STA138 Final Project

Jingyuan Liang

12/9/2019

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]

And my task to this project 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.

2. Data Preperation

Looking at this data, the outcome variable is in counts in the last two columns. Even though we can process data in this way, it's more convenient for further analysis that we expand the data to a full list of 5,419 rows with the outcome variable named "ill", with levels "1" and "0" representing whether the Byssinosis occurs. Also the workplace variable is a dummy variable with 3 levels, the 3-levels dummy variable can confuse R and it may be recognized as a continuous variable. So I use "most dusty", "less dusty", "least dusty" as three levels instead of just "1", "2", "3".

Note that the AICs computed afterwards may seem kinda large, that's due to the expansion of the dataset. But the best model results are the same as using the original dataset.

After doing so, the new data looks like this:

```
byss = read.csv("/Users/owen/Downloads/byssss.csv")
head(byss)
```

```
##   X Employment Smoking Sex Race workspace. ill
## 1 0        <10     Yes   M    W most dusty   1
## 2 1        <10     Yes   M    W most dusty   1
## 3 2        <10     Yes   M    W most dusty   1
## 4 3        <10     Yes   M    W most dusty   0
## 5 4        <10     Yes   M    W most dusty   0
## 6 5        <10     Yes   M    W most dusty   0
```

3. Data Analysis & Model Selection

```
library(bestglm)
library(LogisticDx)
library(olsrr)
library(car)
library(pROC)
model=glm(formula = ill ~ (Employment+Smoking+Sex+Race+workspace.), family = binomial(),
data = byss)
summary(model)
```

```
##
## Call:
## glm(formula = ill ~ (Employment + Smoking + Sex + Race + workspace.),
##     family = binomial(), data = byss)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8080  -0.1980  -0.1535  -0.1362   3.2269
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -5.0769     0.2624 -19.349  < 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 0.587982
## RaceW                 -0.1163     0.2072  -0.562 0.574425
## workspace.less dusty   0.1507     0.2860   0.527 0.598193
## workspace.most dusty   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: 1477.2  on 5418  degrees of freedom
## Residual deviance: 1197.9  on 5411  degrees of freedom
## AIC: 1213.9
##
## Number of Fisher Scoring iterations: 7
```

```
fullModel1 = model
nullModel1 = glm(ill ~ 1,data=byss,family = binomial())
```

The first step we do is to fit a full additive logistic model to get a basic understanding of both the data and the full model. Then we will use multiple model selection techniques to determine which is the best additive model.

```
bestForwardAIC=step(nullModel1,scope = list(lower = nullModel1, upper = fullModel1),dire
ction = "forward")
```

```
## Start:  AIC=1479.19
## ill ~ 1
##
##              Df Deviance    AIC
## + workspace.  2   1225.1 1231.1
## + Sex         1   1435.8 1439.8
## + Smoking     1   1455.8 1459.8
## + Employment  2   1467.0 1473.0
## + Race        1   1471.2 1475.2
## <none>            1477.2 1479.2
##
## Step:  AIC=1231.08
## ill ~ workspace.
##
##              Df Deviance    AIC
## + Employment  2   1210.0 1220.0
## + Smoking     1   1213.1 1221.1
## + Race        1   1222.6 1230.6
## <none>            1225.1 1231.1
## + Sex         1   1224.4 1232.4
##
## Step:  AIC=1219.98
## ill ~ workspace. + Employment
##
##           Df Deviance    AIC
## + Smoking  1   1198.5 1210.5
## <none>         1210.0 1220.0
## + Race     1   1209.7 1221.7
## + Sex      1   1210.0 1222.0
##
## Step:  AIC=1210.55
## ill ~ workspace. + Employment + Smoking
##
##        Df Deviance    AIC
## <none>      1198.5 1210.5
## + Race  1   1198.2 1212.2
## + Sex   1   1198.2 1212.2
```

The first technique we use is the forward selection with AIC. From the result we know that the best model from forward selection includes variable “Workspace”, “Employment”, and “Smoking”, and the AIC for this model is 1210.55.

```
bestBackwardAIC=step(fullModel1,scope = list(lower = nullModel1, upper = fullModel1),dir
ection = "backward")
```

```
## Start:  AIC=1213.94
## ill ~ (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
## ill ~ 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
## ill ~ 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
```

The second technique we use is the backward selection with AIC. From the result we know that the best model from backward selection includes variable “Workspace”, “Employment”, and “Smoking”, and the AIC for this model is 1210.55. The same result as forward selection.

```
bestBidirectionAIC=step(fullModel1,scope = list(lower = nullModel1, upper = fullModel1),
direction = "both")
```

```
## Start:  AIC=1213.94
## ill ~ (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
## ill ~ Employment + Smoking + Race + workspace.
##
##              Df Deviance    AIC
## - Race        1   1198.5 1210.5
## <none>            1198.2 1212.2
## + Sex         1   1197.9 1213.9
## - Employment  2   1210.8 1220.8
## - Smoking     1   1209.7 1221.7
## - workspace.  2   1418.5 1428.5
##
## Step:  AIC=1210.55
## ill ~ Employment + Smoking + workspace.
##
##              Df Deviance    AIC
## <none>            1198.5 1210.5
## + Race        1   1198.2 1212.2
## + Sex         1   1198.2 1212.2
## - Smoking     1   1210.0 1220.0
## - Employment  2   1213.1 1221.1
## - workspace.  2   1445.6 1453.6
```

The third technique we use is the bidirectional selection with AIC. From the result we know that the best model from bidirectional selection includes variable “Workspace”, “Employment”, and “Smoking”, and the AIC for this model is 1210.55. The same result as forward selection and backward selection.

```
bestSubsetAIC = bestglm(Xy = byss, family = binomial(),IC = "AIC",method = "exhaustive")
bestSubsetAIC
```

```
## AIC
## Best Model:
##               Df Sum Sq Mean Sq F value Pr(>F)
## X              1   4.13   4.128 152.581 <2e-16 ***
## Employment     2   2.82   1.408  52.055 <2e-16 ***
## Smoking        1   0.05   0.048   1.773  0.183
## workspace.     2   6.56   3.278 121.165 <2e-16 ***
## Residuals   5412 146.43   0.027
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The final technique we use is the all subset selection with AIC. From the result we know that the best model from all subset selection includes variable “Workspace”, “Employment”, and “Smoking”.

For now, we have a consistant best model with all four techniques. Therefore, we will continue to use this model as the current best model for further analysis.

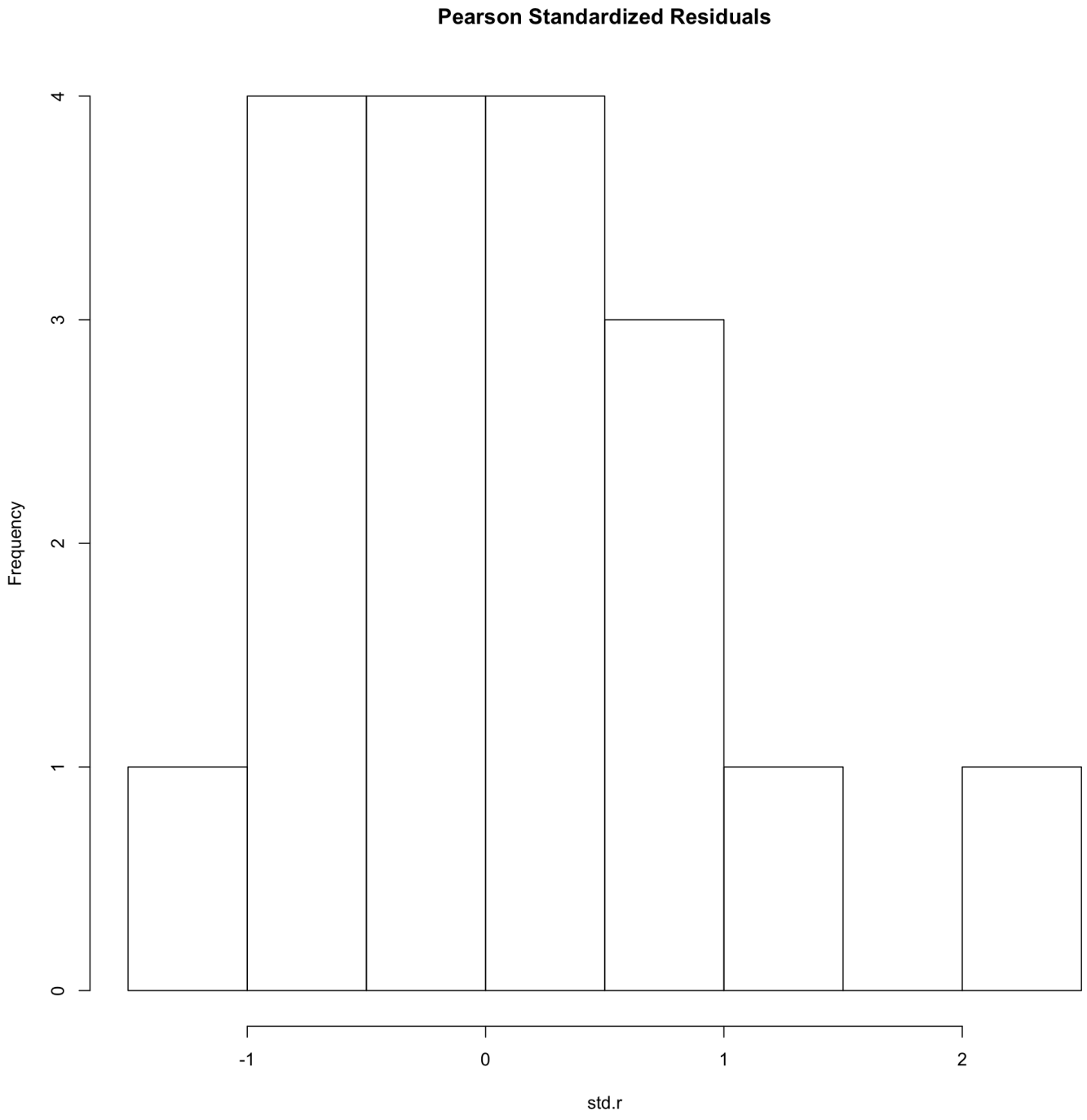
```
summary(glm(formula = ill ~ (Employment+Smoking+workspace.+Employment:workspace.), famil
y = binomial(),data = byss))
```

```
##
## Call:
## glm(formula = ill ~ (Employment + Smoking + workspace. + Employment:workspace.),
##     family = binomial(), data = byss)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7609  -0.1976  -0.1639  -0.1406   3.2356
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       -4.92040    0.27261 -18.049  < 2e-16 ***
## Employment>=20                     0.37729    0.32720   1.153  0.24887
## Employment10~19                   -0.30870    0.55558  -0.556  0.57846
## SmokingYes                         0.61718    0.19123   3.227  0.00125 **
## workspace.less dusty               0.07513    0.42756   0.176  0.86051
## workspace.most dusty               2.32317    0.29433   7.893 2.95e-15 ***
## Employment>=20:workspace.less dusty  -0.07900    0.61568  -0.128  0.89790
## Employment10~19:workspace.less dusty  0.85771    0.88086   0.974  0.33019
## Employment>=20:workspace.most dusty   0.51148    0.40481   1.263  0.20641
## Employment10~19:workspace.most dusty  1.13667    0.63775   1.782  0.07470 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1477.2  on 5418  degrees of freedom
## Residual deviance: 1193.9  on 5409  degrees of freedom
## AIC: 1213.9
##
## Number of Fisher Scoring iterations: 7
```

```
model2 = glm(formula = ill ~ (Employment+Smoking+workspace.), family = binomial(),data =
byss)
```

Just by looking at the predictor variable, I have an intuition that “employment” and “workspace” may have an interaction effect, since it’s reasonable that people tend to absorb more harmful substance if they work a long time in a very dusty place than they work few years in a less dusty place. So I added this interaction term to see if it can improve the model. Unfortunately, the AIC goes higher after adding this interaction, which means the model is not improving. Moreover, most of the p-values of this interaction are also insignificant. Therefore, we will keep the previous best model for now.

```
good.stuff = dx(model2)
good.stuff = as.data.frame(good.stuff)
pear.r = good.stuff$Pr
std.r = good.stuff$sPr
df.beta = good.stuff$dBhat
change.pearson = good.stuff$dChisq
hist(std.r, main = "Pearson Standardized Residuals")
```



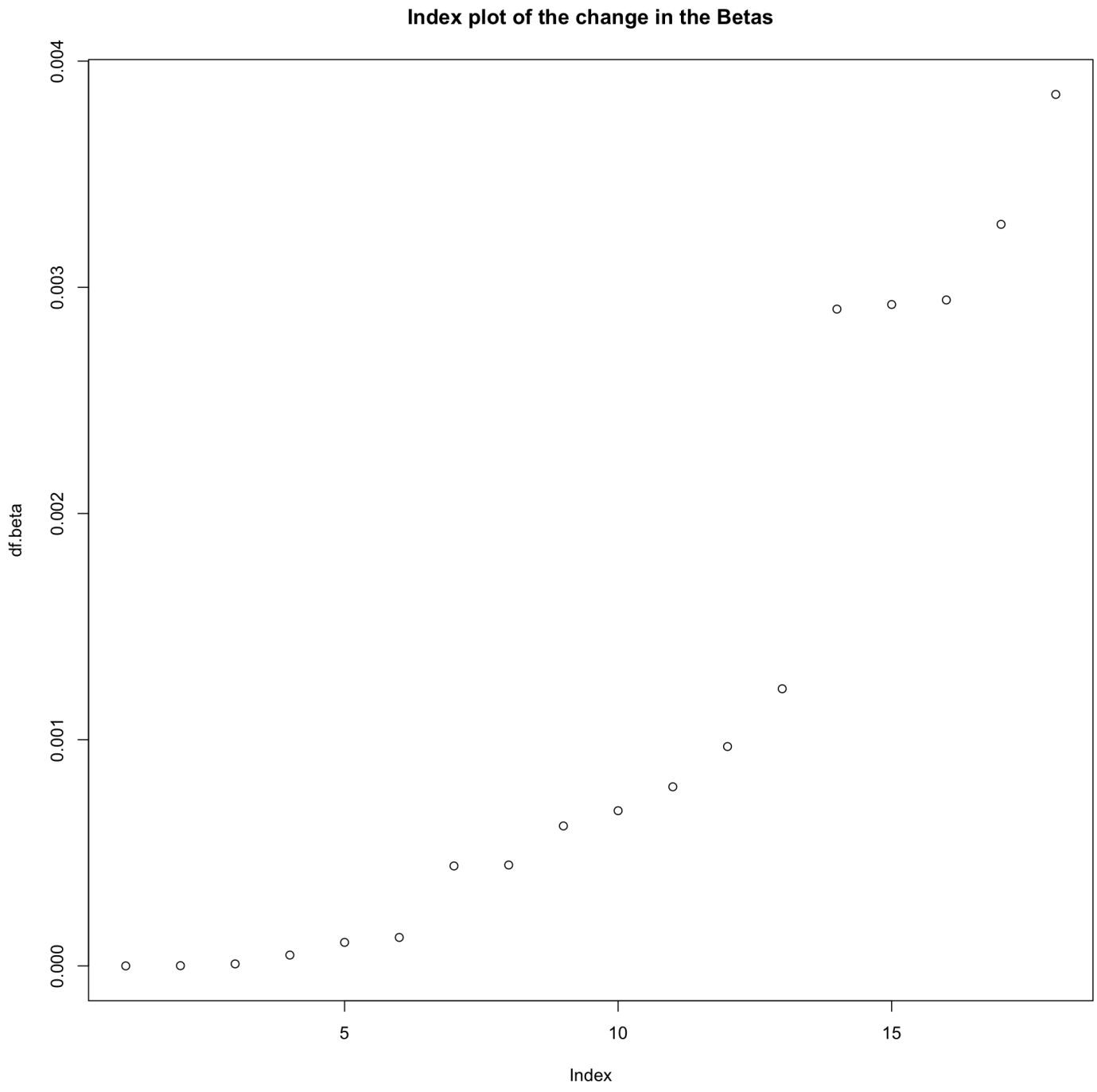

```
cutoff.std = 2.0
std.r[std.r > cutoff.std]
```

```
## [1] 2.274912
```

```
good.stuff[std.r > cutoff.std,c(1:(length(model2$coefficients)+1),which(names(good.stuf
f) == "sPr"))]
```

```
##    (Intercept) Employment>=20 Employment10~19 SmokingYes workspace.less dusty
## 14           1              0               0          0                     1
##    workspace.most dusty y      sPr
## 14                    0 5 2.274912
```

```
plot(df.beta,main = "Index plot of the change in the Betas")
```



From the plot of Pearson Standardized Residuals and Dfbeta, we find an outlier, which is the data point of row 14. We will remove it in the new model.

```
model3=glm(formula = ill ~ Employment+Smoking+workspace., family = binomial(),data = bys
s[-(14),])
summary(model3)
```

```
##
## Call:
## glm(formula = ill ~ Employment + Smoking + workspace., family = binomial(),
##     data = byss[-(14), ])
##
## 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)           -5.1715     0.2349 -22.016  < 2e-16 ***
## Employment>=20         0.6706     0.1813   3.698 0.000217 ***
## Employment10~19        0.5036     0.2490   2.022 0.043129 *
## SmokingYes             0.6222     0.1908   3.262 0.001108 **
## workspace.less dusty   0.1681     0.2841   0.592 0.554100
## workspace.most dusty   2.7188     0.1898  14.322  < 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: 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
```

The AIC is reduced by 0.2. And we can see that there's an insignificant level of "workspace", which is the "less dusty" level.

```
byss2 = byss
byss2$workspace. <- factor(c("less dusty", "less dusty", "most dusty")[byss2$workspac
e.])
model4=glm(formula = ill ~ Employment+Smoking+workspace., family = binomial(),data = bys
s2[-14,])
summary(model4)
```

```
##
## Call:
## glm(formula = ill ~ Employment + Smoking + workspace., family = binomial(),
##     data = byss2[-14, ])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7382  -0.2070  -0.1486  -0.1399   3.2022
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -5.1211     0.2170 -23.604  < 2e-16 ***
## Employment>=20         0.6684     0.1813   3.687 0.000227 ***
## Employment10~19        0.4995     0.2489   2.007 0.044747 *
## SmokingYes             0.6206     0.1907   3.254 0.001140 **
## workspace.most dusty   2.6712     0.1696  15.749  < 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: 1477.1  on 5417  degrees of freedom
## Residual deviance: 1198.6  on 5413  degrees of freedom
## AIC: 1208.6
##
## Number of Fisher Scoring iterations: 7
```

Since we see there's a insignificant level ("less dusty") of the variable "workspace", we have two ways to deal with it: first, carry out a LR test and consider to drop the entire varible; second, combining the levels. I choose the latter, since the "most dusty" level is extremely significant for our model, while it's also making sense that we combine the levels "less dusty" to "least dusty". These two levels can just be "not that dusty" compared to "most dusty".

After combining levels, the AIC dropped to 1208.6. We will keep using this model.

4. Interpretation

```
b0=summary(model4)$coefficients[1,1]
b1=summary(model4)$coefficients[2,1]
b2=summary(model4)$coefficients[3,1]
b3=summary(model4)$coefficients[4,1]
b4=summary(model4)$coefficients[5,1]

exp(b0)
```

```
## [1] 0.005969551
```

```
#odds of having the disease Byssinosis for non-smoking people that have been employed fo
r less than 10 years in the less dusty/least dusty workplace is 0.005969551.

exp(b1)
```

```
## [1] 1.951167
```

```
#the estimated odds of having the disease Byssinosis for people that are employed for mo
re than 20 years are 1.951167 times that of people who are employed for less than 10 yea
rs. Holding all other variables constant.

exp(b2)
```

```
## [1] 1.647828
```

```
#the estimated odds of having the disease Byssinosis for people that are employed for be
tween 10 and 19 years are 1.647828 times that of people who are employed for less than 1
0 years. Holding all other variables constant.

exp(b3)
```

```
## [1] 1.860046
```

```
#the estimated odds of having the disease Byssinosis for smoking people are 1.843 times
 that of non-smoking people is 1.860046, holding all other variables constant.

exp(b4)
```

```
## [1] 14.45741
```

```
#the estimated odds of having the disease Byssinosis for people that are working in most
dusty workplace are 14.45741 times that of people who are working in less dusty/least du
sty workplace. Holding all other variables constant.
```

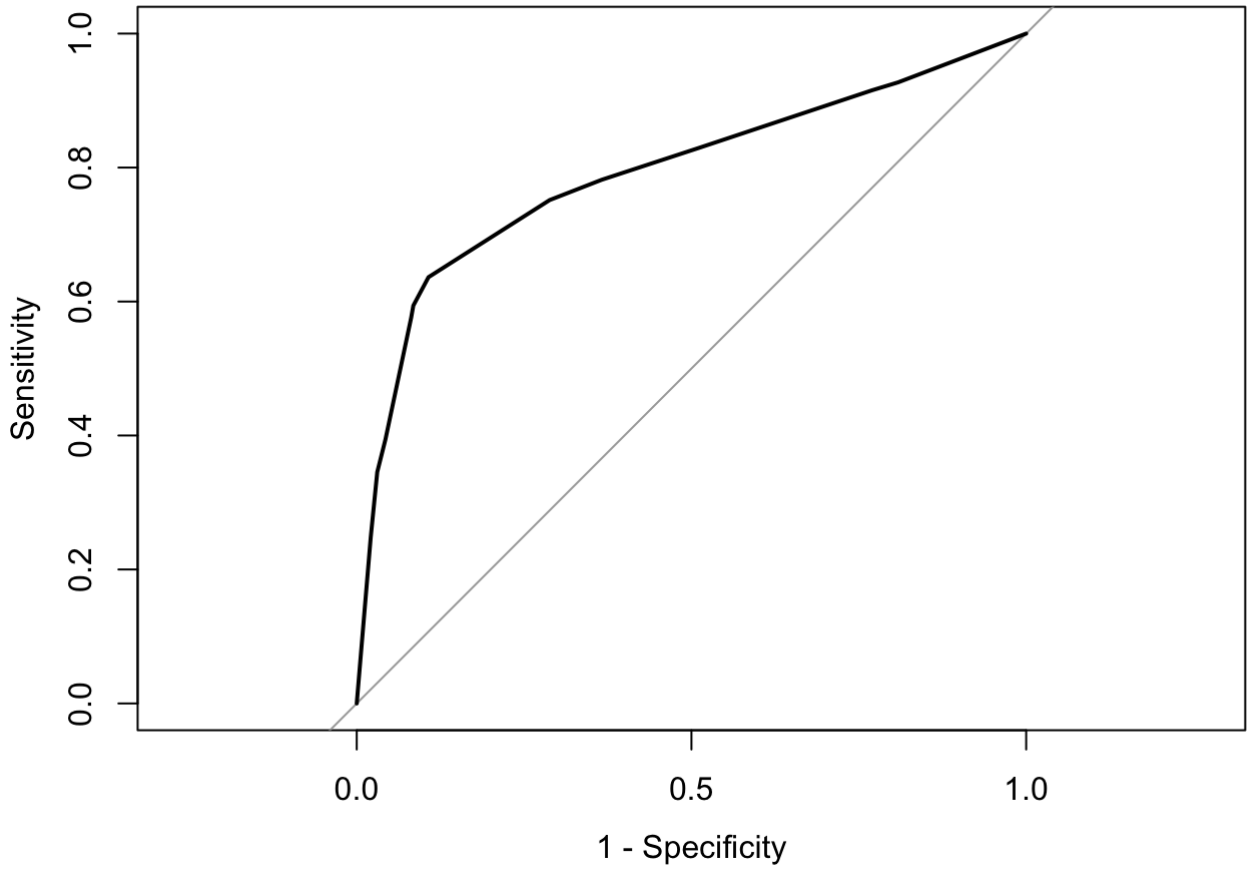
```
estimates =  summary(model4)$coefficients[,1]
SE =  summary(model4)$coefficients[,2]

alpha = 0.05
z.a.2 = qnorm(1-alpha/2)
upper.bounds = estimates +z.a.2*SE
lower.bounds = estimates -z.a.2*SE
Wald.CI = cbind(lower.bounds,upper.bounds)
Wald.CI
```

```
##                       lower.bounds upper.bounds
## (Intercept)           -5.54631057   -4.6958564
## Employment>=20         0.31305370    1.0238015
## Employment10~19        0.01171123    0.9872044
## SmokingYes             0.24674359    0.9944589
## workspace.most dusty   2.33876519    3.0036483
```

From the 95% Confidence Intervals for βs, no CI covers 0, which means all βs are significant and we should keep all of them.

4. Predictions

```
my.auc = auc(byss2[-14,]$ill,fitted(model4),plot = TRUE,legacy.axes = TRUE)
```




```
my.auc
```

```
## Area under the curve: 0.7925
```

```
auc.CI = ci(my.auc,level = 1-0.05)
auc.CI
```

```
## 95% CI: 0.7497-0.8353 (DeLong)
```

Th AUC indicates that our fit does a reasonable job, since AUC is around 0.8. (The more AUC close to 1, the better the model does.)

```
predict(model4, newdata = data.frame(Employment = ">=20", Smoking = "Yes", workspace. =
"most dusty"),type = "response")
```

```
##         1
## 0.2385132
```

```
predict(model4, newdata = data.frame(Employment = "<10", Smoking = "Yes", workspace. =
"most dusty"),type = "response")
```

```
##         1
## 0.1383246
```

Here I try to do the prediction of the model we've selected. There's 0.2385132 probability that a smoking person who worked in the most dusty place for more than 20 years. There's 0.1383246 probability that a smoking person who worked in the most dusty place for less than 10 years.

5. Conclusion

Finally, with the best model founded, three predictors variables are included. They are "Employment", "Smoking", and "Workspace". One interesting thing about this dataset is the predictor variables are all categorical. And for the selected ones, they all positively affect the outcome (Byssinosis), there's a very little chance to get Byssinosis if people do not smoke, not working in dusty place, and work less than 10 years.