Module 3 Project

Link to GitHub Repository

https://github.com/elizaennis/Module3

Quarto

Quarto enables you to weave together content and executable code into a finished document. To learn more about Quarto see https://quarto.org.

```
#! label: load-packages
#! include: false
#renv::restore()
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
```

```
library(palmerpenguins)
library(arsenal)
```

```
Attaching package: 'arsenal'

The following object is masked from 'package:lubridate':

    is.Date
```

```r
library(dplyr)
library(readr)
library(table1)
```

```
Attaching package: 'table1'

The following objects are masked from 'package:base':

    units, units<-
```

```r
library(quarto)
library(tinytex)
#tinytex::install_tinytex()
```

Introduction

We begin with a simulated data set of 5000 observations, each assigned 5 characteristics (smoker status, sex, age, cardiac condition, and cost).

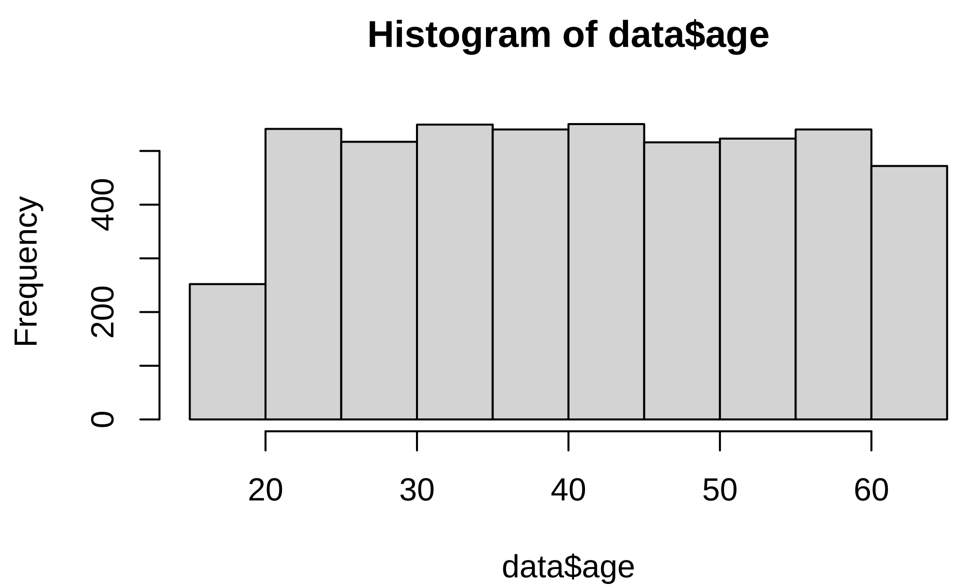
We established cost as our response / dependent variable and female, smoke, age, and cardiac as our predictor variables. Smoke, female, and cardiac are binary, while age and cost are continuous. Based on our initial look at the data, we can see that 10.2% of the observations are of non-smokers, 48.7% are female, 3.8% have a cardiac condition, and the average age is 41.5 with a standard deviation of 13.5 years and an approximately uniform distribution between ages 18-65. For costs, we can see that costs are approximately normally distributed, and the mean cost is $9,670. When cost is made into a categorical variable, we find that 3.8% of observations fall below $9,000, 76.5% between ~$9,000-$10,000, and the remaining 19.9% are above $10,000.

Using this data set, we will then use several different methods to identify the association between each of the predictor variables and costs.

```r
#! label: load-data-and-make-table-1
#! include: false

# Read in simulated data
current_dir <- getwd()
data_path <- "cohort.csv"
output_path <- "output.csv"
data <- read.csv(data_path)

# Get information about data
str(data)
```

```
'data.frame':	5000 obs. of  5 variables:
 $ smoke  : int  1 0 0 0 0 0 0 0 0 0 ...
 $ female : int  0 1 0 0 0 0 1 0 0 0 ...
 $ age    : int  44 46 56 35 49 64 46 60 31 35 ...
 $ cardiac: int  0 0 0 0 0 0 0 0 0 0 ...
 $ cost   : int  10566 9668 9889 9780 10200 10082 9461 9737 9779 9758 ...
```

```r
hist(data$age)
```



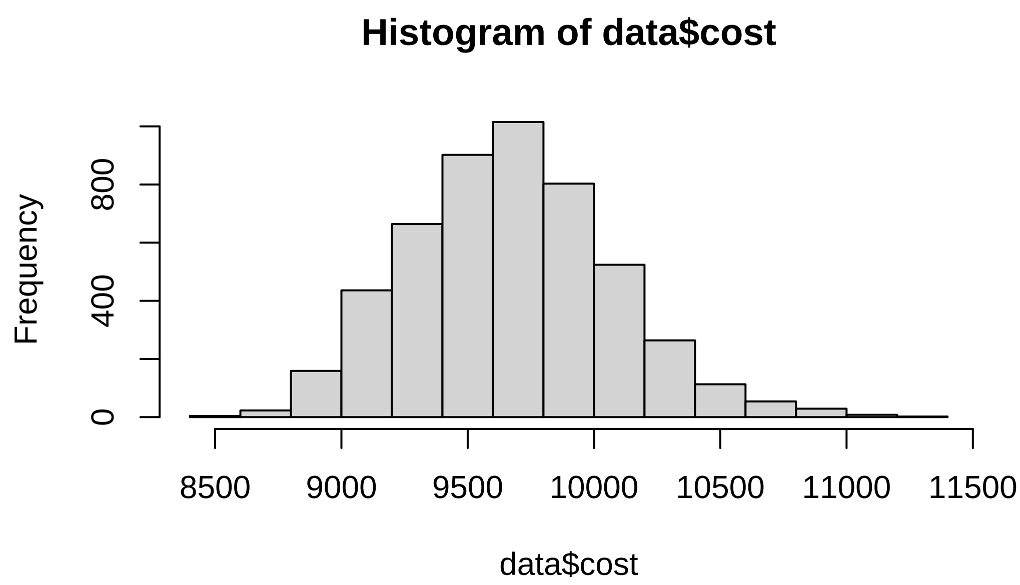
```r
min(data$age)
```

```
[1] 18
```

```r
max(data$age)
```

```
[1] 65
```

```r
hist(data$cost)
```




```r
#Reformat data to work for table 1
max(data$cost)
```

```
[1] 11326
```

```r
median(data$cost)
```

```
[1] 9664
```

```r
min(data$cost)
```

```
[1] 8478
```

```r
data$cost_cat <- 0
data$cost_cat <- ifelse(data$cost > 10000, "C: >$10,000",
                        ifelse(data$cost <= 10000 & data$cost >= 9000, "B: $9000-$9999",
                               ifelse(data$cost < 9000 & data$cost >= 8000, "A: $8000-$8999"
table(data$cost_cat)
```

```

A: $8000-$8999 B: $9000-$9999    C: >$10,000
           181           3825            994
```

```r
data$cost_cat <- as.factor(data$cost_cat)
data <- data %>%
  mutate(smoke = case_when(
    smoke == 1 ~ "smoker",
    smoke == 0 ~ "non-smoker",
    TRUE ~ NA_character_
  )) %>%
  mutate(sex = case_when(
    female == 1 ~ "female",
    female == 0 ~ "male",
    TRUE ~ NA_character_
  )) %>%
  mutate(cardiac = case_when(
    cardiac == 1 ~ "cardiac_condition",
    cardiac == 0 ~ "no_condition",
    TRUE ~ NA_character_
  )) %>%
  select (smoke,sex,age,cardiac,cost,cost_cat)

#Make Table 1
(Table1 <- table1(~ smoke + sex + age + cardiac + cost + cost_cat, data=data))
```

```
Get nicer `table1` LaTeX output by simply installing the `kableExtra` package
```

	Overall
	(N=5000)
smoke	
non-smoker	4492 (89.8%)
smoker	508 (10.2%)
sex	
female	2435 (48.7%)
male	2565 (51.3%)
age	
Mean (SD)	41.5 (13.5)
Median [Min, Max]	41.0 [18.0, 65.0]
cardiac	
cardiac_condition	190 (3.8%)
no_condition	4810 (96.2%)
cost	
Mean (SD)	9670 (403)
Median [Min, Max]	9660 [8480, 11300]
cost_cat	
A: $8000-$8999	181 (3.6%)
B: $9000-$9999	3825 (76.5%)
C: >$10,000	994 (19.9%)

Methods

My exploration of the association between smoking, sex, age, and history of cardiac condition and costs began with getting a general understanding of the data by calculating means, medians, standard deviations, and distribution types for continous variables and the percentage of observations fitting each characteristic for categorical variables. I then looked at the proportion of each predictor variable that fell into cost categories to get a sense of potential associations. Then, I used a linear regression model with cost as a continuous outcome variable to identify the dollar increases associated with a change in each predictor variable. To better understand the odds ratios and relative impact of each, I also ran a generalized linear model (glm) using "high" and "low" cost categories divided at the median cost. We can use lm and glm methods because we can assume a linear correlation between the variables.

Results

We find that smoking, male sex, older age, and history of cardiac condition are all associated with higher costs. More specifically, using a basic linear regression model, we find that one additional year of age is associated with an $18 increase in costs, being a smoker is associated with a $593 increase in costs, being male is associated with a $294 increase in costs, and having a cardiac condition is associated with a $289 increase in costs. Accounting for interactions between the predictor variables, the increase in costs associated with smoking is just $504 while being male is associated with increasing costs by $308 and having a cardiac condition is associated with increased costs of $309. Among the predictor variables in the model, smoking status exhibited the highest odds ratio with categorical high/low cost. Between predictor variables, the greatest correlation is between smoking and cardiac history and the second is between male sex and cardiac history.

```r
#! label: analyze_data
#! include: false

#Relevel data
data$cardiac <- as.factor(data$cardiac)
data$cardiac <- relevel(data$cardiac, ref = "no_condition")
data$sex <- as.factor(data$sex)
data$sex <- relevel(data$sex, ref = "male")
data$smoke <- as.factor(data$smoke)
data$smoke <- relevel(data$smoke, ref = "non-smoker")

#Build table demonstrating differences in predictor values by cost
(Table2 <- table1(~ smoke + sex + age + cardiac | cost_cat, data=data))
```

```
Get nicer `table1` LaTeX output by simply installing the `kableExtra` package
```

	A: $8000-$8999	B: $9000-$9999	C: >$10,000	Overall
	(N=181)	(N=3825)	(N=994)	(N=5000)
smoke				
non-smoker	181 (100%)	3692 (96.5%)	619 (62.3%)	4492 (89.8%)
smoker	0 (0%)	133 (3.5%)	375 (37.7%)	508 (10.2%)
sex				
male	11 (6.1%)	1790 (46.8%)	764 (76.9%)	2565 (51.3%)
female	170 (93.9%)	2035 (53.2%)	230 (23.1%)	2435 (48.7%)
age				
Mean (SD)	25.4 (5.62)	39.5 (12.7)	52.1 (10.9)	41.5 (13.5)
Median [Min, Max]	24.0 [18.0, 44.0]	39.0 [18.0, 65.0]	55.0 [18.0, 65.0]	41.0 [18.0, 65.0]
cardiac				
no_condition	181 (100%)	3757 (98.2%)	872 (87.7%)	4810 (96.2%)
cardiac_condition	0 (0%)	68 (1.8%)	122 (12.3%)	190 (3.8%)

```r
#Build linear regression model to determine variable relationships with cost as continuous ou
(model1 <- lm(cost ~ age + smoke + sex + cardiac, data = data))
```

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

Coefficients:
            (Intercept)                       age               smokesmoker
                8988.80                     18.21                    592.76
              sexfemale  cardiaccardiac_condition
                -293.65                    289.22
```

```r
(model1_interaction <- lm(cost ~ age * smoke * sex * cardiac, data = data))
```

```
Call:
lm(formula = cost ~ age * smoke * sex * cardiac, data = data)

Coefficients:
                          (Intercept)
                            8995.7344
                                  age
                              18.1066
                          smokesmoker
                             503.6356
                            sexfemale
                            -308.4684
             cardiaccardiac_condition
                             309.4004
                      age:smokesmoker
                               1.6694
                        age:sexfemale
                               0.2512
                smokesmoker:sexfemale
                             152.3952
         age:cardiaccardiac_condition
                              -1.4400
 smokesmoker:cardiaccardiac_condition
                             112.7956
   sexfemale:cardiaccardiac_condition
```

```
                                         28.9362
                       age:smokesmoker:sexfemale
                                         -2.9258
          age:smokesmoker:cardiaccardiac_condition
                                         -0.9863
            age:sexfemale:cardiaccardiac_condition
                                          2.8248
    smokesmoker:sexfemale:cardiaccardiac_condition
                                       -151.9435
age:smokesmoker:sexfemale:cardiaccardiac_condition
                                          1.3594
```

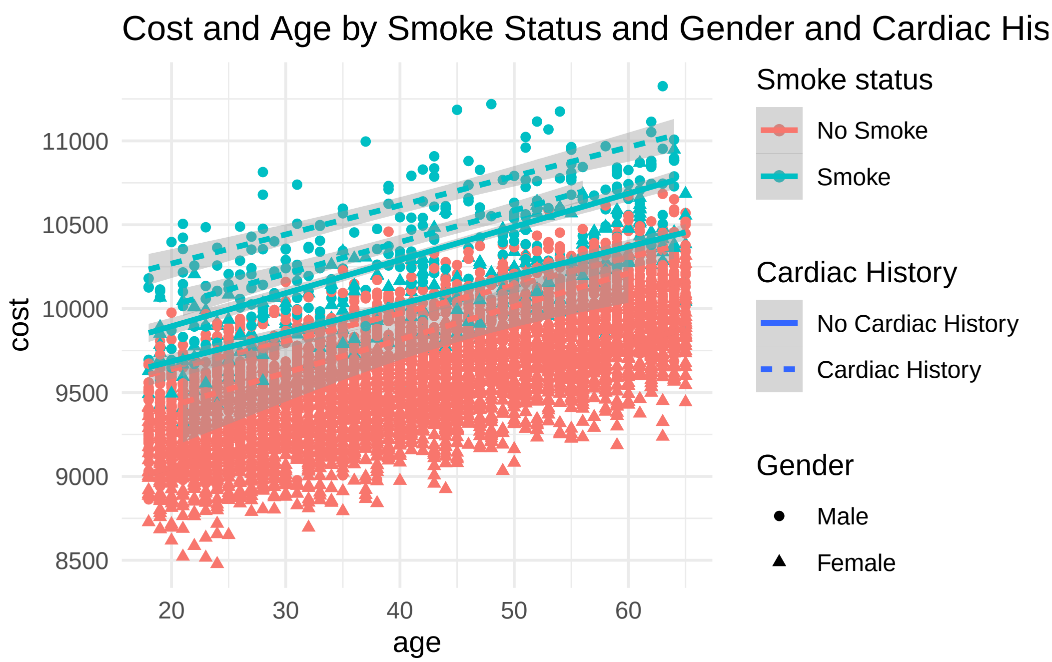
```
summary(model1_interaction)
```

```
Call:
lm(formula = cost ~ age * smoke * sex * cardiac, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-703.78 -136.56   -1.57  136.64  757.11

Coefficients:
                                                    Estimate Std. Error t value
(Intercept)                                        8995.7344    13.6880 657.199
age                                                  18.1066     0.3146  57.556
smokesmoker                                         503.6356    47.2480  10.659
sexfemale                                          -308.4684    19.4634 -15.849
cardiaccardiac_condition                            309.4004    61.8140   5.005
age:smokesmoker                                       1.6694     1.0905   1.531
age:sexfemale                                         0.2512     0.4456   0.564
smokesmoker:sexfemale                               152.3952    64.2570   2.372
age:cardiaccardiac_condition                         -1.4400     1.5175  -0.949
smokesmoker:cardiaccardiac_condition                112.7956   105.1050   1.073
sexfemale:cardiaccardiac_condition                   28.9362   200.9880   0.144
age:smokesmoker:sexfemale                            -2.9258     1.4673  -1.994
age:smokesmoker:cardiaccardiac_condition             -0.9863     2.5107  -0.393
age:sexfemale:cardiaccardiac_condition                2.8248     4.3739   0.646
smokesmoker:sexfemale:cardiaccardiac_condition     -151.9435   384.7106  -0.395
age:smokesmoker:sexfemale:cardiaccardiac_condition    1.3594     8.7141   0.156
                                                   Pr(>|t|)
(Intercept)                                         < 2e-16 ***
```

```
age                                                < 2e-16 ***
smokesmoker                                        < 2e-16 ***
sexfemale                                          < 2e-16 ***
cardiaccardiac_condition                          5.77e-07 ***
age:smokesmoker                                     0.1259
age:sexfemale                                       0.5730
smokesmoker:sexfemale                               0.0177 *
age:cardiaccardiac_condition                        0.3427
smokesmoker:cardiaccardiac_condition                0.2832
sexfemale:cardiaccardiac_condition                  0.8855
age:smokesmoker:sexfemale                           0.0462 *
age:smokesmoker:cardiaccardiac_condition            0.6944
age:sexfemale:cardiaccardiac_condition              0.5184
smokesmoker:sexfemale:cardiaccardiac_condition      0.6929
age:smokesmoker:sexfemale:cardiaccardiac_condition  0.8760
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 199 on 4984 degrees of freedom
Multiple R-squared:  0.7563,    Adjusted R-squared:  0.7556
F-statistic:  1031 on 15 and 4984 DF,  p-value: < 2.2e-16
```

```r
#Build a figure with all variables
(Figure1 <- ggplot(data, aes(x = age, y = cost, color = smoke, shape = sex,linetype = cardiac
  geom_point() +
  geom_smooth(method = "lm") +
  labs(
    title = "Cost and Age by Smoke Status and Gender and Cardiac History",
    color = "Smoke status",
    shape = "Gender",
    linetype = "Cardiac History"
  ) +
  scale_color_discrete(labels = c("No Smoke", "Smoke")) +
  scale_shape_discrete(labels = c("Male", "Female")) +
  scale_linetype_discrete(labels = c("No Cardiac History", "Cardiac History")) +
  theme_minimal())
```

```
`geom_smooth()` using formula = 'y ~ x'
```

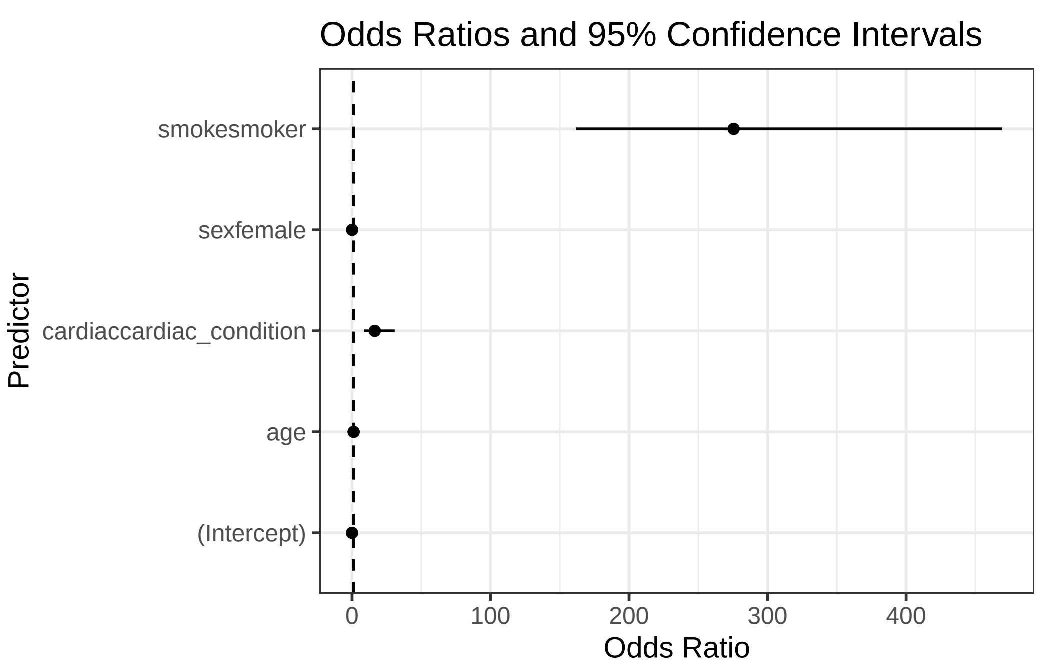



```
#Analyze with cost as binary outcome (above/below median) to calculate odds ratios
data$cost_highlow <- 0
data$cost_highlow <- ifelse(data$cost > median(data$cost), 1, 0)
(model2 <- glm(cost_highlow ~ sex + age + smoke + cardiac, data = data, family = binomial(li
```

```
Call:  glm(formula = cost_highlow ~ sex + age + smoke + cardiac, family = binomial(link = "lo
    data = data)

Coefficients:
             (Intercept)                 sexfemale                       age
                 -5.8680                   -2.6189                    0.1605
             smokesmoker  cardiaccardiac_condition
                  5.6186                    2.7987

Degrees of Freedom: 4999 Total (i.e. Null);  4995 Residual
Null Deviance:       6931
Residual Deviance: 3514      AIC: 3524
```

```
coef_summary <- summary(model2)$coefficients

# Calculate odds ratios and their confidence intervals
```

```r
odds_ratios <- exp(coef_summary[, "Estimate"])
ci_lower <- exp(coef_summary[, "Estimate"] - 1.96 * coef_summary[, "Std. Error"])
ci_upper <- exp(coef_summary[, "Estimate"] + 1.96 * coef_summary[, "Std. Error"])

# Combine results into a data frame
odds_ratios_df <- data.frame(
  OddsRatio = odds_ratios,
  LowerCI = ci_lower,
  UpperCI = ci_upper,
  Predictor = rownames(coef_summary)
)

# Plot odds ratios and confidence intervals
(Figure2 <- ggplot(odds_ratios_df, aes(x = OddsRatio, y = Predictor)) +
  geom_point() +
  geom_errorbarh(aes(xmin = LowerCI, xmax = UpperCI), height = 0) +
  geom_vline(xintercept = 1, linetype = "dashed") +
  labs(x = "Odds Ratio", y = "Predictor", title = "Odds Ratios and 95% Confidence Intervals")
  theme_bw())
```




```r
#quarto_render("Module3Project.qmd", output_format = "pdf")
```