STA 138 FINAL PROJECT

John Tran 916232852
Collaborators for the code: Grant Smith, Jaymie Tam, and Hasan Rahman

Introduction:
In this report, we were given a data set about Byssinosis, a rare lung disease that workers contracted due to the cotton dust in 1973. This data set consisted of 5,419 subjects based on their type of work, employment years, smoking, sex, race, and whether they had byssinosis or not. Our task was to discover if there was any association between these factors and byssinosis.

Materials and Methods:
In this section, we will discuss which methods to interpret this data and which statistical tools used to analyze any relationship with these categories towards the disease. Basically this is the procedures in a lab report. The output of these methods will be discussed in the Results section.

Subsetting the Categories for observations

Our group decided to sum all the certain categories of the data since it is categorical data. Initially, it would be confusing to treat this dataset as any other data set. So we filtered out the data based on the category and subset its data in that respective category. Here is a summary of all these subsets. Initial observations that we discovered were that workspace and and employment categories have ordinal values, while smoking, sex, and race are binary. There are 165 subjects with Byssinosis and 5254 subjects without Byssinosis. With this summary, we decided to use logistic regression.

Using logistic regression

We created a full model that included all the factors and the response is Byssinosis. We had to bind NonByssinosis and Byssinosis in one category to make it the response. Our full model is full.gl<-glm(cbind(Byssinosis,Non.Byssinosis)~Employment + Smoking+Sex+Race +Workspace ,family=binomial(link=logit),data=Byn). Then we find the summary of the full model which will be discussed in the Results section.

Likelihood Ratio Test

Based on the summary of the full model we created, we decided to verify the results of by doing likelihood ratio test. The likelihood ratio test help us test which parameters have any significance on Byssinosis. The results will be discussed in the Results section.

Confidence Interval for the full model

Since we are using logistic regression we decided to the log confidence intervals to find the interval and test whether our odds intervals were greater than 1. In this section, we had to use the bonferroni correction based on how many comparisons we had in data. We check which parameter that influences the effect of disease based on the odds. We did it at level of significance = .05.

Tests for analyzing certain predictors (Workspace, employment, Smoking,race,sex)

A package we used for these factors is vcdExtra since Workspace and employee are both ordinal data. We used the Mantel Haenzel Test for employment and for workspace we created table to show the relationship it had individually with Byssinosis. Results are in the results section. In R it is the CMHtest function to use the Mantel Haenzel test.

Since smoking,race and sex are binary data we used a different statistical tool to help us analyze its effects on the disease.Similar to workspace parameter, we created a contingency table for We used the chisq.test built in R, which is the Pearson Chi Square test. We used this test for independence on the response variable.

Model Selection

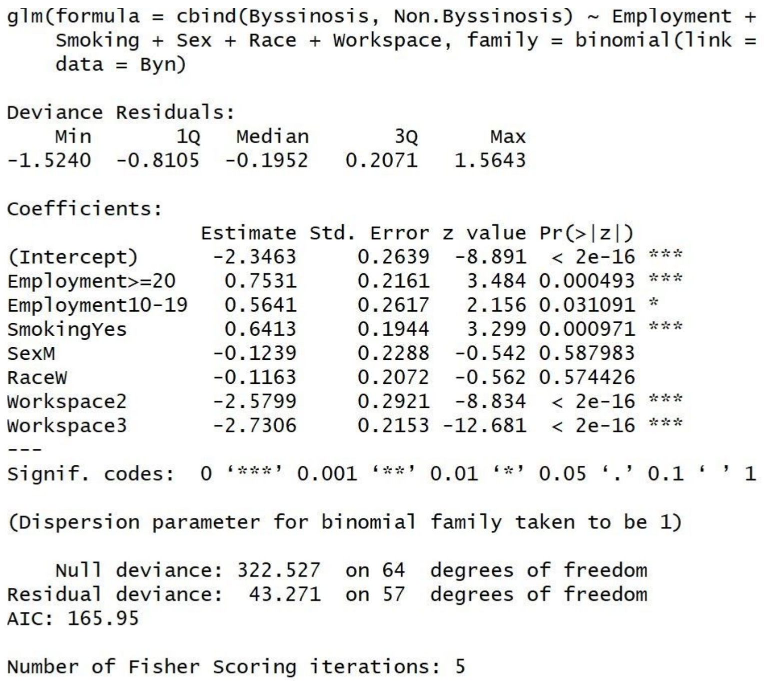
To determine our best model for this dataset, we used drop function and the step function to determine the best model. This is based on our AIC and BIC. With the lower AIC it helps us penalize certain models with too many parameters or factors making it easier to choose the right model.We used the bidirectional process to find the best AIC.

Diagnostics

We created binned plot for the residuals of our best model using the arm package in R and also plot predict and fitted residuals. These residuals will provide us which has the most influence on the best model and disease. Also we created a histogram to visualize the residuals to determine if we have a poor model.

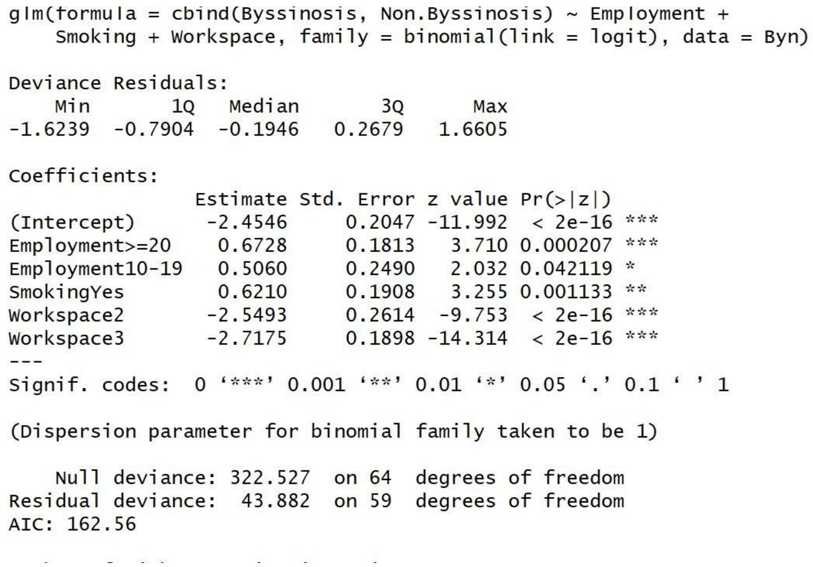
Results and Analysis:

Based on our full model, there are certain variables that hold significance. According to this table that was outputted from R.



```
glm(formula = cbind(Byssinosis, Non.Byssinosis) ~ Employment +
    Smoking + Sex + Race + Workspace, family = binomial(link = logit),
    data = Byn)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5240  -0.8105  -0.1952   0.2071   1.5643

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      -2.3463     0.2639  -8.891  < 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 0.587983
RaceW            -0.1163     0.2072  -0.562 0.574426
Workspace2       -2.5799     0.2921  -8.834  < 2e-16 ***
Workspace3       -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: 322.527  on 64  degrees of freedom
Residual deviance:  43.271  on 57  degrees of freedom
AIC: 165.95

Number of Fisher Scoring iterations: 5
```

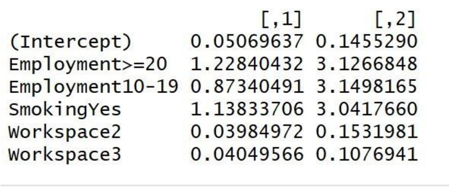
As we can see sex and race are not significant, so we should drop it from our model.
To verify this summary we run a reduced model without sex and race since its not significant.



```
glm(formula = cbind(Byssinosis, Non.Byssinosis) ~ Employment +
    Smoking + Workspace, family = binomial(link = logit), data = Byn)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6239  -0.7904  -0.1946   0.2679   1.6605

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      -2.4546     0.2047 -11.992  < 2e-16 ***
Employment>=20    0.6728     0.1813   3.710 0.000207 ***
Employment10-19   0.5060     0.2490   2.032 0.042119 *
SmokingYes        0.6210     0.1908   3.255 0.001133 **
Workspace2       -2.5493     0.2614  -9.753  < 2e-16 ***
Workspace3       -2.7175     0.1898 -14.314  < 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: 322.527  on 64  degrees of freedom
Residual deviance:  43.882  on 59  degrees of freedom
AIC: 162.56
```

The output came out to be like this. This resulted in smaller AIC then before and different estimates for the parameters. All the parameters that were not removed are still significant in this reduced model. Then to further validate this result we did a likelihood ratio test for model reduction. Our null hypothesis is sex and race has no effect while our alternative hypothesis says it has an effect. H_o: $B_1 = 0$, $B_2 = 0$, H_a: Any are zero. The resulting test statistic was .610 and the p- value was .736. Based on the p- value, we can conclude that dropping sex and race should give us a better model.

Next we created a log odds confidence interval to determine which parameters in our reduced model that alter the probability of having Byssinosis.

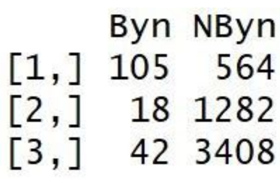


```
                      [,1]      [,2]
(Intercept)     0.05069637 0.1455290
Employment>=20  1.22840432 3.1266848
Employment10-19 0.87340491 3.1498165
SmokingYes      1.13833706 3.0417660
Workspace2      0.03984972 0.1531981
Workspace3      0.04049566 0.1076941
```

Looking at our results, the intervals for both employment parameters and smoking are greater than 1, that tells us that the probability of having byssinosis is much greater. While workspace seems to have lower odds based on these results, meaning that they decrease the probability of having Byssinosis.

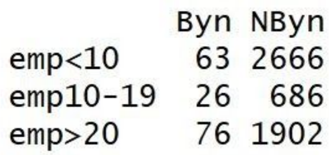
Since workspace and employment is ordinal data, we ran the Mantel Haenzel Test. Here are two separate contingency tables for both employment and workspace.

Workspace



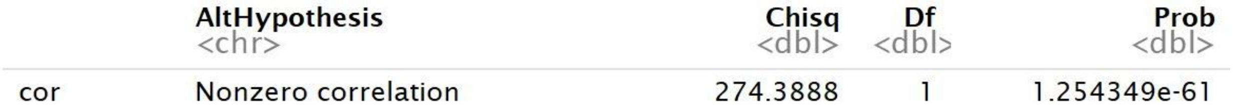
```
     Byn NByn
[1,] 105  564
[2,]  18 1282
[3,]  42 3408
```

Employment



```
          Byn NByn
emp<10     63 2666
emp10-19   26  686
emp>20     76 1902
```

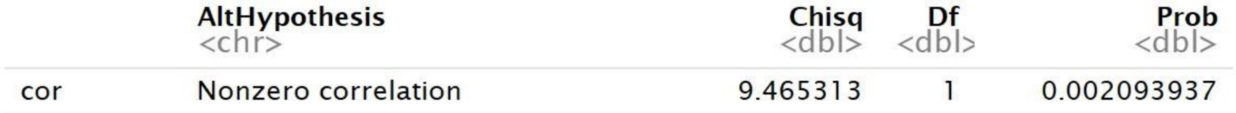
Here are the results for workspace.



```
     AltHypothesis               Chisq    Df           Prob
     <chr>                       <dbl> <dbl>          <dbl>
cor  Nonzero correlation      274.3888     1   1.254349e-61
```

The p-value from this test is extremely small, we can conclude that there is some association between Byssinosis and Workspace.

Here are the results for employment.

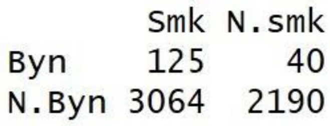


```
     AltHypothesis               Chisq    Df           Prob
     <chr>                       <dbl> <dbl>          <dbl>
cor  Nonzero correlation      9.465313     1    0.002093937
```

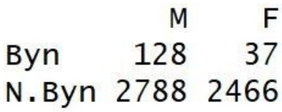
The p-value from this test is small at any level of significance, we can conclude that there is some association between Byssinosis and employment.

Next section, we tested for smoking, sex, and race using Pearson Chi Square test. Here are the contingency table for each variable.

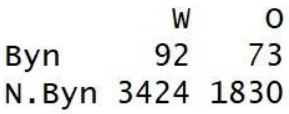
Smoke Sex Race



```
        Smk N.smk
Byn     125    40
N.Byn  3064  2190
```



```
          M    F
Byn     128   37
N.Byn  2788 2466
```



```
          W    O
Byn      92   73
N.Byn  3424 1830
```

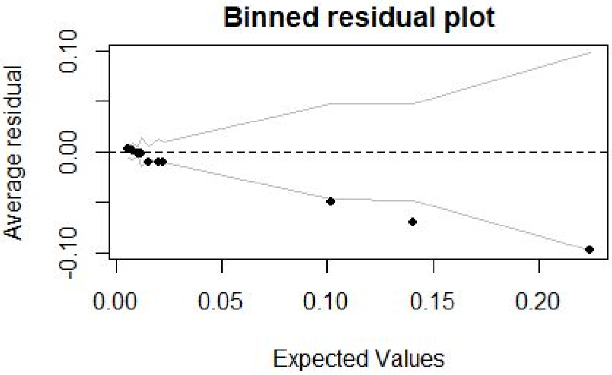
Here are the results from the Pearson Chi Square tests for each variable.

	Test statistic	P-value
Smoke	20.092	7.379e-06
Sex	38.671	5.017e-10
Race	6.2195	.01264

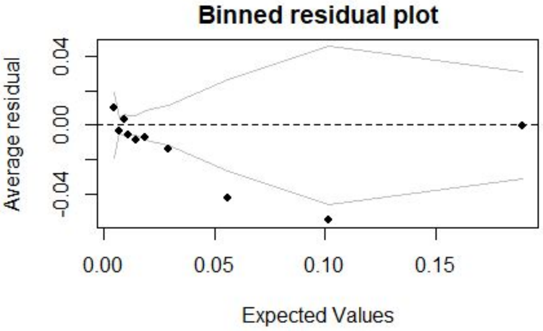
Based on the p-values for these variables being less than .05 level of significance, we can conclude that there is an association between Byssinosis and each of these variables.(Not to be confuse it is just between byssinosis and smoke, sex, race independently). However at a level of .01, we can conclude that race band byssinosis are independent.

Even though we have a decent model with a low AIC, we know we can find a better model through model selection. In this case, we indeed found a better model with a lower AIC than the reduced model. It resulted in this model: glm(formula = cbind(Byssinosis, Non.Byssinosis) ~ Employment + Smoking + Sex + Workspace + Sex:Workspace, family = binomial(link = logit), data = Byn). The AIC for this model is 160.7 which was lower than the reduced model which had AIC of 162.56. The reduced model was not bad, but it just the new model with an interaction term is better. So the significance between the sex and workspace created a better model in the process.

Once we got the best model, we decided to find the diagnostics of the residuals to validate that we have a good model and compare to our original reduce model. Here is binned plot of our residuals of our original reduced to the new model.



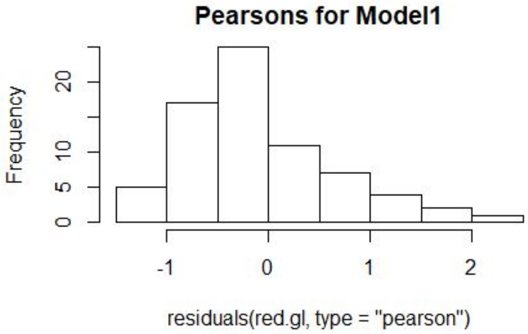
Original Reduced model



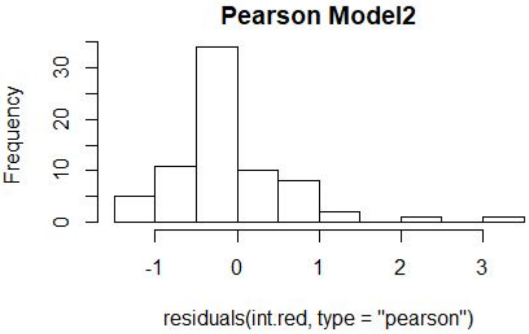
New Model

Compared to Original Reducel Model, the residuals of the New Model are similar but excludes the excess residuals out of the confidence band and the confidence band of the New Model is much smaller but more of the residuals are within the band. However, some still fall out of the confidence band suggesting the New Model might not be the best model. Here is two histograms of the pearson residuals.

Original Reduced



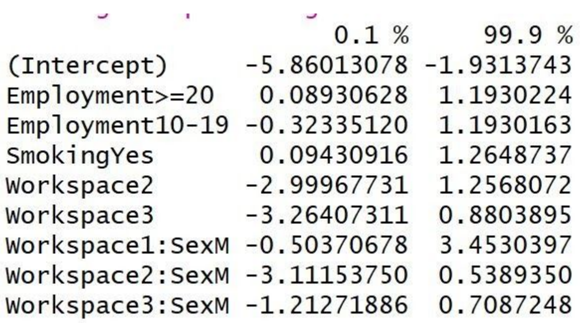
New



The New Model is actually worse than the Original Reduced Model, the residuals are greater than 2 or 3 stating that this is a lack of fit. However, this shows that AIC is all everything.

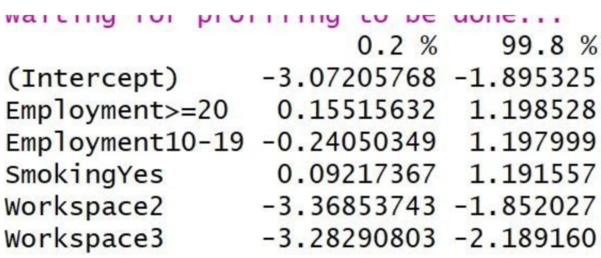
After we tested confidence interval using bonferroni corrections to test. One of the New Model had 9 comparisons while the Original Reduced model had 6. Here are the results.

New



```
                      0.1 %     99.9 %
(Intercept)     -5.86013078 -1.9313743
Employment>=20   0.08930628  1.1930224
Employment10-19 -0.32335120  1.1930163
SmokingYes       0.09430916  1.2648737
Workspace2      -2.99967731  1.2568072
Workspace3      -3.26407311  0.8803895
Workspace1:SexM -0.50370678  3.4530397
Workspace2:SexM -3.11153750  0.5389350
Workspace3:SexM -1.21271886  0.7087248
```

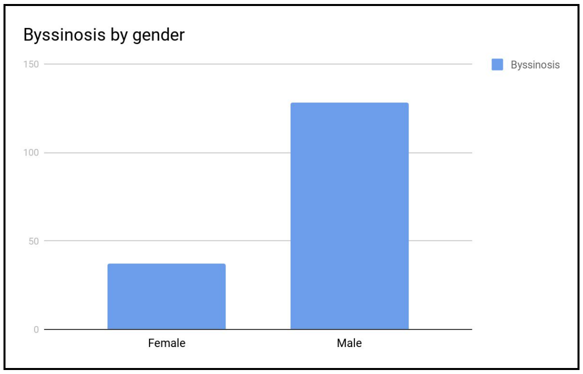
Old Reduced

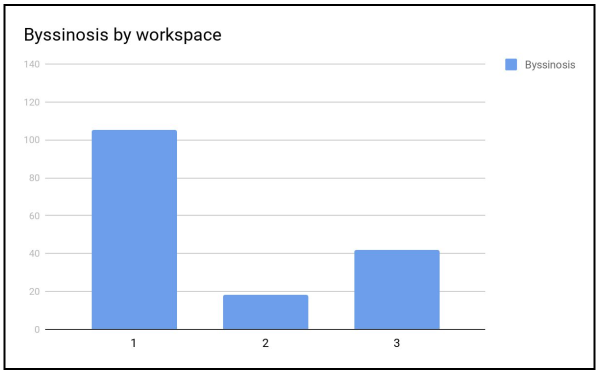


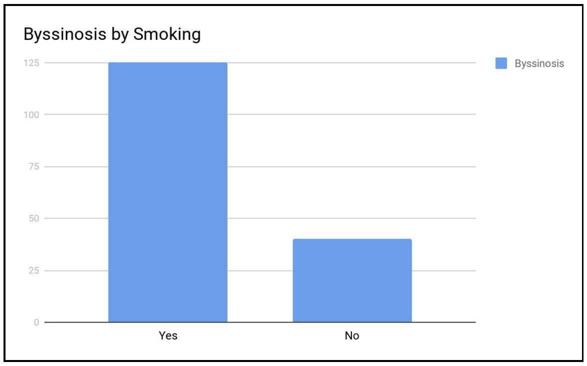
```
waiting for profiling to be done...
                      0.2 %    99.8 %
(Intercept)     -3.07205768 -1.895325
Employment>=20   0.15515632  1.198528
Employment10-19 -0.24050349  1.197999
SmokingYes       0.09217367  1.191557
Workspace2      -3.36853743 -1.852027
Workspace3      -3.28290803 -2.189160
```

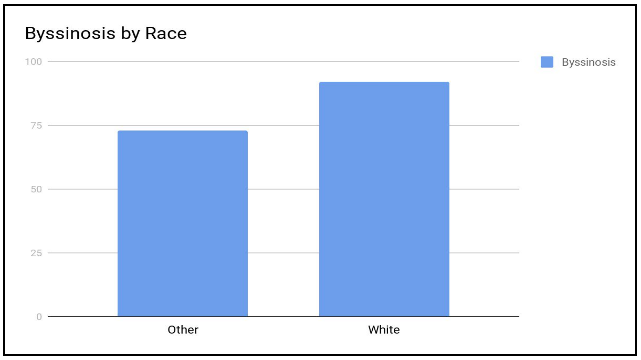
This New model has a lot of intercepts that necessarily have any effect on Byssinosis, while the Old Reduced Model has some effect to to Byssinosis.

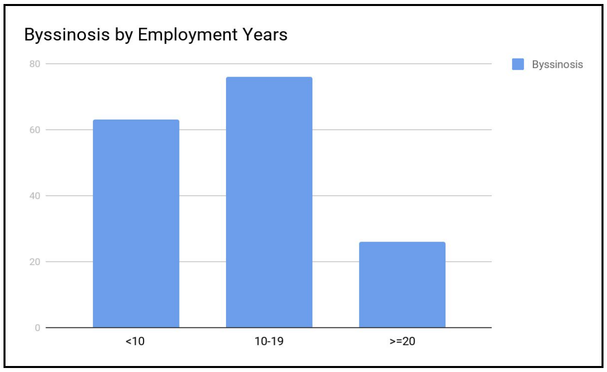
Here are individual histogram of each parameter that has byssinosis.











Conclusion:

After using many statistical tools to analyze whether these parameters have any association with Byssinosis, we found through multiple test of independence that employment, workspace, sex, race, and smoke have some association with Byssinosis. There is definitely a certain effect. However when we found the log odd confidence interval between these parameters it shows that workspace actually decrease the probability of having Byssinosis compared to the others. Even we used model selection to find a better model to fit this data, but the model lacked the fit surprisingly with its lower AIC. In conclusion, there is meaning association that causes Byssinosis.

Appendix:

```
knitr::opts_chunk$set(echo = TRUE)
Byn<-read.csv('Byssinosis.csv')
Byn$Workspace<-as.factor(Byn$Workspace)
(Byn)

#sum(Byn$Byssinosis,Byn$Non.Byssinosis) 5419

#Summary Stats
sum(Byn$Byssinosis[Byn$Smoking=='Yes'],Byn$Non.Byssinosis[Byn$Smoking=='Yes'])
sum(Byn$Byssinosis[Byn$Smoking=='No'],Byn$Non.Byssinosis[Byn$Smoking=='No'])

sum(Byn$Byssinosis[Byn$Smoking=='Yes'],Byn$Non.Byssinosis[Byn$Smoking=='Yes'])
sum(Byn$Byssinosis[Byn$Smoking=='No'],Byn$Non.Byssinosis[Byn$Smoking=='No'])

sum(Byn$Byssinosis[Byn$Sex=='M'],Byn$Non.Byssinosis[Byn$Sex=='M'])
sum(Byn$Byssinosis[Byn$Sex=='F'],Byn$Non.Byssinosis[Byn$Sex=='F'])


sum(Byn$Byssinosis[Byn$Employment=='<10'],Byn$Non.Byssinosis[Byn$Employment=='<10'])
sum(Byn$Byssinosis[Byn$Employment=='10-19'],Byn$Non.Byssinosis[Byn$Employment=='10-19'])
sum(Byn$Byssinosis[Byn$Employment=='>=20'],Byn$Non.Byssinosis[Byn$Employment=='>=20'])

sum(Byn$Byssinosis[Byn$Workspace==1],Byn$Non.Byssinosis[Byn$Workspace==1])
sum(Byn$Byssinosis[Byn$Workspace==2],Byn$Non.Byssinosis[Byn$Workspace==2])
sum(Byn$Byssinosis[Byn$Workspace==3],Byn$Non.Byssinosis[Byn$Workspace==3])
n<-165+5254

full.gl<-glm(cbind(Byssinosis,Non.Byssinosis)~Employment + Smoking+Sex+Race +Workspace
,family=binomial(link=logit),data=Byn)
summary(full.gl)

library(car)
vif(full.gl) #no multicollinearity
red.gl<-glm(cbind(Byssinosis,Non.Byssinosis)~Employment + Smoking + Workspace
,family=binomial(link=logit),data=Byn)
summary(red.gl)

#Likelihood Test for model reduction
L0 = logLik(red.gl)
L1 = logLik(full.gl)
LR.test = as.numeric(-2*(L0 - L1))
LR.pval = pchisq(LR.test, df = length(coefficients(full.gl)) - length(coefficients(red.gl)),lower.tail = F )
LR.test
LR.pval
```

```
log.ci<-cbind(red.gl$coefficients-2.576*summary(red.gl)$coef[,2],red.gl$coefficients+2.576*summary(red.gl)$coef[,2])
exp(log.ci)
library(vcdExtra)

Byn[,c(5:7)]
sum(Byn$Byssinosis[Byn$Workspace==1]);sum(Byn$Byssinosis[Byn$Workspace==2]);sum(Byn$Byssinosis[Byn$Workspace==3])

sum(Byn$Non.Byssinosis[Byn$Workspace==1]);sum(Byn$Non.Byssinosis[Byn$Workspace==2]);sum(Byn$Non.Byssinosis[Byn$Workspace==3])


wk.sp<-matrix(c(105,18,42,564,1282,3408),ncol=2)
colnames(wk.sp)<-c('Byn','NByn');wk.sp
library(vcdExtra)
CMHtest(wk.sp,rscores=c(1,2,3))

#sum(Byn$Byssinosis[Byn$Employment=='<10']);sum(Byn$Byssinosis[Byn$Employment=='10-19']);sum(Byn$Byssinosis[Byn$Employment=='>=20'])

#sum(Byn$Non.Byssinosis[Byn$Employment=='<10']);sum(Byn$Non.Byssinosis[Byn$Employment=='10-19']);sum(Byn$Non.Byssinosis[Byn$Employment=='>=20'])

emp.mat<-matrix(c(63,26,76,2666,686,1902),ncol=2)
colnames(emp.mat)<-c('Byn','NByn');emp.mat
CMHtest(emp.mat,rscores=c(1,2,3))

sum(Byn$Byssinosis[Byn$Smoking=='Yes']);sum(Byn$Non.Byssinosis[Byn$Smoking=='Yes'])
sum(Byn$Byssinosis[Byn$Smoking=='No']);sum(Byn$Non.Byssinosis[Byn$Smoking=='No'])
sm.mat<-matrix(c(125,3064,40,2190),ncol=2);rownames(sm.mat)<-c('Byn','N.Byn');colnames(sm.mat)<-c('Smk','N.smk')
sm.mat
chisq.test(sm.mat,correct = FALSE)
#sum(Byn$Byssinosis[Byn$Sex=='M']);sum(Byn$Non.Byssinosis[Byn$Sex=='M'])
#sum(Byn$Byssinosis[Byn$Sex=='F']);sum(Byn$Non.Byssinosis[Byn$Sex=='F'])

sex.m<-matrix(c(128,2788,37,2466),ncol=2);rownames(sex.m)<-c('Byn','N.Byn');colnames(sex.m)<-c('M','F')
sex.m

chisq.test(sex.m,correct = FALSE)
sum(Byn$Byssinosis[Byn$Race=='W']);sum(Byn$Non.Byssinosis[Byn$Race=='W'])
sum(Byn$Byssinosis[Byn$Race=='O']);sum(Byn$Non.Byssinosis[Byn$Race=='O'])

race.m<-matrix(c(92,3424,73,1830),ncol=2);rownames(race.m)<-c('Byn','N.Byn');colnames(race.m)<-c('W','O')
race.m
chisq.test(race.m,correct = FALSE)
names(Byn)
summary(full.gl) #AIC 165.95
drop1(full.gl,test = 'Chisq') #AIC 164.24
drop1(full.gl,test = 'Chisq',k=log(dim(Byn[1]))) #BIC 166.18

summary(red.gl)

#both tests model selection methods agree that sex and race should be dropped from the model
step(full.gl,scope=~Employment*Smoking*Workspace*Sex*Race,direction='both',trace = FALSE)
step(red.gl,scope=~Employment*Smoking*Workspace*Sex*Race,direction = 'both',trace = FALSE)


#the interaction model includes the following and provides the lowest AIC 160.7

int.red<-glm(cbind(Byssinosis, Non.Byssinosis) ~ Employment + Smoking + Sex +
    Workspace + Sex:Workspace,family=binomial(link=logit),data=Byn)

summary(int.red)


#install.packages('arm')
library(arm)

binnedplot(fitted(red.gl),residuals(red.gl,type="response"))
length(predict(red.gl))
length(dffits(red.gl))


plot(fitted(red.gl),predict(red.gl))


#pearson residuals (Reduced GLM before AIC/BIC)

binnedplot(fitted(red.gl),residuals(red.gl,type="response"))

pearson_vs_std<- cbind(rstandard(red.gl,type="pearson"), residuals(red.gl,type="pearson"),
residuals(red.gl,type="deviance"), rstandard(red.gl,type="deviance"))

colnames(pearson_vs_std) <-c("standardized","pearson","deviance","std. dev. residuals")
#head(pearson_vs_std)


summary(pearson_vs_std)
hist(residuals(red.gl,type="pearson"),main='Pearsons for Model1') #based on the pearson residual for the
reduced model before AIC selection, we can see that our residuals do not imply a poor model.
```

```
#pearson residuals (GLM w/ interaction and AIC/BIC)

binnedplot(fitted(int.red),residuals(int.red,type="response"))

pearson_vs_std.r<- cbind(rstandard(int.red,type="pearson"), residuals(int.red,type="pearson"),
residuals(int.red,type="deviance"), rstandard(int.red,type="deviance"))

colnames(pearson_vs_std.r) <-c("standardized","pearson","deviance","std. dev. residuals")
head(pearson_vs_std.r)

summary(pearson_vs_std.r)
hist(residuals(int.red,type="pearson"),main='Pearson Model2') #based on the pearson residual for the
reduced interaction model after AIC selection, contains
?sort

sort(pearson_vs_std.r[,2],decreasing = T)[1] #point 37 is a potential outlier of influence
confint(int.red,level=1-(0.05/18))
confint(red.gl,level=1-(0.05/12))
```