STA 138 Project

Yunhao Yang, Tianyi Feng, Chan Woong Joo

December 7th, 2021

Abstract

This project aims to address byssinosis, a form of pneumoconiosis to which workers are exposed to cotton dust. In the project, using the Byssinosis data, we will consider five variables which may be related to the illness. Most of these variables may show relationships with the illness, and yet we try to find that some of them are not independent and we will use logistic regression models to show which variables are significant.

Introduction

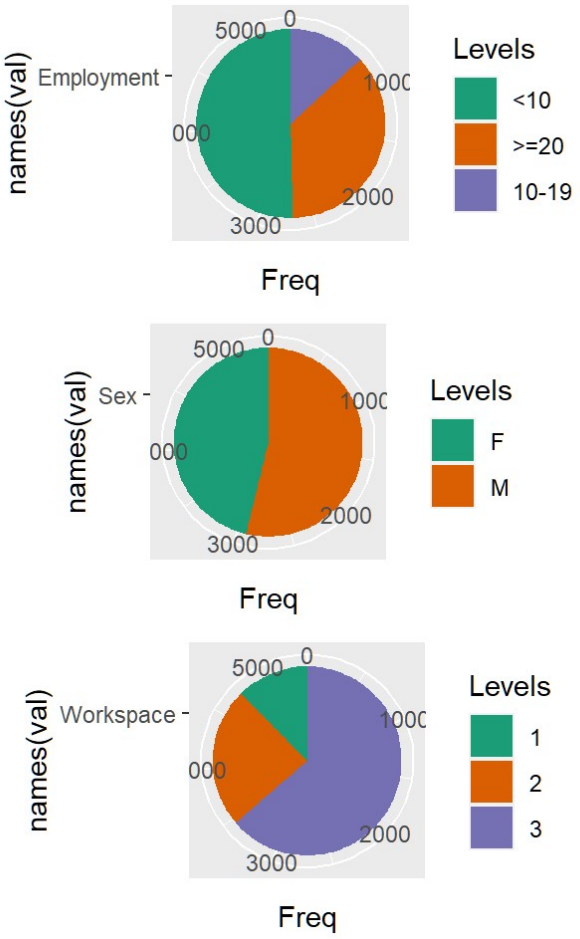
In 1973, a large cotton textile company in North Carolina participated in a study to investigate the prevalence of byssinosis, a form of pneumoconiosis to which workers exposed to cotton dust are subject. Data was collected on 5,419 workers, including:

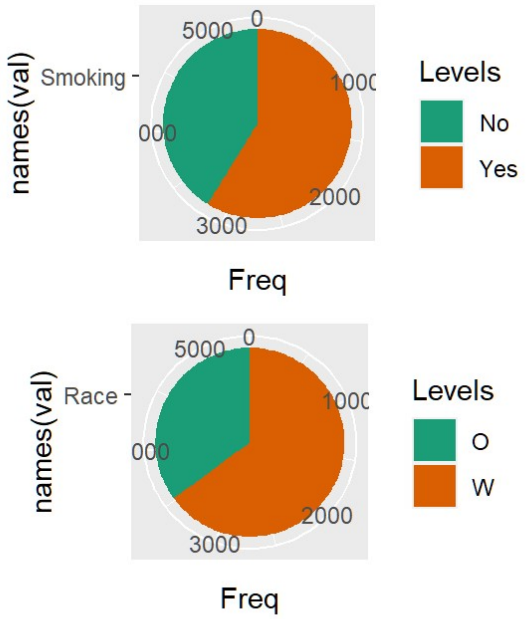
- Type of work place [1 (most dusty), 2 (less dusty), 3 (least dusty)]
- Employment, years [< 10, 10–19, 20–]
- Smoking [Smoker, or not in last 5 years]
- Sex [Male, Female]
- Race [White, Other]
- Byssinosis [Yes, No]

Through visualization, logistic regression, and modeling, we aim to investigate the relationship between the presence of byssinosis on different variables. The analysis hereafter was conducted using R and its built-in functions.

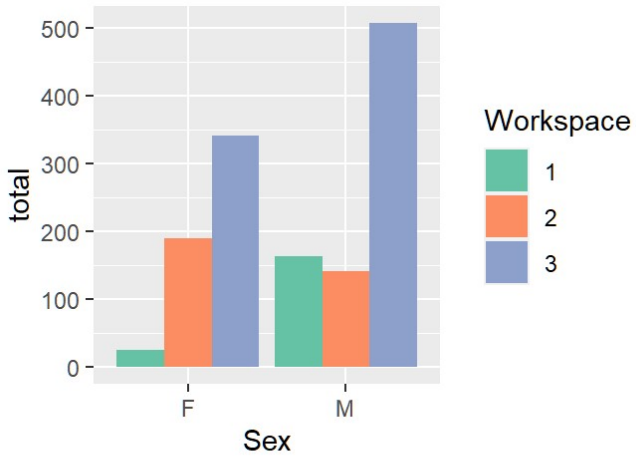
Visualization

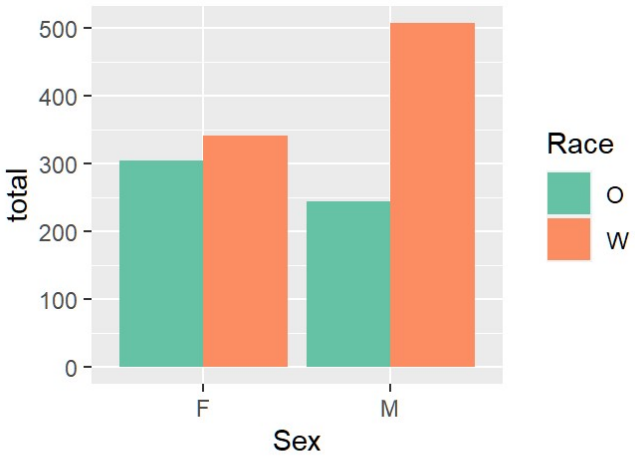
We have implemented pie charts to first visualize the presence of byssinosis across five different variables: Employment years, smoking, sex, race, and workplace. The resulting pie charts are as follows:

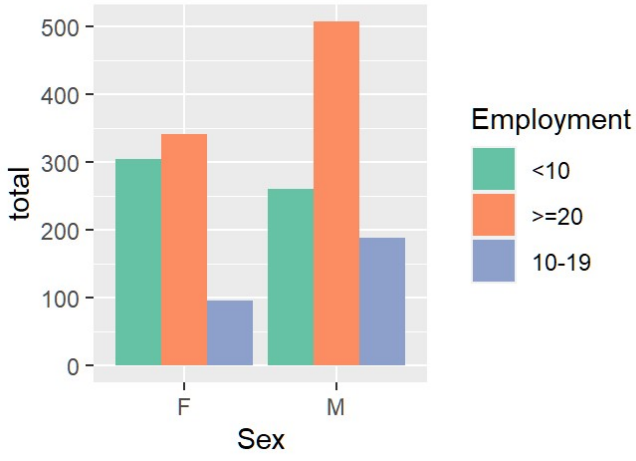




Pie charts imply that there are some disproportionalities among different genders. The following bar charts suggest that there clearly are some differences.





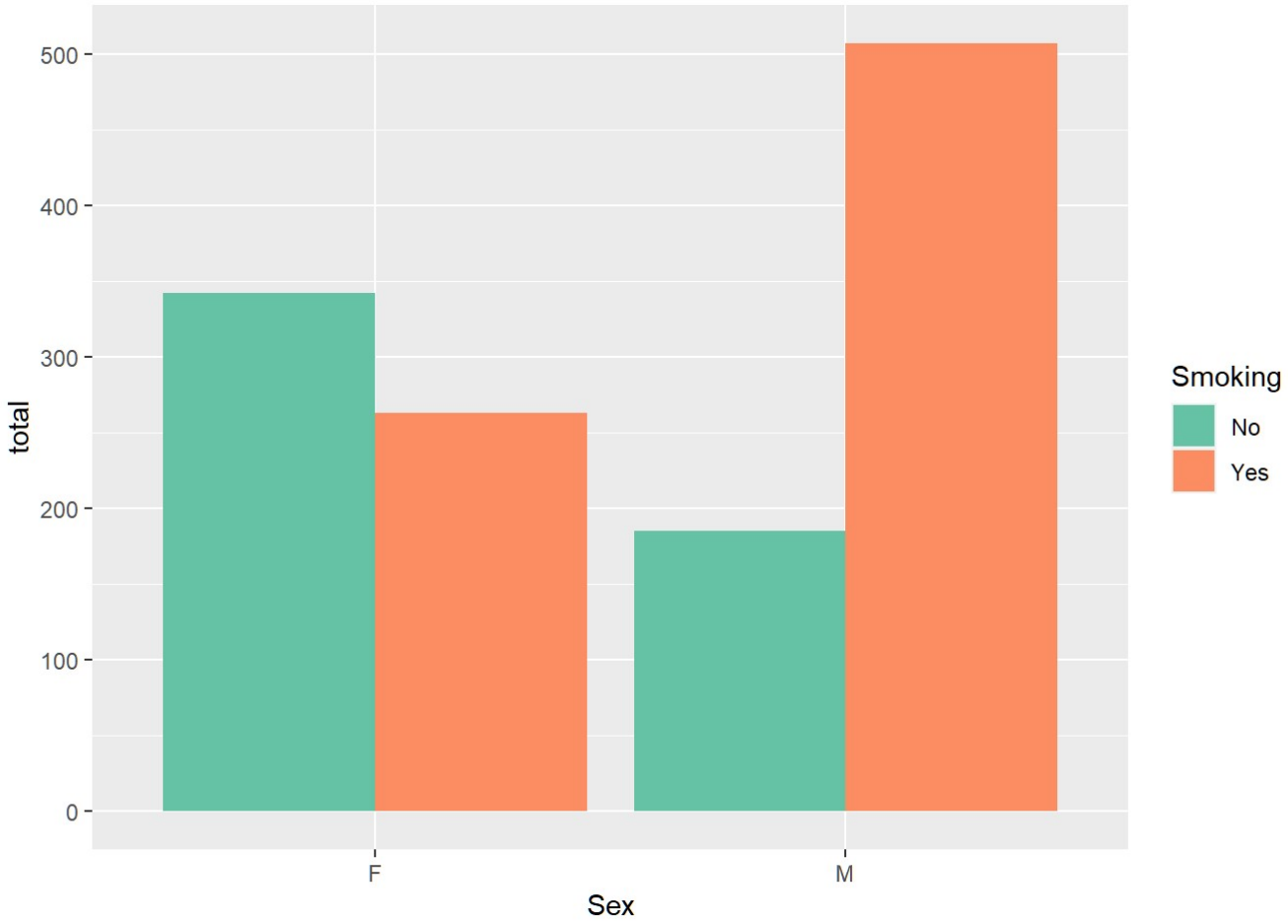


Male workers are more dominant in some areas than females and thus the distribution of some variables are disproportional, i.e. workspace and employment. This fact was taken into consideration in logistic regression modeling and variable selections in the Modeling section.

The presence of byssinosis (Y, response variable) is clearly more prevalent in some variables than the other. For example, the response variable seems to have more proportions among male than female, smokers than non-smokers, whites (race) than non-whites, and those who work in least dusty than in most dusty, surprisingly.

To better understand these variables, we further analyzed by creating contingency tables on different sex and smoking,and visualizing them through box plots:

	No	Yes
F	**1373**	**1130**
M	**857**	**2059**



Independence of these variables are then considered and tested using Chi-square Test of Independence. The test-statistic and corresponding p-value are:

variables	Statistic	P.value
Sex and Smoking	**359.5959**	**0**

The Chi-square Test of Independence suggests that smoking and sex are not independent and the proportion of smoking in Male is significantly higher than that in Female.

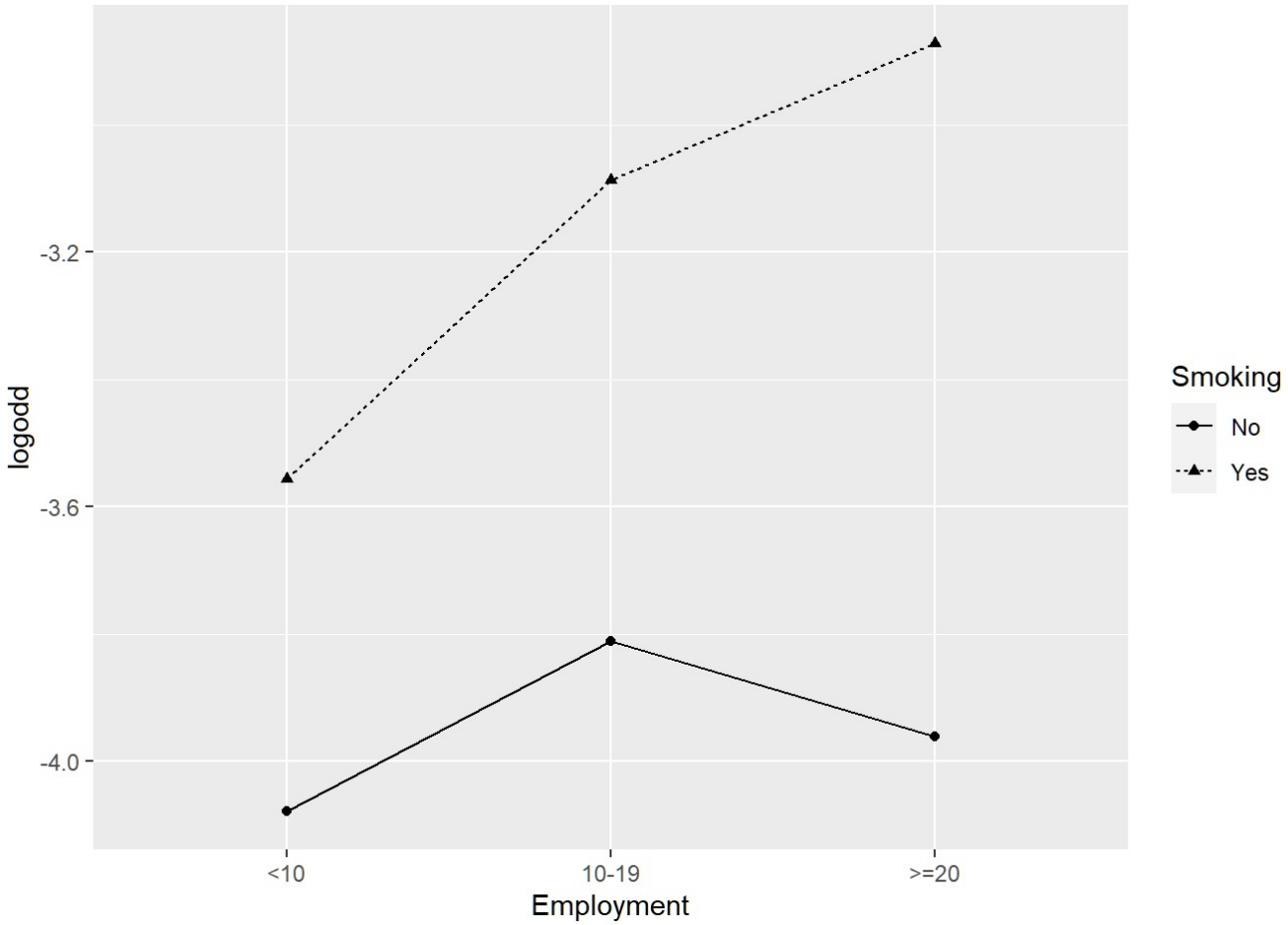
Furthermore, we analyzed by again creating a contingency table for Sex and Byssinosis, and the corresponding Chi-square Test for independence is as follows:

	0	1
F	**2466**	**37**
M	**2788**	**128**

variables	Statistic	P.value
Sex and Byssinosis	**37.69079**	**0**

The Chi-square Test of Independence also suggests that sex and byssinosis are not independent, and Male has higher proportion of Byssinosis present. Sex and experience of smoking may influence the presence of Byssinosis. To figure out this problem, we aim to formulate logistic regression models and henceforth select the significant variables to explain it in the modeling section.

Modeling



The first model we built is an additive model that contains all the variables, named model 1. All other variables are significant except Sex and Race. This indicates sex and race have little effect on the response. So we decided to delete them.

The second model we built is an additive model that drops Sex and Race, named model 2. This time all variables are significant and the corresponding AIC is 122.66. This suggests the model is good.

To better fit the data, a third model is fitted using all interaction terms from variables excluding Sex and Race, and a AIC stepwise procedure is performed, named model 3. The residual variables are similar to model 2 except for the interaction term for Employment and Smoking is kept. The AIC is 122.01 and is better than model 2.

In order to choose from model 2 and model 3, we can compare log-odds of the response in different categories of Smoking and Employment. For those who smokes, there is clear interaction effect between employment year and smoking. The longer they work, the more possible they get the disease. But for people who do not smoke, the interaction relationship is vague. This is reasonable because people with longer employment years are typically older. And their years of smoking are also potentially longer, which poses a positive interaction relationship.

Appendix

All R codes:

```
library(ggplot2)
library(knitr)
library(kableExtra)
library(gridExtra)
library(dplyr)

setwd('D:\\Documents\\课程\\2021 Fall\\STA 138\\final-project')
data = read.csv("Byssinosis.csv")
total = data$Byssinosis+data$Non.Byssinosis
data$total = total
data$Workspace = as.factor(data$Workspace)
data$Employment = as.factor(data$Employment)
data$Race = as.factor(data$Race)

data2 = c()
for (i in 1:(dim(data)[1])){
  if(data[i,6]!=0){
    data2 = rbind(data2,cbind(data[i,1:5],data.frame(y = rep(1,data[i,6]))))
  }
  if(data[i,7]!=0){
    data2 = rbind(data2,cbind(data[i,1:5],data.frame(y = rep(0,data[i,7]))))
  }
}

p = function(i){
  val = data2[i]
  table =as.data.frame(table(val))
  colnames(table) = c("Levels","Freq")
   ggplot(table, aes(x= names(val), y = Freq, fill = Levels))+geom_bar(width = 1, s
tat = "identity") +   coord_polar("y", start=0) + scale_fill_brewer(palette="Dark
2")
}

grid.arrange(p(1), p(2),p(3),p(4),p(5), ncol=2,nrow = 3)
p1 = ggplot(data=data, aes(x=Sex,y = total, fill =Workspace)) +geom_bar(stat="ident
ity", position=position_dodge())+ scale_fill_brewer(palette = "Set2")


p2 = ggplot(data=data, aes(x=Sex,y = total, fill =Race)) +geom_bar(stat="identity",
position=position_dodge())+ scale_fill_brewer(palette = "Set2")

p3 = ggplot(data=data, aes(x=Sex,y = total, fill =Employment)) +geom_bar(stat="iden
tity", position=position_dodge())+ scale_fill_brewer(palette = "Set2")

grid.arrange(p1,p2,p3, ncol=2,nrow = 2)
table1 = table(data2$Sex, data2$Smoking)
form1 = table1 %>%  kable() %>% kable_styling("responsive",full_width = F, position
= "center")%>%column_spec(2:3, bold = T,background = "#F5DEB3") %>% column_spec(1,
bold = T)


p1 = ggplot(data=data, aes(x=Sex,y = total, fill = Smoking)) +geom_bar(stat="identi
ty", position=position_dodge())+ scale_fill_brewer(palette = "Set2")
```

```
form1
print(p1)
test1 = chisq.test(table1)
form2 = data.frame(variables = c("Sex and Smoking"), Statistic = c(test1$statistic
[[1]]), P.value = c(test1$p.value[[1]]))
form2 %>% kable() %>% kable_styling("responsive",full_width = F, position = "cente
r")%>%column_spec(1:3, bold = T,background = "#F5DEB3")

table1 = table(data2$Sex, data2$y)
form1 = table1 %>%  kable() %>% kable_styling("responsive",full_width = F, position
= "center")%>%column_spec(2:3, bold = T,background = "#F5DEB3") %>% column_spec(1,
bold = T)



form1

test1 = chisq.test(table1)
form2 = data.frame(variables = c("Sex and Byssinosis"), Statistic = c(test1$statist
ic[[1]]), P.value = c(test1$p.value[[1]]))
form2 %>% kable() %>% kable_styling("responsive",full_width = F, position = "cente
r")%>%column_spec(1:3, bold = T,background = "#F5DEB3")

rm(list = ls())
setwd('D:\\Documents\\课程\\2021 Fall\\STA 138\\final-project')
data = read.csv('Byssinosis.csv')
data$Employment = factor(data$Employment,levels = c('<10','10-19','>=20'))

data1 = data
set.seed(1)
for (i in 1:5){
  data[,i] = as.factor(data[,i])
  data1[,i] = as.numeric(data[,i])
}
chisq.test(data1,correct = F)
n = sum(data[,6:7])

ind = sample(n,round(3*n/4))

train = data[ind,]
test = data[-ind,]

lm1 = glm(cbind(Byssinosis,Non.Byssinosis)~.,train,family = 'binomial')

lm2 = glm(cbind(Byssinosis,Non.Byssinosis)~.-Sex-Race,train,family = 'binomial')

lm3 = glm(cbind(Byssinosis,Non.Byssinosis)~(.-Sex-Race)^2,train,family = 'binomial
')

lm4 = step(lm3,direction = 'both',k = 2)

lm5 = step(lm3,direction = 'both',k = log(n))

intdata = data %>% group_by(Employment,Smoking)%>%summarise(logodd = log(sum(Byssin
osis)/sum(Non.Byssinosis)))
ggplot(intdata, aes(x=Employment, y=logodd, group=Smoking)) +
```

```
  geom_line(aes(linetype=Smoking))+
  geom_point(aes(shape=Smoking))
```