**STA 138** Final Project

Prevalence of Byssinosis

12/9/22

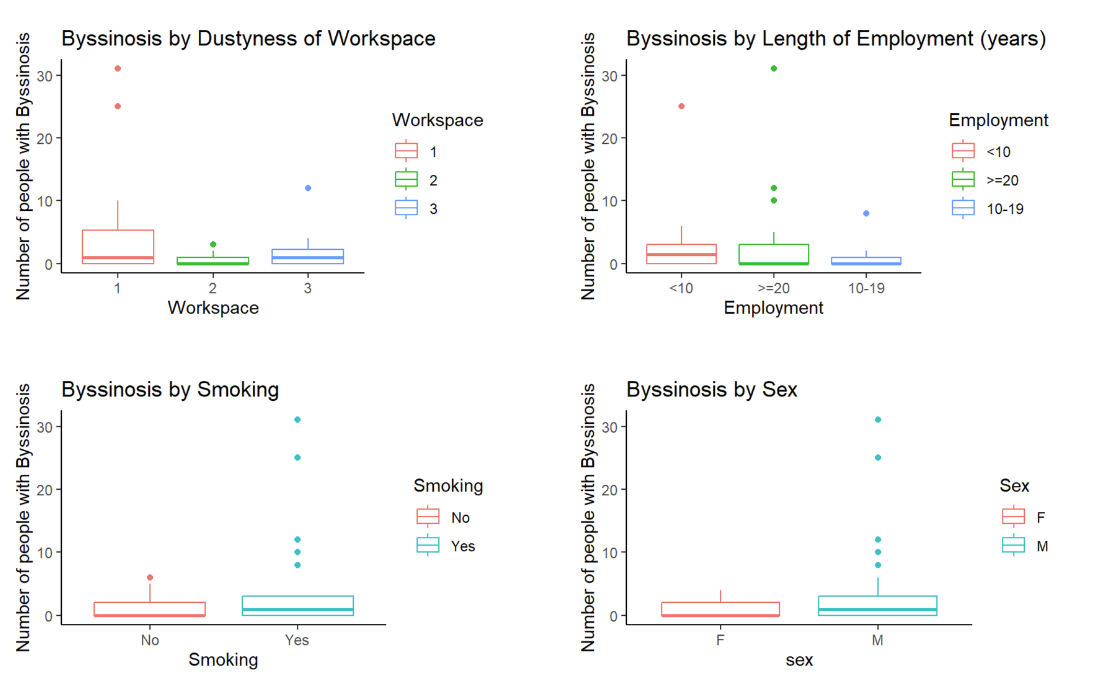
Aurora Travers, Cormac O’Sullivan, Diego Alatorre, Kayla Pham, Rachel How

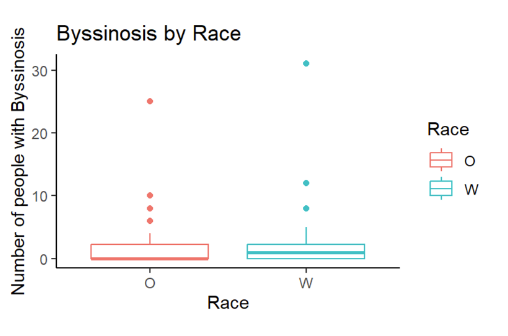
# Introduction

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. Our task is to investigate relationships between this disease on the one hand and smoking status, sex, race, length of employment, smoking, and dustiness of workplace (on a scale of 1 to 3) on the other. In addition, we want to create a statistical model that best fits Byssinosis and find which of these possible predictor variables have a meaningful association and could possibly contribute to the likelihood of having Byssinosis.

# Data Preprocessing

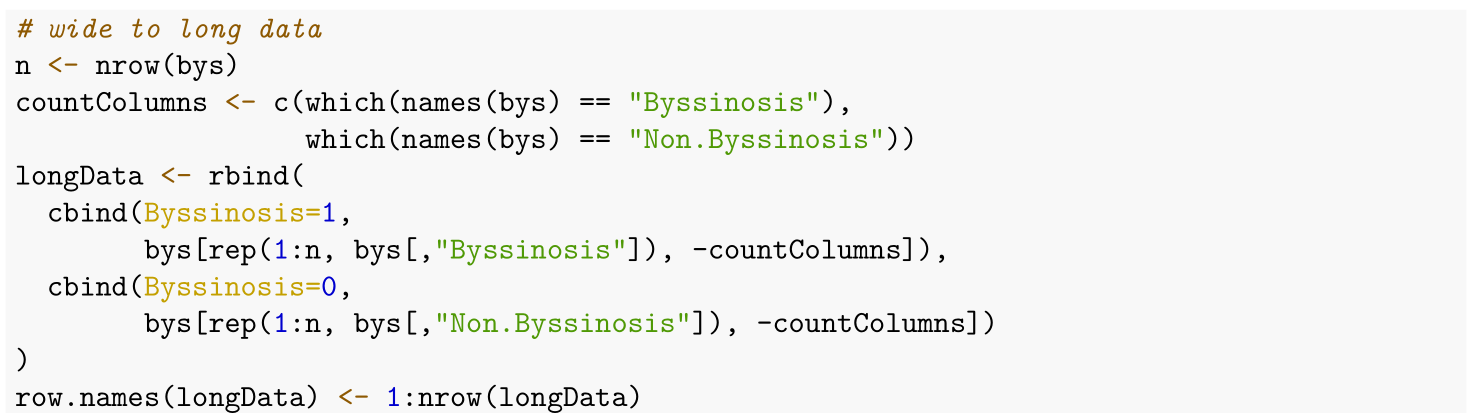
In order to visualize our data, we first made a boxplot of workers with Byssinosis against all predictor variables, giving us five separate boxplots. As we can see in the boxplots, there is a clear difference in the type of workspace and length of employment. We can also tell there is a slight difference in smoking and sex, but we can clearly see that there is negligible difference in race. From only looking at the boxplot, we can predict the variable race wouldn’t affect the model. Also, it is important to note that although we have a dataset of 5,419 workers, only 164 of those have Byssinosis. Therefore, these boxplots only represent a sample of 164 workers with Byssinosis.



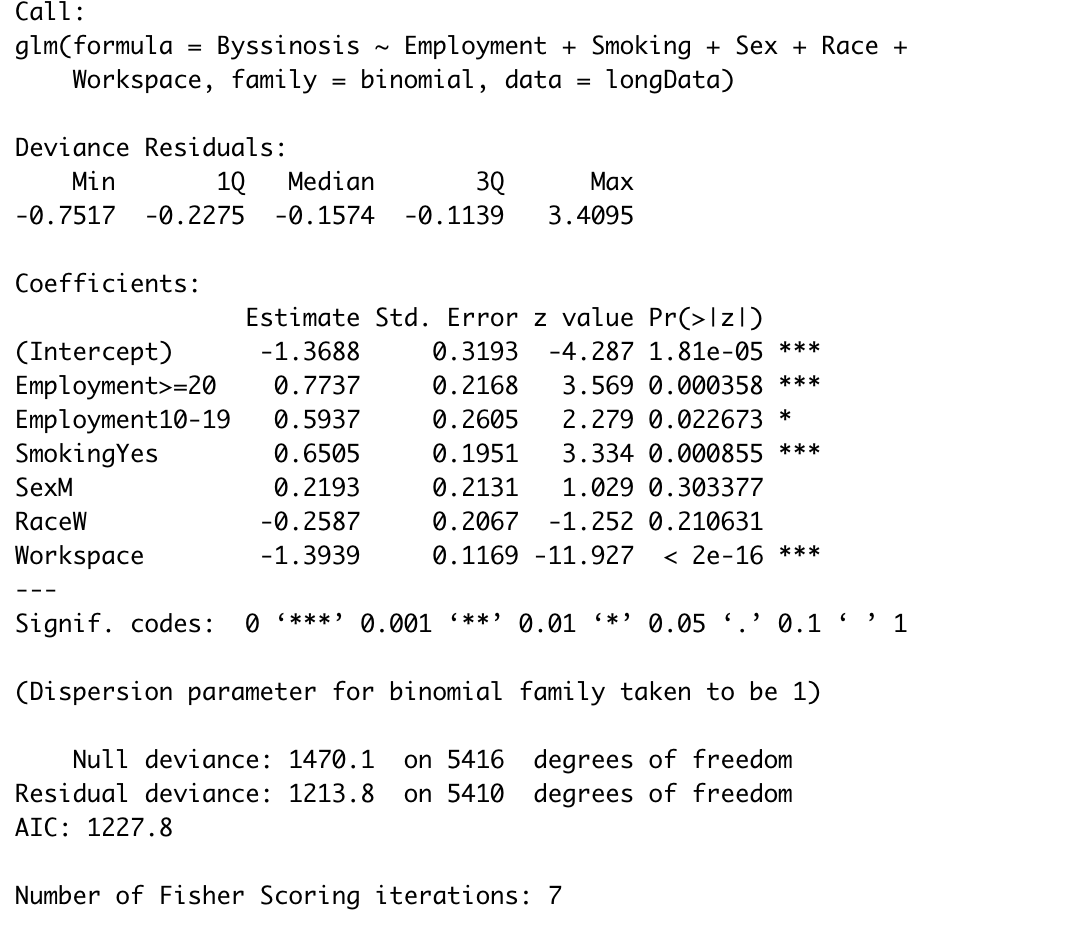


**Figure 1**

We went on and changed the data format. Byssinosis is intended to be a binary data, but it was given as a numerical data. We changed this wide data into a long data format, with 5,417 rows. The code we used is shown below.

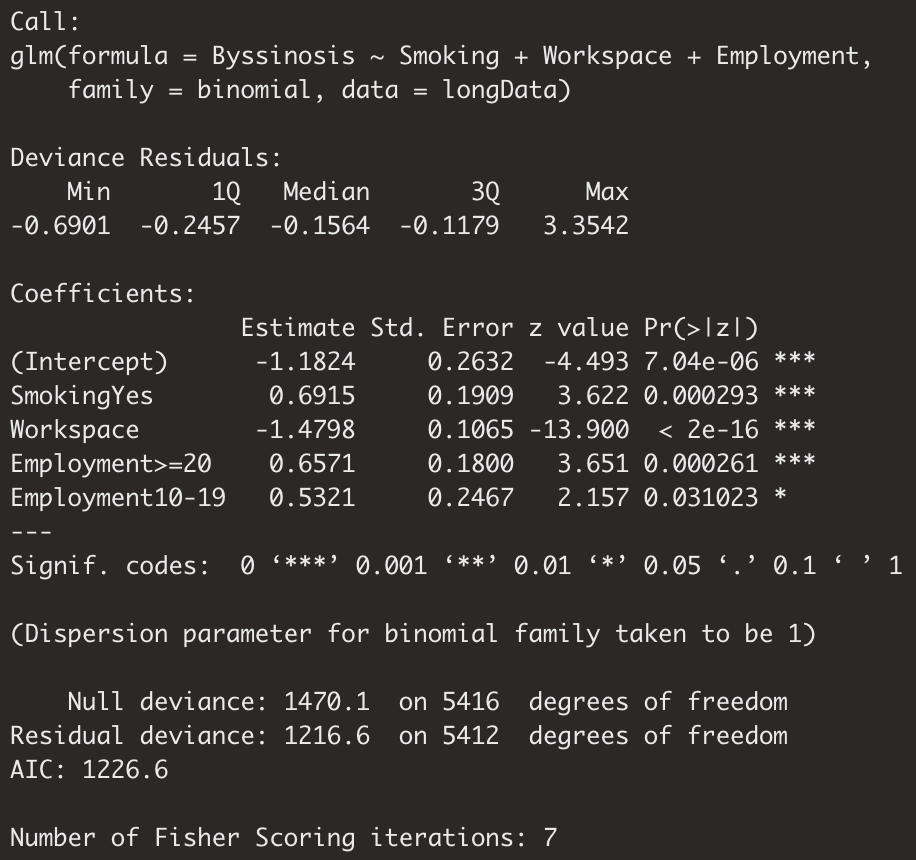


1. Exploratory Data Analysis

To better understand the relationship between byssinosis and the other variables given in the dataset, we want to fit a statistical model. First, we conducted Wald test, using a significance level of 0.05, to see what variables are significant in our model for predicting Byssinosis. From the Wald test, we concluded that Sex and Race were insignificant to our model. The variable Sex had a p-value of 0.303377 and race had a p-value of 0.210631, so we fail to reject the null hypothesis at a significance level of 0.05. In other words, we cannot rule out the possibility of having estimated coefficients equal to zero for these two variables. Therefore, we dropped those two variables and used Workspace, Smoking, and Employment to fit our model. The results are shown below.

In order to find a good predictive model, we conducted subset selection. We started with a model without predictors, only the intercept. We then added more predictors by calculating the AIC. We added the variables that had the smallest AIC across the models with combinations of our predictor variables. We decided to use AIC over BIC due to the nature of our data. AIC tends to be less harsh in penalizing a more complex model and will more often leave more variables and their interactions within the model. While AIC does have a chance of overfitting, it is preferable that we do not leave out any significant terms because doing so could put people in harm's way. For example, not including something like workplace dustiness and concluding it doesn't have an effect on contracting byssinosis is much more harmful than if we included it.

The AIC corroborates our results from the Wald test in that the best fitted model uses Workspace, Smoking, and Employment as predictor variables. After getting the significant variables, we can find the coefficients that best fit our model.



From the information given above, we can create a fitted model:

The fitted model suggests that the chance of presence of Byssinosis increases if the worker smokes and the longer the worker has been employed. Since Employment>=20 has a coefficient larger than Employment 10-19, that means that workers that have been employed for 20 years or greater have a slightly higher chance of Byssinosis than workers employed between 10 to 19 years. In addition, a decrease in workspace score (meaning the more dusty the workspace is) can also increase the chance of presence of Byssinosis.

# Conclusion

We concluded that workplace dustiness, smoking status, length of employment are all significant variables that contribute to the chance of Byssinosis. When predicting our model, we found that Sex and Race are not significant variables that contributes to Byssinosis. Workspaces that are most dusty contribute to the chance of Byssinosis, and workspaces that are the least dusty are less likely to contribute to the chance of Byssinosis. In conclusion, in order for workers to reduce risk of having Byssinosis in the workspace, workers should not smoke and prevent working in a dusty workspace.