Analysis

Genna Campain

2024-05-16

Load Data and Libraries

```
library(dplyr)
# skewness function
library(e1071)
# coefficient plots
library(coefplot)
# set wd
cohort <- read.csv("~/Documents/GitHub/Assignment7/raw-data/cohort.csv") %>%
  select(smoke, female, cardiac, age, cost)
```

Table Describing the Variables

Include mean, sd, min/max, and skewness for each variable

```
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
```

Regression-based Approach

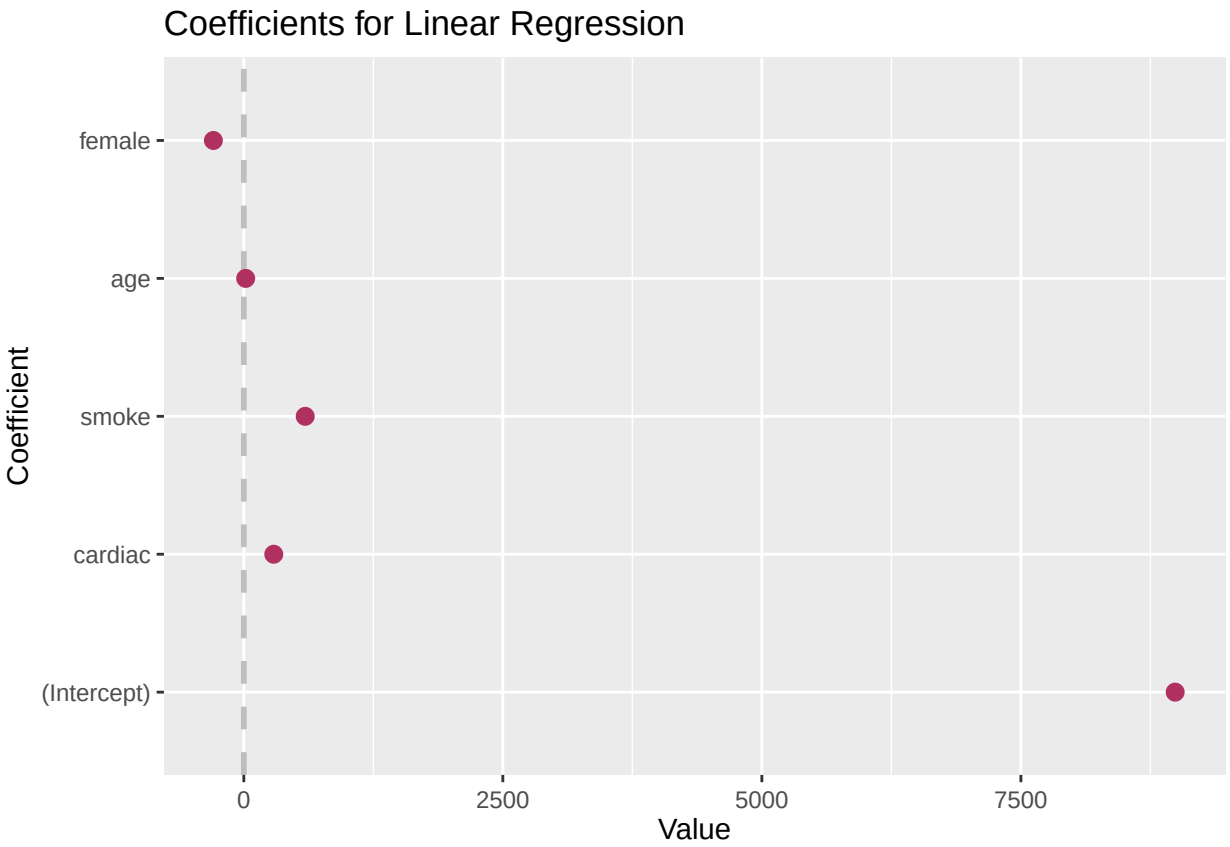
Run regression We don't have descriptions of what exactly the variables mean, but I am going to assume the following: - "Cost" is the cost of a healthcare visit for the individual. This will be my dependent variable of interest. - "Cardiac" indicates whether the individual has previously been seen for a cardiac-related complaint. - "Smoke" indicates whether the person smokes regularily. - "Female" and "Age" are self-explanatory. I am interested in Cost ~ Cardiac + Smoke + Age + Female

```
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
```

Figure

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


Coefficients for Linear Regression
female
age
smoke
cardiac
(Intercept)
Coefficient
0
2500
5000
7500
Value