HRP 203 Final Project

Genna Campain

2024-05-28

Assignment Description

Create a 4 to 6 page Quarto reproducible document to synthesize concepts learned throughout the course and particularly from Module 3. Create the report in a repository on your GitHub account based on the cohort simulated data. Include the following features: introduction section, methods section with notation, results section with data summary table and at least 2 figures, discussion. The assignment is due at 11am Pacific time on Monday, June 10th.

Introduction

Methods

For the data analysis, I used the `cohort` simulated data set provided through the course GitHub repository. A copy of this data set can be found in the `Data` folder in the GitHub repository for this project.

```
  smoke female cardiac age  cost
1     1      0       0  44 10566
2     0      1       0  46  9668
3     0      0       0  56  9889
4     0      0       0  35  9780
5     0      0       0  49 10200
6     0      0       0  64 10082
```

I included all of the variables from the data set in my analysis. Since documentation was not provided, I made assumptions about the meanings of the variables. These assumptions are outlined in Table 1.

Table 1: Definitions for five variables included in analysis

Variable Name	Description
`smoke`	indicator variable, equal to 1 if the patient smokes regularly (more than three times per week) and 0 if not
`female`	indicator variable, equal to 1 if the patient's sex is designated as female and 0 if not
`cardiac`	indicator variable, equal to 1 if the patient has a diagnosis of a cardiac-related problem and 0 if not
`age`	numeric variable, indicates the patient's age in years
`cost`	numeric variable, indicates the cost (in USD) of all healthcare visits for a given patient in a given year

I started the analysis with descriptive statistics for the five variables of interest.

For the next part of the analysis, I ran a regression to examine how an individual's smoking habits, sex, cardiac history, and age can be used to predict their yearly healthcare costs. I used the `lm()` function with the following equation:

$$cost = \beta_0 + \beta_1 smoke + \beta_2 female + \beta_3 cardiac + \beta_4 age$$

Finally, I generated predicted spending amounts for

Results

Data Summary Table

```
varnames <- as.matrix(names(cohort), nrow = 5, ncol = 1)
meanmat <- matrix(data = 0, nrow = 5, ncol = 1)
sdmat <- matrix(data = 0, nrow = 5, ncol = 1)
minmaxmat <- matrix(data = NA, nrow = 5, ncol = 2)
skewmat <- matrix(data = NA, nrow = 5)
for(i in 1:5){
  meanmat[i] <- round(mean(cohort[,i]), digits = 5)
  sdmat[i] <- round(sd(cohort[,i]), digits = 5)
  minmaxmat[i,1] <- round(min(cohort[,i]), digits = 5)
  minmaxmat[i,2] <- round(max(cohort[,i]), digits = 5)
  skewmat[i] <- round(skewness(cohort[,i]), digits = 5)
}
table <- cbind(varnames, minmaxmat, meanmat, sdmat, skewmat)
colnames(table) <- list("Variable", "Min", "Max", "Mean", "SD", "Skewness")
as.data.frame(table)
```

```
  Variable  Min   Max      Mean         SD Skewness
1    smoke    0     1    0.1016    0.30215  2.63656
2   female    0     1     0.487    0.49988    0.052
3  cardiac    0     1     0.038    0.19122  4.83128
4      age   18    65   41.4702    13.5407  0.01173
5     cost 8478 11326 9672.2744  402.63168  0.32417
```

Figure 1

```
# cost regression
reg1 <- lm(cost ~ cardiac + smoke + age + female, data = cohort)
summary(reg1)
```

```

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

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 ***
cardiac      289.2236    15.2189   19.00   <2e-16 ***
smoke        592.7583     9.5149   62.30   <2e-16 ***
age           18.2124     0.2081   87.50   <2e-16 ***
female      -293.6548     5.7041  -51.48   <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
```

```
coeffig <- coefplot(reg1,
                    title = "Coefficients for Linear Regression",
                    color = "Maroon")
coeffig
```

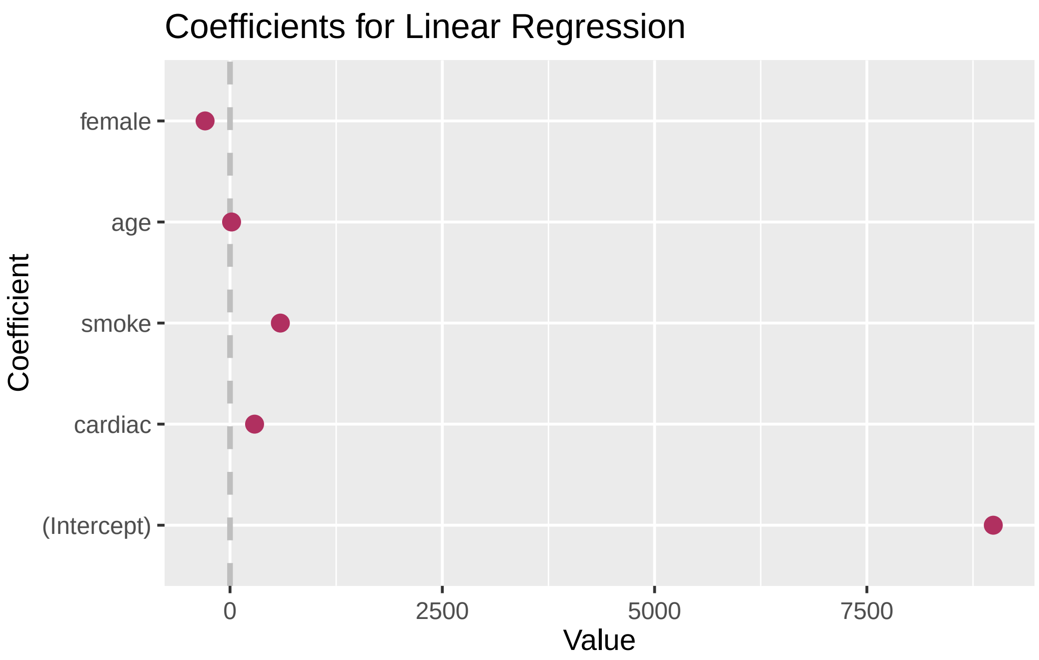
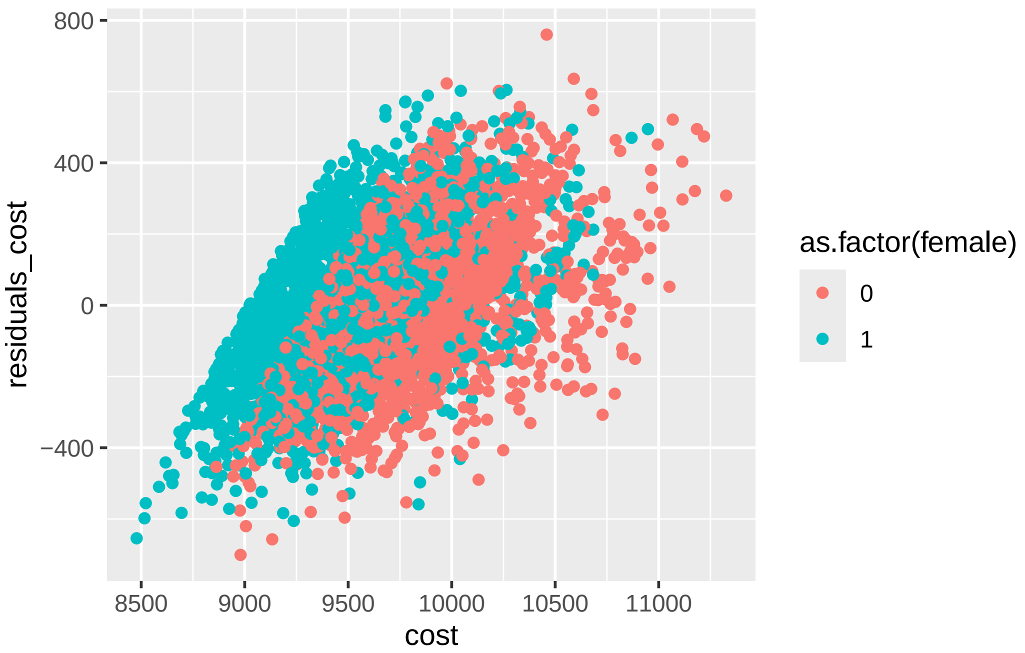


Figure 2

```
# make predictions
cohort$predict_cost <- predict(reg1, cohort)
cohort$residuals_cost <- cohort$cost - cohort$predict_cost
```

```
# plot residuals by sex
ggplot(cohort, aes(cost, residuals_cost)) +
  geom_point(aes(color = as.factor(female)))
```



Discussion

```
1 + 1
```

```
[1] 2
```