¶ Final Project:

STA 138 Analysis of Categorical Data

Ryandeep Chawla, Mick Hashimoto, Justin Luong, Shelly Sagy, and Hanah Shih

December 09, 2022

I. Abstract

II. Introduction

The dataset that we will be analyzing for this project was sampled in 1973 from a large cotton textile company in North Carolina. The purpose of this study is to investigate the prevalence of Byssinosis, which is 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]

From this data we will investigate relationships between this disease as the dependent variable and smoking status, sex, race, length of employment, and dustiness of workplace as our predictor variables. Essentially this will determine which, if any, of these predictor variables has any degree of influence on Byssinosis infection in workers exposed to cotton dust.

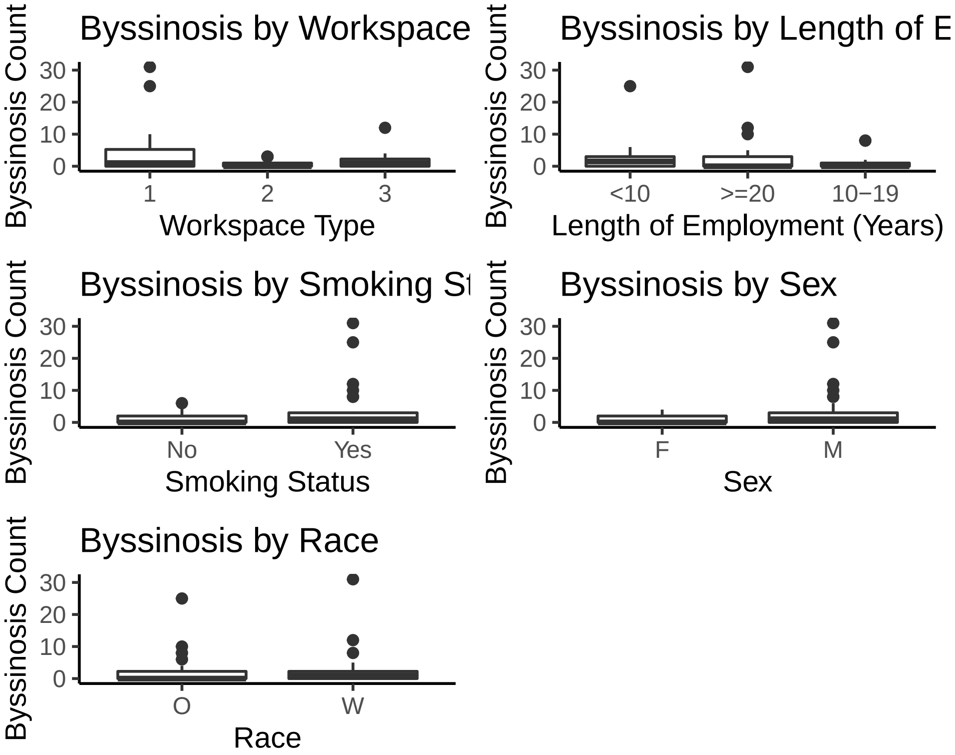
This question might be of interest to any companies that require their workers to be exposed to cotton dust, because this data may suggest that they may or may not need to take precautions against Byssinosis infection in their workspaces. Learning about the demographics of workers who are most likely to be affected by Byssinosis could also allow for pharmaceutical companies and Biotech companies to develop treatments that are most suitable for certain demographics, and focus future studies on them. The results of this study will likely lead to further research on why a certain demographic is more or less affected by Dyssinosis, with this data simply helping to narrow the scope of future studies.

III. Methods

```
##   Employment Smoking Sex Race Workspace Byssinosis Non.Byssinosis
## 1        <10     Yes   M    W         1          3             37
## 2        <10     Yes   M    O         1         25            139
## 3        <10     Yes   F    W         1          0              5
## 4        <10     Yes   F    O         1          2             22
## 5        <10      No   M    W         1          0             16
## 6        <10      No   M    O         1          6             75
```

IV. Exploratory Analysis

Box Plots



We conducted a preliminary analysis of the distribution of the data through creating box plots of byssinosis prevalence by various categorical variables (type of work place, employment length, smoking status, sex, and race). These plots will be used as as descriptive summary to aid our understanding and support our conclusions.

The distribution of the data appears to be relatively even besides for few exceptions. Those that have been employed for 10 to 19 years appear to have a smaller risk of getting byssinosis. Although, work spaces with the most dust appear to have more cases of byssinosis, it is the middle level of dustiness that has the least subjects with byssinosis. Males also have a slightly larger representation of byssinosis cases. There also are many more outliers for smokers and males, while their respective counterparts have little to none. There appears to be not much of a byssinosis distributional difference between ethnicities. However, there is a larger mean for Caucasian individuals and subjects with less than 10 years of experience. This suggests that there may be slight significant differences in the distribution of subjects more prone to byssinosis.

Odds Ratio

```
## [1] 0.02308538
```

```
## [1] 0.03651685
```

```
## [1] 0.03795547
```

```
## [1] 2.239268
## [1] 2.948311
```

```
## [1] 0.6750805
```

```
## [1] 0.1569507
```

```
## [1] 0.01384615
```

```
## [1] 0.01189095
```

The odds ratio of smoking vs. not smoking is 2.239 so smokers are more 2.239 times more likely to have Byssinosis than nonsmokers.

The odds ratio of Males vs. Females having Byssinosis is 2.948 so males are 2.948 times more likely to have Byssinosis than females. The odds ratio of being White vs. not is 0.675, meaning white people are 0.675 times more likely to get Byssinosis. There is a strong relationship between whether or not people smoke and gender but this is not shown in race.

The odds of having Byssinosis based on employment length is [0.023, 0.037, 0.038] respectively. Employment length lead to a higher odds of having byssinosis, but there is a very small difference.

The odds of having Byssinosis based on workspace is [0.016, 0.014, 0.012] ordered from Most Dusty to Least Dusty. This shows that dustier workspaces lead to higher odds of Byssinosis, but there is also a very small difference.

Model Investigation

Additive Model

```
##     Employment        Smoking            Sex           Race      Workspace
##           TRUE           TRUE           TRUE           TRUE           TRUE
##     Byssinosis Non.Byssinosis
##           TRUE           TRUE
```

Remark: We use cbind() here for combining Byssinonsis and Non.Byssinosis because the representation of the data is in wide data format, which is where the y variable is separated into two.

```
##
## Call:
## glm(formula = cbind(Byssinosis, Non.Byssinosis) ~ ., family = "binomial",
##     data = b)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.27443  -0.40948  -0.00353   0.46145   1.91690
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -1.2692     0.2239  -5.669 1.43e-08 ***
## Employment>=20         0.6206     0.1954   3.176  0.00150 **
## Employment10-19        0.5812     0.2119   2.743  0.00609 **
## SmokingYes             0.2205     0.1622   1.360  0.17388
## SexM                  -0.1983     0.1760  -1.127  0.25976
## RaceW                 -0.3525     0.1741  -2.025  0.04288 *
## WorkspaceLess Dusty   -1.2827     0.2184  -5.872 4.31e-09 ***
## WorkspaceLeast Dusty  -1.3701     0.1920  -7.138 9.50e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 113.204  on 71  degrees of freedom
## Residual deviance:  29.239  on 64  degrees of freedom
## AIC: 218.97
##
## Number of Fisher Scoring iterations: 4
```

From the summary of our additive model, we can interpret what the model represents. First, we observe that smoking and sex in the additive model have a p-value which is greater than the 0.05 level of significance, suggesting that they are not statistically significant. For the significant variables, we see that the most significant predictor variables are workspace least dusty and workspace less dusty, suggesting that there is a strong association of the type of workspace and if they had byssinosis or not. A negative coefficient for these two predictor variables suggest that all other variables equal, we can see that if the type of workspace is less dusty/least dusty they are less likely to have byssinosis. We also see that employment is statistically significant as well, with a positive coefficient, hence we can infer that as the number of years of employment increases there is a more likely chance that a person has byssinosis. Race is also another significant variable that is observed which also has a negative coefficient.

ANOVA

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: cbind(Byssinosis, Non.Byssinosis)
##
## Terms added sequentially (first to last)
##
##
##            Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                          71    113.204
## Employment  2    8.760        69    104.444 0.0125245 *
## Smoking     1    2.678        68    101.766 0.1017239
## Sex         1    2.212        67     99.554 0.1369395
## Race        1   12.726        66     86.828 0.0003606 ***
## Workspace   2   57.589        64     29.239 3.125e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

By looking at the table above we can see that the deviances increase as each variable is added, we see the largest increase in the deviance when the Workspace predictor variable is added and it also significantly reduces the residual deviance. We see the second largest increase in the deviances when Race is also added to the model. We can also see that smoking and sex decrease the deviances and also have a high p-value, suggesting that the model without these variables explains more or less the same amount of variation.

Interaction Model

```
##
## Call:
## glm(formula = cbind(Byssinosis, Non.Byssinosis) ~ .^2, family = binomial(),
##     data = b)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.32525  -0.28202   0.02489   0.37120   1.46877
##
## Coefficients:
##                                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)                          -1.494913   0.416895  -3.586 0.000336 ***
## Employment>=20                        0.681730   0.568032   1.200 0.230077
## Employment10-19                       1.052178   0.606102   1.736 0.082568 .
## SmokingYes                            0.152742   0.435308   0.351 0.725676
## SexM                                 -0.099485   0.429962  -0.231 0.817019
## RaceW                                -0.423018   0.504755  -0.838 0.401994
## WorkspaceLess Dusty                  -0.624078   0.490349  -1.273 0.203116
## WorkspaceLeast Dusty                 -0.909192   0.470160  -1.934 0.053139 .
## Employment>=20:SmokingYes             0.184867   0.403404   0.458 0.646760
## Employment10-19:SmokingYes           -0.025425   0.446863  -0.057 0.954627
## Employment>=20:SexM                  -0.024307   0.442233  -0.055 0.956167
## Employment10-19:SexM                 -0.214683   0.491610  -0.437 0.662333
## Employment>=20:RaceW                 -0.227668   0.414695  -0.549 0.583005
## Employment10-19:RaceW                 0.011424   0.452254   0.025 0.979848
## Employment>=20:WorkspaceLess Dusty    0.290516   0.574060   0.506 0.612806
## Employment10-19:WorkspaceLess Dusty   0.004005   0.598043   0.007 0.994657
## Employment>=20:WorkspaceLeast Dusty  -0.171802   0.465088  -0.369 0.711833
## Employment10-19:WorkspaceLeast Dusty -0.928899   0.545014  -1.704 0.088314 .
## SmokingYes:SexM                       0.004233   0.363080   0.012 0.990697
## SmokingYes:RaceW                      0.321670   0.359691   0.894 0.371163
## SmokingYes:WorkspaceLess Dusty       -0.558346   0.452017  -1.235 0.216744
## SmokingYes:WorkspaceLeast Dusty      -0.090863   0.404421  -0.225 0.822233
## SexM:RaceW                            0.140224   0.402907   0.348 0.727816
## SexM:WorkspaceLess Dusty             -0.164101   0.488109  -0.336 0.736722
## SexM:WorkspaceLeast Dusty            -0.192677   0.447243  -0.431 0.666606
## RaceW:WorkspaceLess Dusty            -0.565935   0.501657  -1.128 0.259264
## RaceW:WorkspaceLeast Dusty           -0.020978   0.412545  -0.051 0.959444
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 113.20  on 71  degrees of freedom
## Residual deviance:  19.09  on 45  degrees of freedom
## AIC: 246.82
##
## Number of Fisher Scoring iterations: 4
```

The model above is an interaction model where the variables are squared. None of the variables are statistically significant as the p-values are greater than 0.05, so we will not consider this model.

ANOVA

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: cbind(Byssinosis, Non.Byssinosis)
##
## Terms added sequentially (first to last)
##
##
##                      Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                                   71    113.204
## Employment            2    8.760        69    104.444 0.0125245 *
## Smoking               1    2.678        68    101.766 0.1017239
## Sex                   1    2.212        67     99.554 0.1369395
## Race                  1   12.726        66     86.828 0.0003606 ***
## Workspace             2   57.589        64     29.239 3.125e-13 ***
## Employment:Smoking    2    0.801        62     28.438 0.6700500
## Employment:Sex        2    0.025        60     28.414 0.9877577
## Employment:Race       2    0.170        58     28.244 0.9185325
## Employment:Workspace  4    3.826        54     24.418 0.4300415
## Smoking:Sex           1    0.132        53     24.285 0.7159380
## Smoking:Race          1    0.917        52     23.368 0.3381663
## Smoking:Workspace     2    2.143        50     21.225 0.3424946
## Sex:Race              1    0.357        49     20.868 0.5503276
## Sex:Workspace         2    0.289        47     20.579 0.8655883
## Race:Workspace        2    1.489        45     19.090 0.4749839
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Although the model does not have statistically significant variables, the ANOVA table has three statistically significant deviances with p-values less than 0.05: Employment, Race, and Workspace. This suggests that these deviances may be very significant than as well as in other models that are more accurate.

Statistically Significant Model

```
##
## Call:
## glm(formula = cbind(Byssinosis, Non.Byssinosis) ~ Race + Employment +
##     Workspace, family = "binomial", data = b)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.24930  -0.40432   0.08309   0.49014   1.94416
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -1.2864     0.1608  -7.998 1.26e-15 ***
## RaceW                 -0.3453     0.1735  -1.990  0.04657 *
## Employment>=20         0.5923     0.1933   3.064  0.00218 **
## Employment10-19        0.5680     0.2100   2.705  0.00683 **
## WorkspaceLess Dusty   -1.2155     0.2065  -5.886 3.96e-09 ***
## WorkspaceLeast Dusty  -1.3208     0.1838  -7.184 6.77e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 113.204  on 71  degrees of freedom
## Residual deviance:  32.077  on 66  degrees of freedom
## AIC: 217.8
##
## Number of Fisher Scoring iterations: 4
```

The model above uses only the statistically significant variables for the x values. This includes race, employment lengths of greater than 20 years and between 10 and 19 years, and less dusty and least dusty workspaces. All of the variable continue to be significant as their p-values are less than 0.05. The least dusty and less dusty workspaces have the strongest association as the p-values are still the lowest by far, which reinforces its significance. A negative coefficient for these two predictor variables suggest that all other variables equal, we can see that if the race is Caucasian or the type of workspace is less dusty/least dusty they are less likely to have byssinosis. However, when the employment length is greater than 20 years or between 10 and 19 years, then there is a positive coefficient which suggests that the subjects' chances of byssinosis would increase.

ANOVA

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: cbind(Byssinosis, Non.Byssinosis)
##
## Terms added sequentially (first to last)
##
##
##            Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                          71    113.204
## Race        1    5.725        70    107.479 0.0167234 *
## Employment  2   16.917        68     90.562 0.0002121 ***
## Workspace   2   58.486        66     32.077 1.995e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

By looking at the table above, we see the deviances of the significant model is similar to the additive model. The deviances increase as each variable is added. We see the largest increase in the deviance when the Workspace predictor variable is added, and it also significantly reduces the residual deviance. We see the second largest increase in the deviances, when employment is also added to the model. We can also see that race increases the deviances. All of the p-values are lower than 0.05, so these deviances are statistically significant.

AIC and BIC

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

```
##
## Call:  glm(formula = cbind(Byssinosis, Non.Byssinosis) ~ ., family = "binomial",
##     data = b)
##
## Coefficients:
##          (Intercept)        Employment>=20       Employment10-19
##              -1.2692                0.6206                0.5812
##           SmokingYes                  SexM                 RaceW
##               0.2205               -0.1983               -0.3525
##  WorkspaceLess Dusty  WorkspaceLeast Dusty
##              -1.2827               -1.3701
##
## Degrees of Freedom: 71 Total (i.e. Null);  64 Residual
## Null Deviance:       113.2
## Residual Deviance: 29.24     AIC: 219
##
## $AIC
## [1] -210.4089
##
## $BIC
## [1] -192.1955
```

```
## [[1]]
##
## Call:  glm(formula = cbind(Byssinosis, Non.Byssinosis) ~ Race + Employment +
##     Workspace, family = "binomial", data = b)
##
## Coefficients:
##          (Intercept)                 RaceW        Employment>=20
##              -1.2864               -0.3453                0.5923
##      Employment10-19   WorkspaceLess Dusty  WorkspaceLeast Dusty
##               0.5680               -1.2155               -1.3208
##
## Degrees of Freedom: 71 Total (i.e. Null);  66 Residual
## Null Deviance:       113.2
## Residual Deviance: 32.08     AIC: 217.8
##
## $AIC
## [1] -214.4089
##
## $BIC
## [1] -200.7489
```

```
## [[1]]
##
## Call:  glm(formula = cbind(Byssinosis, Non.Byssinosis) ~ .^2, family = binomial(),
##     data = b)
##
## Coefficients:
##                          (Intercept)                        Employment>=20
##                            -1.494913                              0.681730
##                      Employment10-19                            SmokingYes
##                             1.052178                              0.152742
##                                 SexM                                 RaceW
##                            -0.099485                             -0.423018
##                  WorkspaceLess Dusty                  WorkspaceLeast Dusty
##                            -0.624078                             -0.909192
##            Employment>=20:SmokingYes            Employment10-19:SmokingYes
##                             0.184867                             -0.025425
##                  Employment>=20:SexM                  Employment10-19:SexM
##                            -0.024307                             -0.214683
##                 Employment>=20:RaceW                 Employment10-19:RaceW
##                            -0.227668                              0.011424
##   Employment>=20:WorkspaceLess Dusty   Employment10-19:WorkspaceLess Dusty
##                             0.290516                              0.004005
##  Employment>=20:WorkspaceLeast Dusty  Employment10-19:WorkspaceLeast Dusty
##                            -0.171802                             -0.928899
##                      SmokingYes:SexM                      SmokingYes:RaceW
##                             0.004233                              0.321670
##       SmokingYes:WorkspaceLess Dusty       SmokingYes:WorkspaceLeast Dusty
##                            -0.558346                             -0.090863
##                           SexM:RaceW              SexM:WorkspaceLess Dusty
##                             0.140224                             -0.164101
##            SexM:WorkspaceLeast Dusty             RaceW:WorkspaceLess Dusty
##                            -0.192677                             -0.565935
##           RaceW:WorkspaceLeast Dusty
##                            -0.020978
##
## Degrees of Freedom: 71 Total (i.e. Null);  45 Residual
## Null Deviance:       113.2
## Residual Deviance: 19.09     AIC: 246.8
##
## $AIC
## [1] -172.4089
##
## $BIC
## [1] -110.9389
```

AIC is a measure of the relative quality of a statistical model. It considers the fit of the model, as well as the number of parameters the model has. Comparing the 3 models using AIC, the model with the lower AIC value is better of the three. This is because a lower AIC value indicates that the model has a better fit to the data, while also having a simpler structure. Thus, sig_model performs the best for AIC with a score of -214.4089 where full_model is -210.4089 and interaction_model is -171.4089.

BIC also produces the same results as AIC with the sig_model being the best because a lower BIC value indicates that the model has a better fit to the data, while also having a simpler structure and considers the length n. Therefore when comparing the 3 models with BIC, sig_model is -200.7489, full_model is -192.1955, and interaction_model is -110.9389. Through BIC, we know that sig_model is the best fit model.

V. Discussion

The data is spread even for the most part, but there are a lot of outliers for race, smokers, and men. This suggested that the data would have marginal differences among the statistically significant variables. As we delved into the odds ratio, we learned which group is more likely to have byssinosis. The odd ratio suggests that men and smokers are more likely to get byssinosis. For the model investigation, we tested three models: additive, interaction, and statistically significant. The statistically significant model had the most accurate model, which indicated that the Caucasian race, employment lengths of greater than 20 years and between

10 and 19 years, and less dusty and least dusty workspaces significantly increase the risk of byssinosis. The model with only statistically significant variables had the best fit as it had the lowest AIC and BIC value. This supports the assumption that the aforementioned statistically significant variables are good indicators of byssinosis.

VI. Conclusion

After completing our model analysis, we discovered that Caucasian race, employment lengths of greater than 20 years and between 10 and 19 years, and less dusty and least dusty workspaces are likely to increase the risks of byssinosis. The significance levels, AIC and BIC values, and box plots support this conclusion. With this result, it may be a good idea for Caucasian people with weaker immune systems to be cautious of the the dustiness of their workspaces especially after 10 years of employment. Even so, the risk of byssinosis does not increase by a lot as supported by the spread of the box plots, so people within this demographic should not worry much.

VII. Appendix

```
# for box plots
byssinosis <- read.csv("Byssinosis.csv")
head(byssinosis)
byssinosis$Workspace = as.factor(byssinosis$Workspace)
box1 = ggplot(data = byssinosis, mapping = aes(x = Workspace, y = Byssinosis)) +
  labs(title = 'Byssinosis by Workspace Dustiness', x = 'Workspace Type', y = 'Byssinosis Count' )+
    theme_classic() +
    geom_boxplot()
box2 <- ggplot(data = byssinosis, mapping = aes(x = Employment, y = Byssinosis),
               main = "Bysinosis status by Employment length") +
  labs(title="Byssinosis by Length of Employment",x="Length of Employment (Years)",
       y = "Byssinosis Count") +
  theme_classic() +
  geom_boxplot()
box3 <- ggplot(data = byssinosis, mapping = aes(x = Smoking, y = Byssinosis)) +
  labs(title="Byssinosis by Smoking Status",x="Smoking Status", y = "Byssinosis Count") +
  theme_classic() +
  geom_boxplot()
box4 <- ggplot(data = byssinosis, mapping = aes(x = Sex, y = Byssinosis)) +
  labs(title="Byssinosis by Sex",x="Sex", y = "Byssinosis Count") +
  theme_classic() +
  geom_boxplot()
box5 <- ggplot(data = byssinosis, mapping = aes(x = Race, y = Byssinosis)) +
  labs(title="Byssinosis by Race",x="Race", y = "Byssinosis Count") +
  theme_classic() +
  geom_boxplot()
grid.arrange(box1,box2,box3,box4,box5, ncol=2)
# odds of byssinosis by employment length
sum(byssinosis$Byssinosis[byssinosis$Employment == '<10']) /
  (sum(byssinosis$Byssinosis[byssinosis$Employment == '<10']) + sum(byssinosis$Non.Byssinosis[byssinosis
sum(byssinosis$Byssinosis[byssinosis$Employment == '10-19']) /
  (sum(byssinosis$Byssinosis[byssinosis$Employment == '10-19']) + sum(byssinosis$Non.Byssinosis[byssinos
sum(byssinosis$Byssinosis[byssinosis$Employment == '>=20']) /
```

```
  (sum(byssinosis$Byssinosis[byssinosis$Employment == '>=20']) + sum(byssinosis$Non.Byssinosis[byssinosi
#odds by smoking status
smoker = sum(byssinosis$Byssinosis[byssinosis$Smoking == "Yes"]) /
  ( sum(byssinosis$Byssinosis[byssinosis$Smoking == "Yes"]) + sum(byssinosis$Non.Byssinosis[byssinosis$S
nonsmoker = sum(byssinosis$Byssinosis[byssinosis$Smoking == "No"]) /
  ( sum(byssinosis$Byssinosis[byssinosis$Smoking == "No"]) + sum(byssinosis$Non.Byssinosis[byssinosis$Sm
smoker/nonsmoker
#odds by sex#odds by sex
male = sum(byssinosis$Byssinosis[byssinosis$Sex == 'M']) /
  (sum(byssinosis$Byssinosis[byssinosis$Sex == 'M']) + sum(byssinosis$Non.Byssinosis[byssinosis$Sex ==
female = sum(byssinosis$Byssinosis[byssinosis$Sex == 'F']) /
  (sum(byssinosis$Byssinosis[byssinosis$Sex == 'F']) + sum(byssinosis$Non.Byssinosis[byssinosis$Sex ==
male/female
#odds of byssinosis by race
white = sum(byssinosis$Byssinosis[byssinosis$Race == 'W']) /
  (sum(byssinosis$Byssinosis[byssinosis$Race == 'W']) + sum(byssinosis$Non.Byssinosis[byssinosis$Race ==
poc = sum(byssinosis$Byssinosis[byssinosis$Race == 'O']) /
  (sum(byssinosis$Byssinosis[byssinosis$Race == 'O']) + sum(byssinosis$Non.Byssinosis[byssinosis$Race ==
white/poc
#odds of byssinosis by workspace
sum(byssinosis$Byssinosis[byssinosis$Workspace == '1']) /
  (sum(byssinosis$Byssinosis[byssinosis$Workspace == '1']) + sum(byssinosis$Non.Byssinosis[byssinosis$Wo
sum(byssinosis$Byssinosis[byssinosis$Workspace == '2']) /
  (sum(byssinosis$Byssinosis[byssinosis$Workspace == '2']) + sum(byssinosis$Non.Byssinosis[byssinosis$Wo
sum(byssinosis$Byssinosis[byssinosis$Workspace == '3']) /
  (sum(byssinosis$Byssinosis[byssinosis$Workspace == '3']) + sum(byssinosis$Non.Byssinosis[byssinosis$Wo
# for model
library(bestglm)
library(plyr)
library(gridExtra)
b<- read.csv('Byssinosis.csv', stringsAsFactors = T)
b$Workspace = as.factor(b$Workspace)
b$Byssinosis=as.factor(b$Byssinosis)
b$Non.Byssinosis=as.factor(b$Non.Byssinosis)
b$Workspace = revalue(b$Workspace, c('1' = 'Most Dusty', '2' = 'Less Dusty',
                                     '3' = 'Least Dusty'))
sapply(b,is.factor)
library('gridExtra')
b<- read.csv("Byssinosis.csv",stringsAsFactors = T)
b$Workspace = as.factor(b$Workspace)
b$Byssinosis=as.factor(b$Byssinosis)
b$Non.Byssinosis=as.factor(b$Non.Byssinosis)
b$Workspace = revalue(b$Workspace, c("1" = "Most Dusty", "2" = "Less Dusty",
                                     "3" = "Least Dusty"))
#sapply(b,is.factor)
add_model<- glm(cbind(Byssinosis,Non.Byssinosis)~., family="binomial", data=b)
summary(add_model)
anova(add_model, test="Chisq")
interaction_model = glm(cbind(Byssinosis,Non.Byssinosis) ~ .**2, data = b,
                        family = binomial())
summary(interaction_model)
anova(interaction_model, test= "Chisq")
sig_model <- glm(cbind(Byssinosis,Non.Byssinosis)~Race+Employment+Workspace, family="binomial", data=b)
```

```
summary(sig_model)
anova(sig_model, test= "Chisq")
aikaike_bayesian_info_criterion <- function(model) {
n <- length(model$fitted.values)
k <- length(model$coefficients)
ll <- -2 * model$null.deviance
aic <- ll + 2 * k
bic<-(log(n)*k)+aic-(2*k)
return(list(model,AIC = aic,BIC=bic))
}
aikaike_bayesian_info_criterion(add_model)
aikaike_bayesian_info_criterion(sig_model)
aikaike_bayesian_info_criterion(interaction_model)
```