STA 138 Final Project

Brian Le, 916111559, (no collaborator)

2023-03-24

Introduction and Dataset

Byssinosis is a type of respiratory disease and we are interested in identifying if it should be a concern in the workplace. Data was collected from cotton textile company employees in North Carolina, where they are exposed to cotton dust. The dataset we are working with has 72 rows with 7 variables. The data is aggregated based on byssinosis cases for a total sample size of 5417.

- Employment: takes one of 3 values - "<10", "10-19", ">=20", indicating the length of an employee's employment
- Smoking: a binary variable which takes a "No" or "Yes" value and indicates whether an employee has smoked in the last 5 years
- Sex: "M" or "F" depending on the sex of the employee
- Race: "W" or "O" for white and other race
- Workspace: 1, 2, or 3, as most, less, and least dusty
- Byssinosis: Count of byssinosis cases for a combination of the above variables
- Non.Byssinosis: Same as above but for non-byssinosis cases

Exploring the Data

First of all, what proportion of employees have Byssinosis?

```
sum(byssinosis$Byssinosis)/(sum(byssinosis$Byssinosis)+sum(byssinosis$Non.Byssinosis))
```

```
## [1] 0.03027506
```

164 out of 5417 employees have Byssinosis. This may seem low, but since Byssinosis can be a serious condition, we need to look further.

Maybe some groups of employees are more susceptible to Byssinosis? Since the data is aggregated by unique groups, we can look at the range of Byssinosis cases to roughly see how large the difference is between the group with the lowest number of cases and the group with the lowest number of cases.

```
range(byssinosis$Byssinosis)
```

```
## [1]  0 31
```

0 to 31 cases seems like a large difference. A quick glance at the data suggests that some groups may be more likely to contract Byssinosis. It's common knowledge that smoking is bad for your lungs, so let's see if smokers are more likely to get Byssinosis. We'll take the sample odds ratio for Byssinosis for smokers vs non-smokers.

```
##           byssinosis non.byssinosis
## smoker           125           3064
## nonsmoker         39           2189
```

```
#calculating odds ratio (a/c)/(b/d)
smoker_OR = (smoke_table[1,1]/smoke_table[1,2])/(smoke_table[2,1]/smoke_table[2,2])
smoker_OR
```

```
## [1] 2.289826
```

The odds ratio of byssinosis for smokers vs non-smokers is 2.289826, which means that smokers are a bit more than twice as likely to get Byssinosis. It's reasonable to conclude that smoking has a contribution to Byssinosis. Are a significant proportion of employees smokers?

```
#calculating proportion of smokers
num_employees = sum(smoke_table)
num_smokers = smoke_table[1,1] + smoke_table[1,2]
num_smokers/num_employees
```

```
## [1] 0.5887022
```

About 58.9% of employees smoke, so we're looking at just over half of the company's employees.

What about length of employment? Does the time an employee accumulates at work contribute to Byssinosis?

```
##       byssinosis non.byssinosis
## <10           63           2666
## 10-19         26            686
## >=20          75           1901
```

We can then calculate the odds ratio for all the combinations of employment groups.

```
##               odds ratio
## >=20 vs 10-19   1.040950
## 10-19 vs <10    1.603869
## >=20 vs <10     1.669547
```

Basically, this all comes to tell us that going from <10 years of employment to 10-19 years of employment increases the likelihood of byssinosis by about 1.6 times, and further going to >20 years of employment does not further increase the likelihood of byssinosis by a significant amount. It's important to note here that the groups of employment years are fairly wide. We cannot say that at exactly 10 years of employment is where the likelihood for byssinosis increases. It's possible that there is a large increase in likelihood for byssinosis earlier in employment, but all we can do here is compare the groups we have.

How does dustiness of workspace relate to byssinosis?

```
##            byssinosis non.byssinosis
## Most dust         105            564
## Less dust          18           1282
## Least dust         41           3407
```

We can again calculate the odds ratios between the groups.

```
##                     odds ratio
## most vs less dust    13.259456
## less vs least dust    1.166736
## most vs least dust   15.470291
```

Based on these odds ratios, it seems that workspaces with most dust are much more likely to have workers get byssinosis than less or least dusty workspaces. Employees working in workspaces considered most dusty are about 15.5 times and 13.3 times more likely to get byssinosis than employees working in least and less dusty workspaces respectively. It looks like there is a slight difference between less and least dusty workspaces, as employees in less dusty workspaces are 1.17 times more likely to get byssinosis than employees in least dusty workspaces.

Logistic Regression Model

Since byssinosis is a binary variable in the data, a logistic regression model is a good way to model having byssinosis in terms of all the predictors.

A logistic regression model is:

$log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + .. + \beta_p x_p$

Starting with a model with all the predictors:

```
model1 = glm(cbind(Byssinosis, Non.Byssinosis) ~ Employment + Smoking + Sex +
               Race + as.factor(Workspace), family = "binomial", data = byssinosis)
summary(model1)
```

```
##
## Call:
## glm(formula = cbind(Byssinosis, Non.Byssinosis) ~ Employment +
##     Smoking + Sex + Race + as.factor(Workspace), family = "binomial",
##     data = byssinosis)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5076  -0.8147  -0.1998   0.1887   1.5709
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -2.3411     0.2651  -8.832  < 2e-16 ***
## Employment>=20          0.7444     0.2169   3.432 0.000600 ***
## Employment10-19         0.5678     0.2619   2.168 0.030177 *
## SmokingYes              0.6709     0.1961   3.421 0.000625 ***
## SexM                   -0.1503     0.2296  -0.655 0.512777
```

```
## RaceW                   -0.1176     0.2077  -0.566 0.571270
## as.factor(Workspace)2   -2.5904     0.2926  -8.854  < 2e-16 ***
## as.factor(Workspace)3   -2.7617     0.2172 -12.713  < 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: 324.114  on 64  degrees of freedom
## Residual deviance:  42.679  on 57  degrees of freedom
## AIC: 164.98
##
## Number of Fisher Scoring iterations: 5
```

Looking at the p-values in the summary, Sex and Race predictors have high p-values so they may not contribute a significant amount to the log-odds of getting byssinosis.

Using stepwise selection using AIC, we can narrow down the predictors that don't contribute a large amount to byssinosis.

```
model2$coefficients
```

```
##           (Intercept)        Employment>=20       Employment10-19
##            -2.4684777             0.6595540             0.5054502
##            SmokingYes as.factor(Workspace)2 as.factor(Workspace)3
##             0.6463926            -2.5467780            -2.7392668
```

As we can see here, Sex and Race were dropped from the model. Looking at these coefficients, the workspace predictors stand out as having quite large (negative) coefficients. The coefficients are log-odds ratios that compare the variable to a baseline, which in the case of employment is being employed less than 10 years. Taking the exponential of these makes them easier to interpret.

- Employment (10-19): $exp(0.5055) = 1.657814$, meaning that having been employed 10-19 years will increase risk of byssinosis.
- Employment (>=20): $exp(0.6596) = 1.934019$, meaning that having been employed >=20 years will increase risk of byssinosis even further.
- Smoking: $exp(0.6464) = 1.908657$, meaning that being a smoker will contribute to risk of byssinosis.
- Less dusty workspaces: $exp(-2.5468) = 0.07833193$, meaning that byssinosis risk for less dusty workspaces is quite a bit lower than "most dusty" workspaces.
- Least dusty workspaces: $exp(-2.7393) = 0.06461556$, meaning that byssinosis risk for least dusty workspaces is quite a bit lower than "most dusty" workspaces.

Conclusion

Based on the data analysis, we can conclude that byssinosis is associated with employment length, smoking, and workspace. Workspace was the predictor that had the most effect on byssinosis due to the large coefficients in the logistic regression model. This is followed by employment length and if the employee is a smoker. Based on the data, we can see that the longer an employee stays, the more likely they are to get byssinosis. The same can be said for if they are a smoker. Notable issues with the data include the classification of employment and workspace. Employment was divided into three 10 year groups, which is no

small amount of time, and may impact the performance of the model. The same can be said for workspace, which was divided into three vague groups of “most”, “less”, and “least”. Although some association with the predictors and byssinosis are shown here, more data should be collected for further analysis. Suggestions include providing the exact number of years employees are employed and providing exact measurements of dust in the air.

Code Appendix

```
library(dplyr)
byssinosis = read.csv("Byssinosis.csv")
sum(byssinosis$Byssinosis)/(sum(byssinosis$Byssinosis)+sum(byssinosis$Non.Byssinosis))
range(byssinosis$Byssinosis)
smoke_table = matrix(nrow = 2, ncol = 2,
                     dimnames = list(c("smoker", "nonsmoker"),c("byssinosis", "non.byssinosis")))
smoke_table[1,1] = sum((byssinosis %>% filter(Smoking == "Yes"))$Byssinosis)
smoke_table[1,2] = sum((byssinosis %>% filter(Smoking == "Yes"))$Non.Byssinosis)
smoke_table[2,1] = sum((byssinosis %>% filter(Smoking == "No"))$Byssinosis)
smoke_table[2,2] = sum((byssinosis %>% filter(Smoking == "No"))$Non.Byssinosis)
smoke_table
#calculating odds ratio (a/c)/(b/d)
smoker_OR = (smoke_table[1,1]/smoke_table[1,2])/(smoke_table[2,1]/smoke_table[2,2])
smoker_OR
#calculating proportion of smokers
num_employees = sum(smoke_table)
num_smokers = smoke_table[1,1] + smoke_table[1,2]
num_smokers/num_employees
employ_table = matrix(nrow = 3, ncol = 2,
                      dimnames = list(c("<10", "10-19", ">=20"),c("byssinosis", "non.byssinosis")))
employ_table[1,1] = sum((byssinosis %>% filter(Employment == "<10"))$Byssinosis)
employ_table[1,2] = sum((byssinosis %>% filter(Employment == "<10"))$Non.Byssinosis)
employ_table[2,1] = sum((byssinosis %>% filter(Employment == "10-19"))$Byssinosis)
employ_table[2,2] = sum((byssinosis %>% filter(Employment == "10-19"))$Non.Byssinosis)
employ_table[3,1] = sum((byssinosis %>% filter(Employment == ">=20"))$Byssinosis)
employ_table[3,2] = sum((byssinosis %>% filter(Employment == ">=20"))$Non.Byssinosis)
employ_table
#odds ratios of byssinosis for employment category
OR_table_employ = matrix(nrow = 3, ncol = 1,
                    dimnames = list(c(">=20 vs 10-19", "10-19 vs <10", ">=20 vs <10"), c("odds ratio")))
OR_table_employ[1,1] = (employ_table[3,1]/employ_table[3,2])/(employ_table[2,1]/employ_table[2,2])
OR_table_employ[2,1] = (employ_table[2,1]/employ_table[2,2])/(employ_table[1,1]/employ_table[1,2])
OR_table_employ[3,1] = (employ_table[3,1]/employ_table[3,2])/(employ_table[1,1]/employ_table[1,2])
OR_table_employ
#odds ratios of byssinosis for workspace category
dust_table = matrix(nrow = 3, ncol = 2,
                     dimnames = list(c("Most dust", "Less dust", "Least dust"),c("byssinosis", "non.byss
dust_table[1,1] = sum((byssinosis %>% filter(Workspace == "1"))$Byssinosis)
dust_table[1,2] = sum((byssinosis %>% filter(Workspace == "1"))$Non.Byssinosis)
dust_table[2,1] = sum((byssinosis %>% filter(Workspace == "2"))$Byssinosis)
dust_table[2,2] = sum((byssinosis %>% filter(Workspace == "2"))$Non.Byssinosis)
dust_table[3,1] = sum((byssinosis %>% filter(Workspace == "3"))$Byssinosis)
dust_table[3,2] = sum((byssinosis %>% filter(Workspace == "3"))$Non.Byssinosis)
dust_table
OR_table_dust = matrix(nrow = 3, ncol = 1,
                    dimnames = list(c("most vs less dust", "less vs least dust", "most vs least dust"), c
OR_table_dust[1,1] = (dust_table[1,1]/dust_table[1,2])/(dust_table[2,1]/dust_table[2,2])
OR_table_dust[2,1] = (dust_table[2,1]/dust_table[2,2])/(dust_table[3,1]/dust_table[3,2])
OR_table_dust[3,1] = (dust_table[1,1]/dust_table[1,2])/(dust_table[3,1]/dust_table[3,2])
OR_table_dust
model1 = glm(cbind(Byssinosis, Non.Byssinosis) ~ Employment + Smoking + Sex +
```

```
              Race + as.factor(Workspace), family = "binomial", data = byssinosis)
summary(model1)
model2 = step(model1)
model2$coefficients
```