STA 138 - Final Project - 2019 FALL

Hao Luo, 912423597

Contents

INTRODUCTION

Byssinosis is a form of occupational disease, it also know as brown lung disease. According to **Healthline.com**, it is often caused by inhaling particles from flax, hemp or cotton. The symptoms of byssinosis are similar to asthma and include tightness in the chest, wheezing, and coughing. In worst case senario, subject may experience flu-like symptons. Therefore, we want to investigate relationshops between the disease and the factors of smkoing status, sex, race, length of employment, and workspace.

In this investigation, wether a person has byssinosis or not, is a categorical variable and it only has two levels, and the explanatory variables are either nominal or binary (see the table at **EXPLANATORY VARIABLES** below). Furthermore, suppose that eash subject in this investigation is independent and notice that the probability of someone getting Byssinosis is $\frac{165}{5419} \approx 3.04\%$, then the response variable is distributed Bernoulli

First, we will conduct an exploratory data analysis to check the integrity of our dataset and the relationship between the explanatory and response variables. Then, we will use Backward selection method based on the Akaike Information Criterion (AIC) to select the best model. Then, removing influential outliers from our dataset based on the p-values of standardized residuals and the DFBeta plot. Finally, we want a model that can predict the probability that wether a person gets Byssinosis reasonably well.

EXPLORATORY ANAlYSIS ON DATASET

An exploratory data analysis is carried out on the given data set to determine the integrity of our dataset.

EXPANATORY VARIABLES

Total sample size of this dataset is 5,419 workers with 6 explanatory variables.

```
##   Employment Smoking Sex Race Workspace Bys
## 1        <10     Yes   M    W         1   0
## 2        <10     Yes   M    W         1   0
## 3        <10     Yes   M    W         1   0
## 4        <10     Yes   M    W         1   0
## 5        <10     Yes   M    W         1   0
```

Employment

An ordinal variable that split the years working in the industry into three groups **<10**, **10-19**, and **>=20**. Majority of the subjects had been in the industry less than 10 years.

years	<10	10-19	>=20
counts	2729	712	1978

Smoking

A catagorical variable with two levels of **yes**, or **NO**. the smoking status on a subject wether they are currently smoking or not smoking in the last 5 years. We can see that 58.85% of the subjects which is the majority of the workers that who are smokers.

Smoking	No	Yes
counts	2230	3189

Sex

A catagorical variables indicates **Female** or **Male**. The proportion of the gender is 2916/(2503+2916) = 53.81%, which male is slightly higher by 3.81%. A

Gender	F	M
counts	2503	2916

Race

A catagorical variable indicates the race **White** or **Other**. The majority of the workers are white. Race sometimes plays a role on certain diseases.

Race	Other	White
counts	1903	3516

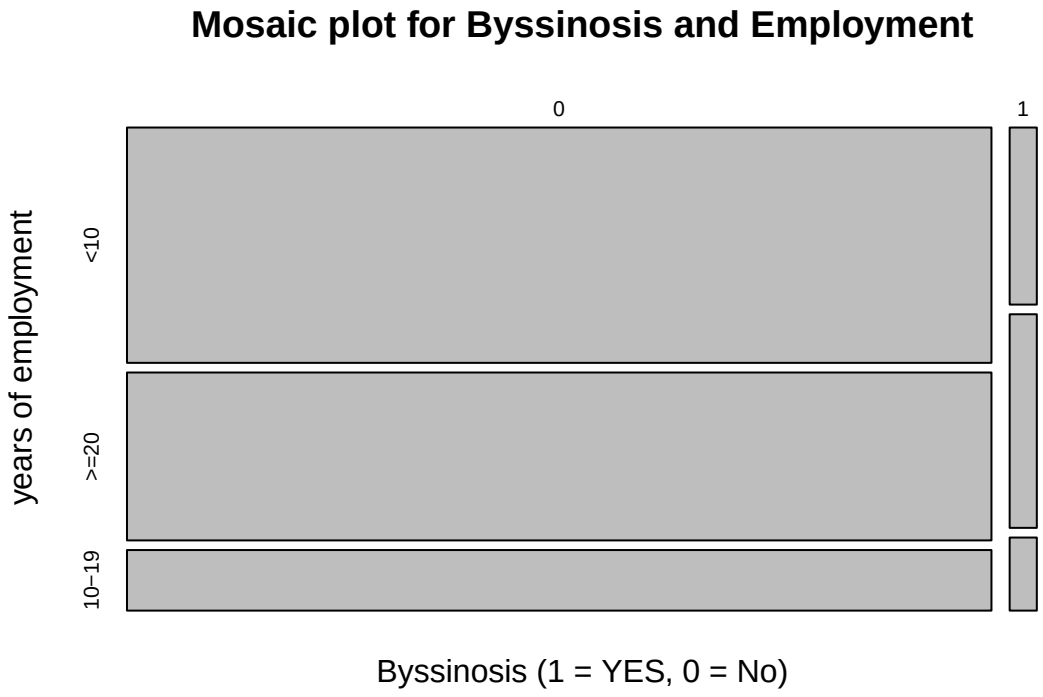
Workspace

A catagorical variable with levels of thee: **[1,(most dusty)]**, **[2,(less dusty)]**, and **[3,(least dusty)]**.

Workspace	[1,(most dusty)]	[2,(less dusty)]	[3,(least dusty)]
counts	669	1300	3450

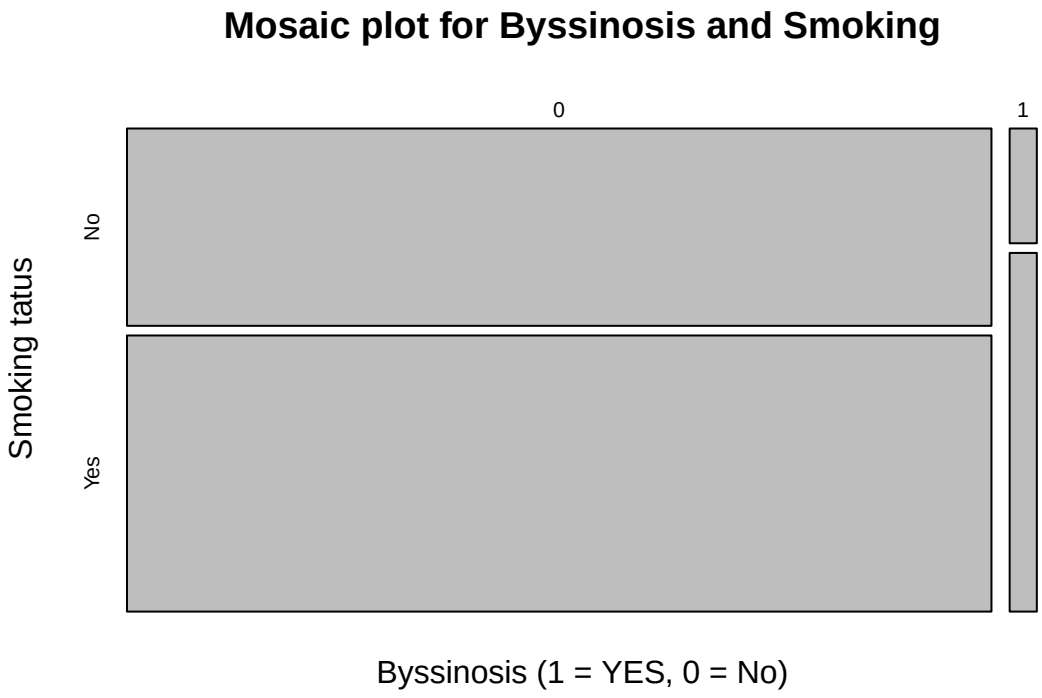
RESPONSE VARIABLE

Byssinosis and Employment

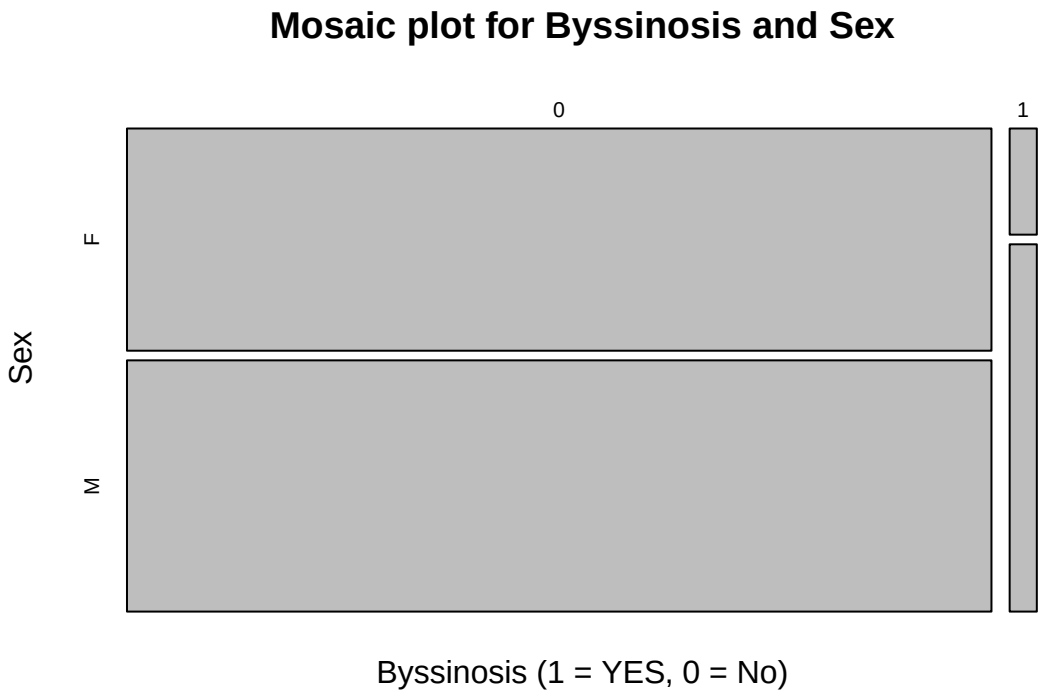


From the mosaic plot above,

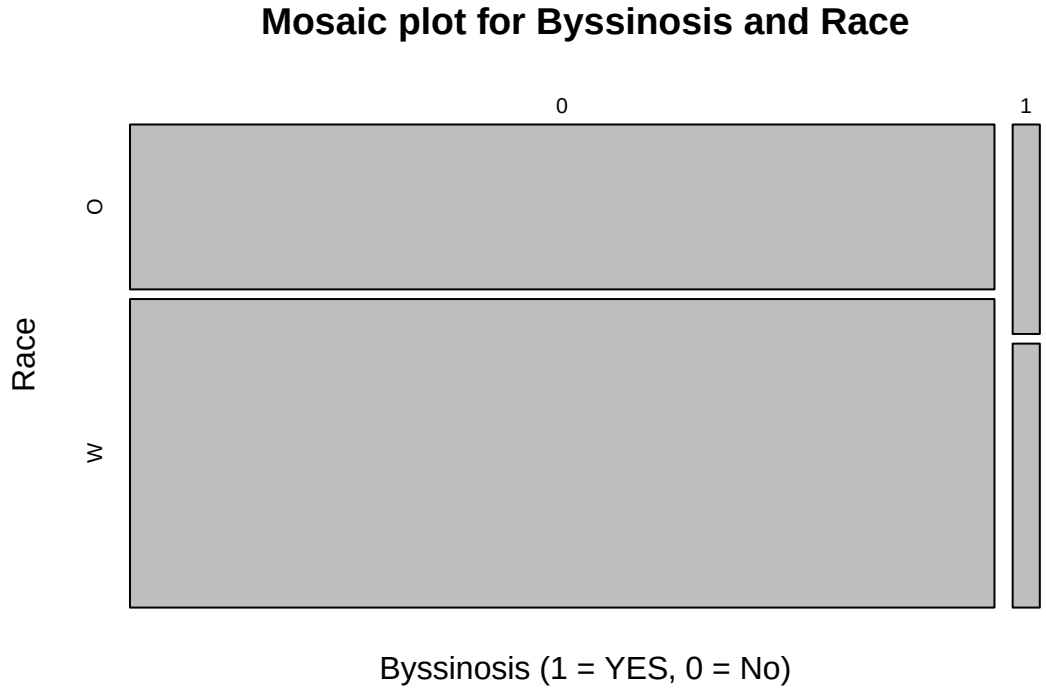
Byssinosis and Smoking



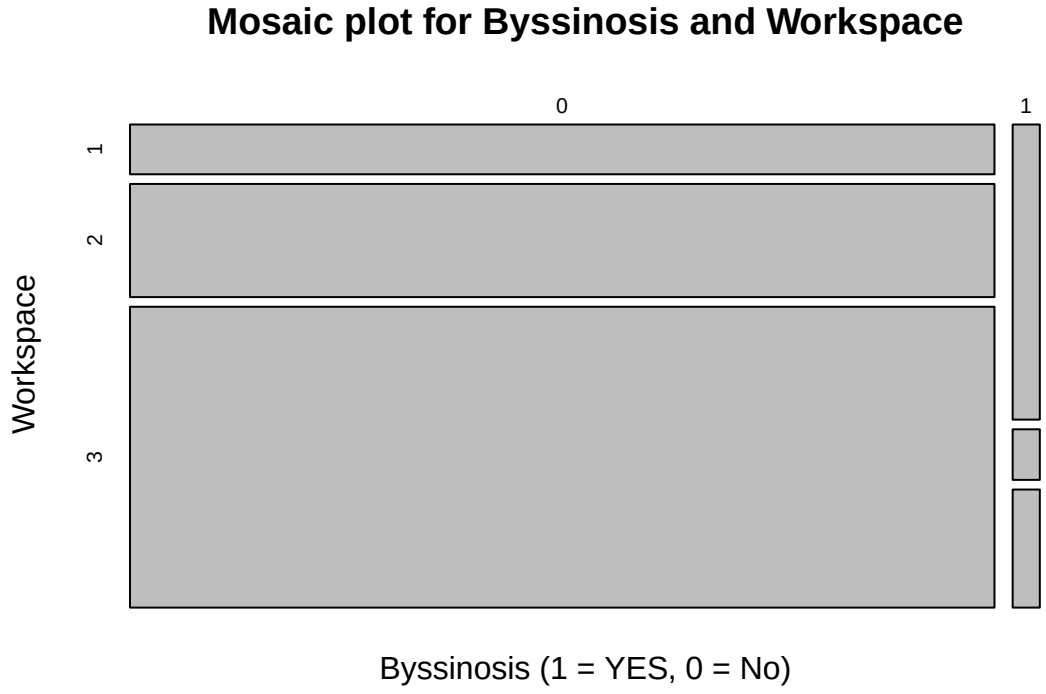
Byssinosis and Sex



Byssinosis and Race



Byssinosis and Workspcace



MODEL SELECTION

Method

For this investivation, our goal is to predict the chance of someone getting byssinosis based on our variables. Therefore, our main focous is not on making inferences of our logistic regression model. Thus being able to predict the probability of the occurrence of bissinosis is more important than getting a "best" model. With having this mind, we will use the backward selection with the Akaike Information Criterion (AIC) in the model selection process.

The reason for using backward selection method and AIC is that they both thend to over-fit the data. Benefit in doing so will provide us a better prediction on such disease.

We conducted the backward model selection process with AIC in R. The package we used for this procedure is the built-in {stats} package in R. Specifically, we used the step function. Here are the coefficients of the best model and the backward selection process:

Logit(π) = -2.4546 + 0.6728 $X_{Employment>=20}$ + 0.5060 $X_{Employment10-19}$ + 0.6210 $X_{Smoking,Yes}$ - 2.5493 $X_{Workspace,2}$ - 2.7175 $X_{Workspace,3}$

```
## Start:  AIC=1213.94
## Bys ~ Employment + Smoking + Sex + Race + Workspace
##
```

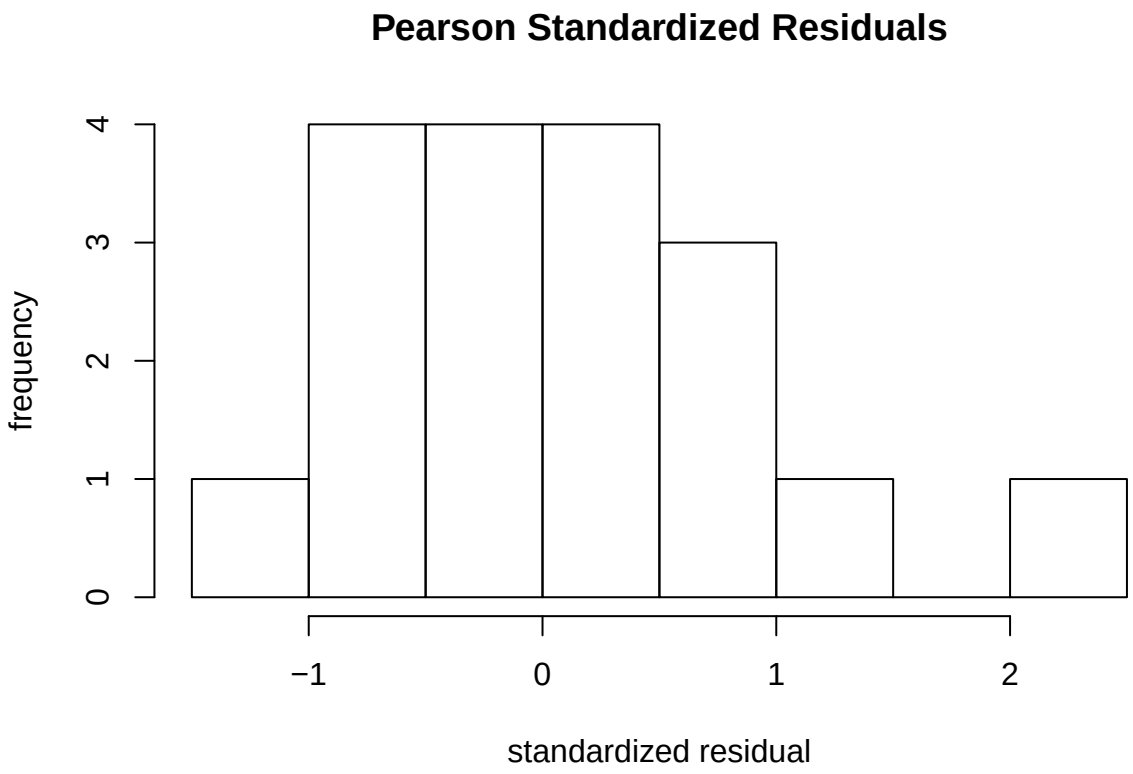
```
##              Df Deviance    AIC
## - Sex         1   1198.2 1212.2
## - Race        1   1198.2 1212.2
## <none>            1197.9 1213.9
## - Employment  2   1210.8 1222.8
## - Smoking     1   1209.7 1223.7
## - Workspace   2   1396.6 1408.6
##
## Step:  AIC=1212.23
## Bys ~ Employment + Smoking + Race + Workspace
##
##              Df Deviance    AIC
## - Race        1   1198.5 1210.5
## <none>            1198.2 1212.2
## - Employment  2   1210.8 1220.8
## - Smoking     1   1209.7 1221.7
## - Workspace   2   1418.5 1428.5
##
## Step:  AIC=1210.55
## Bys ~ Employment + Smoking + Workspace
##
##              Df Deviance    AIC
## <none>            1198.5 1210.5
## - Smoking     1   1210.0 1220.0
## - Employment  2   1213.1 1221.1
## - Workspace   2   1445.6 1453.6
```

Checking For Influential Outliers

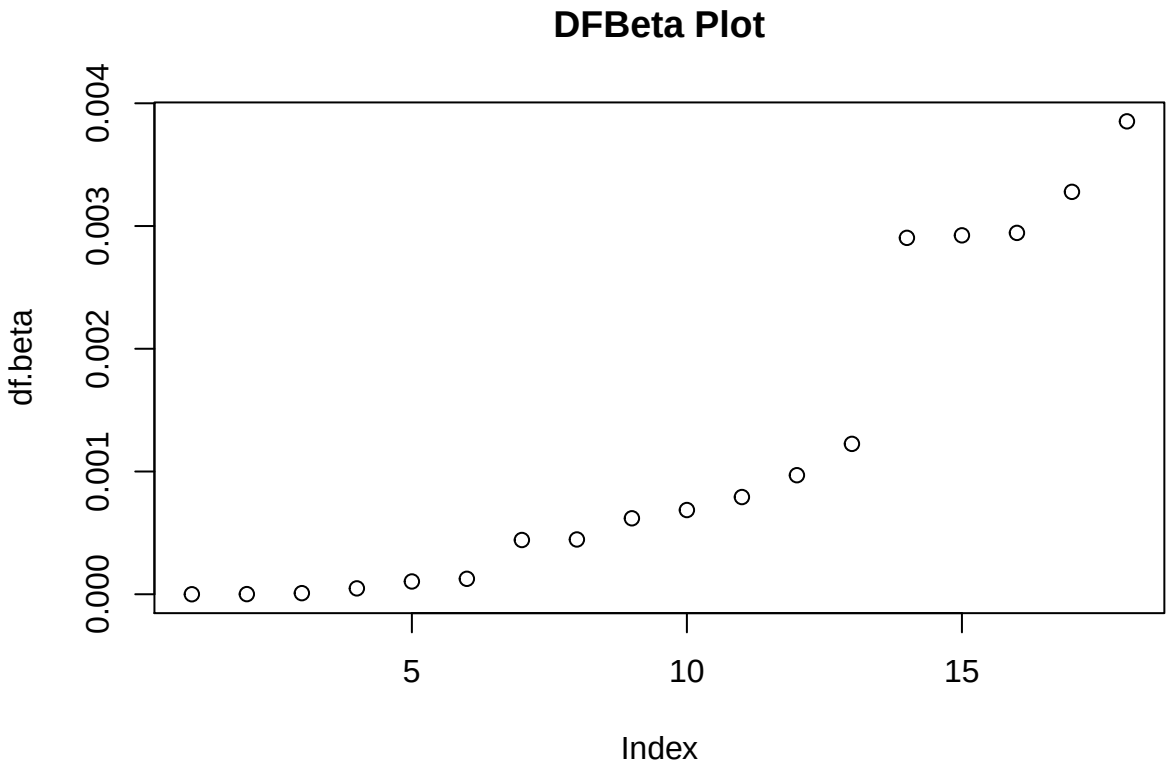
Method

it is crucial to check if there is any point that does not follow our model. Outliers are found using the standardized residual method. In addition, the DFBeta of each sample is also calculated to see its influence. Only the influential outliers are removed.

Result



Based on the graph above, we see that there is one outlier and influential observation in this data according to our logistic regression model. The Pearson Standardized residual for logistic regression approximately follows a skewed standard normal distribution. Therefore, we could interpret the plot with Z-score. Since the corresponding p-value for $Z = \pm 1$, thus, we decided to use standardized residual = 1 as a threshold to determine whether the subject is an outlier.



See the DFbeta Plot, it sugguets that the change in β cuased by most observations that are below 0.002. Therefore, we picked 0.002 as the threshold influential observation, and we got the outlier and had it removed. See the chart below.

```
##    Employment Smoking Sex Race Workspace Bys
## 14        <10     Yes   M    W         1   0
```

After removing the outlier, we have a new logistic regression model which will be introduced later.

Interaction Term

Model Selection

To study if the interaction terms may affect the goodness of the fit, we fitted all interactions between two variables to see if we should add interaction term to our model. See the chart below. By AIC, we have **Smoking*Workspace** as the interaction term. We do not know if the interaction term has effect to the model or not. Therefore, we will conduct a Likelihood Ratio test.

	LL	p	n	AIC	BIC
Bys ~ Employment + Smoking + Workspace	-599.125	6	5418	1210.250	1249.835
Bys ~ Employment + Smoking + Workspace + Employment*Smoking	-598.233	8	5418	1212.465	1265.245
Bys ~ Employment + Smoking + Workspace + Employment*Workspace	-596.844	10	5418	1213.689	1279.663

	LL	p	n	AIC	BIC
Bys ~ Employment + Smoking + Workspace + Smoking*Workspace	-596.689	8	5418	1209.377	1262.157

Likelihood Ratio test

H_o: The interaction term can be dropped H_A: The interaction term cannot be dropped

Test stats: $G^2 = -2(L_o - L_1) = -2(-599.274 + 596.848) = 4.852$ df $= 1$

P-value: $p(\chi^2 > 4.852) = 0.0276141$

Conclusion: Reject H_o, conclude that the interaction term cannot be dropped

Therefore, our final model is:

$\ln(\frac{\hat{\pi}}{1-\hat{\pi}})$ = -2.7132 + 0.6614 $X_{Employment>=20}$ + .4964 $X_{Employment10-19}$ + 0.9595 $X_{Smoking,Yes}$ - 1.7767 $X_{Workspace,2}$ - 2.3459 $X_{Workspace,3}$ -1.1830 $X_{Smoking,Yes} * X_{Workspace,2}$ - 0.4886 $X_{Smoking,Yes} * X_{Workspace,3}$

Interpretation of the Logistic Regression Model

Since the odds are much more meaningful than the log odds, we will exponentiate the model coefficients before interpret.

exp(β) = exp(-2.7132) = 0.06632423 here stands for the log odds of getting Byssinosis for who is employed less than ten years, do not smoke, work in the most dusty environment.

- $X_{Employment>=20}$

  While holding other variables constant, the odds of getting Byssinosis for people who has been employed greater or equal 20 years is 1.937503 times of who has been emplyed less than ten years.

- $X_{Employment10-19}$

  While holding other variables constant, the odds of getting Byssinosis for people who has been employed between 10 and 19 years is 1.642797 times of who has been emplyed less than ten years.

- $X_{Smoking,Yes}$ While holding other variables constant, the odds of getting byssinosis for who smokes is 2.610391 times of who does not smoke or did not smoke in the past 5 years.

- $X_{Workspace,2}$ While holding other variables constant, the odds of getting byssinosis for who work in less dusty environment is 0.1691956 times of who work in the most dusty enviroment.

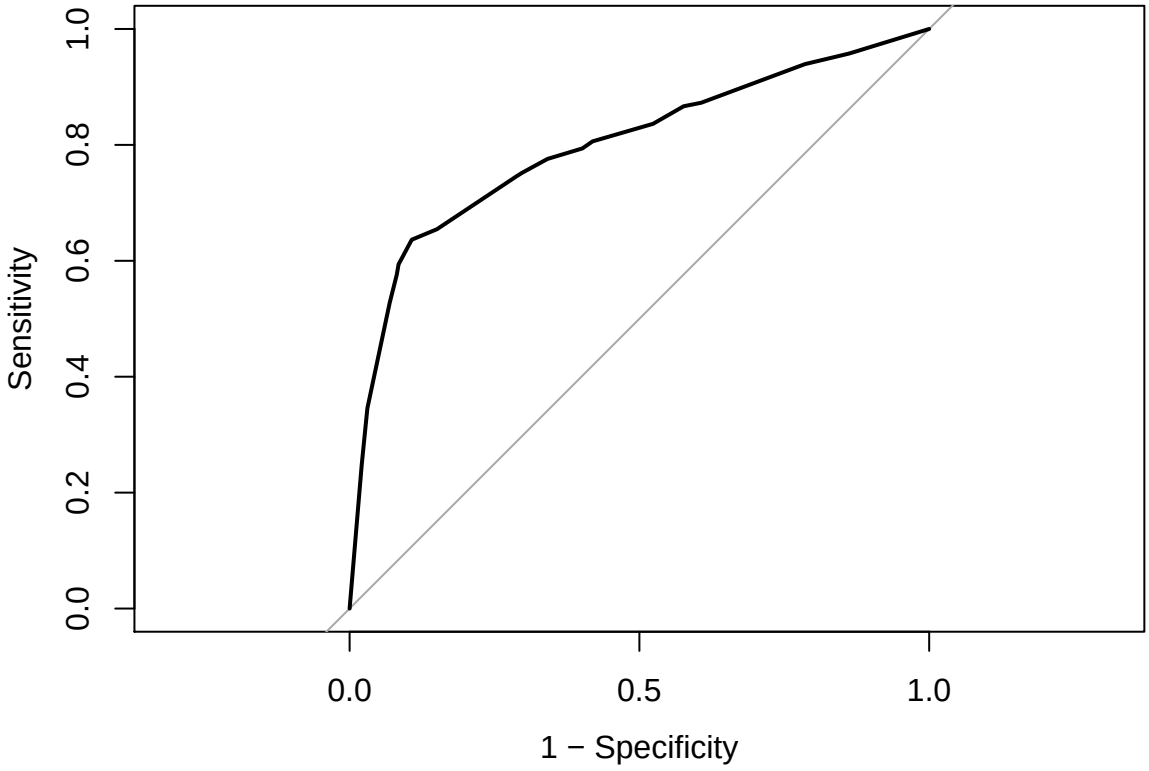
- $X_{Workspace,3}$ While holding other variables constant, the odds of getting byssinosis for who work in the least dusty enviroment is 0.09576098 times of who work in the most dusty enviroment.

- $X_{Smoking,Yes} * X_{Workspace,2}$ While holding other variables constant, the odds of getting byssinosis for who smokes and work in less dusty environment is 0.3063583 times of who smokes and work in the most dusty environment.

- $X_{Smoking,Yes} * X_{Workspace,3}$ While holding other variables constant, the odds of getting byssinosis for who smokes and work in the least dusty environment is 0.6134847 times of who smokes and work in the most dusty environment.

  From above, the employment time has most effects in this model. It explains effects of texposure to dusty environment over time. Also, dusty level in working environment plays a big role as well. So is smoking that increase the chance of having such disease.

Model Goodness of Fit Check with ROC Plot and Error Matrix



To check the overall Goodness of fit, refer to the ROC plot above. According to R, the AUC (area under the curve) is 0.7977 with 95% confidence interval (0.7564 , 0.839), it indicagtes that our model fits very good with to the data.

Furthermore, we chose $\pi_o = 0.0304$ as a cutoff and built an error matrix in R. If $\hat{\pi} > 0.0304$, then it sugguest that the subject has byssinosis and vice versa. Please refer to the tables below.

ERROR MATRIX	$\hat{y} = 0$	$\hat{y} = 1$	total
y = 0	4690	563	5253
y = 1	60	105	165
total	4750	668	5418

Sensitivity	Specificity	Error-Rate
0.6363636	0.8928231	0.1149871

The result is not bad for predicting who acturally has such disease. However, the sensitivity is a bit quite low for predicting diseases. Such a low Sensitivity might be due to the overffiting from the model, small data set. For what we have so far, it should be acceptable for disease prevention. For the model to be medical grade, we need more data.

Application of the Logistic Model

Assume we want to predict the probability of Byssinosis for an individual who has been working in the industry for more than 20 years, smokes, and work in the most dusty envirment.

$\ln(\frac{\hat{\pi}}{1-\hat{\pi}})$ = -2.7132 + 0.6614 $X_{Employment>=20}$ + 0.9595 $X_{Smoking,Yes}$

$\hat{\pi}_1 = 0.2511854$

Based on the model, the probability for such an indivual to have Byssinosis is approximately 25.12%

Conclusion

```
##
## Call:
## glm(formula = Bys ~ Employment + Workspace + Smoking, family = binomial(link = "logit"),
##     data = new_df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7385  -0.2023  -0.1486  -0.1450   3.2178
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -2.4527     0.2047 -11.984  < 2e-16 ***
## Employment>=20    0.6706     0.1813   3.698 0.000217 ***
## Employment10-19   0.5036     0.2490   2.022 0.043129 *
## Workspace2       -2.5507     0.2614  -9.759  < 2e-16 ***
## Workspace3       -2.7188     0.1898 -14.322  < 2e-16 ***
## SmokingYes        0.6222     0.1908   3.262 0.001108 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1477.1  on 5417  degrees of freedom
## Residual deviance: 1198.3  on 5412  degrees of freedom
## AIC: 1210.3
##
## Number of Fisher Scoring iterations: 7
```

Based on the p-value of each explanatory variable, the variable with the smallest p-value is less than 2e-16, which indicates that this variable has the significant effect on the response variable. Then followed by the workspaces that indicates the most important variable. Then, the rest of the variables are important because they are be low 0.05. Based on the variable, it could contribute a positive or negative effect on our response variable.

R Appendix

```
library(ggplot2)
library(LogisticDx)
```

```r
library(pROC)

dat_bys = read.csv("Byssinosis.csv")

# begin ------funtion to reasemble the data frame for Byssinosis to be [yex,no] variable---------
dat_new_bys_yes = data.frame()
NEW_FRAME = function( dat = y, v = x) {
  a = dat[1,]
  for (i in (1: (length(v))) ){
    if ( v[i] != 0) {
      for ( j in (1: v[i] )){
        a = rbind(a, dat[i,])
      }
    }
  }
  a = a[-c(1),]
  return(a)
}

a = NEW_FRAME(dat = dat_bys, v = dat_bys$Non.Byssinosis)
b = NEW_FRAME(dat = dat_bys, v = dat_bys$Byssinosis)
#non-BYS
a = a[,1:5]
a$Bys = 0
#BYS
b = b[,1:5]
b$Bys = 1

#conbine Data frame
df = rbind(a,b)
df$Workspace = factor(df$Workspace)
row.names(df) = 1:nrow(df)
# -------------- End ------------------------------------------------
# begin ----------Expanatory---------------------------------
head(df,5)
#table(df$Employment)
#table(df$Smoking)
#table(df$Sex)
#table(df$Race)
#table(df$Workspace)
# ========== response variable ----------
mosaicplot(table(df$Bys, df$Employment),
           main = 'Mosaic plot for Byssinosis and Employment',
           xlab = "Byssinosis (1 = YES, 0 = No)",
           ylab = "years of employment")


mosaicplot(table(df$Bys, df$Smoking),
           main = 'Mosaic plot for Byssinosis and Smoking',
           xlab = "Byssinosis (1 = YES, 0 = No)",
           ylab = "Smoking tatus")

mosaicplot(table(df$Bys, df$Sex),
```

```
           main = 'Mosaic plot for Byssinosis and Sex',
           xlab = "Byssinosis (1 = YES, 0 = No)",
           ylab = "Sex")


mosaicplot(table(df$Bys, df$Race),
           main = 'Mosaic plot for Byssinosis and Race',
           xlab = "Byssinosis (1 = YES, 0 = No)",
           ylab = "Race")



mosaicplot(table(df$Bys, df$Workspace),
           main = 'Mosaic plot for Byssinosis and Workspace',
           xlab = "Byssinosis (1 = YES, 0 = No)",
           ylab = "Workspace")


# end---------explanatory variable----------
#---MODEL selection---

all_lm = glm( Bys ~ Employment + Smoking + Sex + Race + Workspace, family = binomial(link = logit), data

empty_lm = glm(Bys ~ 1 , data = df, family = binomial(link = logit))

best_lm = step(all_lm ,scope = list(lower = empty_lm, upper = all_lm), direction = "backward")

best_lm = glm(Bys ~ Employment + Smoking + Workspace, family = binomial(link = logit), data = df)

#all_lm
# ---------------same result as above from book method (digital page 114 of 490)------------------
#lg_model_og = glm( (Byssinosis/(Byssinosis + Non.Byssinosis)) ~ Employment + Smoking + Sex + Race + fa
#summary(lg_model_og)
#-------------------------------------------------------------------------------------------------
# explore outliers

good.stuff = dx(best_lm)
good.stuff = as.data.frame(good.stuff) #Convert to dataframe because of annoying things
pear.r = good.stuff$Pr #Pearsons Residuals
std.r = good.stuff$sPr #Standardized residuals (Pearson)
df.beta = good.stuff$dBhat #DF Beta for removing each observation
change.pearson = good.stuff$dChisq #Change in pearson X^2 for each observation

hist(std.r, main = "Pearson Standardized Residuals", xlab = 'standardized residual', ylab = 'frequency'
plot(df.beta,main = "DFBeta Plot")
# Find number of outliers
outliers = as.data.frame(good.stuff[abs(good.stuff$sPr) > 1,])


cutoff.beta = 0.002
remove = outliers[outliers$dBhat > cutoff.beta, ]
```

```r
new_df = df[-c(14),]

df[14,]

# test for interaction term
All_models = c("Bys ~ Employment + Smoking + Workspace",
               "Bys ~ Employment + Smoking + Workspace + Employment*Smoking",
               "Bys ~ Employment + Smoking + Workspace + Employment*Workspace",
               "Bys ~ Employment + Smoking + Workspace + Smoking*Workspace")

All.Criteria = function(the.model){
  p = length(the.model$coefficients)
  n = length(the.model$residuals)
  the.LL = logLik(the.model)
  the.BIC =  -2*the.LL + log(n)*p
  the.AIC =  -2*the.LL + 2*p
  the.results = c(the.LL,p,n,the.AIC,the.BIC)
  names(the.results) = c("LL","p","n","AIC","BIC")
  return(the.results)
}

all_model_crit = t(sapply(All_models, function(M){
  current.model = glm(M, data = new_df, family = binomial)
  All.Criteria(current.model)
}))

RC = round(all_model_crit,3)
#RC

best_lm = glm(Bys ~ Employment + Smoking + Workspace + Smoking*Workspace, data = new_df, family = binom
# ROC plot
my.auc = auc(best_lm$y, fitted(best_lm),plot = TRUE,legacy.axes = TRUE)
#my.auc

# predicted value
auc.CI = ci(my.auc,level = 1-0.05)
#auc.CI
# error matrix
pi.0 = 165/5419
truth = new_df$Bys #The true values of Bys
predicted = ifelse(fitted(best_lm)>pi.0,1,0) #The predicted values of y based on pi.0
my.table = table(truth,predicted)
sens = sum(predicted == 1 & truth == 1 )/sum(truth == 1)
spec = sum(predicted == 0 & truth == 0 )/sum(truth == 0)
error = sum(predicted != truth)/length(predicted)
results = c(sens,spec,error)
names(results) = c("Sensitivity", "Specificity", "Error-Rate")
#results
summary(glm(Bys~Employment + Workspace + Smoking , data = new_df, family = binomial(link = "logit")))
```