STA_138_Final_Project

Edwin Que, Charles Chien

12/14/2020

1. 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]

Our task is to investigate relationships between this disease on the one hand and smoking status, sex, race, length of employment, smoking, and dustiness of workplace on the other.

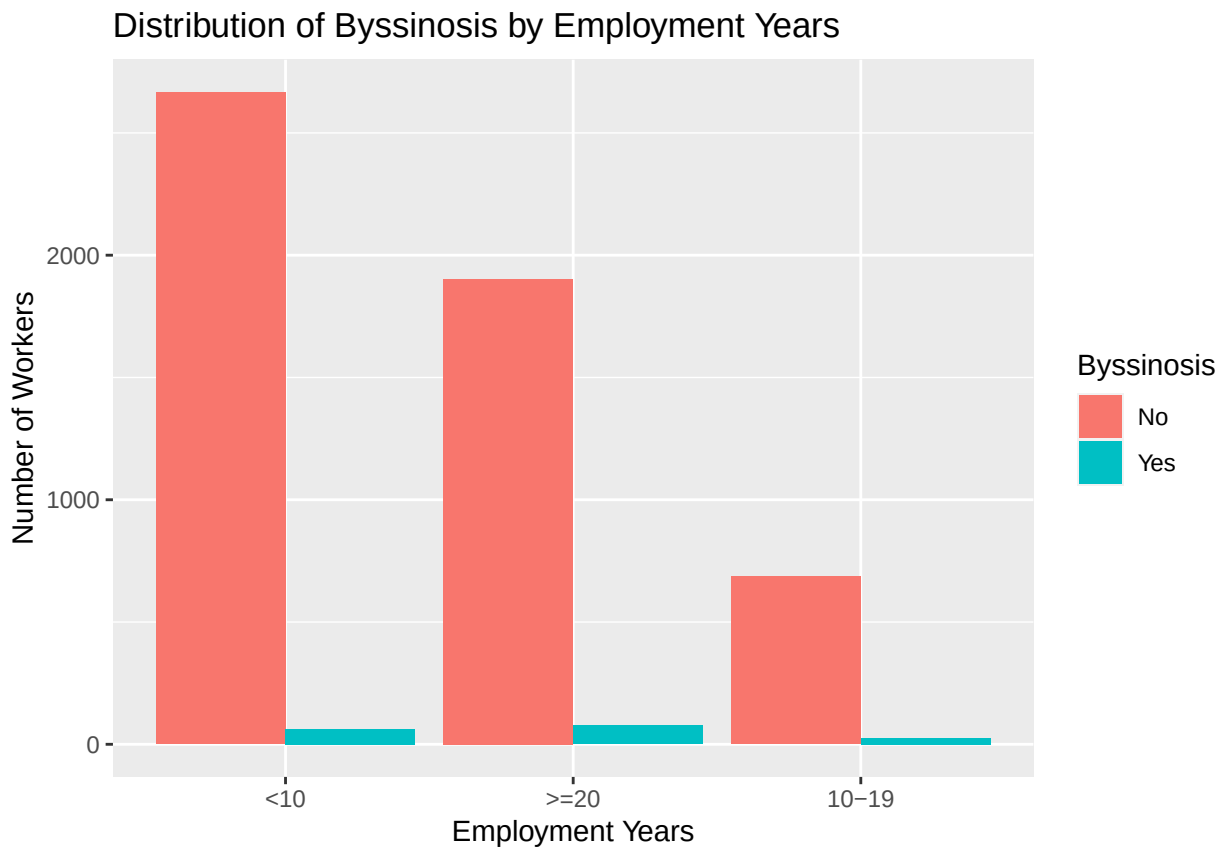
2. Exploratory Data Analysis

Here, we would establish preliminary assumptions on variable effects, and validate them afterwards with proper tests.

2.1 Employment

From our results, it seem like the proportion of having byssinosis is roughly the same across all employment years, with >=20 being the highest and <10 the lowest.

- $P(Byssinosis| < 10) = 63/(2666 + 63) = 0.02308538$
- $P(Byssinosis| >= 20) = 76/(1902 + 76) = 0.03842265$
- $P(Byssinosis|10 - 19) = 26/(26 + 686) = 0.03651685$

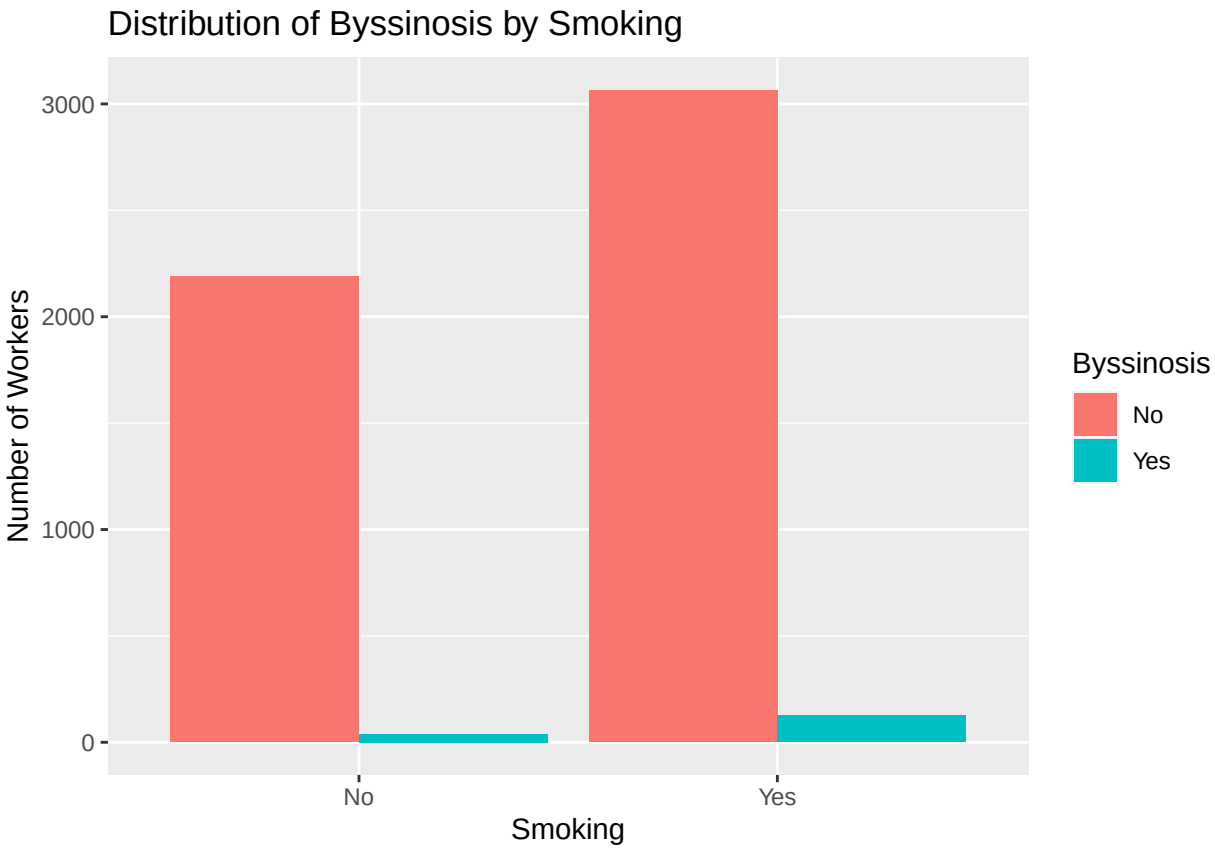



```
##
##           No  Yes
##   <10   2666   63
##   >=20  1902   76
##   10-19  686   26
```

2.2 Smoking

From our results, it seem like the proportion of having byssinosis is higher for smokers than for nonsmokers.

- $P(Byssinosis|NonSmokers) = 40/(2190 + 40) = 0.01793722$
- $P(Byssinosis|Smokers) = 125/(125 + 3064) = 0.03919724$




```
##
##         No  Yes
##   No  2190   40
##   Yes 3064  125
```

2.3 Sex

From our results, it seem like the proportion of having byssinosis is higher for male than for female.

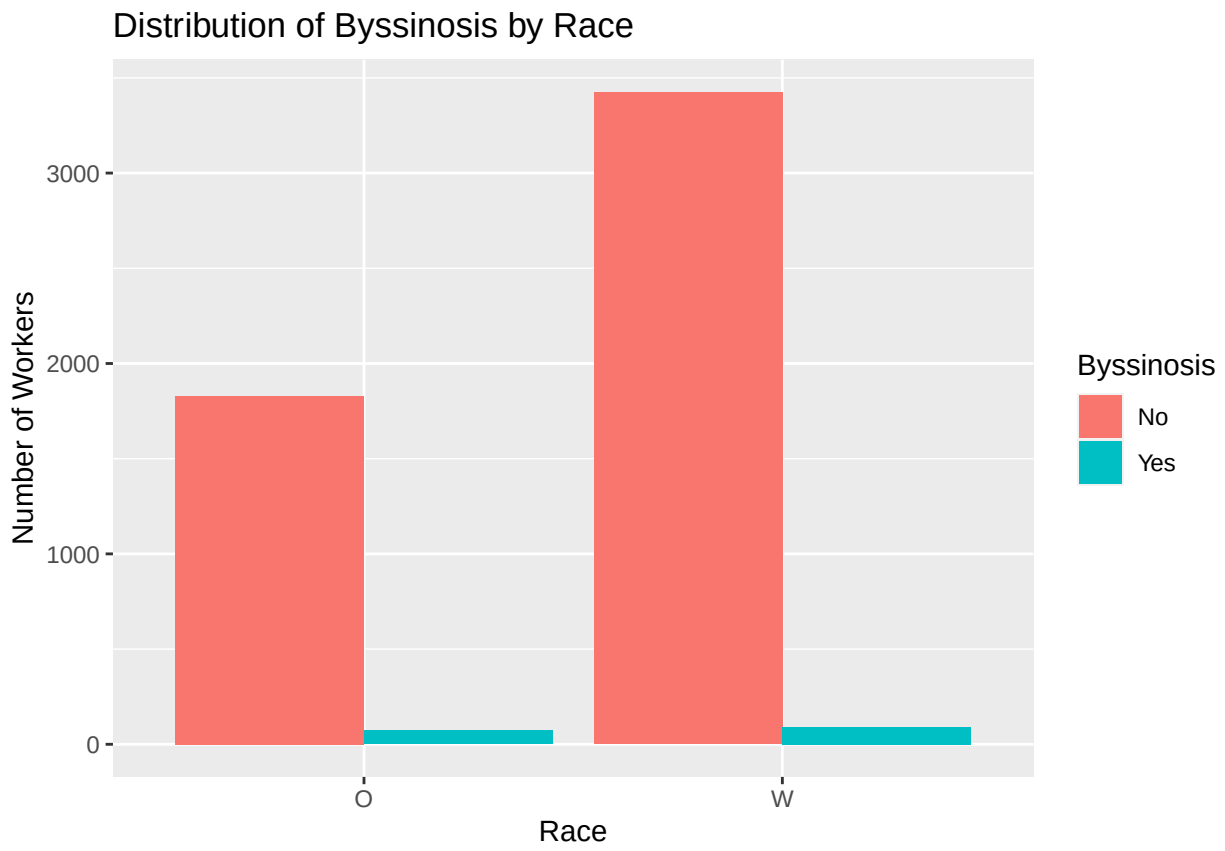
- $P(Byssinosis|Female) = 37/(2466 + 37) = 0.01478226$
- $P(Byssinosis|Male) = 128/(2788 + 128) = 0.04389575$


```
##
##       No  Yes
##   F 2466   37
##   M 2788  128
```

2.4 Race

From our results, it seem like the proportion of having byssinosis is slightly higher for other races than white people.

- $P(Byssinosis|OtherRace) = 73/(1830 + 73) = 0.03836048$
- $P(Byssinosis|White) = 92/(3424 + 92) = 0.0261661$

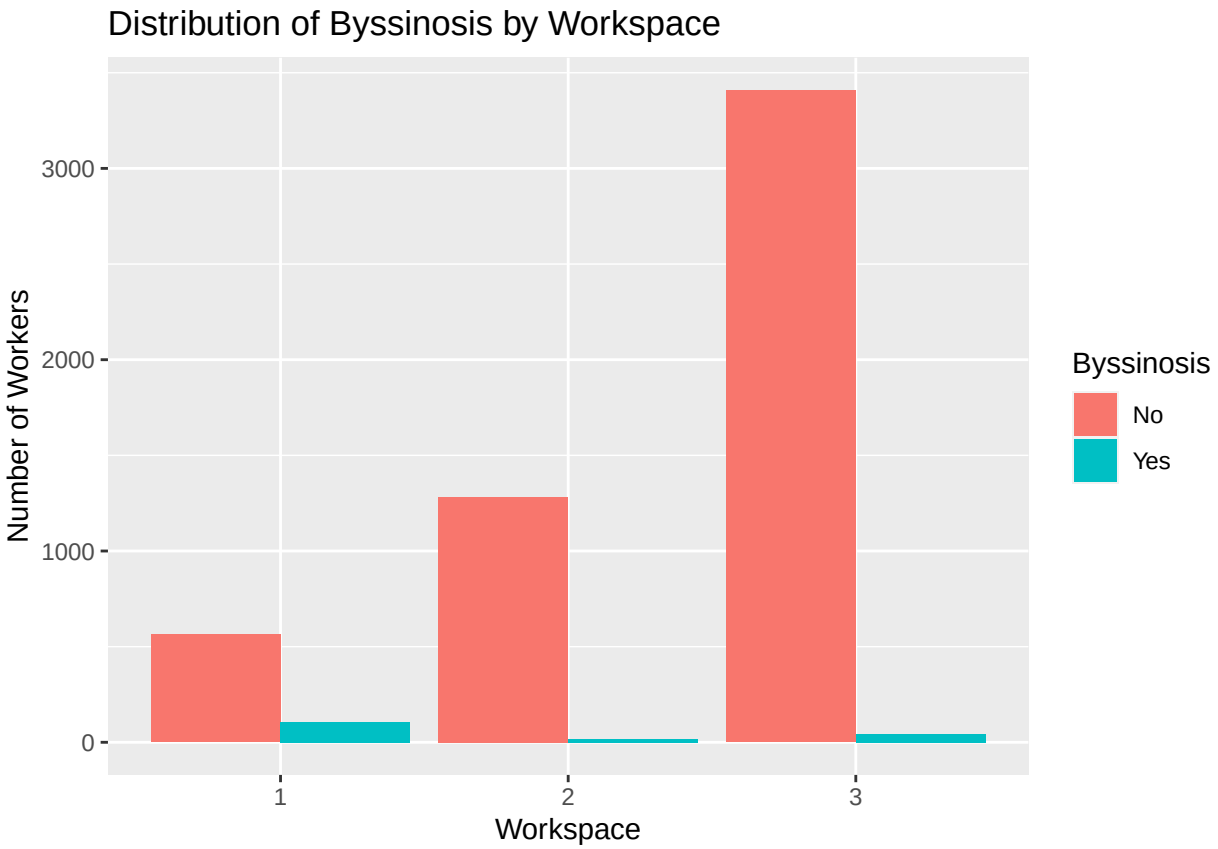



```
##
##       No  Yes
##   O 1830   73
##   W 3424   92
```

2.5 Workspace

From our results, it seem like the proportion of having byssinosis is the significantly higher where the workspace is the most dusty.

- $P(Byssinosis|MostDusty) = 105/(564 + 105) = 0.1569507$
- $P(Byssinosis|LessDusty) = 18/(1282 + 18) = 0.01384615$
- $P(Byssinosis|LeastDusty) = 42/(3408 + 42) = 0.01217391$




```
##
##       No  Yes
##   1  564  105
##   2 1282   18
##   3 3408   42
```

Overall, from the exploratory data analysis, 3 factors appear to contribute to having Byssinosis, particularly:

1. Smoking
2. Being a Male
3. Most Dusty Workspace

Now, we will build a logistic regression model and evaluate these variables to test our hypothesis and observations.

3. Variable Selection

Here, since our main goal is to build a logistic model focusing on investigating relationships and inference, we choose to use BIC over AIC as our criteria since it is easier for interpretations. AIC is more complicated and is only better if we want a better prediction result.

3.1 Model Selected Using Forward BIC

```
## Start:  AIC=743.07
## y ~ 1
##
##              Df Deviance    AIC
## + Workspace   2   613.45 637.17
## + Sex         1   710.59 726.40
## + Smoking     1   725.68 741.49
## <none>            735.16 743.07
## + Race        1   731.59 747.40
## + Employment  2   727.57 751.29
##
## Step:  AIC=637.17
## y ~ Workspace
##
##              Df Deviance    AIC
## <none>            613.45 637.17
## + Smoking     1   608.14 639.76
## + Race        1   612.05 643.67
## + Sex         1   612.21 643.83
## + Employment  2   605.08 644.60

##
## Call:
## glm(formula = y ~ Workspace, family = binomial, data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.5707  -0.1686  -0.1686  -0.1348   3.0676
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.7323     0.1504 -11.515  < 2e-16 ***
## Workspace2   -2.9636     0.4368  -6.784 1.17e-11 ***
## Workspace3   -2.5144     0.2547  -9.870  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 735.16  on 2709  degrees of freedom
## Residual deviance: 613.45  on 2707  degrees of freedom
## AIC: 619.45
##
## Number of Fisher Scoring iterations: 7
```

Here, workspace is our only input variable.

3.2 Model Selected Using bidirectional BIC

```
## Start:  AIC=743.07
## y ~ 1
```

```
##
##              Df Deviance    AIC
## + Workspace   2   613.45 637.17
## + Sex         1   710.59 726.40
## + Smoking     1   725.68 741.49
## <none>            735.16 743.07
## + Race        1   731.59 747.40
## + Employment  2   727.57 751.29
##
## Step:  AIC=637.17
## y ~ Workspace
##
##              Df Deviance    AIC
## <none>            613.45 637.17
## + Smoking     1   608.14 639.76
## + Race        1   612.05 643.67
## + Sex         1   612.21 643.83
## + Employment  2   605.08 644.60
## - Workspace   2   735.16 743.07


##
## Call:
## glm(formula = y ~ Workspace, family = binomial, data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.5707  -0.1686  -0.1686  -0.1348   3.0676
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.7323     0.1504 -11.515  < 2e-16 ***
## Workspace2   -2.9636     0.4368  -6.784 1.17e-11 ***
## Workspace3   -2.5144     0.2547  -9.870  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 735.16  on 2709  degrees of freedom
## Residual deviance: 613.45  on 2707  degrees of freedom
## AIC: 619.45
##
## Number of Fisher Scoring iterations: 7
```

Here, we see that both bidirectional and forward stepwise selection with BIC yielded the same model with workspace as the only input variable.

3.3 Testing for nonzero coefficients

Now, we want to test this on the test dataset (50% of the original) that we separated out at the beginning.

```
##
```

```
## Call:
## glm(formula = both_BIC$model, family = binomial, data = test)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.5987  -0.1950  -0.1438  -0.1438   3.0254
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.6281     0.1502  -10.84  < 2e-16 ***
## Workspace2   -2.3247     0.3279   -7.09 1.34e-12 ***
## Workspace3   -2.9380     0.2805  -10.47  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 742.02  on 2708  degrees of freedom
## Residual deviance: 608.03  on 2706  degrees of freedom
## AIC: 614.03
##
## Number of Fisher Scoring iterations: 7

## [1] 1
```

Using the Wald test with the test dataset, We see that the p values are close to zero, meaning that they are very significant and non zero. To briefly interpret the estimates, the magnitude of the estimates suggest the chance of having Byssinosis decreases less in workspace 1 (most dusty) than in workspaces 2 and 3 (less and least dusty.)

Additionally, we have conducted a deviance of goodness of fit test to examine the quality of our model. From the p value that we have obtained which is 1, we know that we have rejected the null hypothesis and the model we have developed is quite robust.

Code Appendix

```
knitr::opts_chunk$set(echo = TRUE)
#libraries and setseed
library(readr)
library(ISLR)
library(ggplot2)
set.seed(123)

#import dataset
Byssinosis <- read_csv("Byssinosis.csv")

#turn numerical variable into factor
Byssinosis$Workspace <- as.factor(Byssinosis$Workspace)

#make long data from wide data
yes <- Byssinosis$Byssinosis
```

```
no <- Byssinosis$`Non-Byssinosis`
yesInd <- c(
  rep(0,sum(no)),
  rep(1,sum(yes))
  )
longdata <- rbind(
  Byssinosis[rep(1:72,no),1:5],
  Byssinosis[rep(1:72,yes),1:5]
)
longdata$y <- yesInd

#create train and test datasets using 50-50 split on long data
trainIndex <- sample(5419, 2710)
train = longdata[trainIndex,]
test = longdata[-trainIndex,]

##create "yes","no" response for EDA, change name of column y
longdataeda <- longdata
longdataeda$y <- ifelse(longdata$y==1,"Yes","No")
names(longdataeda)[names(longdataeda) == "y"] <- "Byssinosis"
ggplot(longdataeda, aes(x= Employment, fill = Byssinosis)) + geom_bar(position = position_dodge())+
  labs(y= "Number of Workers", x="Employment Years", title = "Distribution of Byssinosis by Employment

table(longdataeda$Employment,longdataeda$Byssinosis)
ggplot(longdataeda, aes(x= Smoking, fill = Byssinosis)) + geom_bar(position = position_dodge())+
  labs(y= "Number of Workers", x="Smoking", title = "Distribution of Byssinosis by Smoking")

table(longdataeda$Smoking,longdataeda$Byssinosis)
ggplot(longdataeda, aes(x= Sex, fill = Byssinosis)) + geom_bar(position = position_dodge())+
  labs(y= "Number of Workers", x="Sex", title = "Distribution of Byssinosis by Sex")

table(longdataeda$Sex,longdataeda$Byssinosis)
ggplot(longdataeda, aes(x= Race, fill = Byssinosis)) + geom_bar(position = position_dodge())+
  labs(y= "Number of Workers", x="Race", title = "Distribution of Byssinosis by Race")

table(longdataeda$Race,longdataeda$Byssinosis)
ggplot(longdataeda, aes(x= Workspace, fill = Byssinosis)) + geom_bar(position = position_dodge())+
  labs(y= "Number of Workers", x="Workspace", title = "Distribution of Byssinosis by Workspace")

table(longdataeda$Workspace,longdataeda$Byssinosis)
forward_BIC <- step(glm(y~1, binomial, train),
     scope = ~Employment*Smoking*Sex*Race*Workspace,
     direction = "forward",
     k = log(dim(train)[1]))
summary(forward_BIC)
both_BIC <- step(glm(y~1, binomial, train),
     scope = ~Employment*Smoking*Sex*Race*Workspace,
     direction = "both",
     k = log(dim(train)[1]))
summary(both_BIC)
testFit <- glm(both_BIC$model, binomial, test)
summary(testFit)
1-pchisq(608.03, 2706)
```