STA138 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:

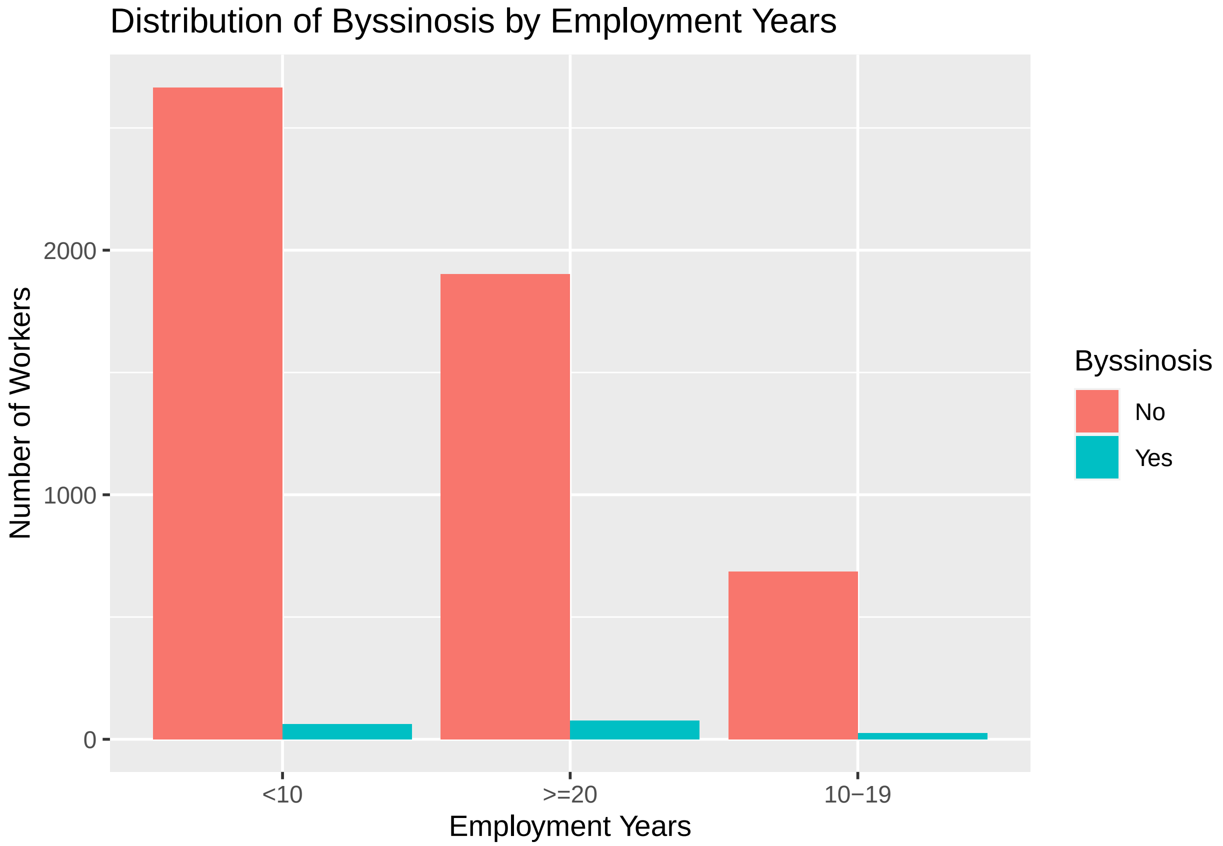
- Type of work place [1 (most dusty), 2 (less dusty), 3 (least dusty)]
- Employment, years [< 10, 10–19, 20–]
- Smoking [Smoker, or not in last 5 years]
- Sex [Male, Female]
- Race [White, Other]
- Byssinosis [Yes, No]

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, Pearson's Chi-Square Test of Independence, Likelihood Ratio Test

Here, we would establish preliminary assumptions on variable effects from ggplots, and validate them afterwards with proper tests such as Pearson's Chi-Square Test of Independence and the Likelihood Ratio Test.

2.1 Employment



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

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.

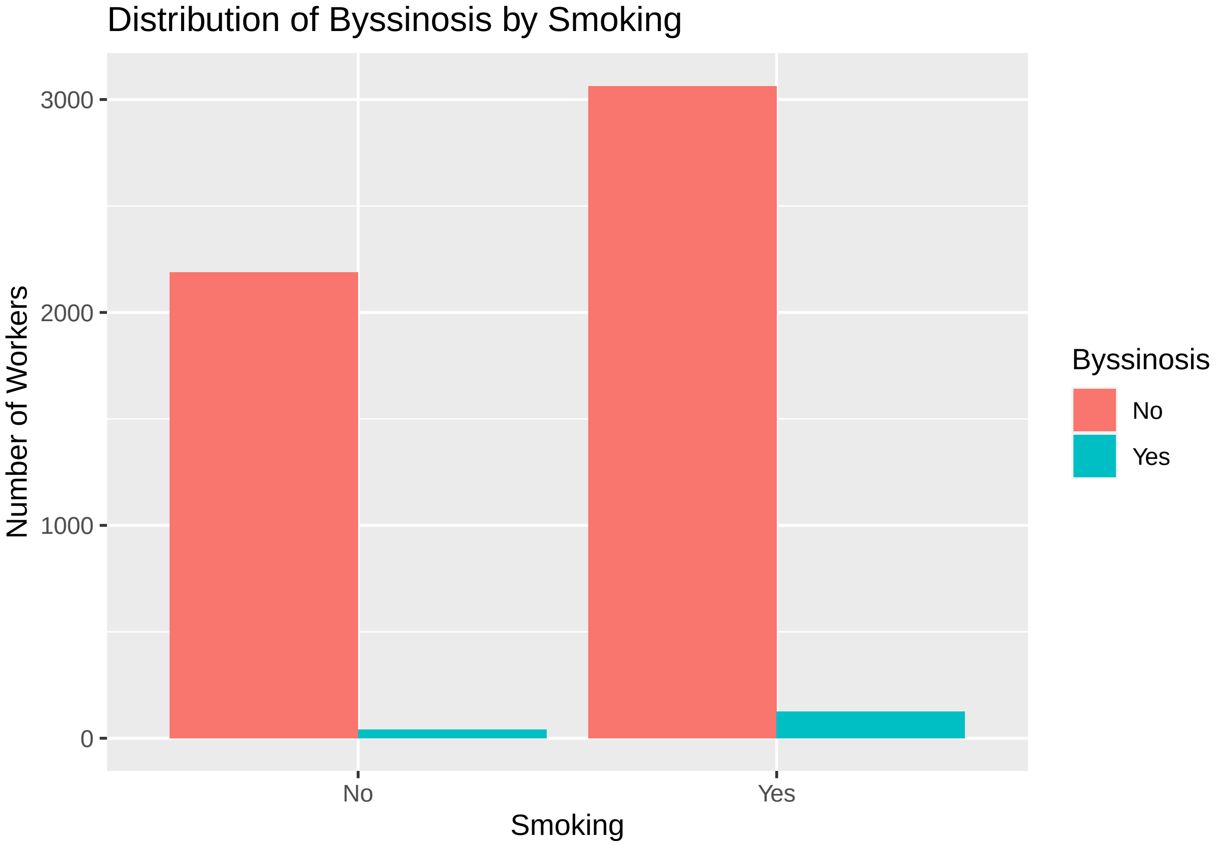
- $P(Byssinosis| < 10) = 63/(2666 + 63) = 0.02308538$
- $P(Byssinosis| >= 20) = 76/(1902 + 76) = 0.03842265$
- $P(Byssinosis|10 - 19) = 26/(26 + 686) = 0.03651685$

After conducting Pearson's chi squared test of independence and the likelihood ratio test, we have obtained the following p values:

- Pearson's p-value: 0.0062186
- Likelihood Ratio Test p-value: 0.0059884

At 5% significance level, we reject the null hypothesis for both Pearson's chi squared test of independence as well as likelihood ratio test. There is sufficient evidence to conclude definitively that there is an effect of employment on byssinosis.

2.2 Smoking



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

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

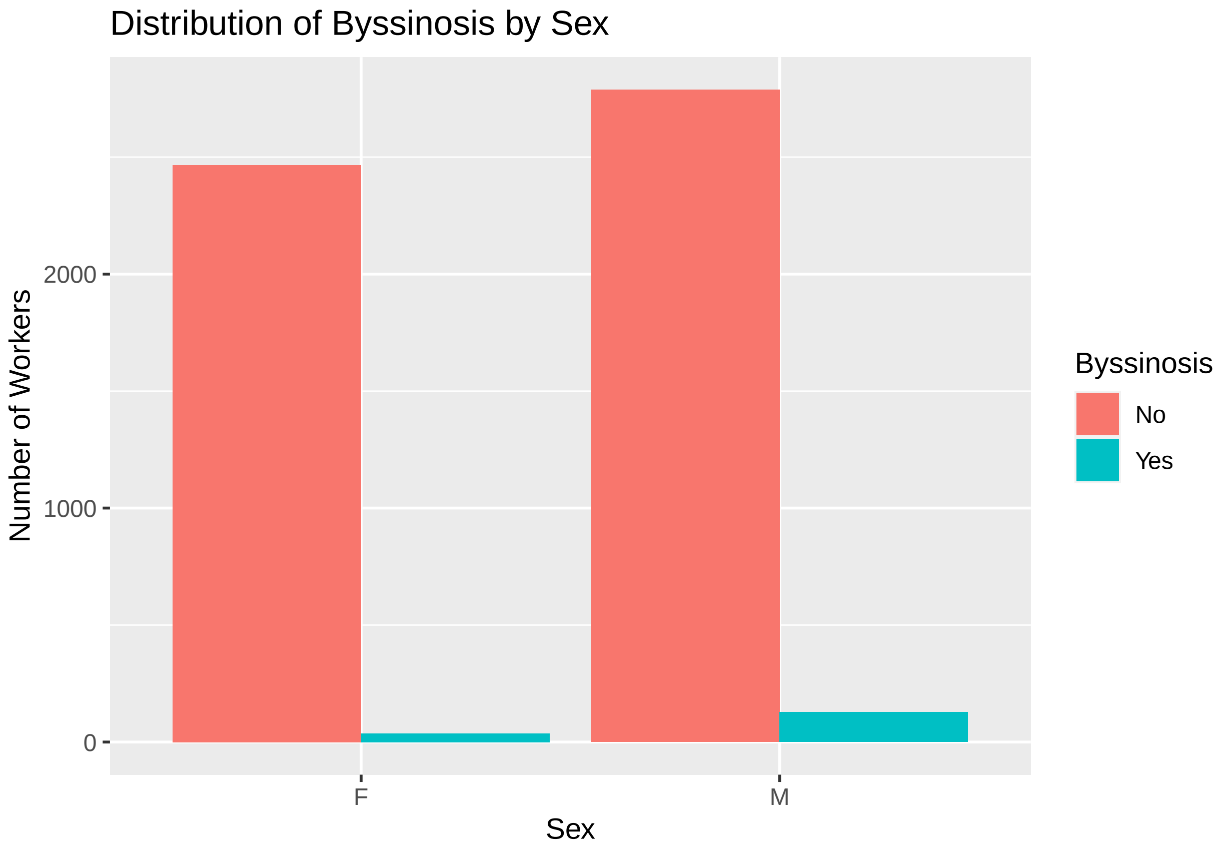
- $P(Byssinosis|NonSmokers) = 40/(2190 + 40) = 0.01793722$
- $P(Byssinosis|Smokers) = 125/(125 + 3064) = 0.03919724$

After conducting Pearson's chi squared test of independence and the likelihood ratio test, we have obtained the following p values:

- Pearson's p-value: 7.3789182×10^{-6}
- Likelihood Ratio Test p-value: 3.6860635×10^{-6}

At 5% significance level, we reject the null hypothesis for both Pearson's chi squared test of independence as well as likelihood ratio test. There is sufficient evidence to conclude definitively that there is an effect of smoking on byssinosis.

2.3 Sex



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

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

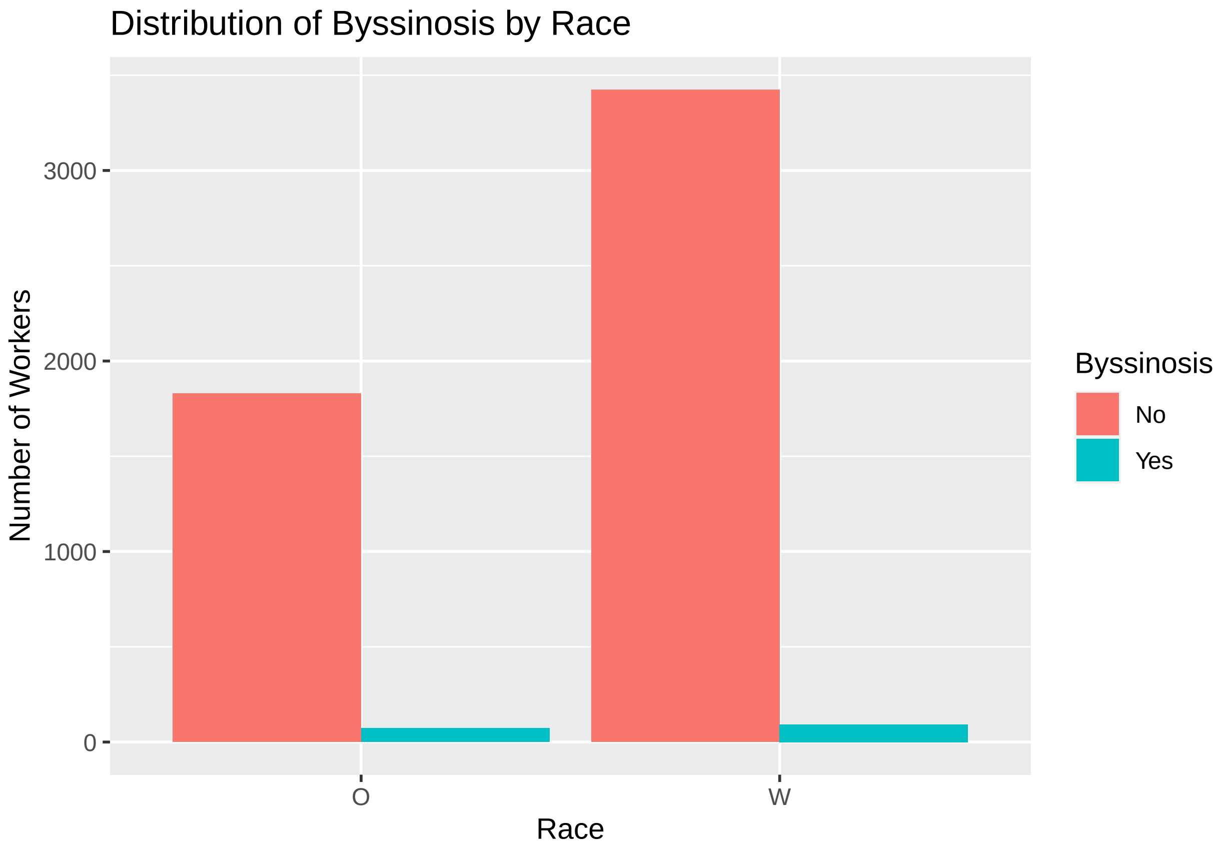
- $P(Byssinosis|Female) = 37/(2466 + 37) = 0.01478226$
- $P(Byssinosis|Male) = 128/(2788 + 128) = 0.04389575$

After conducting Pearson's chi squared test of independence and the likelihood ratio test, we have obtained the following p values:

- Pearson's p-value: $5.0168592 \times 10^{-10}$
- Likelihood Ratio Test p-value: $1.2764567 \times 10^{-10}$

At 5% significance level, we reject the null hypothesis for both Pearson's chi squared test of independence as well as likelihood ratio test. There is sufficient evidence to conclude definitively that there is an effect of sex on byssinosis.

2.4 Race



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

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

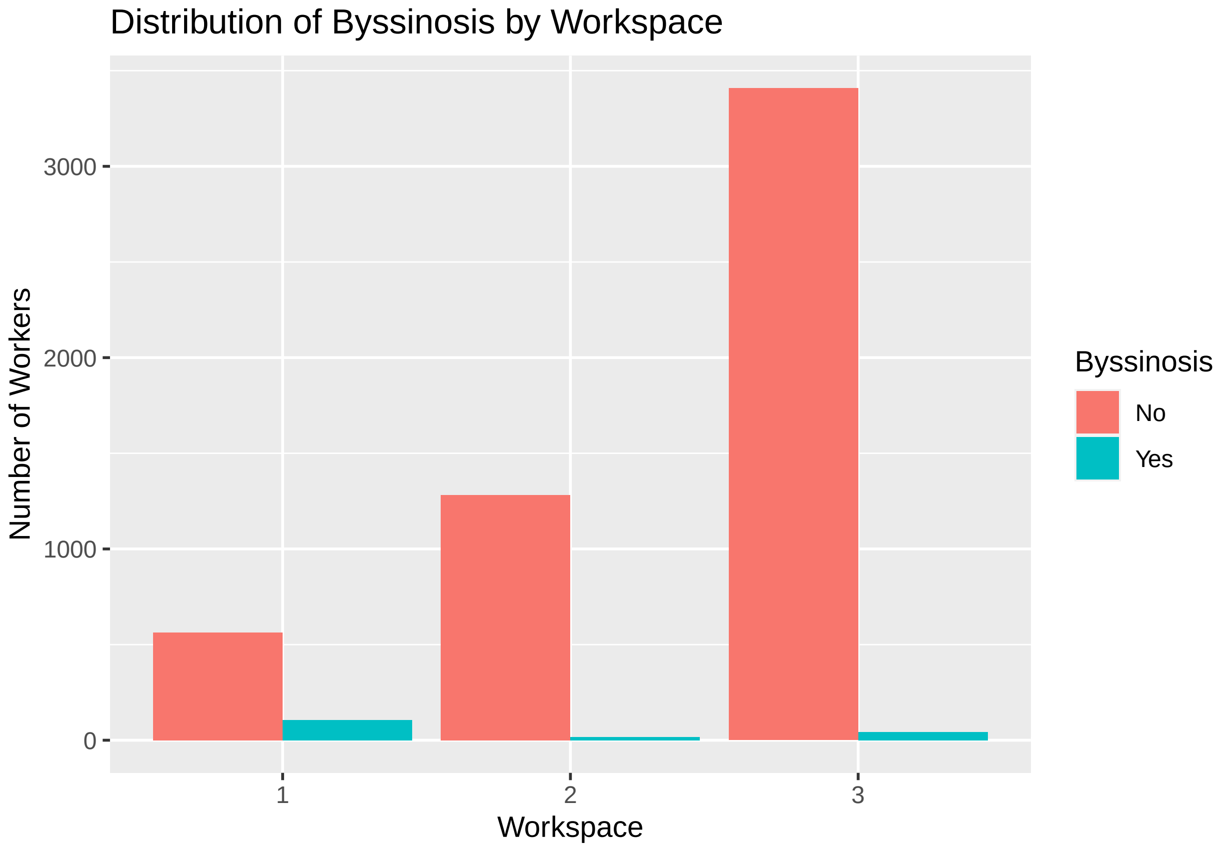
- $P(Byssinosis|OtherRace) = 73/(1830 + 73) = 0.03836048$
- $P(Byssinosis|White) = 92/(3424 + 92) = 0.0261661$

After conducting Pearson's chi squared test of independence and the likelihood ratio test, we have obtained the following p values:

- Pearson's p-value: 0.0126354
- Likelihood Ratio Test p-value: 0.0140976

At 5% significance level, we reject the null hypothesis for both Pearson's chi squared test of independence as well as likelihood ratio test. There is sufficient evidence to conclude definitively that there is an effect of race on byssinosis.

2.5 Workspace



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

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

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

After conducting Pearson's chi squared test of independence and the likelihood ratio test, we have obtained the following p values:

- Pearson's p-value: 0
- Likelihood Ratio Test p-value: 0

At 5% significance level, we reject the null hypothesis for both Pearson's chi squared test of independence as well as likelihood ratio test. There is sufficient evidence to conclude definitively that there is an effect of workspace on byssinosis.

2.6 Observations

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

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

However, from our Pearson's test and Likelihood Ratio Tests, we have somehow concluded that all the provided variables have a significant effect on byssinosis at 5% significance level.

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

3. Variable Selection, Logistic Model

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.

To briefly interpret the model and the estimates, here workspace1 would be our baseline case.

All other things equal, the estimated log-odds for workspace1 according to this model are just the intercept alpha hat, which corresponds to estimated odds of 0.1768, and estimated probablity of 0.15 of having byssinosis for a person working at the most dusty workspace, holding everything else constant.

All other things equal, the estimated log-odds ratio of having byssinosis for workspace1 vs. workspace2 is Beta1 hat. Thus, estimated odds of having byssinosis under this model are 0.0516 times those of workspace1, holding everything else constant.

All other things equal, The estimated log-odds ratio of having byssinosis for workspace1 vs. workspace3 is Beta2 hat. Thus, estimated odds of having byssinosis under this model are 0.0809 time those of workspace1, holding everything else constant.

So, this model suggests that people working in the most dusty workspace are more likely to have byssinosis.

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 = quasibinomial, data = test)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.5987  -0.1950  -0.1438  -0.1438   3.0254
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.6281     0.1503 -10.831  < 2e-16 ***
## Workspace2   -2.3247     0.3281  -7.086 1.75e-12 ***
## Workspace3   -2.9380     0.2807 -10.467  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for quasibinomial family taken to be 1.001117)
```

```
##
##     Null deviance: 742.02  on 2708  degrees of freedom
## Residual deviance: 608.03  on 2706  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 7
```

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.

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. Our dispersion parameter that is 1.00 also shows that our model doesn't suffer from overdispersion.

4. Conclusion

At the beginning, we have subjectively assumed that smoking, being a male, and being in the most dusty workspace contribute the most to byssinosis from the table and plots we have observed.

After conducting both the Pearson's Chi-Squared Test of Independence and the Likelihood Ratio Test, we have observed some dependencies between byssinosis and all of our provided variables. Note, our choice of significance at 5% might be one of the factors why we have rejected all of our tests. If we were to chance it to 1%, say, then we would actually fail to reject our hypothesis test for the variable race.

Lastly, from our logistic regression built by BIC stepwise selection, we have seen that the only factor that contributed significantly to byssinosis is the presence of dust. Workspace which are the most dusty contribute more as compared to the other ones. However, as a caveat here, we might have built a model that is not the "most" useful for real world predictions. If we were to build our model using AIC as our variable selection criteria, there is a likely chance that we will have included more variables and build a model that is even better at predicting than the model provided here. For the purposes of this report, we would stick with BIC.

5. 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(
```

```r
  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"
#Visualization
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

#Make Table
O1 <- table(longdataeda$Employment,longdataeda$Byssinosis);O1
E1 <- rowSums(O1)%*%t(colSums(O1))/sum(O1)

#Pearson's chisq test of independence
pearsonStatistic1 <- sum((O1-E1)^2/E1)
pearsonpVal1 <- 1-pchisq(pearsonStatistic1,2)

#Likelihood Ratio Test
LRstatistic1 <- -2*sum(O1*log(E1/O1))
LrpVal1 <- 1-pchisq(LRstatistic1,2)
#Visualization
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")

#Make Table
O2 <- table(longdataeda$Smoking,longdataeda$Byssinosis);O2
E2 <- rowSums(O2)%*%t(colSums(O2))/sum(O2)

#Pearson's chisq test of independence
pearsonStatistic2 <- sum((O2-E2)^2/E2)
pearsonpVal2 <- 1-pchisq(pearsonStatistic2,1)

#Likelihood Ratio Test
LRstatistic2 <- -2*sum(O2*log(E2/O2))
LrpVal2 <- 1-pchisq(LRstatistic2,1)
#Visualization
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")

#Make Table
```

```
O3 <- table(longdataeda$Sex,longdataeda$Byssinosis);O3
E3 <- rowSums(O3)%*%t(colSums(O3))/sum(O3)

#Pearson's chisq test of independence
pearsonStatistic3 <- sum((O3-E3)^2/E3)
pearsonpVal3 <- 1-pchisq(pearsonStatistic3,1)

#Likelihood Ratio Test
LRstatistic3 <- -2*sum(O3*log(E3/O3))
LrpVal3 <- 1-pchisq(LRstatistic3,1)
#Visualization
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")

#Make Table
O4 <- table(longdataeda$Race,longdataeda$Byssinosis);O4
E4 <- rowSums(O4)%*%t(colSums(O4))/sum(O4)

#Pearson's chisq test of independence
pearsonStatistic4 <- sum((O4-E4)^2/E4)
pearsonpVal4 <- 1-pchisq(pearsonStatistic4,1)

#Likelihood Ratio Test
LRstatistic4 <- -2*sum(O4*log(E4/O4))
LrpVal4 <- 1-pchisq(LRstatistic4,1)
#Visualization
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")

#Make Table
O5 <- table(longdataeda$Workspace,longdataeda$Byssinosis);O5
E5 <- rowSums(O5)%*%t(colSums(O5))/sum(O5)

#Pearson's chisq test of independence
pearsonStatistic5 <- sum((O5-E5)^2/E5)
pearsonpVal5 <- 1-pchisq(pearsonStatistic5,2)

#Likelihood Ratio Test
LRstatistic5 <- -2*sum(O5*log(E5/O5))
LrpVal5 <- 1-pchisq(LRstatistic5,2)
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, quasibinomial, test)
summary(testFit)
dev <- 1-pchisq(608.03, 2706)
```