Investigating Byssinosis

Joseph Gonzalez

12/6/2020

Introduction

Byssinosis(brown lung) is a lung disease that affects people who work in the textile industry. This disease is rare and is associated with inhaling industrial particles, like cotton dust and raw flax. For this project, we investigate Byssinosis' occurrence within a large cotton textile company in North Carolina. We are interested in Byssinosis' relationship with other working categorical characteristics, such as smoking and race. We also construct a logistic model that can predict whether a worker has Byssinosis or not.

Project Format

This project has three main sections:

- **Definitions and Properties:** This section contains short summaries for terms that occur throughout the paper.
- **Exploratory Data Analysis**: This section contains a description of the data's format and contents. We also use contingency tables to identify each field's relationship with Byssinosis(independence or dependence), which may identify the important factors for prediction.
- **Modeling Byssinosis**: In this section, we find the best fitting logistic model. This includes forming models, selecting models based on selection criteria, and evaluating each model's performance(error rate)

Definitions and Formulas:

- **Contingency Table:** A table of values that summarize the relationship between categorical data. Shows the variables' frequencies.

  We use contingency tables to compare Bynossis to the other categorical variables. Our goal is to identify any present relationships.

- **Likelihood Ratio Test For Independence:** Hypothesis test that evaluates whether variables are independent.

$$H_0 : \text{Variables are independent}$$

$$H_a : \text{Variables are not independent}$$

$$x = -2\log\Big(\frac{L(\hat{\pi}_I, y)}{L(\hat{\pi}, y)}\Big)$$

$$= -2\sum_{ij} O_{ij}\log\Big(\frac{E_{ij}}{O_{ij}}\Big)$$

Where O_{ij} are the observed counts and E_{ij} are the expected counts.

We will use the likelihood ratio test for independence with the contingency tables to test whether there is evidence that Bynossis is dependent on the other variables. If the result is significant, this means that there is evidence that the variable may influence the subject's chance of getting Bynossis.

- **Logistic Regression:** Predictive model that describes odds/probability that an event will or will not occur(binary).

$$\log(\frac{\pi}{1-\pi}) = \alpha + B_p X_p$$

Where p is an indicator for the available variables

We will use a logistic model to further identify the variables that are important for predicting byssinosis. It will also reveal how each important variables affects the log odds of a subject contracting byssinosis.

Logistic Model Assumptions:

1. The relationship between the predictors and the response is linear.
2. The response term is i.i.d binomial. Observations are independent.

Exploratory Analysis

Data Description:

The data set consists of 7 columns and 72 rows. The names and values for the 7 columns are:

- **Type of work place:** X_1 →1(most dusty), 2(less dusty), 3(least dusty).
- **Year Employed:** X_2 → <10, 10-19, >=20
- **Smoking:**X_3 → Yes or No in last 5 years
- **Sex:**X_4 → Male, Female
- **Race:** X_5 → White, Other
- **Byssinosis:** Yes, No

Variable Analysis With Contingency Tables

In this section, we will investigate Byssinosis' relationship with the other variables. We will build the observed/expected contingency tables and conduct tests of independence. This section helps us identify the variables that might be important in the logistic model. For the independence tests, we will conduct likelihood ratio tests. Each test is conducted at a $\alpha = 0.05$ significance level.

Type of Work Place:

```
## [1] "Observed Counts"
```

```
##
##        1    2    3
##   0  564 1282 3408
##   1  105   18   42
```

```
## [1] "Expected Counts"
```

```
##              1          2         3
## [1,] 648.63001 1260.41705 3344.9529
## [2,]  20.36999   39.58295  105.0471
```

From the observed and expected contingency tables, we can see that the observed values for the *infected with byssinosis* row(2nd row) are much different from the corresponding values in the expected contingency table. The likelihood test produced a 252.1082 test statistic and a p-value that is very close to 0. Therefore, this evidence suggests that the observed counts are too far away from the expected counts to be due to chance alone and we reject that the type of workplace and Byssinosis are independent. Furthermore, this indicates that the Type of Work Place may be an important variable to use in the logistic model.

Employment:

```
## [1] "Observed Counts"
```

```
##
##      <10 >=20 10-19
##   0 2666 1902   686
##   1   63   76    26
```

```
## [1] "Expected Counts"
```

```
##             <10       >=20     10-19
## [1,] 2645.90626 1917.77302 690.32072
## [2,]   83.09374   60.22698  21.67928
```

From the observed and expected contingency tables, we can see that the observed values for the *infected with byssinosis* row(2nd row) are relatively different from the corresponding values in the expected contingency table. The likelihood test produced a 10.23586 test statistic and a p-value that is 0.005988414. Therefore, this evidence suggests that the distance the observed counts are from the expected counts may not be due to chance alone and we reject that employment years and Byssinosis are independent. Furthermore, the test statistic and p-value show that the employment years variable is valid to use in the logistic model, but may not be as important as the type of workplace.

Smoking:

```
## [1] "Observed Counts"
```

```
##
##       No  Yes
##   0 2190 3064
##   1   40  125
```

```
## [1] "Expected Counts"
```

```
##              No        Yes
## [1,] 2162.10002 3091.89998
## [2,]   67.89998   97.10002
```

From the observed and expected contingency tables, we can see that the observed values for the *infected with byssinosis* row(2nd row) are different from the corresponding values in the expected contingency table. The likelihood test produced a 21.42154 test statistic and a p-value that is 0.000003686064. Therefore, this evidence suggests that the distance the observed counts are from the expected counts may not be due to chance alone and we reject that smoking and Byssinosis are independent. Furthermore, the test statistic and p-value show that the smoking variable is valid to use in the logistic model, but may be more important than employment years.

Sex:

```
## [1] "Observed Counts"
```

```
##
##        F    M
##   0 2466 2788
##   1   37  128
```

```
## [1] "Expected Counts"
```

```
##              F         M
## [1,] 2426.7876 2827.2124
## [2,]   76.2124   88.7876
```

From the observed and expected contingency tables, we can see that the observed values for the *infected with byssinosis* row are different from the corresponding values in the expected contingency table. The likelihood test produced a 41.34426 test statistic and a p-value that is 1.276457e-10. Therefore, this evidence suggests that the distance the observed counts are from the expected counts may not be due to chance alone and we reject that sex and Byssinosis are independent. Furthermore, the test statistic and p-value show that the sex variable is valid to use in the logistic model, but may be more important than employment years and smoking.

Race:

```
## [1] "Observed Counts"
```

```
##
##        O    W
##   0 1830 3424
##   1   73   92
```

```
## [1] "Expected Counts"
```

```
##                O         W
## [1,] 1845.05665 3408.9433
## [2,]   57.94335  107.0567
```

From the observed and expected contingency tables, we can see that the observed values for the *infected with byssinosis* row are slightly different from the corresponding values in the expected contingency table. The likelihood test produced a 6.025885 test statistic and a p-value that is 0.01409757. While this p-value suggests significance under $\alpha = 0.05$, it is larger than the p-values for the other predictors. In my opinion, this evidence suggests that race may not be an important predictor to include in the logistic regression.

Building A Model

In this section, we will form a few models using forwards and backward step-wise selection. All tests will use a $\alpha = 0.05$ significance level. The following sections describe the method to obtain the model, the model's residual deviance, and the model's AIC. We will select the best model based on residual deviance, parameter significance and AIC.

Full Model(No Interactions)

Model 1:

Included Predictors: employment, smoking, sex, race, workspace

$AIC = 135.53$

Residual deviance $= 49.16$

Forward Stepwise(No Interactions)

Model 2:

Included Predictors: workspace, employment, smoking

$AIC = 131.6$

Residual deviance $= 49.26$

Backwards(With interactions)

Model 3:

Included Predictors: employment, smoking, sex, workspace, employment:sex interaction, sex:workspace interaction, smoking:workspace interaction

$AIC = 131.7$

Residual deviance $= 35.29$

After selection, we have 3 models to test. From a first look, model 3 has the best AIC and residual deviance on the training set. We could assume that this may be the best fitting model, but we first check parameter significance and AIC values on the test set.

AIC, Parameter Significance, and diagnostics:

In this section, we investigate the parameters and AIC on the test set. we will take steps to fit the best model with the most significant parameters and reduce overfitting.

Model 1: Full Model With No Interactions

Insignificant Parameters: Sex, Race

AIC = 166.93

Residual deviance = 44.605

Goodness of Fit Deviance Test: 0.8837523

Dispersion parameter: 0.6852865

From the output, it appears that there are a few insignificant parameters in the full model with no interactions. If we conduct hypothesis testing, we will find that the "better" model will eventually look like model 2. Therefore, we can remove model 1 from consideration.

Model 2: Workspace, Employment, Smoking

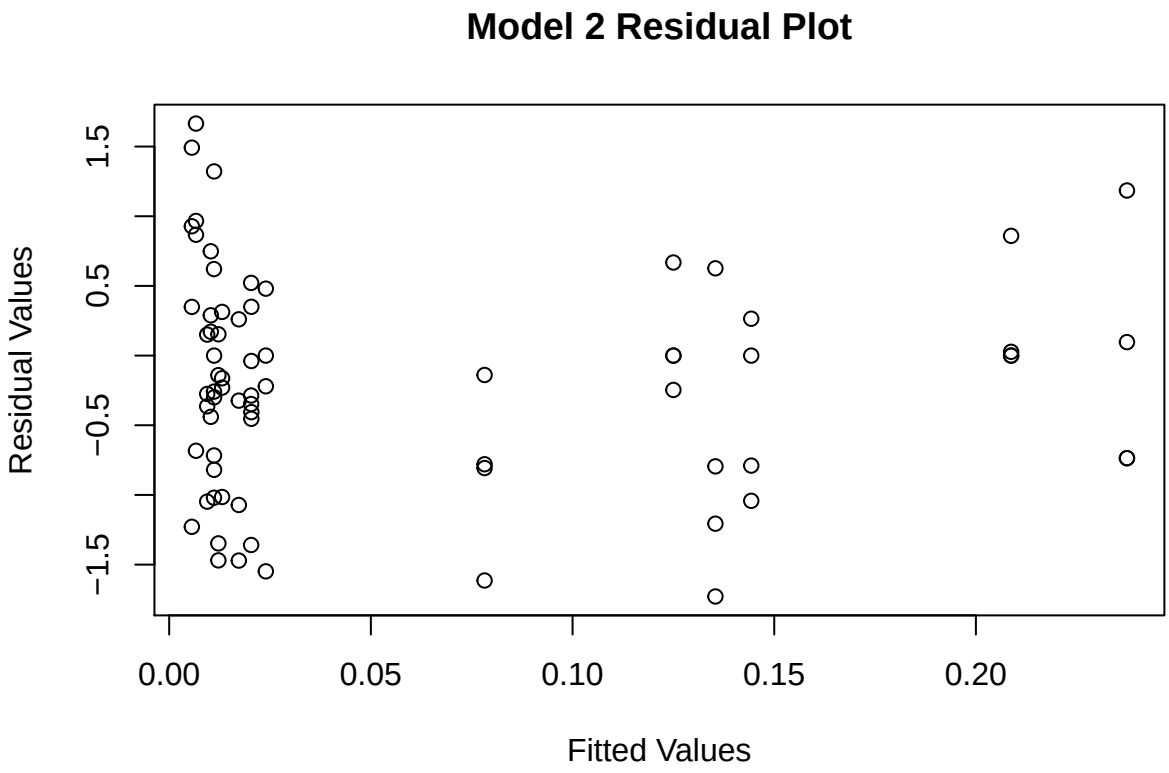
Insignificant Parameters: None

AIC = 163.76

Residual deviance = 45.434

Goodness of Fit Deviance Test: 0.9027212

Dispersion parameter: 0.6652218



Compared to Model 1 and Model 3, Model 2 has the most significant parameters and a relatively low AIC value. The Goodness of Fit Deviance Test and dispersion parameter shows that the model fits well with no evidence for breaks in assumptions. All residuals are less than 2 and more than negative 2.

Model 3: Employment, Smoking, Sex, Workspace

Insignificant Parameters: Various

$AIC = 131.67$

Residual deviance $= 35.294$

Goodness of Fit Deviance Test: 0.9426631

Dispersion parameter: 1

Compared to Model 1 and Model 2, Model 3 has the lowest AIC value and smallest residual deviance. The Goodness of Fit Deviance Test and Dispersion parameter shows that the model fits well with no evidence for breaks in assumptions. However, there are many insignificant parameters in the model. This model may overfit the data.

Ultimately, Model 2 seems to be the "best" model for the byssinosis data. This project's goal is to investigate the relationship between the disease and the predictors. Therefore, we will use model 2 to further analyze these relationships. If the goal is prediction, we should investigate larger models in the selection process.

Final Model and Describing Coefficients:

The final model is:

$$\log(\frac{\hat{\pi}}{1-\hat{\pi}}) = -2.4546 - 2.5493X_{1,2} - 2.7175X_{1,3} + 0.6728X_{2,>=20} + 0.5060X_{2,10-19} + 0.6210X_{3,\text{yes}}$$

Parameter Interpretation

Baseline Case: The baseline case is when the workplace is most dusty, less than 10 years employed, and no smoking in the last 5 years. The estimated log-odds for the baseline case is -2.4546. The odds of byssinosis for the baseline case is 0.0859 times those of the other.

Workspace:

All other things equal, the log-odds ratio of byssinosis for less dusty workspace vs most dusty workspace is -2.5493. The estimated odds of byssinosis for a less dusty workspace are 0.0781 times those of most dusty workspaces, holding all other variables constant.

All other things equal, the log-odds ratio of byssinosis for least dusty workspace vs most dusty workspace is -2.7175. The estimated odds of byssinosis for a least dusty workspace are 0.066 times those of most dusty workspaces, holding all other variables constant.

Employment:

All other things equal, the log-odds ratio of byssinosis for employment more than 20 years vs employment less than 10 years is 0.6728. The estimated odds of byssinosis for employment more than 20 years are 1.959 times those of employment less than 10 years, holding all other variables constant.

All other things equal, the log-odds ratio of byssinosis for employment between 10 and 19 years vs employment less than 10 years is 0.5060. The estimated odds of byssinosis for employment between 10 and 19 years are 1.659 times those of employment less than 10 years, holding all other variables constant.

Smoking:

All other things equal, the log-odds ratio of byssinosis for smoking in the last 5 years vs not smoking in the last 5 years is 0.6728. The estimated odds of byssinosis for employment between 10 and 19 years are 1.960 times those of not smoking in the last 5 years, holding all other variables constant.

Final Statement

From this analysis, it appears that workspace, employment, and smoking provide the most information or best fit for a worker contracting byssinosis. We found that the selection of these variables matches their significance in the independence test with byssinosis. This study shows how important statistical analysis can be to correctly identify components that influence diseases.

Code Appendix

```
knitr::opts_chunk$set(echo = TRUE)

library(dplyr)
library(magrittr)
library(knitr)

Bynoss_data = read.csv("Byssinosis.csv", header = TRUE)
attach(Bynoss_data)
names(Bynoss_data)
sapply(Bynoss_data, class)
i= 1
Byss_data_exp = data.frame()

repeat{
  binos_count = Bynoss_data$Byssinosis[i]
  if(binos_count==0){
    temp_data = data.frame()
  }else{temp_data = data.frame(Bynoss_data[i, -c(6, 7)],
                               Byss = rep(1,binos_count))}
  Byss_data_exp = rbind(Byss_data_exp , temp_data)
  i = i + 1
  if(i > length(Bynoss_data$Byssinosis)){break}
}


i=1
repeat{
  binos_count = Bynoss_data$Non.Byssinosis[i]
  if(binos_count==0){
    temp_data = data.frame()
  }else{temp_data = data.frame(Bynoss_data[i, -c(6, 7)],
                               Byss = rep(0,binos_count))}
  Byss_data_exp = rbind(Byss_data_exp, temp_data)
  i = i + 1
  if(i > length(Bynoss_data$Byssinosis)){break}
}
sum(Bynoss_data$Non.Byssinosis)
detach(Bynoss_data)
names(Byss_data_exp)
attach(Byss_data_exp)
#Observed Values:
obs_wp = as.matrix(table(Byss_data_exp$Byss, Byss_data_exp$Workspace))
print("Observed Counts")
print(obs_wp)
#Expected Values:
Exp_wp = rowSums(obs_wp) %*% t(colSums(obs_wp))/sum(obs_wp)
print("Expected Counts")
print(Exp_wp)

#Likelihood ratio test:
lrt_wp = -2*sum(obs_wp*log(Exp_wp/obs_wp))
```

```r
#p-value
LRpVal = 1-pchisq(lrt_wp, (2-1)*(3-1))
#Observed Values:
obs_e = as.matrix(table(Byss_data_exp$Byss, Byss_data_exp$Employment))
print("Observed Counts")
print(obs_e)

#Expected Values:
Exp_e = rowSums(obs_e) %*% t(colSums(obs_e))/sum(obs_e)
print("Expected Counts")
print(Exp_e)

#Likelihood ratio test:
lrt_e = -2*sum(obs_e*log(Exp_e/obs_e))

#p-value
LRpVal = 1-pchisq(lrt_e, (2-1)*(3-1))
#Observed:
obs_smok = as.matrix(table(Byss_data_exp$Byss, Byss_data_exp$Smoking))
print("Observed Counts")
print(obs_smok)

#Expected:
Exp_smok = rowSums(obs_smok) %*% t(colSums(obs_smok))/sum(obs_smok)
print("Expected Counts")
print(Exp_smok)

#Likelihood ratio test:
lrt_smok = -2*sum(obs_smok*log(Exp_smok/obs_smok))

#p-value
LRpVal = 1-pchisq(lrt_smok, (2-1)*(2-1))


#Observed:
obs_sex = as.matrix(table(Byss_data_exp$Byss, Byss_data_exp$Sex))
print("Observed Counts")
print(obs_sex)

#Expected:
Exp_sex  = rowSums(obs_sex) %*% t(colSums(obs_sex))/sum(obs_sex)
print("Expected Counts")
print(Exp_sex)

#Likelihood ratio test:
lrt_sex = -2*sum(obs_sex*log(Exp_sex/obs_sex))

#p-value
LRpVal = 1-pchisq(lrt_sex, (2-1)*(2-1))

#Observed:
obs_race = as.matrix(table(Byss_data_exp$Byss, Byss_data_exp$Race))
print("Observed Counts")
```

```r
print(obs_race)

#Expected:
Exp_race = rowSums(obs_race) %*% t(colSums(obs_race))/sum(obs_race)
print("Expected Counts")
print(Exp_race)

#Likelihood ratio test:
lrt_race = -2*sum(obs_race*log(Exp_race/obs_race))

#p-value
LRpVal = 1-pchisq(lrt_race, (2-1)*(2-1))
#Divide the long data set into a train and test set
n = dim(Byss_data_exp)[1]
set.seed(10)
train_setL = sample(c(TRUE, FALSE), n,replace = TRUE)
train_dataL = Byss_data_exp[train_setL,]
test_dataL = Byss_data_exp[-train_setL,]

#Divide the long data set into a train and test set
train_dataW = aggregate(cbind(Byss = Byss, notByss=1-Byss, total=1)~
                          Employment + Smoking +
                           Sex + Race + Workspace, FUN=sum,
                        data = train_dataL, drop = FALSE)
train_dataW$Byss[is.na(train_dataW$Bys)] = 0
train_dataW$notByss[is.na(train_dataW$notByss)] = 0
train_dataW$total[is.na(train_dataW$total)] = 0

test_dataW = aggregate(cbind(Byss = Byss, notByss=1-Byss, total=1)~
                         Employment + Smoking +
                           Sex + Race + Workspace, FUN=sum,
                       data = test_dataL, drop = FALSE)
test_dataW$Byss[is.na(test_dataW$Bys)] = 0
test_dataW$notByss[is.na(test_dataW$notByss)] = 0
test_dataW$total[is.na(test_dataW$total)] = 0

train_dataW$Workspace = as.factor(train_dataW$Workspace)
test_dataW$Workspace = as.factor(test_dataW$Workspace)
model_full = glm(cbind(Byss, notByss) ~.-total,
                 family = binomial, data = train_dataW)
summary(model_full)
#Forward step-wise on the train:
#step(glm(cbind(Byss, notByss)~1, family = binomial, data = train_dataW),
          #scope = ~Employment*Smoking*Sex*Race*Workspace,
          #direction = "forward")
#4 vars, AIC = 146.2, Null Deviance = 204, Residual Deviance= 65.83
#model_f1 = glm(formula = cbind(Byss, notByss) ~ Workspace + Employment +
   # Smoking + Workspace:Smoking, family = binomial, data = train_dataW)

#step(glm(cbind(Byss, notByss)~1, family = binomial, data = train_dataW),
          #scope = ~Employment*Smoking*Sex*Race*Workspace,
          #direction = "both")
#4 vars, AIC = 146.2, Null Deviance = 204, Residual Deviance= 65.83
```

```r
#Same as forward!

step(glm(cbind(Byss, notByss)~1, family = binomial, data = train_dataW),
          scope = ~Employment+Smoking+Sex+Race+Workspace,
         direction = "forward")
model_f1 = glm(formula = cbind(Byss, notByss) ~ Workspace + Employment +
    Smoking, family = binomial, data = train_dataW)
#Same as above

#step(glm(cbind(Byss, notByss)~1, family = binomial, data = train_dataW),
          #scope = ~Employment+Smoking+Sex+Race+Workspace,
          #direction = "both")
#Same model and specs as above
#Backward Selection
step(glm(cbind(Byss, notByss)~.^2, family = binomial, data = train_dataW[,-8]),
          direction = "backward")
#8 vars, AIC = 140.4, Null Deviance = 204, Residual Deviance = 52.06
model_b1 = glm(formula = cbind(Byss, notByss) ~ Employment + Smoking + Sex +
    Workspace + Employment:Sex + Smoking:Workspace + Sex:Workspace,
    family = binomial, data = train_dataW[, -8])

#step(glm(cbind(Byss, notByss)~.^2, family = binomial, data = train_dataW[,-8]),
        # direction = "both")
#Same model as above



#step(glm(cbind(Byss, notByss)~., family = binomial, data = train_dataW[,-8]),
          #direction = "backward")
#4 vars, AIC = 146.2, Null Deviance = 204, Residual Deviance=  65.83

#model_b3 =  glm(formula = cbind(Byss, notByss) ~ Employment
#+ Smoking + Workspace,
    #family = binomial, data = train_dataW[, -8])

#step(glm(cbind(Byss, notByss)~., family = binomial, data = train_dataW[,-8]),
         # direction = "both")
#Same as above
#Model 1(full model):
model_full = glm(cbind(Byss, notByss) ~.-total, family = binomial,
                 data = test_dataW)
summary(model_full)
1-pchisq(44.605,57)
model_full = glm(cbind(Byss, notByss) ~.-total, family = quasibinomial,
                 data = test_dataW)
summary(model_full)
#Full model turns to model 2

#Model 2:
model_f1_test = glm(formula = cbind(Byss, notByss) ~ Workspace + Employment +
    Smoking, family = binomial, data = test_dataW)

summary(model_f1_test)
```

```r
1-pchisq(45.434,59)

#model_f1_test = glm(cbind(Byss, notByss) ~Workspace + Employment +
    #Smoking, family = quasibinomial, data = test_dataW)

#summary(model_f1_test)

#Model 3:
model_b1_test =glm(formula = cbind(Byss, notByss) ~ Employment + Smoking + Sex +
    Workspace + Employment:Sex + Smoking:Workspace + Sex:Workspace,
    family = binomial, data = train_dataW[, -8])
summary(model_b1_test)

1-pchisq(35.294,50)

#model_b1_test = glm(formula = cbind(Byss, notByss) ~
              #Employment + Smoking + Sex +
    #Workspace + Employment:Sex + Smoking:Workspace + Sex:Workspace,
    #family = binomial, data = train_dataW[, -8])
#summary(model_b1_test)

ry <- residuals(model_f1_test, type="deviance")
rx <- fitted.values(model_f1_test)
rxy <- cbind(fitted=rx, residual=ry)
plot(rxy, main = "Model 2 Residual Plot", xlab = "Fitted Values", ylab = "Residual Values")
final_model =  glm(cbind(Byssinosis, Non.Byssinosis) ~ as.factor(Workspace) + Employment +
                     Smoking, family = quasibinomial, data = Bynoss_data)
summary(final_model)
```