Final Project

STA 138 | Camden Possinger | Winter 2022

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, 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]

Exploratory Data Analysis

Before we start our formal analysis of what potential variables are associated with Byssinosis let's take an exploratory look at our data. The first thing to notice is that a majority of the data are individuals who did not get Byssinosis.

```
bys_data <- read.csv("/home/cam/Documents/STA 138/Project/Byssinosis.csv")

# Create longer data that is easier to plot
plot_data <- bys_data %<>% pivot_longer(cols = c(Byssinosis,Non.Byssinosis)) %>% filter(value != 0)

# Split data based on Byssinosis
plot_data_bys <- plot_data %>% filter(name == "Byssinosis", value != 0)
plot_data_no_bys <- plot_data %>% filter(name == "Non.Byssinosis", value != 0)
```

```
data.frame("Byssinosis" = plot_data_bys$value %>% sum,
           "No_Byssinosis" = plot_data_no_bys$value %>% sum) %>% kbl
```

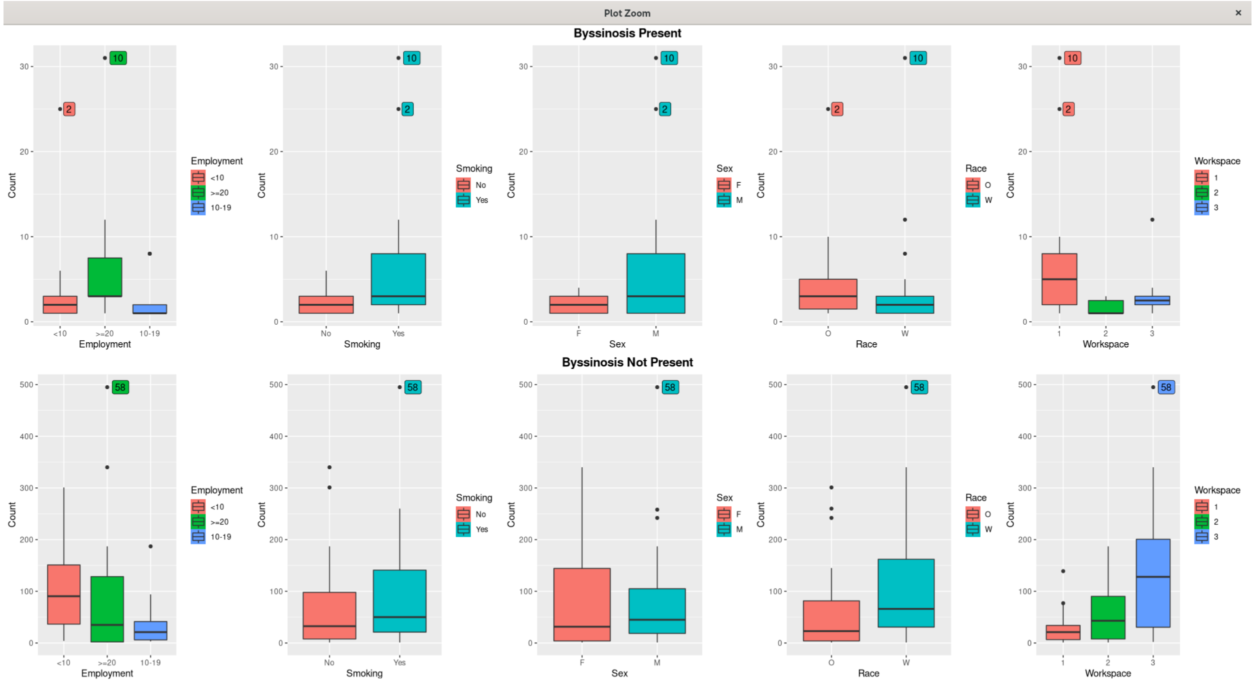
Byssinosis	No_Byssinosis
165	5254

Keeping that in mind let's plot the frequencies of these categorical variables for both individuals who had Byssinosis and those who did not. I'm going to plot them separately because the scale for these two cases is very different. Here we're going to use the purrr package to problematically create these plots and the ggpubr package merge them together.

```
is_outlier <- function(x) {
  return(x < quantile(x, 0.25) - 1.5 * IQR(x) | x > quantile(x, 0.95) + 1.5 * IQR(x))
}

plot_data_bys %<>% mutate(outlier = ifelse(is_outlier(value),
                                           rownames(plot_data_bys),
                                           as.numeric(NA)))
```

```r
plot_data_no_bys %<>% mutate(outlier = ifelse(is_outlier(value),
                                              rownames(plot_data_no_bys),
                                              as.numeric(NA)))
```

```r
plot_list_bys <- colnames(plot_data_bys)[1:5] %>%

  map(~ ggplot(data = plot_data_bys,
               aes(x = factor(!!sym(.x)),
                   y = value,
                   fill = factor(!!sym(.x))))+
        geom_label(aes(label = outlier),
                   na.rm = TRUE,
                   hjust = -0.3)+
        geom_boxplot()+
        xlab(.x)+
        ylab("Count")+
        labs(fill = .x)
      )

plot_list_no_bys <- colnames(plot_data_no_bys)[1:5] %>%

  map(~ ggplot(data = plot_data_no_bys,
               aes(x = factor(!!sym(.x)),
                   y = value,
                   fill = factor(!!sym(.x))))+
        geom_label(aes(label = outlier),
                   na.rm = TRUE,
                   hjust = -0.3)+
        geom_boxplot()+
        xlab(.x)+
        ylab("Count")+
        labs(fill = .x)
      )


bys_plots    <- ggarrange(plotlist = plot_list_bys,ncol = 5) %>%
                        annotate_figure(top = text_grob("Byssinosis Present",
                                                        face = "bold",
                                                        size = 14))

no_bys_plots <- ggarrange(plotlist = plot_list_no_bys,ncol = 5) %>%
                        annotate_figure(top = text_grob("Byssinosis Not Present",
                                                        face = "bold",
                                                        size = 14))


# Don't want to deal with latex and R markdown image size formatting
# Going to include a screenshot of this result
#ggarrange(bys_plots,no_bys_plots,nrow = 2)
```



A couple things in this plot are worth highlighting:

- In the top right plot where Byssinosis is present and Workspace is on the x axis it's clear that the first and dustiest workspace contains more individuals who had Byssinosis than the other less dusty workspaces. We can see in the plot below it where Byssinosis was not present there is an almost increasing linear trend between the different workspaces. We should explore this relationship in our formal analysis.
- A majority of individuals who got Byssinosis had been working for more than 20 years. However we don't see a similar pattern with individuals who did not get Byssinosis. The number of individuals that worked more than 20 years is similar to the number of individuals who had worked less than 10 years. This might have an association with Byssinosis, but it might not it's hard to tell.
- Smoking, Sex, and Race don't have large differences and probably do not have associations with Byssinosis.
- There are 3 outliers that are notable, let's explore those further

```
plot_data_bys[10,1:7] %>% kbl
```

Employment	Smoking	Sex	Race	Workspace	name	value
>=20	Yes	M	W	1	Byssinosis	31

```
plot_data_bys[2,1:7] %>% kbl
```

Employment	Smoking	Sex	Race	Workspace	name	value
<10	Yes	M	O	1	Byssinosis	25

```
plot_data_no_bys[58,1:7] %>% kbl
```

Employment	Smoking	Sex	Race	Workspace	name	value
>=20	Yes	M	W	3	Non.Byssinosis	495

An interesting thing to point out is that these outliers are pretty similar to each other in terms of demographics. Especially the 1st and 3rd rows are exactly the same except for the Workspace. The individuals who worked in the dustier environment contracted Byssinosis and those who worked in the least dusty environment did not contract Byssinosis. It's also interesting to point out that in the 2nd row displayed that individuals who were not white and worked less then 10 years still contracted Byssinosis in this outlier case. These outliers weaken the argument that Employment time and Race play a role in understanding the cause of Byssinosis cases, but strengthens the argument for dusty vs. not dusty workspaces.

Formal Analysis

We're interested in possible associations between individuals who contracted Byssinosis and various demographics such as their race, smoking status, sex, employment time, what workspace they worked in, and any possible interaction effects. To analyze these possible associations we're going to use logistic regression and test for non zero model coefficients.

First let's get our data into long format

```
 model_data <- data.frame()

for(row_num in 1:(plot_data %>% nrow)){

  row <- plot_data[row_num,]

  name  <- row$name
  value <- row$value

  if(name == "Byssinosis"){

  new_row <- row %>% select(-name) %>% mutate(value = 1)

  }else{

  new_row <- row %>% select(-name) %>% mutate(value = 0)

  }

  for(value in 1:row$value){

   model_data %<>% rbind(new_row)

  }

}


model_data %<>%  lapply(factor) %>% as.data.frame

model_data %<>% mutate("Byssinosis" = value) %>% select(-value)

model_data_bys <- model_data %>% filter(Byssinosis == 1)
model_data_no_bys <- model_data %>% filter(Byssinosis == 0)
```

Now we can split our data into training and test sets. In this process I want to make sure that individuals who contracted Byssinosis are in both training and test sets. It is likely that when we split the training and test sets that no individuals who contracted Byssinosis are represented in those sets. Here I'm going to put 50% of the 162 individuals who contracted Byssinosis and put them in the training set and the remaining 50% in the test set. The same is done with the individuals who did not contract Byssinosis. This guarantees that both the training and testing models will be able to pick up the possible signal from the minority class.

```
set.seed(42)

train_ind_bys <- sample(model_data_bys %>% nrow, (model_data_bys %>% nrow) * 0.5)

train_bys <- model_data_bys[train_ind_bys,]
test_bys <- model_data_bys[-train_ind_bys,]


train_ind_no_bys <- sample(model_data_no_bys %>% nrow, (model_data_no_bys %>% nrow) * 0.5)

train_no_bys <- model_data_no_bys[train_ind_no_bys,]
test_no_bys <- model_data_no_bys[-train_ind_no_bys,]


train_set <- train_bys %>% rbind(train_no_bys)
test_set <- test_bys %>% rbind(test_no_bys)
```

```
best_model_train <- step(glm(Byssinosis~1,family = "binomial",data = train_set),
                   scope = ~Employment*Smoking*Sex*Race*Workspace,
                   direction = "forward",
                   k = log(2709),
                   trace = 0)


print(best_model_train$formula)
```

```
## Byssinosis ~ Workspace
```

The step function found that the model with Workspace has the lowest BIC

```
bys_final_model_sum <- glm(Byssinosis ~ Workspace, family = "binomial",data = test_set) %>% summary

bys_final_model_sum
```

```
##
## Call:
## glm(formula = Byssinosis ~ Workspace, family = "binomial", data = test_set)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6140  -0.1522  -0.1522  -0.1476   3.0083
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.5731     0.1468 -10.713  < 2e-16 ***
## Workspace2   -2.9409     0.4074  -7.219 5.25e-13 ***
## Workspace3   -2.8801     0.2686 -10.723  < 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: 742.08  on 2709  degrees of freedom
## Residual deviance: 594.70  on 2707  degrees of freedom
## AIC: 600.7
```

```
##
## Number of Fisher Scoring iterations: 7
```

Workspace is significant when fit with the test set, so we can conclude that these model coefficients are not zero at significance level 0.017 which is $\frac{0.05}{3}$ and that there is an association between the dust level of the workspace and contracting Byssinosis.

Now that we know that there is a significant association between the dust levels of workspaces and contracting Byssinosis let's interpret the model parameter estimates.

```r
data.frame("Dust_Level" = c("Most Dusty", "Less Dusty","Least Dusty"),
           "Odds_of_Byssinosis" = bys_final_model_sum$coefficients[,1] %>% exp %>% unname) %>% kbl
```

Dust_Level	Odds_of_Byssinosis
Most Dusty	0.2074074
Less Dusty	0.0528169
Least Dusty	0.0561284

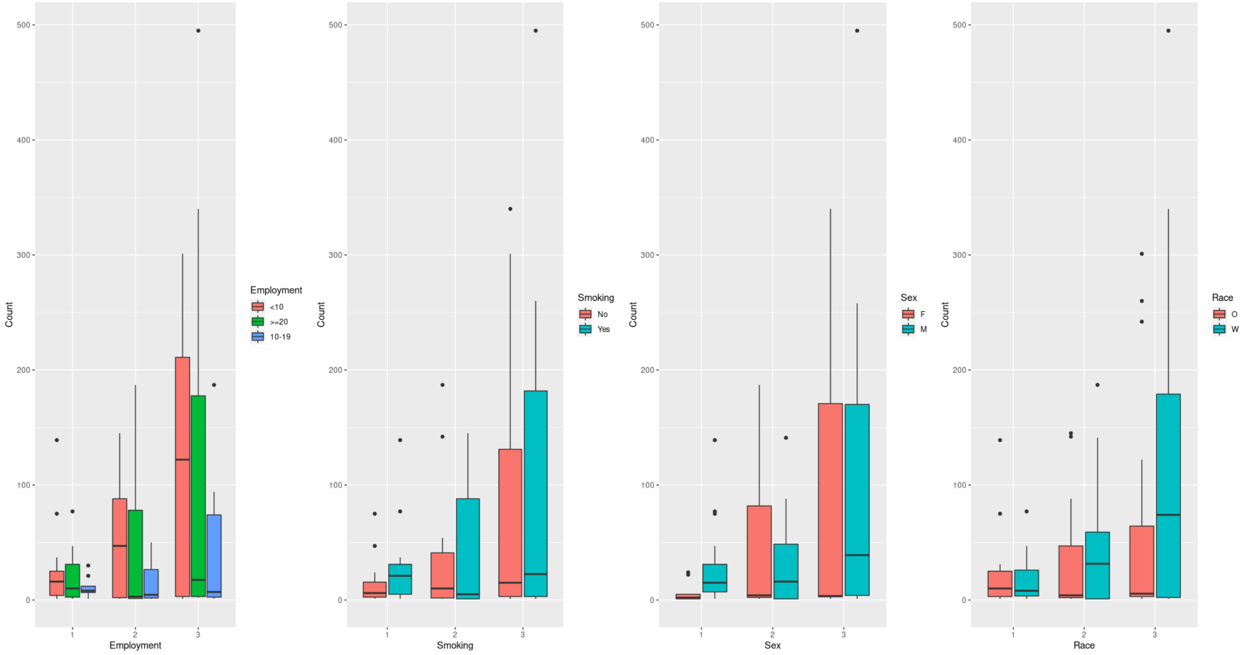
```r
most_dusty <- bys_final_model_sum$coefficients[1,1] %>% exp
less_dusty <- bys_final_model_sum$coefficients[2,1] %>% exp
least_dusty <- bys_final_model_sum$coefficients[3,1] %>% exp



data.frame("Dust_Level" = c("Most Dusty vs. Less Dusty",
                           "Most Dusty vs. Least Dusty",
                           "Less Dusty vs. Least Dusty"),

           "Odds_Ratio_of_Byssinosis" = c(most_dusty / less_dusty,
                                          most_dusty / least_dusty,
                                          less_dusty / least_dusty)) %>% kbl
```

Dust_Level	Odds_Ratio_of_Byssinosis
Most Dusty vs. Less Dusty	3.9269136
Most Dusty vs. Least Dusty	3.6952318
Less Dusty vs. Least Dusty	0.9410016

Here we can see that the odds of Byssinosis increases by about 15% compared to the odds of Byssinosis for the less dusty workplace and least dusty workplace. We can also see that the odds for contracting Byssinosis increases by over 300% for the most dusty workspace compared to the workspaces with less dust. We can also see that there is a 6% decrease in odds of Byssinosis between the less dusty workspace and the least dusty workspace. We can say here that the risk is not linear among the different workspaces the most dusty workspace is where people get Byssinosis.

But who works in these different workspaces? Are there any patterns that indicate who typically worked in these different workspaces? Let's take an initial look at various boxplots for the three different workspaces and the other demographic variables.

```r
workspace_plot_list <- colnames(plot_data)[1:4] %>%
  map(~ ggplot(data = plot_data,
               aes(x = factor(Workspace),
                   y = value,
                   fill = factor(!!sym(.x))))+
        geom_boxplot()+
        xlab(.x)+
        ylab("Count")+
        labs(fill = .x))


workspace_plots <- ggarrange(plotlist = workspace_plot_list,ncol = 4)
```



One observation that stands out to me is that there are considerably more white individuals in the least dusty workspace than individuals of other races.

```
percent_white <- ((model_data %>%

  filter(Workspace == "3", Race == "W") %>% nrow)/(model_data %>% filter(Race == "W") %>% nrow)) %>%
  round(2) %>%
  multiply_by(100) %>%
  as.character

percent_other <- ((model_data %>%

  filter(Workspace == "3",Race == "O") %>% nrow)/(model_data %>% filter(Race == "O") %>% nrow)) %>%
  round(2) %>%
  multiply_by(100) %>%
  as.character


data.frame("Race" = c("Other","White"),
           "Percent_In_Least_Dusty_Workspace" = c(paste0(percent_other,"%"),
                                                  paste0(percent_white,"%"))) %>% kbl
```

Race	Percent_In_Least_Dusty_Workspace
Other	55%
White	68%

Let's see if this relationship is expressed in the following logistic regression model. First we need to resample our training and test sets because we're not concerned with if an individual contracted Byssinosis.

```
set.seed(42)
model_data_dummy <- model_data %>%
  cbind("Workspace3" = model.matrix(~0+.,data=model_data)[,"Workspace3"])


train_ind <- sample(model_data_dummy %>% nrow,
                    (model_data_dummy %>% nrow) * 0.5)

train_set <- model_data_dummy[train_ind,]
test_set <- model_data_dummy[-train_ind,]
```

```
best_model_train <- step(glm(Workspace3~1,family = "binomial",data = train_set),
                   scope = ~Employment*Smoking*Sex*Race,
                   direction = "forward",
                   k = log(2709),
                   trace = 0)


print(best_model_train$formula)
```

```
## Workspace3 ~ Race
```

```
ws_final_model_sum <- glm(Workspace3 ~ Race ,family = "binomial",data = test_set) %>% summary

ws_final_model_sum
```

```
##
## Call:
## glm(formula = Workspace3 ~ Race, family = "binomial", data = test_set)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.522  -1.295   0.868   0.868   1.065
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.27136    0.06501   4.174 2.99e-05 ***
## RaceW        0.51077    0.08298   6.155 7.49e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3529.1  on 2709  degrees of freedom
## Residual deviance: 3491.4  on 2708  degrees of freedom
## AIC: 3495.4
##
## Number of Fisher Scoring iterations: 4
```

```
white <- (ws_final_model_sum$coefficients[1,1] +  ws_final_model_sum$coefficients[2,1]) %>% exp
other <- (ws_final_model_sum$coefficients[1,1]) %>% exp

 data.frame("Individual" = c("White","Other"),
           "Odds_of_Least_Dusty" = c(white,other)) %>% kbl
```

Individual	Odds__of__Least__Dusty
White	2.186131
Other	1.311751

```
data.frame("Individual" = c("White vs. Other"),
           "Odds_Ratio_of_Least_Dusty" = c(white/other)) %>% kbl
```

Individual	Odds__Ratio__of__Least__Dusty
White vs. Other	1.666575

Looking at the above odds ratio of working in the least dusty workspace we can see that the odds increase by about 67% if an individual is white compared with a person of a different race. This shows that white people were more likely to work in the least dusty environment compared with people of other races. This is evidence that the textile company did not evenly allocate individuals to each type of work space potentially disproportionately exposing people of color to dustier environments which leads to the contraction of Byssinosis.