Project

Diana Rueda

12/7/2021

In order to analyze the relationships within the data, we start by looking at the proportions of smoking status, sex, race, lenght of employment and dustiness of workplace in comparison to the presence and absence of byssinosis. Then we continue to test for independance in each case with the assistance of contingency tables and likelihood ratio test. Finally we build a generalized linear model with the assistance of functions such as glm and step.

We should note that the data contained in Byssinosis.csv is in wide format. Thus, there are two columns containing counts of infections of byssinosis (Byssinosis) and non infected employees (Non.Byssinosis).

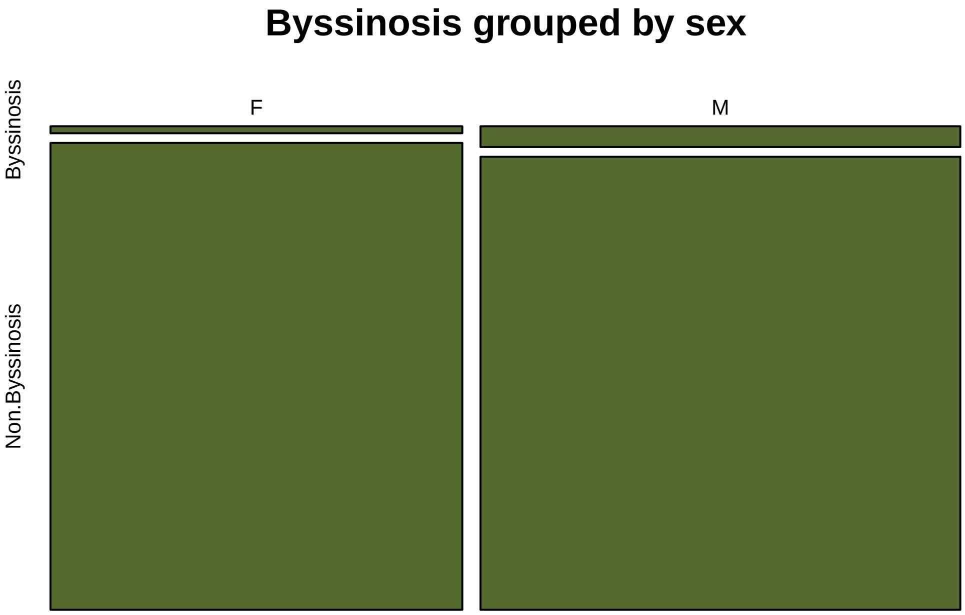
```
bys = read.csv('Byssinosis.csv')
```

Analyzing Proportion Visualizations by Column

Sex

```
# visualize Sex proportions
table_sex = aggregate(cbind(Byssinosis, Non.Byssinosis) ~ Sex, data = bys, sum)
rownames(table_sex) = table_sex$Sex
table_sex = subset(table_sex, select = -Sex)

mosaicplot(table_sex, main = 'Byssinosis grouped by sex', color = 'darkolivegreen')
```

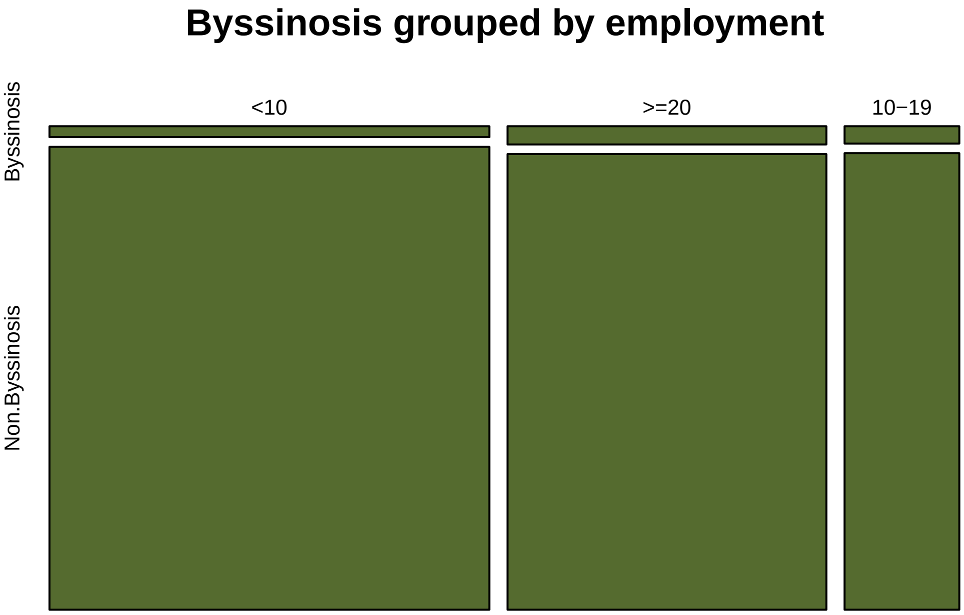


The male employees have a bigger proportion of byssinosis as compared to female. From the mosaic plot we can see it is more than twice the proportion of women with byssinosis. Sex might be a good variable for our model but more analysis is required.

Employment

```r
# visualize Employment proportions
table_employment = aggregate(cbind(Byssinosis, Non.Byssinosis) ~ Employment, data = bys, sum)
rownames(table_employment) = table_employment$Employment
table_employment = subset(table_employment, select = -Employment)

mosaicplot(table_employment, main = 'Byssinosis grouped by employment', color = 'darkolivegreen')
```

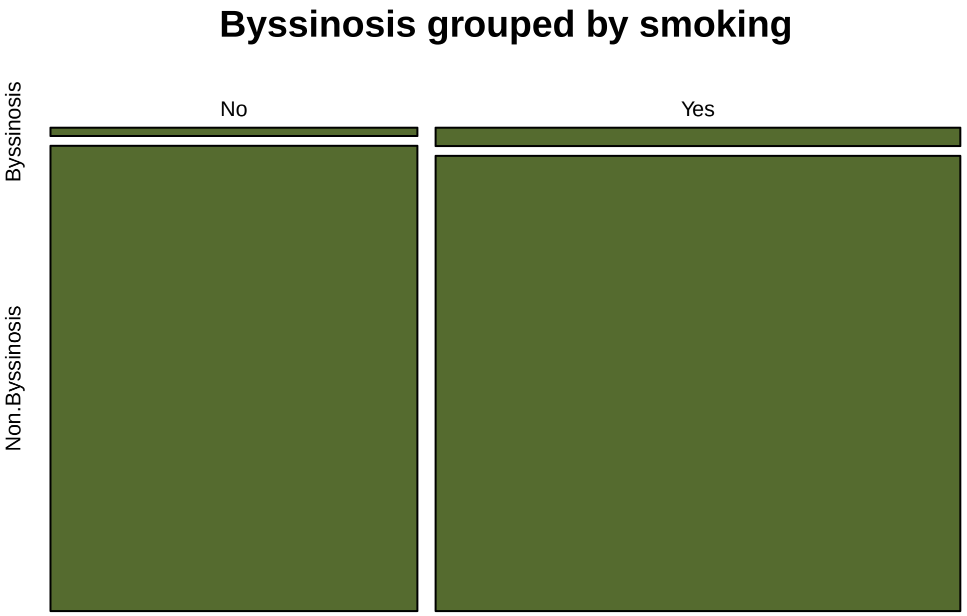


The employees who have been there for more than 20 years have a bigger proportion of byssinosis as compared to those who have been there for a shorter period of time. From the mosaic plot we can see the proportion increases as the amount of years increase. Length of employment might have a linear relationship with the amount of infections among employeesbut more analysis is required.

Smoking

```r
# visualize Smoking proportions
table_smoking = aggregate(cbind(Byssinosis, Non.Byssinosis) ~ Smoking, data = bys, sum)
rownames(table_smoking) = table_smoking$Smoking
table_smoking = subset(table_smoking, select = -Smoking)

mosaicplot(table_smoking, main = 'Byssinosis grouped by smoking', color = 'darkolivegreen')
```

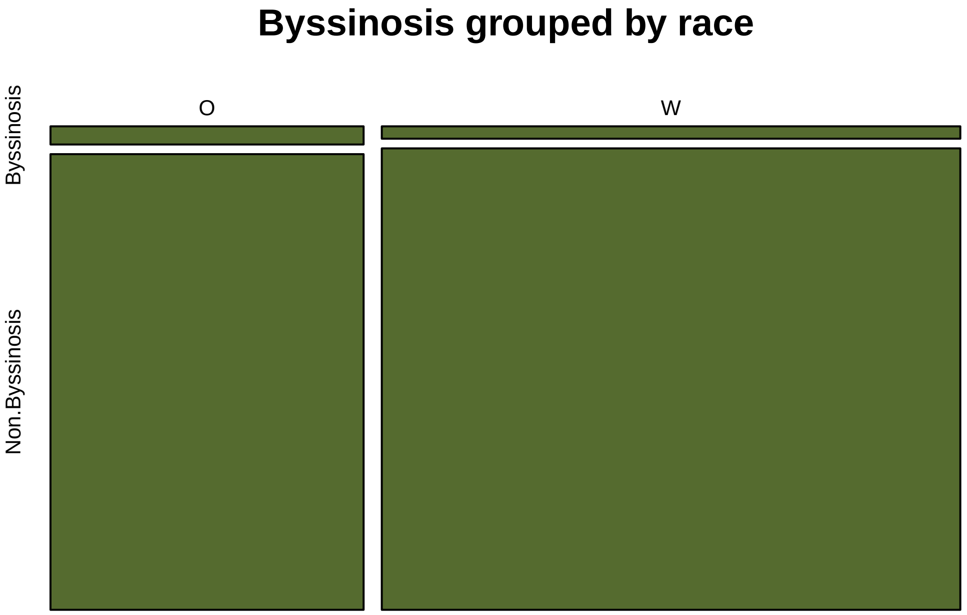


The employees who smoke have a bigger proportion of byssinosis as compared to non smokers. From the mosaic plot we can see it is more than twice the proportion of women with byssinosis. Smoking might be a good variable for our model but more analysis is required.

Race

```
# visualize Race proportions
table_race = aggregate(cbind(Byssinosis, Non.Byssinosis) ~ Race, data = bys, sum)
rownames(table_race) = table_race$Race
table_race = subset(table_race, select = -Race)

mosaicplot(table_race, main = 'Byssinosis grouped by race', color = 'darkolivegreen')
```

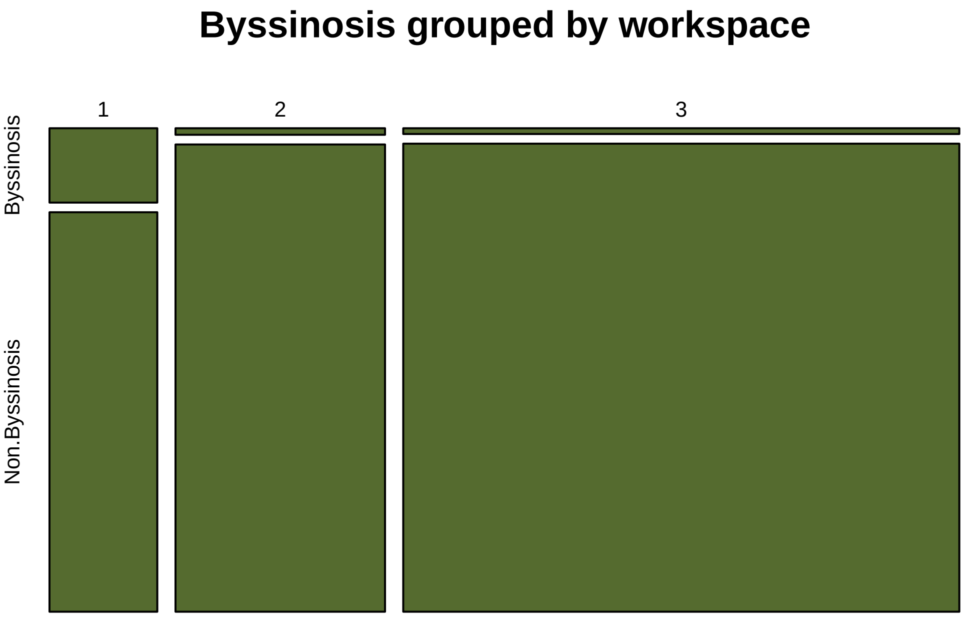


The employees from other races have a bigger proportion of byssinosis as compared to white employees. From the mosaic plot we can see it is almost twice the proportion of white employees with byssinosis. Race might also be a good variable for our model but more analysis is required.

Workplace

```
# visualize Workspace proportions
table_workspace = aggregate(cbind(Byssinosis, Non.Byssinosis) ~ Workspace, data = bys, sum)
rownames(table_workspace) = table_workspace$Workspace
table_workspace = subset(table_workspace, select = -Workspace)

mosaicplot(table_workspace, main = 'Byssinosis grouped by workspace', color = 'darkolivegreen')
```



The employees working in a mostly dusty environment have a bigger proportion of byssinosis as compared to those working in less and least dusty environments. The amount of dust in the work environment might be a great variable for our model.

Analyzing Independance

We are using Likelihood Ratio test to test for independance. The null hypothesis states the following variables are independent from Byssinosis.

Sex

The observe counts for Byssinosis ~ Sex:

```
table_sex
```

```
##   Byssinosis Non.Byssinosis
## F         37           2466
## M        128           2788
```

The expected counts for Byssinosis ~ Sex:

```
E_sex = outer(rowSums(table_sex), colSums(table_sex)) / sum(table_sex)
E_sex
```

```
##   Byssinosis Non.Byssinosis
## F    76.2124       2426.788
## M    88.7876       2827.212
```

The Likelihood ratio test statistic for Byssinosis ~ Sex is:

```
#Likelihood ratio test
LRstat = 2 * sum(table_sex * log(table_sex / E_sex))
LRstat
```

```
## [1] 41.34426
```

The p-value for Byssinosis ~ Sex under the null hypothesis is:

```
# p-value
1 - pchisq(LRstat, 1)
```

```
## [1] 1.276457e-10
```

Since the p-value < 0.05 we reject the null hypothesis of independence. Thus, there is significant evidence to say there exists dependance between these two variables.

Smoking

The observe counts for Byssinosis ~ Smoking:

```
table_smoking
```

```
##     Byssinosis Non.Byssinosis
## No          40           2190
## Yes        125           3064
```

The expected counts for Byssinosis ~ Smoking:

```
E_smoking = outer(rowSums(table_smoking), colSums(table_smoking)) / sum(table_smoking)
E_smoking
```

```
##     Byssinosis Non.Byssinosis
## No    67.89998         2162.1
## Yes   97.10002         3091.9
```

The Likelihood ratio test statistic for Byssinosis ~ Smoking is:

```
#Likelihood ratio test
LRstat = 2 * sum(table_smoking * log(table_smoking / E_smoking))
LRstat
```

```
## [1] 21.42154
```

The p-value for Byssinosis ~ Smoking under the null hypothesis is:

```
# p-value
1 - pchisq(LRstat, 1)
```

```
## [1] 3.686064e-06
```

Since the p-value < 0.05 we reject the null hypothesis of independence. Thus, there is significant evidence to say there exists dependance between these two variables.

Employment

The observe counts for Byssinosis ~ Employment:

```
table_employment
```

```
##       Byssinosis Non.Byssinosis
## <10           63           2666
## >=20          76           1902
## 10-19         26            686
```

The expected counts for Byssinosis ~ Employment:

```
E_employment = outer(rowSums(table_employment), colSums(table_employment)) / sum(table_employment)
E_employment
```

```
##       Byssinosis Non.Byssinosis
## <10     83.09374      2645.9063
## >=20    60.22698      1917.7730
## 10-19   21.67928       690.3207
```

The Likelihood ratio test statistic for Byssinosis ~ Employment is:

```
#Likelihood ratio test
LRstat = 2 * sum(table_employment * log(table_employment / E_employment))
LRstat
```

```
## [1] 10.23586
```

The p-value for Byssinosis ~ Employment under the null hypothesis is:

```r
# p-value
1 - pchisq(LRstat, 1)
```

```
## [1] 0.001377366
```

Since the p-value < 0.05 we reject the null hypothesis of independence. Thus, there is significant evidence to say there exists dependance between these two variables.

Race

The observe counts for Byssinosis ~ Race:

```r
table_race
```

```
##   Byssinosis Non.Byssinosis
## O         73           1830
## W         92           3424
```

The expected counts for Byssinosis ~ Race:

```r
E_race = outer(rowSums(table_race), colSums(table_race)) / sum(table_race)
E_race
```

```
##   Byssinosis Non.Byssinosis
## O   57.94335       1845.057
## W  107.05665       3408.943
```

The Likelihood ratio test statistic for Byssinosis ~ Race is:

```r
#Likelihood ratio test
LRstat = 2 * sum(table_race * log(table_race / E_race))
LRstat
```

```
## [1] 6.025885
```

The p-value for Byssinosis ~ Race under the null hypothesis is:

```r
# p-value
1 - pchisq(LRstat, 1)
```

```
## [1] 0.01409757
```

Since the p-value < 0.05 we reject the null hypothesis of independence. Thus, there is significant evidence to say there exists dependance between these two variables.

Workspace

The observe counts for Byssinosis ~ Workspace:

```
table_workspace
```

```
##   Byssinosis Non.Byssinosis
## 1        105            564
## 2         18           1282
## 3         42           3408
```

The expected counts for Byssinosis ~ Workspace:

```
E_workspace = outer(rowSums(table_workspace), colSums(table_workspace)) / sum(table_workspace)
E_workspace
```

```
##   Byssinosis Non.Byssinosis
## 1   20.36999        648.630
## 2   39.58295       1260.417
## 3  105.04706       3344.953
```

The Likelihood ratio test statistic for Byssinosis ~ Workspace is:

```
#Likelihood ratio test
LRstat = 2 * sum(table_workspace * log(table_workspace / E_workspace))
LRstat
```

```
## [1] 252.1082
```

The p-value for Byssinosis ~ Workspace under the null hypothesis is:

```
# p-value
1 - pchisq(LRstat, 1)
```

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

Since the p-value < 0.05 we reject the null hypothesis of independence. Thus, there is significant evidence to say there exists dependance between these two variables.

Model Selection

We have found there is dependence between all of these variables and the presence of byssinosis on employees. Now we proceed to fit a generalized linear model by forward stepwise regression.

```
library(tidyr)
```

```
# transform wide to long data
temp_count = gather(bys, Byssinosis, count, Byssinosis:Non.Byssinosis ,factor_key = TRUE)

temp_rows = subset(temp_count,  select = -count)
long_data = data.frame()

for (i in 1:dim(temp_rows)[1]) {
  n = temp_count[[’count’]][i]
  temp = temp_rows[rep(i, n), ]
  long_data = rbind(long_data, temp)

}

long_data[’Byssinosis’] = ifelse(long_data$Byssinosis == ’Byssinosis’, 1,0)
```

```
# build the model
result <- step(glm(Byssinosis ~ 1, binomial, long_data),
               scope = ~Workspace*Sex*Smoking*Employment*Race,
               test="LRT",
               direction = "forward")
```

```
## Start:  AIC=1479.19
## Byssinosis ~ 1
##
##              Df Deviance    AIC     LRT  Pr(>Chi)
## + Workspace   1   1255.2 1259.2 221.963 < 2.2e-16 ***
## + Sex         1   1435.8 1439.8  41.344 1.276e-10 ***
## + Smoking     1   1455.8 1459.8  21.422 3.686e-06 ***
## + Employment  2   1467.0 1473.0  10.236  0.005988 **
## + Race        1   1471.2 1475.2   6.026  0.014098 *
## <none>            1477.2 1479.2
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Step:  AIC=1259.23
## Byssinosis ~ Workspace
##
##              Df Deviance    AIC     LRT  Pr(>Chi)
## + Smoking     1   1241.4 1247.4 13.7911 0.0002043 ***
## + Employment  2   1240.0 1248.0 15.2009 0.0005002 ***
## + Sex         1   1249.0 1255.0  6.2621 0.0123350 *
## <none>            1255.2 1259.2
## + Race        1   1254.2 1260.2  1.0540 0.3045858
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Step:  AIC=1247.44
## Byssinosis ~ Workspace + Smoking
##
##                     Df Deviance    AIC     LRT  Pr(>Chi)
## + Employment         2   1226.5 1236.5 14.9423 0.0005693 ***
## + Sex                1   1238.2 1246.2  3.2314 0.0722388 .
## + Workspace:Smoking  1   1238.9 1246.9  2.5280 0.1118402
## <none>                   1241.4 1247.4
## + Race               1   1240.5 1248.5  0.9457 0.3308222
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Step:  AIC=1236.5
## Byssinosis ~ Workspace + Smoking + Employment
```

```
##
##                          Df Deviance    AIC    LRT Pr(>Chi)
## + Workspace:Smoking       1   1224.1 1236.1 2.3536   0.1250
## <none>                        1226.5 1236.5
## + Workspace:Employment  2   1222.7 1236.7 3.8395   0.1466
## + Race                    1   1224.8 1236.8 1.6863   0.1941
## + Sex                     1   1225.0 1237.0 1.4759   0.2244
## + Smoking:Employment    2   1224.2 1238.2 2.2511   0.3245
##
## Step:  AIC=1236.14
## Byssinosis ~ Workspace + Smoking + Employment + Workspace:Smoking
##
##                          Df Deviance    AIC    LRT Pr(>Chi)
## <none>                        1224.1 1236.1
## + Sex                     1   1222.2 1236.2 1.8961   0.1685
## + Race                    1   1222.5 1236.5 1.6103   0.2045
## + Workspace:Employment  2   1220.6 1236.6 3.5643   0.1683
## + Smoking:Employment    2   1221.6 1237.6 2.5743   0.2761
```

Using forward step regression with all the variables and its interactions we found that the best model according to the AIC is Byssinosis ~ Workspace + Smoking + Employment + Workspace:Smoking

```
model = glm(Byssinosis ~ Workspace + Smoking + Employment + Workspace:Smoking, binomial, long_data)
summary(model)
```

```
##
## Call:
## glm(formula = Byssinosis ~ Workspace + Smoking + Employment +
##     Workspace:Smoking, family = binomial, data = long_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7081  -0.2455  -0.1471  -0.1130   3.2471
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -1.6739     0.4249  -3.940 8.15e-05 ***
## Workspace              -1.1976     0.2017  -5.938 2.89e-09 ***
## SmokingYes              1.3150     0.4715   2.789 0.005286 **
## Employment>=20          0.6674     0.1797   3.714 0.000204 ***
## Employment10-19         0.5299     0.2471   2.145 0.031989 *
## Workspace:SmokingYes   -0.3665     0.2376  -1.542 0.123001
## ---
## 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: 1224.1  on 5413  degrees of freedom
## AIC: 1236.1
##
## Number of Fisher Scoring iterations: 7
```

The summary of the model selected shows that all but one coefficients are significant. Next we look at the confidence intervals. Below we can see that the only interval containing zero is the one corresponding to the estimate beta for Workspace:SmokingYes which happens to be the same estimate that resulted non significant for our model. Thus, we fail to reject the null hypothesis that this estimate equals zero. In other words we do not have enough evidence to suggest that the estimate for the contribution of the interaction Workspace:SmokingYes is non zero.

```
# confidence intervals
confint.default(model)
```

```
##                            2.5 %      97.5 %
## (Intercept)          -2.50659474 -0.84120750
## Workspace            -1.59286163 -0.80228067
## SmokingYes            0.39091819  2.23917752
## Employment>=20        0.31519785  1.01961967
## Employment10-19       0.04561064  1.01424802
## Workspace:SmokingYes -0.83224782  0.09925006
```

By removing this interaction from the model we obtain the folowing results where all the estimates continue to be significant but the AIC increases from 1236.1 to 1236.5. Thus if your goal requires you to prioritize minimizing AIC The model with formula Byssinosis ~ Workspace + Smoking + Employment + Workspace:Smoking is the best choise.

```
model2 = glm(Byssinosis ~ Workspace + Smoking + Employment, binomial, long_data)
summary(model2)
```

```
##
## Call:
## glm(formula = Byssinosis ~ Workspace + Smoking + Employment,
##     family = binomial, data = long_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6896  -0.2498  -0.1574  -0.1207   3.3432
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -1.1858     0.2618  -4.530  5.9e-06 ***
## Workspace        -1.4663     0.1057 -13.869  < 2e-16 ***
## SmokingYes        0.6670     0.1892   3.526 0.000422 ***
## Employment>=20    0.6699     0.1793   3.735 0.000188 ***
## Employment10-19   0.5328     0.2465   2.162 0.030655 *
## ---
## 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: 1226.5  on 5414  degrees of freedom
## AIC: 1236.5
##
## Number of Fisher Scoring iterations: 7
```