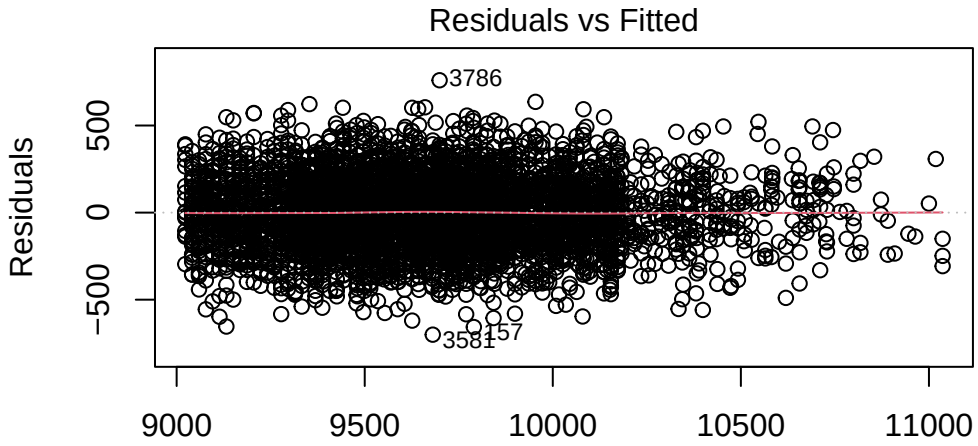
Characteristic	**N = 5,000**
smoke	508 (10%)
female	2,435 (49%)
age	41 (14) [18, 65]
cardiac	190 (3.8%)
cost	9,672 (403) [8,478, 11,326]

Overall we have a dataset with 5,000 patient entries. The dataset was relatively balanced in terms of gender with 49% female and 51% male. About 10% of the cohort were smokers and about 3.8% of the cohort had a cardiac history. The average age was 41. And the Average cost per patient was $9,672.

Next we ran a linear model on the data looking at relationship between cardiac status, gender, and age with cost.

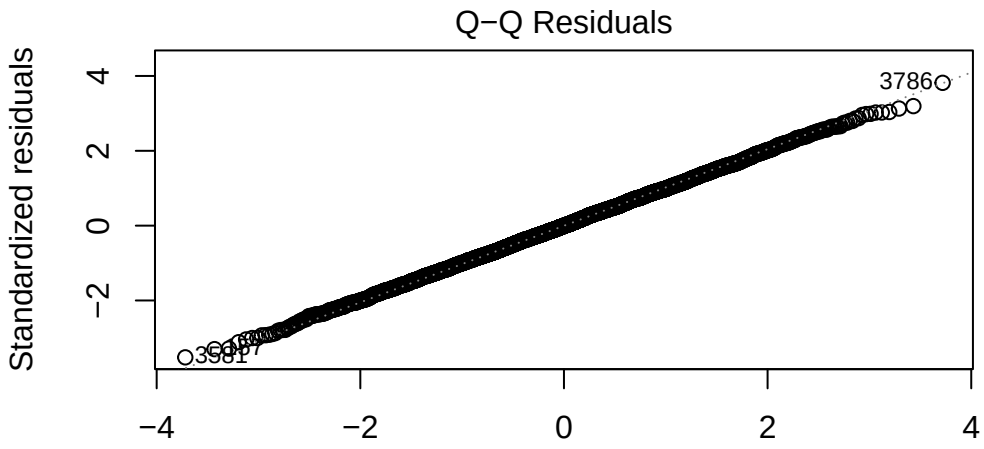
```
model <- lm(cohort_1_$cost ~ cohort_1_$cardiac + cohort_1_$female + cohort_1_$age + cohort_1
summary(model)                                                                          ①

plot(model, 1)
```

① Used lm function to model the fucntion to see how variables are related and to create a residuals vs fitted,a normal q-q, and a scale location plot.

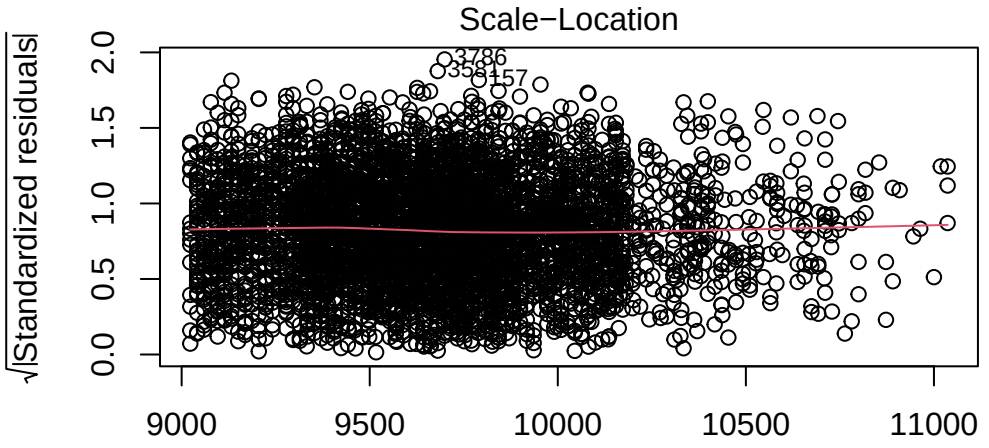


ı(cohort_1_$cost ~ cohort_1_$cardiac + cohort_1_$female + cohort_1_$a(

```
plot(model, 2)
```


Q−Q Residuals
Standardized residuals
3786
3581
−4
−2
0
2
4
Theoretical Quantiles
lm(cohort_1_$cost ~ cohort_1_$cardiac + cohort_1_$female + cohort_1_$ag

```
plot(model, 3)
```


Scale−Location
√|Standardized residuals|
3786
3581
3157
0.0
0.5
1.0
1.5
2.0
9000
9500
10000
10500
11000
Fitted values
lm(cohort_1_$cost ~ cohort_1_$cardiac + cohort_1_$female + cohort_1_$ag

```
Call:
lm(formula = cohort_1_$cost ~ cohort_1_$cardiac + cohort_1_$female +
    cohort_1_$age + cohort_1_$smoke, data = cohort_1_)

Residuals:
    Min      1Q  Median      3Q     Max
-700.87 -137.95   -0.95  136.99  759.92

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       8988.7981     9.5392  942.30   <2e-16 ***
cohort_1_$cardiac  289.2236    15.2189   19.00   <2e-16 ***
cohort_1_$female  -293.6548     5.7041  -51.48   <2e-16 ***
cohort_1_$age       18.2124     0.2081   87.50   <2e-16 ***
cohort_1_$smoke    592.7583     9.5149   62.30   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 199.2 on 4995 degrees of freedom
Multiple R-squared:  0.7555,    Adjusted R-squared:  0.7553
F-statistic:  3859 on 4 and 4995 DF,  p-value: < 2.2e-16
```

<linear model of cardiac status, gender, and age by cost>

When we look estimates for our variables we see there female status is correlated with lower cost. Cardiac status is strongly associated with increased costs, and age is associated with increased costs, but not as much as cardiac status. Our residuals versus fitted plot looks pretty good. Data points are spread out pretty uniformly until cost get to about 10,250.

Link to github: https://github.com/jmyrie312/Assignment7