Byssinosis Project

Yiqi Ren 916181233

2019/12/7

Introduction

*The Goal of the Analysis

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

*Dataset Background

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]

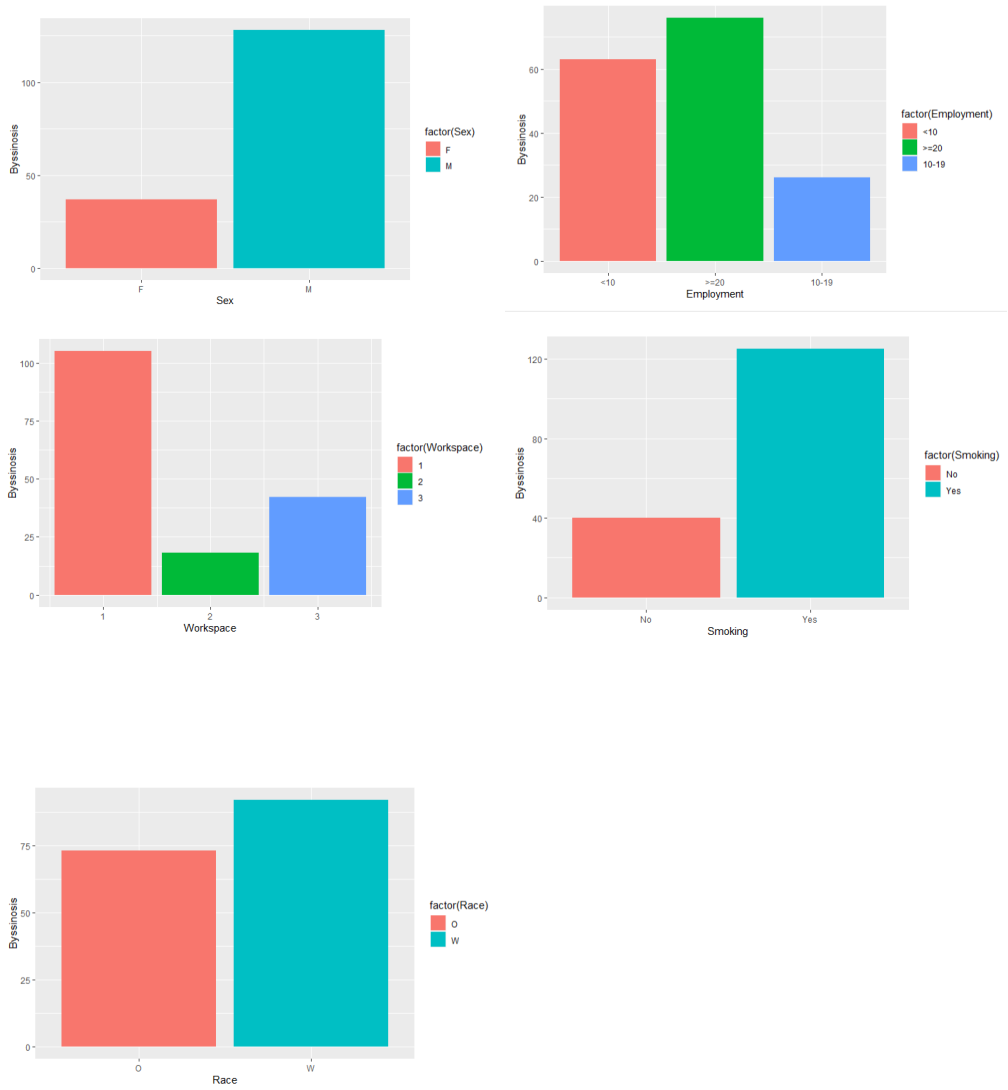
Byssinosis [Yes, No]

*Dataset Summary

	Byssinosis	Non.Byssinosis
Employment<10	63	2666
Employment<10-19	26	686
Employment>=20	76	1902
Race(W)	92	3424
Race(o)	73	1830
Sex(F)	37	2466
Sex(M)	128	2788
Smoking(Yes)	125	3064
Smoking(No)	40	2190
Workspace(1)	105	564
Workspace(2)	18	1282
Workspace(3)	42	3408

Main Analysis

* Visulization



If we visulize the dataset, we have a better understanding of each variable.In the histogram plot, male employees have more Byssinosis than femles. People who work less than 10 years have a lower amount than people who work more than 20 years. the Workspace seems have a big influence of getting Byssinosis since a lot of people getting Byssinsis when they work in a most dusty workspace. Smoking seems also increase the rate of getting Byssinosis because lots of smokers get Byssinosis.

*Independent Test

In this section, I am going to use the Mantel-Haenszel Test to check if each variable is indenpendent to Byssinosis or not.The result is below:

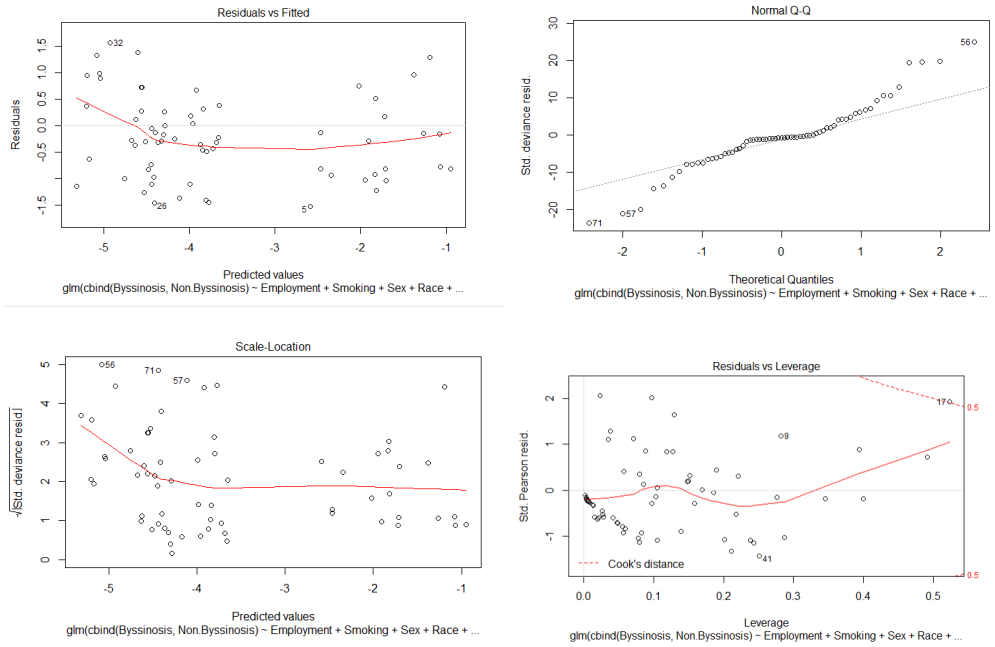
	P-Value	Dision
Employment	0.3541915	Independent
Smoking	0.02292341	Dependent
Sex	0.01626846	Dependent
Race	0.3276745	Independent
workspace	0.03494676	Dependent

From the result table, it is obvious that race and the length of employment have large p-values, and they are larger than any reasonable significance level, so we can conclude that race and the length of employment are independent to the Byssinosis disease. However, Workspace, Smoking and sex are dependent to Byssinosis.

* Model Selection

In ording to see the relationships between each variable and Byssinosis. we first fit the data into a logistic model.

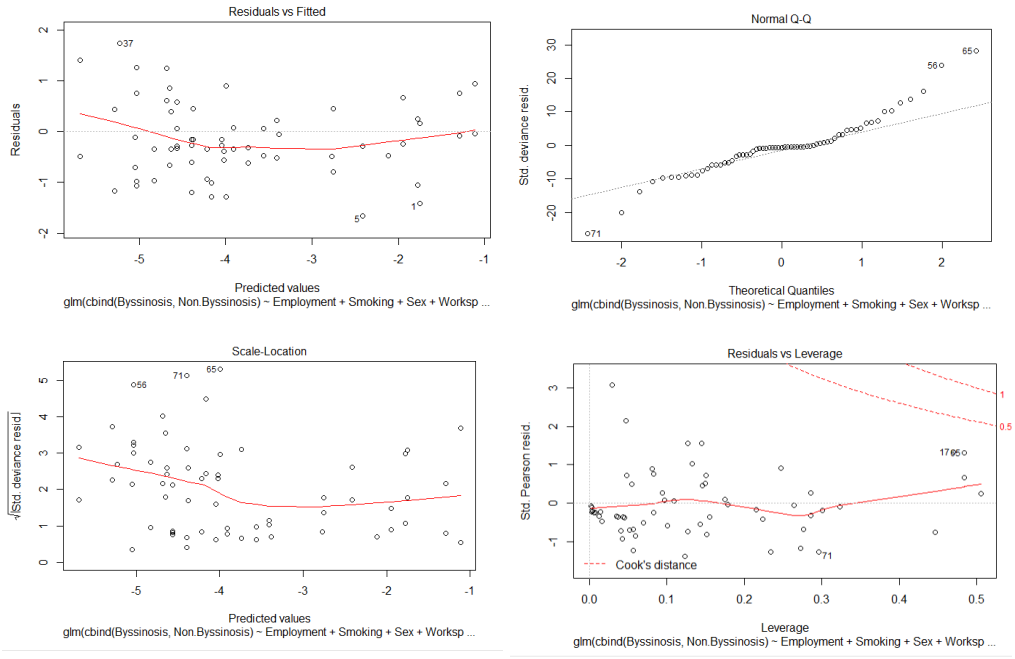
```
##
## Call:
## glm(formula = cbind(Byssinosis, Non.Byssinosis) ~ Employment +
##     Smoking + Sex + Race + Workspace, family = binomial(), data = mydata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5240  -0.8105  -0.1952   0.2071   1.5643
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -2.3463     0.2639  -8.891  < 2e-16 ***
## Employment>=20    0.7531     0.2161   3.484 0.000493 ***
## Employment10-19   0.5641     0.2617   2.156 0.031091 *
## SmokingYes        0.6413     0.1944   3.299 0.000971 ***
## SexM             -0.1239     0.2288  -0.542 0.587983
## RaceW            -0.1163     0.2072  -0.562 0.574426
## Workspace2       -2.5799     0.2921  -8.834  < 2e-16 ***
## Workspace3       -2.7306     0.2153 -12.681  < 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: 322.527  on 64  degrees of freedom
## Residual deviance:  43.271  on 57  degrees of freedom
## AIC: 165.95
##
## Number of Fisher Scoring iterations: 5
```



Based on the summary we get, it is obvious that the p-value of race and sex are larger than any reasonable significance levels, so there are no statistical influences of race and sex on Byssinosis. If we drop these two variables, we will predict a better model. For the diagnostic, if we lookad the the predicted value plot and leverage plot, there are couple outliers and one point outside the cook's distance, which means there is a high influential point in the dataset. Those places are somewhere we have to pay attention after selecting models. However, in order to find a best model, we can directly using Bidirectional selection to find the final "best" model:

```
## Start:  AIC=165.95
## cbind(Byssinosis, Non.Byssinosis) ~ Employment + Smoking + Sex +
##     Race + Workspace
##
##                         Df Deviance    AIC
## + Sex:Workspace          2   35.721 162.40
## - Sex                    1   43.563 164.24
## - Race                   1   43.586 164.26
## + Smoking:Workspace      2   38.679 165.35
## <none>                       43.271 165.95
## + Smoking:Race           1   41.906 166.58
## + Smoking:Sex            1   42.564 167.24
## + Employment:Race        2   41.147 167.82
## + Sex:Race               1   43.237 167.91
## + Employment:Smoking     2   41.442 168.12
## + Employment:Sex         2   41.862 168.54
## + Race:Workspace         2   42.560 169.24
## + Employment:Workspace   4   38.766 169.44
## - Employment             2   56.171 174.85
## - Smoking                1   54.982 175.66
## - Workspace              2  241.941 360.62
##
## Step:  AIC=162.4
## cbind(Byssinosis, Non.Byssinosis) ~ Employment + Smoking + Sex +
##     Race + Workspace + Sex:Workspace
```

```
##
##                         Df Deviance    AIC
## - Race                   1   36.019 160.69
## <none>                       35.721 162.40
## + Smoking:Race           1   33.891 162.57
## + Employment:Sex         2   32.311 162.99
## + Sex:Race               1   34.583 163.26
## + Smoking:Workspace      2   32.807 163.48
## + Smoking:Sex            1   35.024 163.70
## + Employment:Race        2   33.241 163.92
## + Employment:Smoking     2   33.304 163.98
## - Sex:Workspace          2   43.271 165.95
## + Race:Workspace         2   35.442 166.12
## + Employment:Workspace   4   32.384 167.06
## - Employment             2   46.720 169.40
## - Smoking                1   48.040 172.72
##
## Step:  AIC=160.69
## cbind(Byssinosis, Non.Byssinosis) ~ Employment + Smoking + Sex +
##     Workspace + Sex:Workspace
##
##                         Df Deviance    AIC
## <none>                       36.019 160.69
## + Employment:Sex         2   32.493 161.17
## + Smoking:Workspace      2   33.076 161.75
## + Smoking:Sex            1   35.308 161.98
## + Employment:Smoking     2   33.555 162.23
## + Race                   1   35.721 162.40
## - Sex:Workspace          2   43.586 164.26
## + Employment:Workspace   4   32.584 165.26
## - Employment             2   48.578 169.25
## - Smoking                1   48.325 171.00
##
## Call:  glm(formula = cbind(Byssinosis, Non.Byssinosis) ~ Employment +
##     Smoking + Sex + Workspace + Sex:Workspace, family = binomial(),
##     data = mydata)
##
## Coefficients:
##     (Intercept)   Employment>=20  Employment10-19       SmokingYes
##         -3.4097           0.6367           0.4640           0.6578
##            SexM       Workspace2       Workspace3  SexM:Workspace2
##          0.9990          -1.2714          -1.6207          -2.0058
## SexM:Workspace3
##         -1.2576
##
## Degrees of Freedom: 64 Total (i.e. Null);  56 Residual
## Null Deviance:       322.5
## Residual Deviance: 36.02     AIC: 160.7
```



Through comparing the AIC values during each step, the final model we are going to select is

$$Y = \beta_0 + \beta_1 X_{Emp(>=20)} + \beta_2 X_{Emp.(10-19)} + \beta_3 X_{Smok(Yes)} + \beta_4 X_{Sex(M)} + \beta_5 X_{wp(2)} + \beta_6 X_{wp(3)} + \beta_7 X_{Sex(M)} X_{ws(2)} + \beta_8 X_{Sex(M)} X_{ws(3)}$$

If we compare the "residuals vs fitted plot" between the full model and our selected model, more and more points are close to the red line in the graph, which means the outliers are decreasing. Also, there is no more points outside the cook's distance. This means our selected model is getting stable since the high influential point has been removed from our model. Furthermore, we notice that this model contains two interaction terms. Since they have statistical relationships with Byssinosis, we want use them to predict our model. Thus, rhis model will contain the least AIC value (160.7), and it is the best we can reach so far.

Furthermore, If we construct a L-R test to comparing the full model and the final model, the null hypothesis is the full model is better than our selected model, and we get a p-value of 0.007081436 which is smaller than any reasonable significant level, so we reject the null hypothesis and conclude that the selected model is better.

Conclusion

By so far, we have enough information to draw out our conclusion. Race has not significant influence of getting Byssinosis, but the length of employment, Sex, smoking and workspace are all play an important role during the Byssinosis analysis.Furthurmore, we also find out that the interaction terms between workspace and sex are significant.

Appendix code

```
setwd("C:/users/richie/desktop/STA138/Project")
mydata=read.table("Byssinosis.csv",sep=",",header=TRUE)

#Employment
Byssinosis<-mydata
sum(Byssinosis$Byssinosis[Byssinosis$Employment == "<10"])
sum(Byssinosis$Byssinosis[Byssinosis$Employment == "10-19"])
sum(Byssinosis$Byssinosis[Byssinosis$Employment == ">=20"])
sum(Byssinosis$Non.Byssinosis[Byssinosis$Employment == "<10"])
sum(Byssinosis$Non.Byssinosis[Byssinosis$Employment == "10-19"])
sum(Byssinosis$Non.Byssinosis[Byssinosis$Employment == ">=20"])

#Race
sum(Byssinosis$Byssinosis[Byssinosis$Race == "W"])
sum(Byssinosis$Byssinosis[Byssinosis$Race == "O"])
sum(Byssinosis$Non.Byssinosis[Byssinosis$Race == "W"])
sum(Byssinosis$Non.Byssinosis[Byssinosis$Race == "O"])


#Sex
sum(Byssinosis$Byssinosis[Byssinosis$Sex == "F"])
sum(Byssinosis$Byssinosis[Byssinosis$Sex == "M"])
sum(Byssinosis$Non.Byssinosis[Byssinosis$Sex == "F"])
sum(Byssinosis$Non.Byssinosis[Byssinosis$Sex == "M"])

#Smoking
sum(Byssinosis$Byssinosis[Byssinosis$Smoking == "Yes"])
sum(Byssinosis$Byssinosis[Byssinosis$Smoking== "No"])
sum(Byssinosis$Non.Byssinosis[Byssinosis$Smoking == "Yes"])
sum(Byssinosis$Non.Byssinosis[Byssinosis$Smoking == "No"])

#Workspace
sum(Byssinosis$Byssinosis[Byssinosis$Workspace == 1])
sum(Byssinosis$Byssinosis[Byssinosis$Workspace== 2])
sum(Byssinosis$Byssinosis[Byssinosis$Workspace== 3])
sum(Byssinosis$Non.Byssinosis[Byssinosis$Workspace == 1])
sum(Byssinosis$Non.Byssinosis[Byssinosis$Workspace == 2])
sum(Byssinosis$Non.Byssinosis[Byssinosis$Workspace == 3])

#Byssinosis
sum(mydata$Byssinosis)
sum(mydata$Non.Byssinosis)
library(ggplot2)
p <- ggplot2::ggplot(mydata, aes(x = Employment, y = Byssinosis)) + ylab("Byssinosis")
p + geom_col(aes(x = Employment, fill = factor(Employment)))

#SexVSByssinosis
p2 <- ggplot2::ggplot(mydata, aes(x = Sex, y = Byssinosis)) + ylab("Byssinosis")
p2 + geom_col(aes(x = Sex, fill = factor(Sex)))

#Workspace
```

```r
p3 <- ggplot2::ggplot(mydata, aes(x = Workspace, y = Byssinosis)) + ylab("Byssinosis")
p3 + geom_col(aes(x = Workspace, fill = factor(Workspace)))

#Smoking
p3 <- ggplot2::ggplot(mydata, aes(x = Smoking, y = Byssinosis)) + ylab("Byssinosis")
p3 + geom_col(aes(x = Smoking, fill = factor(Smoking)))

#Race
p4 <- ggplot2::ggplot(mydata, aes(x = Race, y = Byssinosis)) + ylab("Byssinosis")
p4 + geom_col(aes(x = Race, fill = factor(Race)))
#independent test
#employment
dec<-mydata
lookup<-data.frame(Employment=levels(dec$Employment),rankNumeric=c(1,3,2))
dec<-merge(dec,lookup,by="Employment")
dec$pickNumeric <-as.numeric(dec$Byssinosis)
MH.test<-sqrt(nrow(dec)-1)*cor(dec$rankNumeric,dec$pickNumeric)
MH.pvalE<-pnorm(-abs(MH.test))
MH.pvalE

#Smoking
dec<-mydata
lookup<-data.frame(Smoking=levels(dec$Smoking),rankNumeric=c(1,2))
dec<-merge(dec,lookup,by="Smoking")
dec$pickNumeric <-as.numeric(dec$Byssinosis)
MH.test<-sqrt(nrow(dec)-1)*cor(dec$rankNumeric,dec$pickNumeric)
MH.pvalS<-pnorm(-abs(MH.test))
MH.pvalS

#Sex
dec<-mydata
lookup<-data.frame(Sex=levels(dec$Sex),rankNumeric=c(1,2))
dec<-merge(dec,lookup,by="Sex")
dec$pickNumeric <-as.numeric(dec$Byssinosis)
MH.test<-sqrt(nrow(dec)-1)*cor(dec$rankNumeric,dec$pickNumeric)
MH.pvalSe<-pnorm(-abs(MH.test))
MH.pvalSe

#Race
dec<-mydata
lookup<-data.frame(Race=levels(dec$Race),rankNumeric=c(1,2))
dec<-merge(dec,lookup,by="Race")
dec$pickNumeric <-as.numeric(dec$Byssinosis)
MH.test<-sqrt(nrow(dec)-1)*cor(dec$rankNumeric,dec$pickNumeric)
MH.pvalR<-pnorm(-abs(MH.test))
MH.pvalR

#Workspace
dec<-mydata
dec$pickNumeric <-as.numeric(dec$Byssinosis)
MH.test<-sqrt(nrow(dec)-1)*cor(dec$Workspace,dec$pickNumeric)
MH.pvalW<-pnorm(-abs(MH.test))
MH.pvalW
```

```r
mydata$Workspace = as.factor(mydata$Workspace)
model=glm(formula = cbind(Byssinosis,Non.Byssinosis) ~ Employment+Smoking+Sex+Race+Workspace,
          family = binomial(),data = mydata)
summary(model)
step(model,scope=~Employment*Smoking*Sex*Race*Workspace)
final<-glm(formula=cbind(Byssinosis,Non.Byssinosis)~Employment+Smoking+Sex+Workspace+Sex:Workspace,famil
plot(final)
final_model = glm(formula = cbind(Byssinosis, Non.Byssinosis) ~ Employment +
                    Smoking + Sex + Workspace + Sex:Workspace, family = binomial(),
                  data = mydata)
L0 = logLik(model)
L0

L1 = logLik(final_model)
L1

L1-L0

G_Square = 2*(L1-L0)
G_Square

p_value = pchisq(G_Square,df = 1,lower.tail = FALSE)
p_value
#0.007081436
```