STA138_Final

Jongwook-Choe,Sungwon-Lee,Hyug-Her

3/18/2022

Tables of Contents

1. Introduction

2. Exploration

- Data set description
- Relationships between variables

3. Analysis

- Testing independency between variables
- Odd ratio testing
- Best fit modeling

4. Interpretation

5. Conclusion

6. Appendix

Introduction

Throughout our analysis, we will be using "Byssinosis.csv" dataset. It is a large data of the cotton textile company in North Carolina participating in a study to investigate the prevalence of byssinosis.

We will formulate a linear regression model to estimate an independent variable and, with statistical analysis, determine how well the variable is explained by the model. From various statistics including AIC and log likelihood, the probability of independent variables under the fitted model and LR test, we would test the fitted model is appropriate to the data. Throughout our research, we will be using Rstudio, a widely used statistical and graphing utility by writing a programming language

Exploration

In 1973, a large cotton textile company in North Carolina participated in a study to investigate the prevalence of byssinosis, a form of pneumoconiosis to which workers exposed to cotton dust are subject. Data was collected on 5,419 workers.

	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
...	...	...	...	...	...	...	...

Data set

There are total of 7 different variables and among them 2 are numerical variables and 5 are Categorical variables.

$$\text{Workspace} = \begin{cases} 3, & \text{least dusty} \\ 2, & \text{less dusty} \\ 1, & \text{most dusty} \end{cases} \tag{1}$$

$$\text{Employment} = \begin{cases} < 10, & \text{less than 10 years of employment} \\ 10 - 19, & \text{10 to 19 years of employment} \\ 20-, & \text{20 or more years of employment} \end{cases} \tag{2}$$

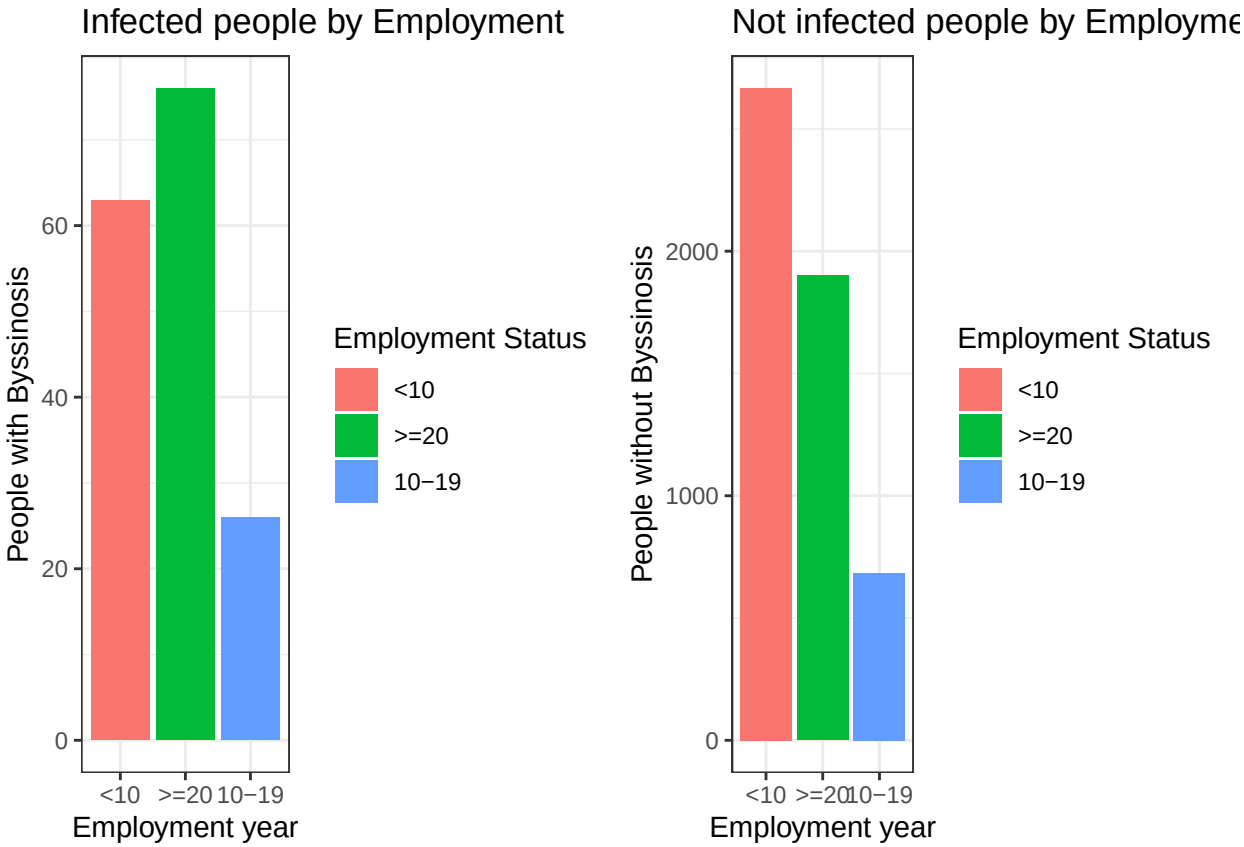
$$\text{Smoking} = \begin{cases} \text{No}, & \text{Not in last 5 years} \\ \text{Yes}, & \text{Smoker} \end{cases} \tag{3}$$

$$\text{Sex} = \begin{cases} \text{F}, & \text{Female} \\ \text{M}, & \text{Male} \end{cases} \tag{4}$$

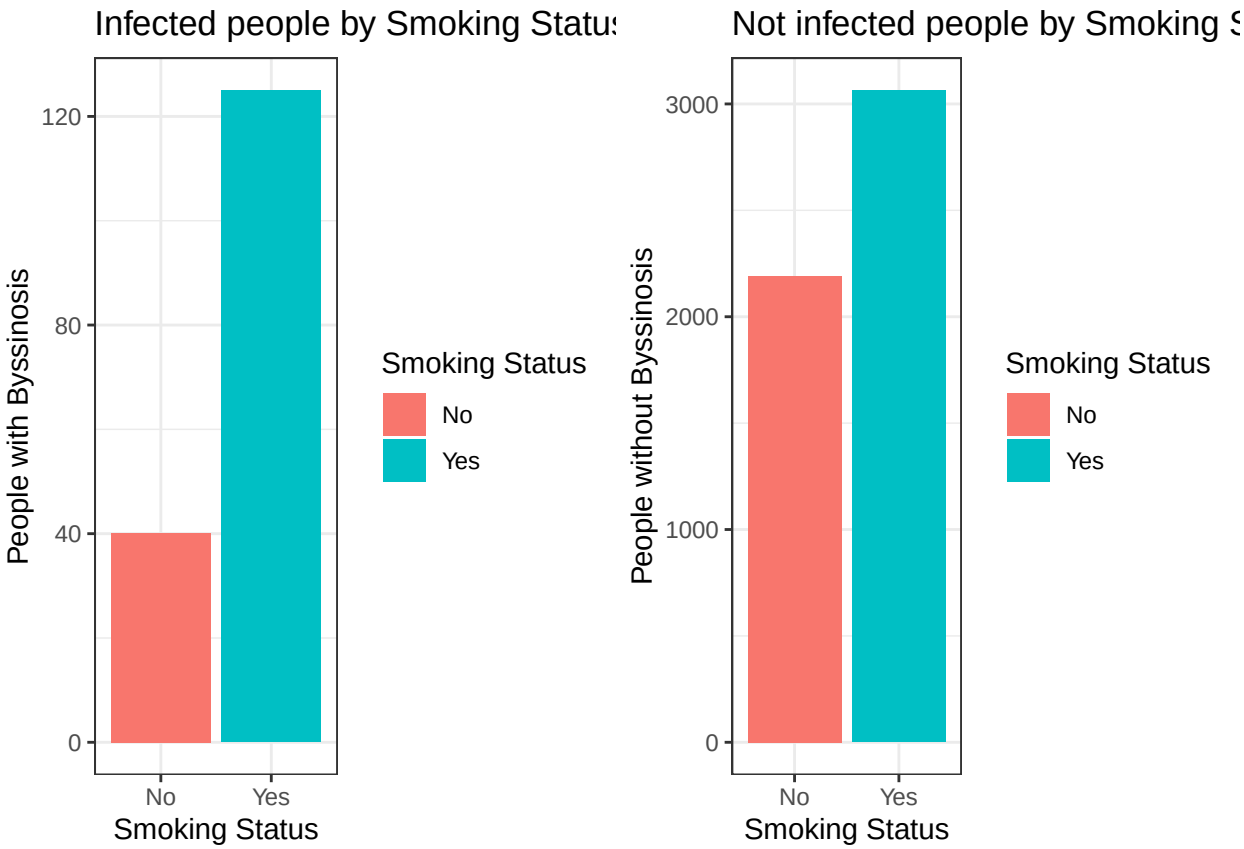
$$\text{Race} = \begin{cases} \text{O}, & \text{Other} \\ \text{W}, & \text{White} \end{cases} \tag{5}$$

Relationships between variables

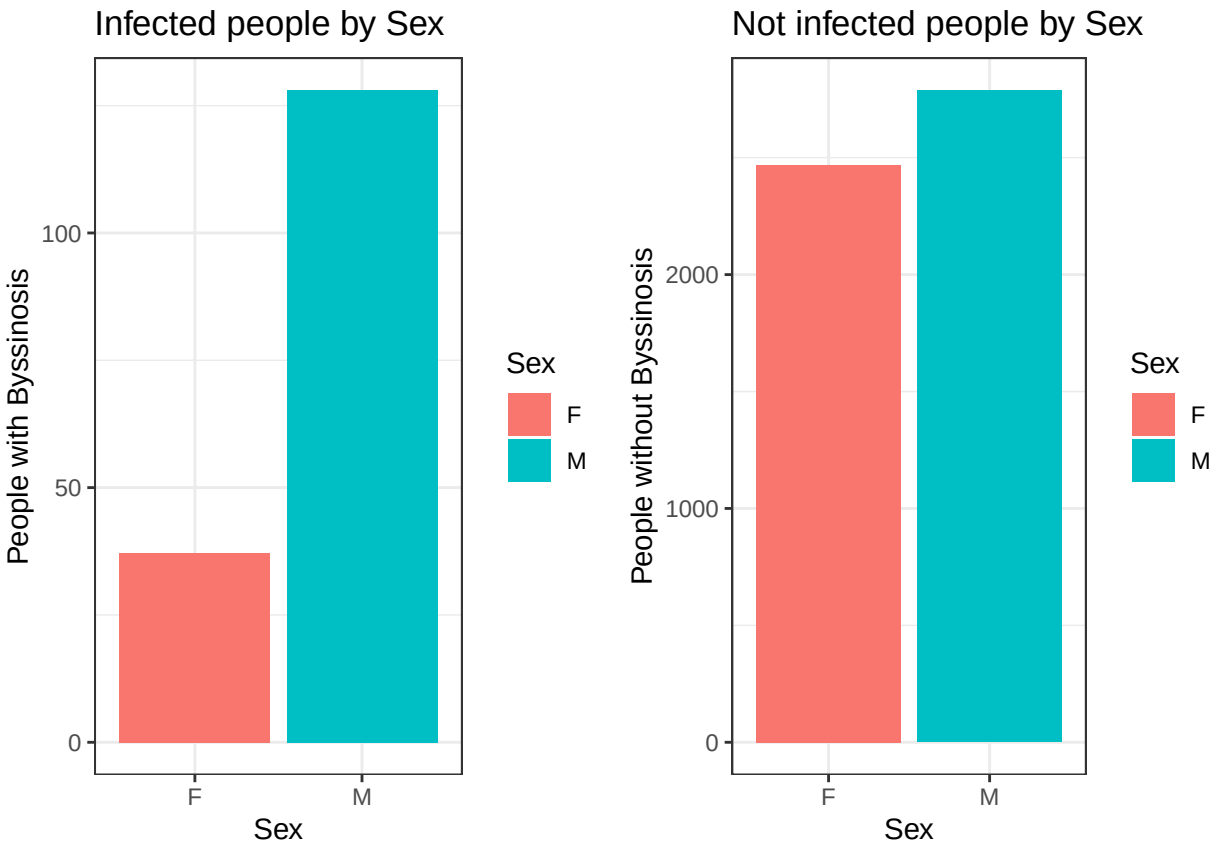
The following bar plots show the data trend of the variables with people who got infected and those who did not get infected.



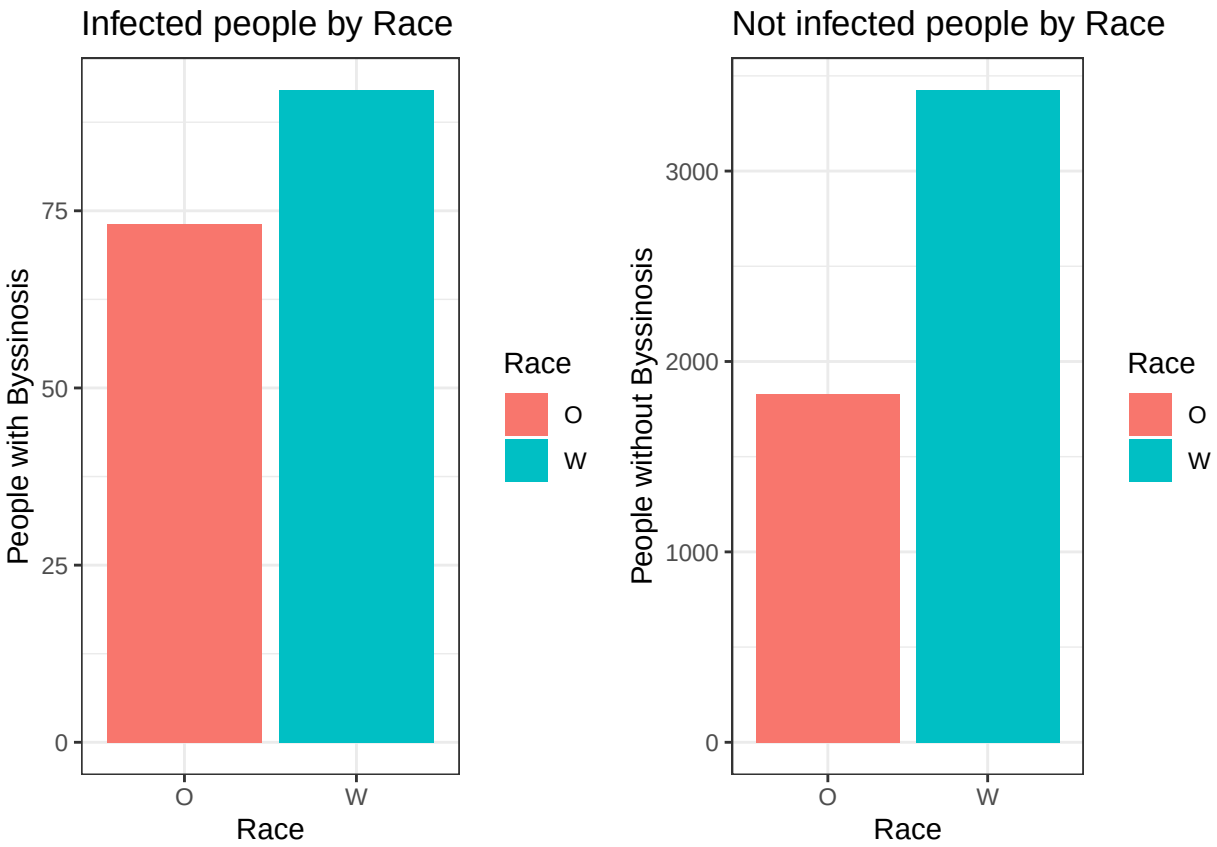
The first plot shows that workers who worked for more than 20 years tend to get more infected. While who worked 10-19 years showed the smallest number of Byssinosis infected cases. On the contrary who worked less than 10 years showed the largest number of not infected cases. The interesting fact is that 10-19 years workers showed the smallest number in both cases.



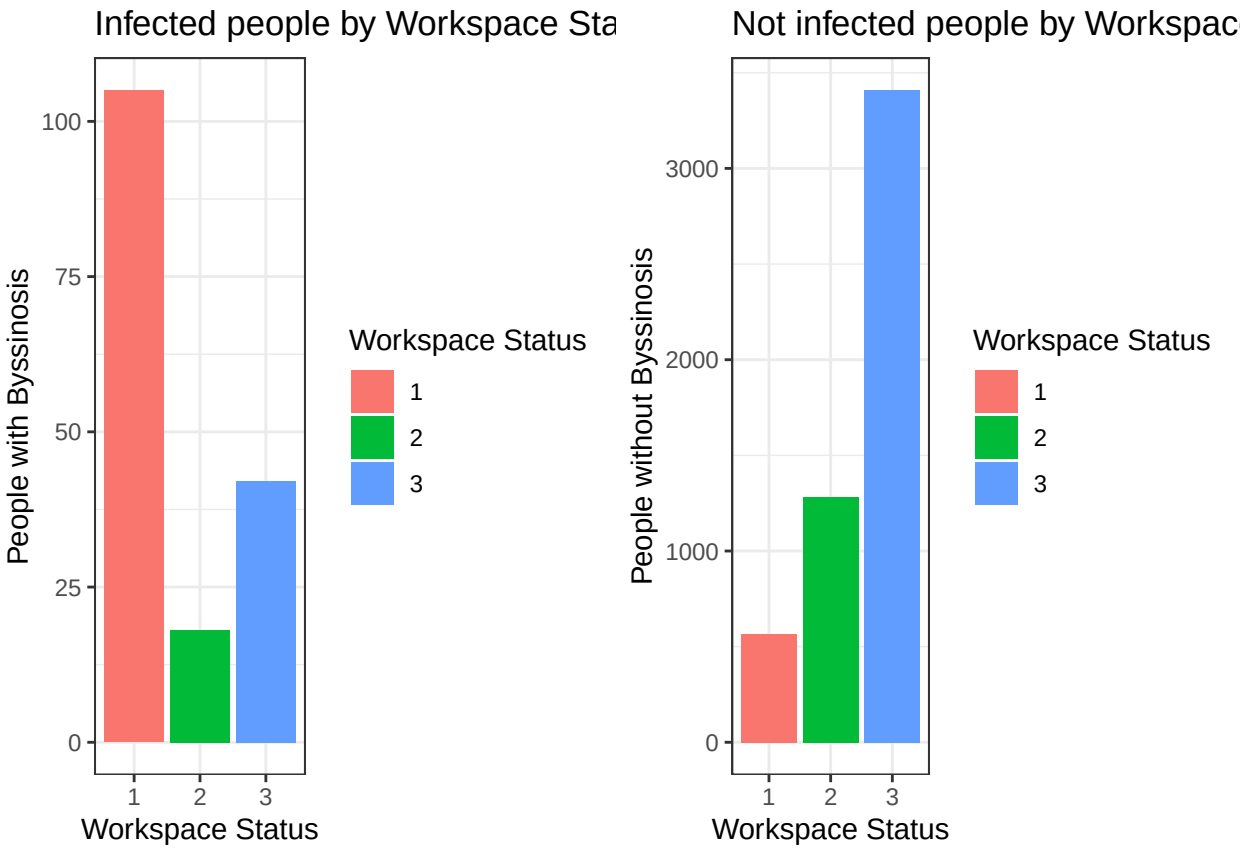
The second plot shows the smoking status of people who get infected and those who are not. Those who get infected had a large number of smokers while there were no big differences in the not-infected group.



Next is the plot of gender by the number of people who get infected or not. Those who are infected had a large number of males while in the not-infected group there was no significant difference. Based on the previous plot we could assume that most of the workers smoke and the majority of smokers are male, while the majority of non-smokers are female.



Not like previous plots, there was no such difference in people's color in the infected group. However, there was a small difference in the not-infected group. There was a larger number of white workers in the not-infected group.



Lastly, workers who work in the dustiest place showed a large number of Byssinosis-infected cases. On the contrary, the largest number of noninfected people is workers who work in the least dusty place. Based on the above bar plot we could assume that **workplace dustiness contributes to the chance of Byssinosis**.

Analysis

First, we will convert data from a wide format to a long format. Each row is a combination of all predictor variables and their counts for each binary category. In analysis rather than actually number we would like to see the relation between variables.

Wide form is useful when looking at multiple lines and series on a graph, or when making tables for quick comparison. On contrary long form is very close to the tidy format. It typically makes the data easy to store, and allows easy transformations to other types.

Testing indepedency between variables

When the sample size is small in 2x2 matrix using Fisher's exact test is more appropriate. However, we have two table that are 3X2 matrix therefore we will use Pearson test of independence.

Stating appropriate null and alternative hypothesis for independence test,

- H0 : Variables are independent
- HA : Variables are dependent

Now compute the value of Pearson χ^2 test statistic for testing the null hypothesis,

$$\chi^2 = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Now compare test-statistic with χ^2 distribution

$$\chi^2 \sim \chi^2_{df}$$

when $df = (\#\text{ columns } - 1) \times (\#\text{ rows } - 1)$

In result, the p-values of testing independence between Byssinosis with type of work place, employment years, smoking, sex, and race is following.

- $\text{p-value}_{\text{Employment}} = 0.734524$
- $\text{p-value}_{\text{Smoking}} = 0.6587742$
- $\text{p-value}_{\text{Sex}} = 0.889139$
- $\text{p-value}_{\text{Race}} = 0.4183977$
- $\text{p-value}_{\text{Workspace}} = 0.8318152$

Since all the p-values are larger than any possible α we fail to reject null. Then we conclude there are no correlations and variables are independent.

Calcilating the Odds Ratio

Following is the contingency table measure of association between an exposure and an outcome,

Also, we will calculate **Odd Ratio** which is used to find the probability of an outcome of an event when there are two possible outcomes and there is plausible casual effect. Since we discuss two possible outcomes for employment year and dustiness of work place we will exclude middle section.

$$\text{Odd ratio} = \frac{\text{Odd ratio}_{A \text{ for } B}}{\text{Odd ratio}_{A \text{ for } B^c}} = \frac{\frac{\frac{\pi_{11}}{\pi_{..}}}{1-\frac{\pi_{11}}{\pi_{..}}}}{\frac{\frac{\pi_{12}}{\pi_{..}}}{1-\frac{\pi_{12}}{\pi_{..}}}}$$

- odd ratio $= 1$ independent,
- odd ratio > 1 have more chance A then B
- odd ratio < 1 have more chance A then B^c

	Byssinosis	Non-Byssinosis
<10	17	24
10-19	10	19
>20	11	22

For Employment year status, the odd ratio of Byssinosis status is 1.4166667. This means workers who worked less than 10 years have more chance of having Byssinosis as compared to who worked more than 20 years which is quite odd after seeing plots.

	Byssinosis	Non-Byssinosis
Not Smoking	17	32
Smoking	21	33

Next, for smoking status, the odd ratio of Byssinosis status is 0.8362889. This means workers who smoke have more chances to have Byssinosis.

	Byssinosis	Non-Byssinosis
Female	17	30
Male	21	35

For gender status, the odd ratio of Byssinosis status is 0.9449986. This means male have more chances to have Byssinosis.

	Byssinosis	Non-Byssinosis
Other	15	31
White	23	34

For race status, the odd ratio of Byssinosis status is 0.7176236. This means white have more chances to have Byssinosis which is also quite odd after seeing plots.

	Byssinosis	Non-Byssinosis
Most dusty	13	19
Less dusty	11	22
Least dusty	14	24

Lastly, for workplace status, the odd ratio of Byssinosis status is 1.1729323. This means the Most dusty workplace workers have more chances of having Byssinosis.

Based on the odds and odds ratio, we can conclude that workers that are male, other races, smoking, and worked for more than 20 years in a most dusty workplace have a high chance of Byssinosis.

Best fit modeling

- (Byssinosis, Non.Byssinosis) ~ 1 → AIC : 442.93
- (Byssinosis, Non.Byssinosis) ~ Workspace → AIC : 224.43
- (Byssinosis, Non.Byssinosis) ~ Workspace + Smoking → AIC : 214.11
- (Byssinosis, Non.Byssinosis) ~ Workspace + Smoking + Employment → AIC : 206.1

Using stepwise forward AIC methods, we find the best fitted linear regression model which includes independent variables as Workspace, Smoking, and Employment . The fitted regression model has AIC as 206.1 which is the lowest among other regression models and this statistics infers that this regression model is well fitted in the data set. This infers that the probability of getting byssinosis would be affected by workspace, smoking, and employment.

Therefore, The best model according to forward step wise selection with selecting lowest AIC is,

$$\log(\frac{\pi}{1-\pi}) = \alpha + \beta_{\text{workspace}} x_{\text{workspace}} + \beta_{\text{smoking}} x_{\text{smoking}} + \beta_{\text{employment}} x_{\text{employment}}$$

Now using the model obtained above with the dataset, to test for evidence of nonzero coefficients

We state appropriate null and alternative hypothesis,

- $H_0 : \beta = 0$
- $H_A : \beta \neq 0$

coefficients:	Estimate	Std. Error	z value	p value
Intercept	-1.1858	0.2618	-4.530	5.9e-0.6
Employment>=20	0.6699	0.1793	3.735	0.000188
Employment10-19	0.5328	0.2465	2.162	0.030655
SmokingYes	0.6670	0.1892	3.526	0.000422
Workspace	-1.4663	0.1057	-13.869	<2e-16

If $\alpha = 0.1$ apply a Bonferroni correction, comparing each p-value to $0.1/3 \approx 0.033$. Even so, p-value is still less than Bonferroni correction. Thus we reject the null hypothesis that $\beta = 0$ and conclude there is a significant relationship between the variables in the linear regression model of the data set.

Therefore, the coefficients are nonzero and the best model according to forward stepwise selection with AIC is

$$\log(\frac{\pi}{1-\pi}) = \alpha + \beta_{\text{workspace}} x_{\text{workspace}} + \beta_{\text{smoking}} x_{\text{smoking}} + \beta_{\text{employment}} x_{\text{employment}}$$

Interpretation

First, we graph the plots of interaction between variables factor levels for visible inferences. From the plots, we observed that For each variable there were trends but some of them were not clear.

From the odds and odds ratio, we assume that workers that are male, of other races, smoking, and worked for more than 20 years in the dustiest workplace have a high chance of byssinosis.

To have the best fit model, we have used stepwise forward regression with AIC and tested this model with the saturated model of LR test, null deviance statistic, Chi-square p-value, and the difference between log-likelihood. This proves that this model is well fitted in the data set.

The fitted model, model proves that workspace, smoking status, and employment period directly affect the chance of Byssinosis. We could conclude that a **smoking person** working **more than 20 years** in a **dusty environment** would have the highest chance to have **Byssinosis** than other workers in the factory.

Conclusion

Relationships between this disease on the one hand and smoking status, sex, race, length of employment, smoking, and dustiness of workplace on the other were not intuitive by examining different plots. Also, some of the odd ratios were not as we expected. Therefore we tested the dependency of the variables and generated best fitted model using valid statistical tools to uncover meaningful associations in the data.

Appendix

```
library(gridExtra)
library(ggplot2)
library(tidyr)
df=read.csv("Byssinosis.csv")
plot1 = ggplot(df, aes(x = factor(Employment), y = Byssinosis, fill = factor(Employment))) +
geom_bar(stat="identity") + ylab('People with Byssinosis') +
labs(title = 'Infected people by Employment', x = 'Employment year')+theme_bw()+
  guides(fill = guide_legend(title="Employment Status"))
plot2 = ggplot(df, aes(x = factor(Employment), y = Non.Byssinosis, fill = factor(Employment))) +
geom_bar(stat="identity") + ylab('People without Byssinosis') +
labs(title = 'Not infected people by Employment', x = 'Employment year')+theme_bw()+
  guides(fill = guide_legend(title="Employment Status"))

plot3 = ggplot(df, aes(x = factor(Smoking), y = Byssinosis, fill = factor(Smoking))) +
geom_bar(stat="identity") + ylab('People with Byssinosis') +
labs(title = 'Infected people by Smoking Status', x = 'Smoking Status')+theme_bw()+
  guides(fill = guide_legend(title="Smoking Status"))
plot4 = ggplot(df, aes(x = factor(Smoking), y = Non.Byssinosis, fill = factor(Smoking))) +
geom_bar(stat="identity") + ylab('People without Byssinosis') +
labs(title = 'Not infected people by Smoking Status', x = 'Smoking Status')+theme_bw()+
  guides(fill = guide_legend(title="Smoking Status"))
```

```
plot5 = ggplot(df, aes(x = factor(Sex), y = Byssinosis, fill = factor(Sex))) +
geom_bar(stat="identity") + ylab('People with Byssinosis') +
labs(title = 'Infected people by Sex', x = 'Sex')+theme_bw()+
  guides(fill = guide_legend(title="Sex"))
plot6 = ggplot(df, aes(x = factor(Sex), y = Non.Byssinosis, fill = factor(Sex))) +
geom_bar(stat="identity") + ylab('People without Byssinosis') +
labs(title = 'Not infected people by Sex', x = 'Sex')+theme_bw()+
  guides(fill = guide_legend(title="Sex"))

plot7 = ggplot(df, aes(x = factor(Race), y = Byssinosis, fill = factor(Race))) +
geom_bar(stat="identity") + ylab('People with Byssinosis') +
labs(title = 'Infected people by Race', x = 'Race')+theme_bw()+
  guides(fill = guide_legend(title="Race"))
plot8 = ggplot(df, aes(x = factor(Race), y = Non.Byssinosis, fill = factor(Race))) +
geom_bar(stat="identity") + ylab('People without Byssinosis') +
labs(title = 'Not infected people by Race', x = 'Race')+theme_bw()+
  guides(fill = guide_legend(title="Race"))

plot9 = ggplot(df, aes(x = factor(Workspace), y = Byssinosis, fill = factor(Workspace))) +
geom_bar(stat="identity") + ylab('People with Byssinosis') +
labs(title = 'Infected people by Workspace Status', x = 'Workspace Status')+theme_bw()+
  guides(fill = guide_legend(title="Workspace Status"))
plot0 = ggplot(df, aes(x = factor(Workspace), y = Non.Byssinosis, fill = factor(Workspace))) +
geom_bar(stat="identity") + ylab('People without Byssinosis') +
labs(title = 'Not infected people by Workspace Status', x = 'Workspace Status')+theme_bw()+
  guides(fill = guide_legend(title="Workspace Status"))

grid.arrange(plot1, plot2, ncol=2)
grid.arrange(plot3, plot4, ncol=2)
grid.arrange(plot5, plot6, ncol=2)
grid.arrange(plot7, plot8, ncol=2)
grid.arrange(plot9, plot0, ncol=2)
longdf=gather(df,Condition,Number,Byssinosis:Non.Byssinosis,factor_key = TRUE)
ndf=longdf[!(longdf$Number==0),]
table1=table(ndf$Employment,ndf$Condition)
table2=table(ndf$Smoking,ndf$Condition)
table3=table(ndf$Sex,ndf$Condition)
table4=table(ndf$Race,ndf$Condition)
table5=table(ndf$Workspace,ndf$Condition)
the.test = chisq.test(table1,correct = FALSE,simulate.p.value=TRUE)
eij = round(the.test$expected,digits=2)
disc = (table1-eij)^2/eij
discsum = sum(disc)
pval1 = 1-pchisq(discsum,2)

the.test = chisq.test(table2,correct = FALSE,simulate.p.value=TRUE)
eij = round(the.test$expected,digits=2)
disc = (table2-eij)^2/eij
discsum = sum(disc)
pval2 = 1-pchisq(discsum,1)

the.test = chisq.test(table3,correct = FALSE,simulate.p.value=TRUE)
eij = round(the.test$expected,digits=2)
```

```
disc = (table3-eij)^2/eij
discsum = sum(disc)
pval3 = 1-pchisq(discsum,1)

the.test = chisq.test(table4,correct = FALSE,simulate.p.value=TRUE)
eij = round(the.test$expected,digits=2)
disc = (table4-eij)^2/eij
discsum = sum(disc)
pval4 = 1-pchisq(discsum,1)

the.test = chisq.test(table5,correct = FALSE,simulate.p.value=TRUE)
eij = round(the.test$expected,digits=2)
disc = (table5-eij)^2/eij
discsum = sum(disc)
pval5 = 1-pchisq(discsum,2)
odd1=(table1[1,1]*table1[2,2])/(table1[2,1]*table1[1,2])
odd2=unname(fisher.test(table2)$estimate)
odd3=unname(fisher.test(table3)$estimate)
odd4=unname(fisher.test(table4)$estimate)
odd5=(table5[1,1]*table5[3,2])/(table5[3,1]*table5[1,2])
result <- step(glm(cbind(Byssinosis, Non.Byssinosis)~1,binomial,df),
               scope = ~Employment*Smoking*Sex*Race*Workspace,
               k=log(32),
               trace=0,
               directioin='forward')
myGlm <- glm(cbind(Byssinosis, Non.Byssinosis)~Employment+Smoking+Workspace, family=binomial, df)
```