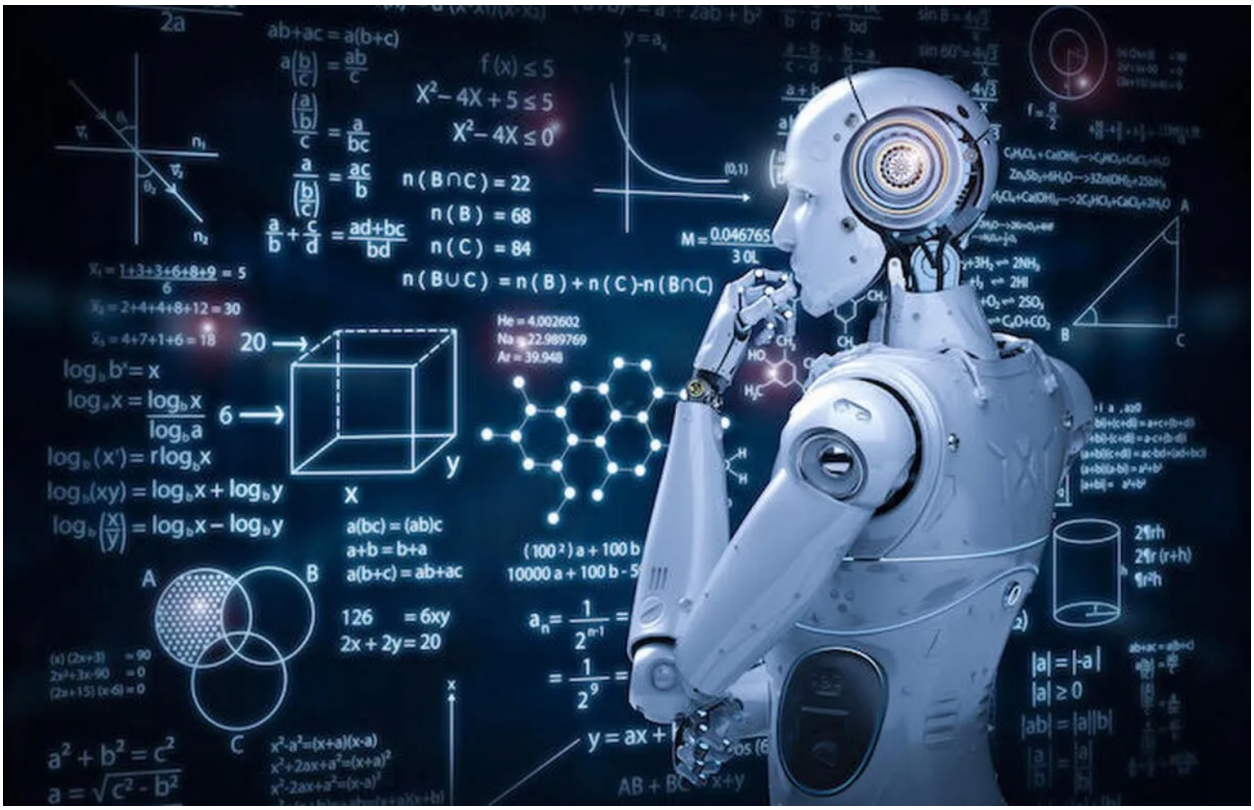
STA 138 Final Project: Model Selection

Masha Omelchak, Chase Robinson, Jack Skinner, Mo Grewal, Karla Cornejo Argueta

December 13, 2024



Introduction

Byssinosis, also known as Monday fever, is a lung disease, caused by inhalation of cotton and other vegetable fibers. CDC has estimated that during the early 70s in the U.S. 20% of cotton workers were evidenced to have byssinosis. In modern days, employers are advised to improve ventilation and provide staff with respiratory protection. In this assignment, we are diving back in time to a 1974 North Carolina cotton factory, in order to explore the dataset and build a logistic regression model with Byssinosis. Our primary goal is to outline risk factors and discover the effects of intersectionality on Byssinosis diagnosis.

Data

The data is provided through the STA138 course on Canvas, in the .csv wide format. It consists of 5,418 observations (employees), with 6 categorical variables: Employment Length, Smoking, Sex, Race, Workplace dustiness level, and Diagnosis of Byssinosis. Variables such as Smoking, Sex, Race, and Diagnosis of Byssinosis are Bernoulli variables, and Employment Length with Workplace dustiness are Multinomial variables, with 3 classification categories each. Since they are categorical,the distribution of each variable has been visualized through pie charts (Figures 1 - 6). In this data set, employees with Byssinosis represent only 3% of the overall population, which is the response variable for our prediction modeling, meaning that in this population $P(Byssinosis) = 0.03$. Some other distributions of interest are Workplace levels: 1 making up 12.35%, level 2 at 24%, and level 3 dustiness with 63.65%. The sex distribution of employees is nearly 50/50, with race being categorized at 64.88% as white and 35.12% as other. Distribution of smokers is also nearly 50/50, which is not surprising for 1970's. Finally, workers employed for under 10 years make up 50.37% of the sample, those working 10-19 years are 13.14% and employees above 20 years are 36.49%.

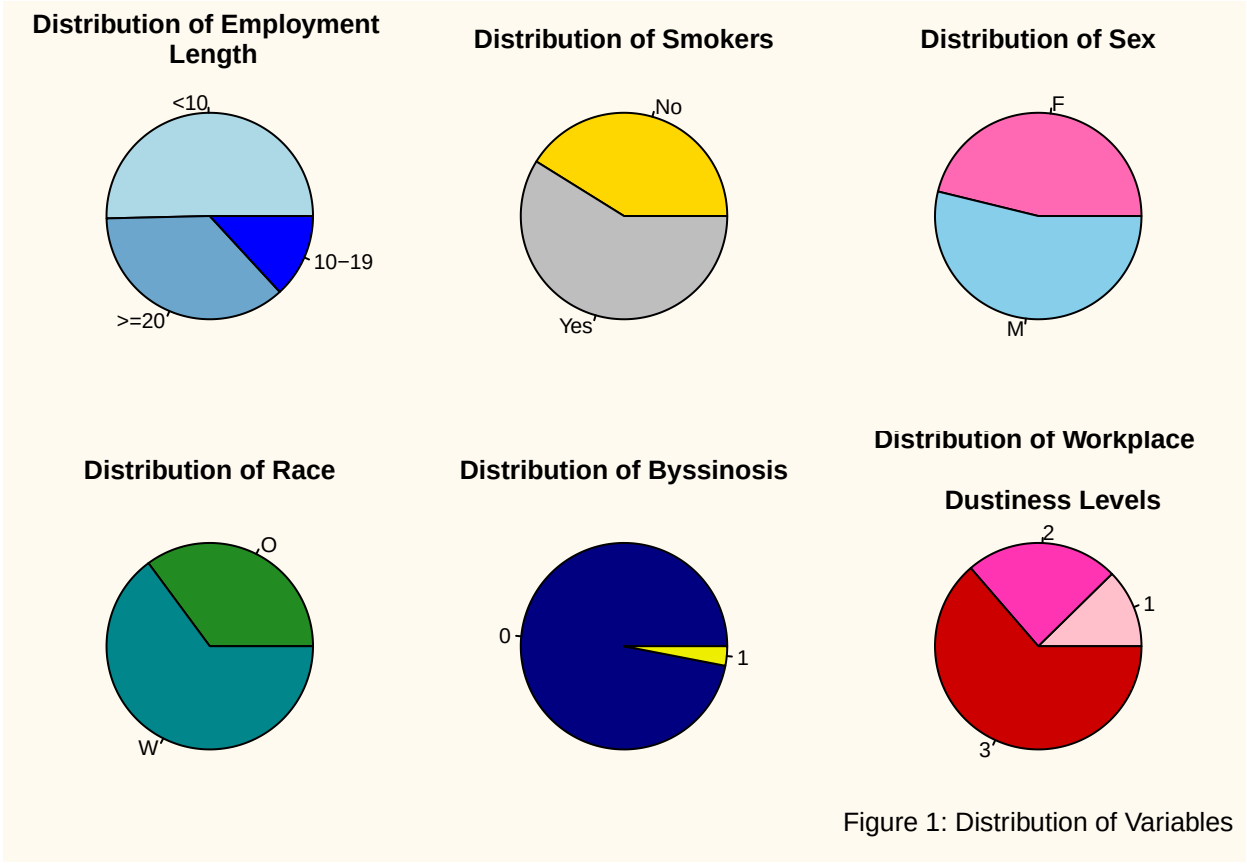


Figure 1: Distribution of Variables

Additionally, we created contingency tables for Byssinosis and each of the right-hand variables and used the Chi-Squared test to find which groups Byssinosis was dependent on. Workplace resulted in the highest X-statistic of 413.71, suggesting that the probability of Byssinosis is heavily dependent on how dusty the workplace is. Sex, Smoking, Employment, and Race, had X-statistics of 37.66, 19.355, 10.177, and 5.804 respectively, which all suggest dependence at a 0.05 significance level.

Table 1: Byssinosis vs Employment

	<10	>=20	10-19
0	2666	1901	686
1	63	76	26

[a] X-squared = 10.177, p-value = 0.006

Table 2: Byssinosis vs Smoking

	No	Yes
0	2189	3064
1	40	125

[a] X-squared = 19.355, p-value = 1.085e-05

Table 3: Byssinosis vs Sex

	F	M
0	2465	2788
1	37	128

[a] X-squared = 37.66, p-value = 8.42e-10

Table 4: Byssinosis vs Race

	O	W
0	1830	3423
1	73	92

[a] X-squared = 5.804, p-value = 0.016

Table 5: Byssinosis vs Workplace

	1	2	3
0	564	1282	3407
1	105	18	42

[a] X-squared = 413.71, p-value = 2.2e-16

Methodology

The given was already cleaned, but required re-formatting from wide to long. This was done with the intention of randomly spitting observations into an 80% training (ntrain = 4334) and 20% testing (ntest= 1084) sets. We also re-formatted Employment so we could have a clean, numeric factor; the new indexing for this variable: 1 = less than 10 years, 2= 10-19 years and 3 = over 19 years of employment at the cotton factory. Since every variable is categorical, quadratic terms will not be included, because they are unidentified, and therefore do not contribute towards the model. This leaves us with a linear additive model and usage of interaction terms. Due to the size of our group, we split into AIC and BIC approaches for binomial and poisson regressions, which allowed us to examine the data from different angles and compare the criteria using an identical training set. Poisson regression was selected as a model due to the skewed distribution of Byssinosis, following the convergence of Binomial distribution to Poisson when p is small (.03) and n is large (5,418). We began the process using forward stepwise regression, starting with an empty model and then comparing its fit using AIC or BIC to each additive term or interaction, stopping the process when the AIC or BIC stopped decreasing.

Model Selection

Stepwise Regression Under AIC

AIC has a factor of $k = 2(p+1)$, where $p+1$ is the number of parameters added to the deviance of the models which accounts for small complexity. It was applied first because BIC always prefers a simpler model, and we needed to observe the possible complexity of the output before simplification. To fit our models using AIC we did forward stepwise regression starting with the first model including only each individual variable: Employment, Smoking, Sex, Race, Workspace or None. From this step we concluded that Workspace leads to a better model, so we continued including Workspace as a variable in our model. Then, we proceeded to include the rest of the variables as additive terms and concluded Smoking is also a factor. Afterward, we compared to other additive variables and concluded Employment is also a factor, but race and sex have no significant relation to Byssinosis. Then, we considered interactions between Workspace and Employment, Workspace and Smoking, and Employment and Smoking. We concluded that there is an interaction between Workspace and Employment but not with Smoking. We obtain the following model:

$$y = -2.0839 - 0.9618x_1 + 0.8687x_2 + 0.4814x_3 - 0.3445x_4$$

Where, x_1 represents workspace which takes on values 1-3 to represent dustiness (very dusty, somewhat dusty, not dusty). x_2 represents employment which takes on values 1-3 to represent years employed at their place of work (<10, 10-19, >19). x_3 represents smoking which takes on values 0 or 1 to represent whether or not someone smokes. x_4 represents the interaction of workspace and employment, increasing the risk of angina conditionally on the increased dustines and length of employment together.

Interpretation and application

Since Workplace and Employment factors cannot be omitted (they're never 0), the base model in our logistic regression model relies on the non-smoking status of the workers. To clarify the impact of the variable, by holding Workplace and Employment constant, we found the odds ratio of byssinosis in a smoker v.s. non-smoker to be 1.618. Our discovery points to an increased chance of lung disease in smokers, keeping all other terms constant.

Since our data converges to Poisson Distribution, we conducted the same steps above under a Poisson Regression model and found similar results. Except we found sex to be a factor and this model included no interactions. Our final model under Poisson is:

$$y = -2.0552 - 0.9697x_1 + 0.7237x_2 + 0.4379x_3 - 0.2871x_4$$

Where, x_1, x_2, x_3, x_4 are the same as under Logistic Regression.

Table 6: Final Selected Variables in Stepwise Regression Models

Variable	AIC	Poisson_AIC	BIC	Poisson_BIC
Workspace	Yes	Yes	Yes	Yes
EmploymentFactor	Yes	Yes	No	No
SmokingYes	Yes	Yes	No	No
Workspace:EmploymentFactor	Yes	Yes	No	No

Stepwise Regression Under BIC

We followed a similar process under BIC, but BIC considers complexity in terms of the data size by adding a factor or $k = log(n)$ where n is the sample size of our training data which is 4,334. We started with the simplest models and compared the BIC of single variables, we concluded that Workspace is a factor. Here we tried other additive terms but concluded that Workspace as a single factor gives the best model.

Our final model is:

$$y = -0.2298 - 1.5762x_1$$

Where x_i is the different types of workspace which takes on values 1-3 to represent dustiness (very dusty, somewhat dusty, not dusty).

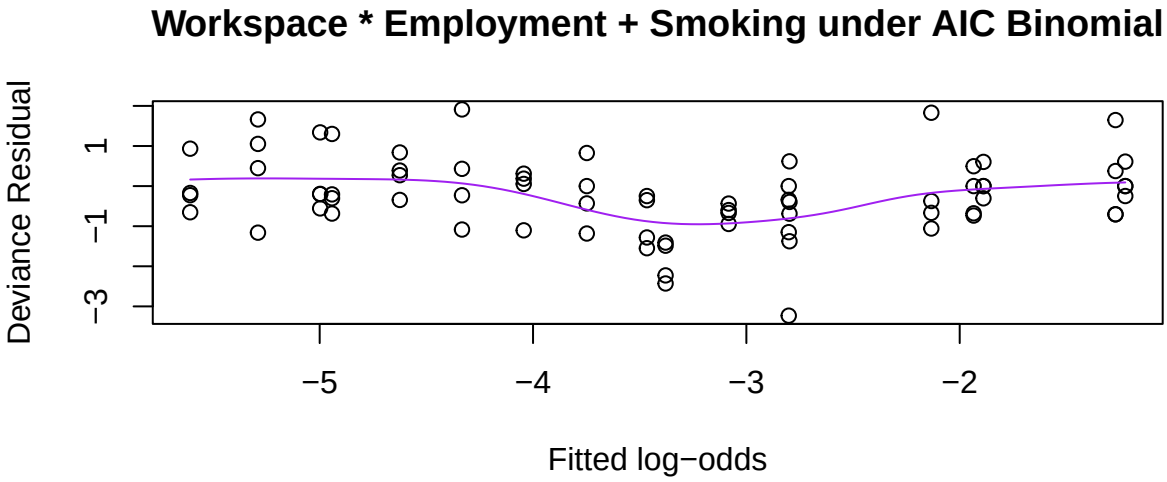
Under Poisson, we get a similar model with $y = -0.4573 - 1.4940x_1$

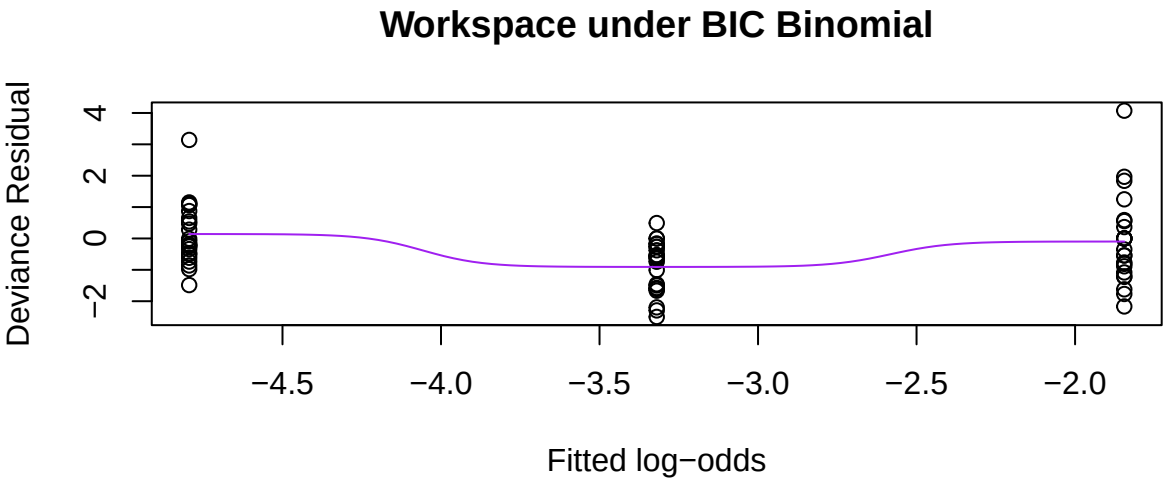
Interpretation and application

Unlike the model fitted using AIC, this model prioritizes simplicity and thus omits terms that do not make large improvements in making predictions. As a result, the odds ratio of Byssinosis between workers is determined by their Workspace only.

Using Residuals

We decided to see the residuals in our model to note anything out of the ordinary, but since residual plots can be ambiguous we do not get any new information. We can see that both models have deviance residuals generally close to 0. The deviations present in their curves are not significant enough to suggest poor model fit.





Prediction

Table 7: AIC Confusion Table

	0	1
0	1047	37
1	0	0

[a] Accuracy: 0.9658672

Table 8: BIC Confusion Table

	0	1
0	1047	37
1	0	0

[a] Accuracy: 0.9658672

Since we split the data, we want to test our model into the new data. Since under both Poisson and Binomial Distribution we got a similar model under AIC and BIC, we decided we would choose one distribution for prediction. We used logistic regression under Binomial distribution to classify the data. Then, we used confusion matrices to evaluate how accurate our models are at predicting Byssinosis. Having a high accuracy would indicate having a reliable model for making predictions. In these tables, having class = 0 means not having Byssinosis, while class = 1 means having Byssinosis. Both models have the same accuracy of 0.9658 which is due to a phenomenon related to Byssinosis having a low probability of occurring.

Discussion

Nonparametric models are used when the assumptions needed for the model aren't met. For generalized linear models, the big assumption that runs the risk of being violated is the link function used is correct. Since the residuals were normally distributed and the variance was homogeneous throughout the model, we can assume that the link function is accurate, and that there is no need to use nonparametric models. We used logistic regression over linear regression because our variables were largely categorical rather than numeric. In the case of workspace, there was an undefined distance between classes meaning we could not use the same techniques. One limitation of our model is that we only used forward stepwise regression. We may have fitted a different model under backward stepwise regression or a combination of both. Additionally, there may be other variables that improve predictions of Byssinosis which were not accounted for by the data. For example, the location of each worker as well as any pre-existing conditions among workers.

Conclusion

Although logistic regression is built on a Binomial model, the low occurrence of the lung disease in the sample population (p = 0.03), conditioned our model on the more accessible factor of absence of Byssinosis (y=0) in workers. We were successfully able to outline risk factors of Byssinosis in cotton factory workers through forward stepwise binomial and poisson regressions using AIC and BIC criteria. Each model overlapped on one factor, Workspace, referring to the level of dustiness in a given factory. BIC for both regression models agreed on Workspace as the only contribution variable, whereas AIC models overlapped on the inclusion of Workspace, Employment and Smoking. Poisson regression additionally included Sex as a contributing variable, where the Binomial regression added an interaction term between Wrokspace and Employment. Each model was trained on the same 80% training data, and upon fitting the models and finding their confusion matrices, every model equally demonstrated an accuracy of 96.59%. This result may be due to the inability of our models to make predictions of class 1, or the relatively low sample size for a prediction problem. Another explanation for this phenomenon could be the convergence of Poisson and Binomial to the sample model due to a constant in Byssinosis probability (p=0.03).

Sources

Corn JK. Byssinosis–an historical perspective. Am J Ind Med. 1981;2(4):331-52. doi: 10.1002/ajim.4700020405. PMID: 7048909.

Code Appendix

Important Variables:

train represents the 80% of data selected for training the model

test represents the 20% of data selected for testing the model

finalboss_aic represents the fitted model using AIC with the binomial family

finalboss_bic represents the fitted model using BIC with the binomial family

finalboss_poisson represents the fitted model formed using AIC with the Poisson family

finalboss_bic_p represents the fitted model formed using BIC with the Poisson family

```
library(tidyverse)
library(dplyr)
library(knitr)
library(psych)
library(gam)
library(corrr)
library(ggplot2)
library(gt)
library(flextable)
library(kableExtra)
library(gridExtra)
library(ggplot2)
library(grid)
library(caret)

## Pivot the wide data to long format to subset test-train data
## using na.omit to clena the data up from possible NAs
widedf <- read.csv("/Users/karlacornejoargueta/Downloads/Byssinosis.csv")
longdf <- pivot_longer(widedf, cols = c('Byssinosis','Non.Byssinosis'),
                       names_to = 'Byssinosis') %>%
  uncount(value) %>%
  mutate(Byssinosis = ifelse(Byssinosis == "Byssinosis",1,0)) %>% na.omit

longdf$EmploymentFactor = as.numeric(as.factor(longdf$Employment))


## doing a 80% training and 20% testing division, because this project focuses
## more on the exploratory position

set.seed(87)
smpl <- floor(.80 * nrow(longdf))
cut_it_up <- sample(seq_len(nrow(longdf)), size = smpl)

train <- longdf[cut_it_up,]
test <- longdf[-cut_it_up,]
#exploring the distribustions of variables
## input plots into 2x3 matrix
par(mfrow=c(2,3), bg = 'floralwhite', mar= c(3,0,3,0), cex.axis = 3,
    cex.lab = 2)

## creating pie charts of each variable
pie(table(longdf$Employment), main = "Distribution of Employment \n Length",
```

```
    col = c('lightblue', 'skyblue3', 'blue'))

pie(table(longdf$Smoking), main = 'Distribution of Smokers',
    col = c('gold', 'gray'))

pie(table(longdf$Sex), main = "Distribution of Sex",
    col = c('hotpink', 'skyblue'))

pie(table(longdf$Race), main = "Distribution of Race",
    col = c('forestgreen', 'turquoise4'))

pie(table(longdf$Byssinosis), main = "Distribution of Byssinosis",
    col = c('navy', 'yellow2'))

pie(table(longdf$Workspace), main = "Distribution of Workplace \n
    Dustiness Levels", col = c('pink', 'maroon1', 'red3'))
## adding a caption to image
mtext("Figure 1: Distribution of Variables", side=1, line=1, cex=0.8)
## computing the chi square for each variable
## testing independence
table.1 <- table(longdf$Byssinosis, longdf$Employment)
chisq.test(table.1)
table.2 <- table(longdf$Byssinosis, longdf$Smoking)
chisq.test(table.2)
table.3 <- table(longdf$Byssinosis, longdf$Sex)
chisq.test(table.3)
table.4 <- table(longdf$Byssinosis, longdf$Race)
chisq.test(table.4)
table.5 <- table(longdf$Byssinosis, longdf$Workspace)
chisq.test(table.5)
## using kbl to output table into pdf format
tab1 = kbl(table.1, caption = "Byssinosis vs Employment", booktabs = T) %>%
  kable_styling(latex_options = c("striped", "hold_position"),
                position = 'center') %>% ## centers table
  column_spec(1, width = "1cm") %>%
  column_spec(2, width = "1cm")  %>%
  column_spec(3, width = "1cm")  %>%
  add_footnote("X-squared = 10.177, p-value = 0.006") ## adds information
##repeating for tables 1:5
tab2 = kbl(table.2, caption = "Byssinosis vs Smoking", booktabs = T) %>%
  kable_styling(latex_options = c("striped", "hold_position"),
                position = 'center') %>%
  column_spec(1, width = "2cm") %>%
  column_spec(2, width = "2cm")  %>%
  add_footnote("X-squared = 19.355, p-value = 1.085e-05")
tab3 = kbl(table.3, caption = "Byssinosis vs Sex", booktabs = T) %>%
  kable_styling(latex_options = c("striped", "hold_position"),
                position = 'center') %>%
  column_spec(1, width = "2cm") %>%
  column_spec(2, width = "2cm")  %>%
  add_footnote("X-squared = 37.66, p-value = 8.42e-10")
tab4 = kbl(table.4, caption = "Byssinosis vs Race", booktabs = T) %>%
  kable_styling(latex_options = c("striped", "hold_position"),
```

```
                position = 'center') %>%
  column_spec(1, width = "2cm") %>%
  column_spec(2, width = "2cm")  %>%
  add_footnote("X-squared = 5.804, p-value = 0.016")
tab5 = kbl(table.5, caption = "Byssinosis vs Workplace", booktabs = T) %>%
  kable_styling(latex_options = c("striped", "hold_position"),
                position = 'center') %>%
  column_spec(1, width = "1cm") %>%
  column_spec(2, width = "1cm")  %>%
  column_spec(3, width = "1cm")  %>%
  add_footnote("X-squared = 413.71, p-value = 2.2e-16")
## outputting each table
tab1
tab2
tab3
tab4
tab5
## Testing the first additive term combinations, as well as a model with only the
## coefficient. -\> AIC(glm(Byssinosis \~ Workspace, family = binomial,
## data = train))
## computing AIC for each binomial model
AIC(glm(Byssinosis ~ EmploymentFactor, family = binomial, data = train))
AIC(glm(Byssinosis ~ Smoking, family = binomial, data = train))
AIC(glm(Byssinosis ~ Sex, family = binomial, data = train))
AIC(glm(Byssinosis ~ Race, family = binomial, data = train))
AIC(glm(Byssinosis ~ Workspace, family = binomial, data = train))
AIC(glm(Byssinosis ~ 1 , family = binomial, data = train))

## AIC(glm(Byssinosis \~ Workspace + Smoking, family = binomial, data = train))
AIC(glm(Byssinosis ~ Workspace, family = binomial, data = train))
AIC(glm(Byssinosis ~ Workspace + Sex, family = binomial, data = train))
AIC(glm(Byssinosis ~ Workspace + Race, family = binomial, data = train))
AIC(glm(Byssinosis ~ Workspace + EmploymentFactor, family = binomial,
        data = train))
AIC(glm(Byssinosis ~ Workspace + Smoking, family = binomial, data = train))

AIC(glm(Byssinosis ~ Workspace + Smoking, family = binomial, data = train))
AIC(glm(Byssinosis ~ Workspace * Smoking, family = binomial, data = train))
AIC(glm(Byssinosis ~ Workspace + Smoking + EmploymentFactor,
        family = binomial, data = train))
AIC(glm(Byssinosis ~ Workspace + Smoking + Sex, family = binomial, data = train))
AIC(glm(Byssinosis ~ Workspace + Smoking + Race, family = binomial, data = train))

## AIC(glm(Byssinosis \~ Workspace + Smoking + Employment, family = binomial,
## data = train))

AIC(glm(Byssinosis ~ Workspace + Smoking + EmploymentFactor,
        family = binomial, data = train))
AIC(glm(Byssinosis ~ Workspace + Smoking * EmploymentFactor,
        family = binomial, data = train))
AIC(glm(Byssinosis ~ Workspace * EmploymentFactor + Smoking ,
        family = binomial, data = train))
AIC(glm(Byssinosis ~ Workspace + Smoking + EmploymentFactor + Sex ,
```

```
        family = binomial, data = train))
AIC(glm(Byssinosis ~ Workspace + Smoking + EmploymentFactor + Race,
        family = binomial, data = train))

## AIC(glm(Byssinosis \~ Workspace \* Employment + Smoking , family =
## binomial, data = train))

AIC(glm(Byssinosis ~ Workspace * EmploymentFactor + Smoking ,
        family = binomial, data = train))
AIC(glm(Byssinosis ~ Workspace * EmploymentFactor + Smoking + Sex ,
        family = binomial, data = train))
AIC(glm(Byssinosis ~ Workspace * EmploymentFactor + Smoking + Race ,
        family = binomial, data = train))

finalboss_aic = glm(Byssinosis ~ Workspace * EmploymentFactor + Smoking ,
                    family = binomial, data = train)
## Trying a poisson model because the p of byssinosis = 0.03 -\> unlikely
## Using AIC under Poisson Distribution
AIC(glm(Byssinosis ~ EmploymentFactor, family = poisson, data = train))
AIC(glm(Byssinosis ~ Smoking, family =poisson, data = train))
AIC(glm(Byssinosis ~ Sex, family = poisson, data = train))
AIC(glm(Byssinosis ~ Race, family = poisson, data = train))
AIC(glm(Byssinosis ~ Workspace, family = poisson, data = train))
AIC(glm(Byssinosis ~ 1 , family = poisson, data = train))

## Same as binomial so far

AIC(glm(Byssinosis ~ Workspace, family = poisson, data = train))
AIC(glm(Byssinosis ~ Workspace + Sex, family = poisson, data = train))
AIC(glm(Byssinosis ~ Workspace + Race, family = poisson, data = train))
AIC(glm(Byssinosis ~ Workspace + EmploymentFactor, family = poisson,
        data = train))

## Still in line with the binomial model

AIC(glm(Byssinosis ~ Workspace + EmploymentFactor, family = poisson,
        data = train))
AIC(glm(Byssinosis ~ Workspace + EmploymentFactor + Sex, family = poisson,
        data = train))
AIC(glm(Byssinosis ~ Workspace + EmploymentFactor + Race, family = poisson,
        data = train))
AIC(glm(Byssinosis ~ Workspace + EmploymentFactor + Smoking, family = poisson,
        data = train))

## Same as binomial

## AIC(glm(Byssinosis \~ Workspace + Employment, family = poisson, data = train))

## AIC(glm(Byssinosis \~ Workspace + Employment + Smoking, family = poisson, data
## = train))

AIC(glm(Byssinosis ~ Workspace + EmploymentFactor + Smoking, family = poisson,
        data = train))
```

```r
AIC(glm(Byssinosis ~ Workspace + EmploymentFactor + Smoking + Sex, family = poisson,
        data = train))
AIC(glm(Byssinosis ~ Workspace + EmploymentFactor + Smoking + Race,
        family = poisson, data = train))
AIC(glm(Byssinosis ~ Workspace * EmploymentFactor + Smoking, family = poisson,
        data = train))
AIC(glm(Byssinosis ~ Workspace * EmploymentFactor * Smoking, family = poisson,
        data = train))
AIC(glm(Byssinosis ~ Workspace * EmploymentFactor * Smoking, family = poisson,
        data = train))

finalboss_poisson = glm(Byssinosis ~ Workspace * EmploymentFactor + Smoking,
                        family = poisson, data = train)

## Final model is the same

## AIC(glm(Byssinosis \~ Workspace * Employment + Smoking, family =
## poisson, data = train)) -\> FINAL MODEL
## Using BIC under the Binomial distribution
AIC(glm(Byssinosis ~ EmploymentFactor, family = binomial, data = train),
    k = log(4334)) ##adding factor of log(4334)
AIC(glm(Byssinosis ~ Smoking, family = binomial, data = train), k = log(4334))
AIC(glm(Byssinosis ~ Sex, family = binomial, data = train), k = log(4334))
AIC(glm(Byssinosis ~ Race, family = binomial, data = train), k = log(4334))
AIC(glm(Byssinosis ~ Workspace, family = binomial, data = train), k = log(4334))
AIC(glm(Byssinosis ~ 1 , family = binomial, data = train), k = log(4334))

## AIC(glm(Byssinosis \~ Workspace, family = binomial, data = train),
## k = log(4334))

AIC(glm(Byssinosis ~ Workspace,
        family = binomial, data = train),
    k = log(4334))

AIC(glm(Byssinosis ~ Workspace + EmploymentFactor,
        family = binomial, data = train),
    k = log(4334))

AIC(glm(Byssinosis ~ Workspace + Sex,
        family = binomial, data = train),
    k = log(4334))

AIC(glm(Byssinosis ~ Workspace + Race,
        family = binomial, data = train),
    k = log(4334))

AIC(glm(Byssinosis ~ Workspace + Smoking,
        family = binomial, data = train),
    k = log(4334))

## AIC(glm(Byssinosis \~ Workspace, family = binomial, data =
## train), k = log(4334))
```

```r
## final model
finalboss_bic =glm(Byssinosis ~ Workspace,
        family = binomial, data = train)
## Using BIC under the Poisson Distribution
AIC(glm(Byssinosis ~ EmploymentFactor, family = poisson, data = train),
    k = log(4334)) ##adding factor of log(4334)
AIC(glm(Byssinosis ~ Smoking, family = poisson, data = train), k = log(4334))
AIC(glm(Byssinosis ~ Sex, family = poisson, data = train), k = log(4334))
AIC(glm(Byssinosis ~ Race, family = poisson, data = train), k = log(4334))
AIC(glm(Byssinosis ~ Workspace, family = poisson, data = train), k = log(4334))
AIC(glm(Byssinosis ~ 1 , family = poisson, data = train), k = log(4334))

## Same as binomial

## AIC(glm(Byssinosis \~ Workspace, family = binomial, data = train),
## k = log(4334))

AIC(glm(Byssinosis ~ Workspace,
        family = poisson, data = train),
    k = log(4334))
AIC(glm(Byssinosis ~ Workspace + EmploymentFactor,
        family = poisson, data = train),
    k = log(4334))

AIC(glm(Byssinosis ~ Workspace + Sex,
        family = poisson, data = train),
    k = log(4334))

AIC(glm(Byssinosis ~ Workspace + Race,
        family = poisson, data = train),
    k = log(4334))

AIC(glm(Byssinosis ~ Workspace + Smoking,
        family = poisson, data = train),
    k = log(4334))

# Same as

## AIC(glm(Byssinosis \~ Workspace, family = binomial, data =
## train), k = log(4334))

## final model
finalboss_bic_p =glm(Byssinosis ~ Workspace,
        family = poisson, data = train)
## exactring the variables of coefficients under each model
aic_vars = names(coef(finalboss_aic))[-1]  ## exclude intercept
poisson_vars = names(coef(finalboss_poisson))[-1]
bic_vars = names(coef(finalboss_bic))[-1]
bic_p_vars = names(coef(finalboss_bic_p))[-1]

## #ombine the results into a data frame
combined_results = data.frame(
  Variable = unique(c(aic_vars, poisson_vars, bic_vars,bic_p_vars)),
```

```
  AIC = ifelse(unique(c(aic_vars, poisson_vars, bic_vars,bic_p_vars))
               %in% aic_vars, "Yes", "No"),
  Poisson_AIC = ifelse(unique(c(aic_vars, poisson_vars, bic_vars,bic_p_vars))
                       %in% poisson_vars, "Yes", "No"),
  BIC = ifelse(unique(c(aic_vars, poisson_vars, bic_vars,bic_p_vars))
               %in% bic_vars, "Yes", "No"),
  Poisson_BIC = ifelse(unique(c(aic_vars, poisson_vars, bic_vars,bic_p_vars))
                       %in% bic_p_vars, "Yes", "No")
)

## using kable to display the table with each variable in each model
kable(combined_results,
      caption = "Final Selected Variables in Stepwise Regression Models")
# AIC binomial model
## Plotting the residual plots for AIC Under Binomial
AICb = glm(cbind(Byssinosis, Non.Byssinosis) ~ Workspace * Employment + Smoking,
                    family = binomial, data = widedf)
ry = residuals(AICb, type="deviance")
rx = logit(fitted.values(AICb))

plot(rx, ry, xlab="Fitted log-odds", ylab="Deviance Residual",
     main = "Workspace * Employment + Smoking under AIC Binomial")
## labeling plot
ks = ksmooth(rx, ry,kernel = "normal", bandwidth=1) ## smoothing the plot
lines(ks, col="purple",lwd=1)
# BIC binomial model
## Computing the residual plot for binomial model under BIC
BICb = glm(cbind(Byssinosis, Non.Byssinosis) ~ Workspace,
                    family = binomial, data = widedf)
ry = residuals(BICb, type="deviance")
rx = logit(fitted.values(BICb))

plot(rx, ry, xlab="Fitted log-odds", ylab="Deviance Residual",
     main = "Workspace under BIC Binomial") ## labels
ks = ksmooth(rx, ry,kernel = "normal", bandwidth=1) ## smoothing function
lines(ks, col="purple",lwd=1)
## creating confusion matrices
# Workspace*Employment + Smoking
prediction <- predict(finalboss_aic, test, type = "response")
prediction2 <- ifelse(prediction > 0.5, 1, 0)
## using confusionMatrix to compute matrix
cm1 = confusionMatrix(factor(prediction2), factor(test$Byssinosis))
## using kbl to style the table
kbl(cm1$table, caption = "AIC Confusion Table", booktabs = T) %>%
  kable_styling(latex_options = c("striped", "hold_position"),
                position = 'center') %>%
  column_spec(1, width = "1cm") %>%
  column_spec(2, width = "1cm")  %>%
  column_spec(3, width = "1cm")  %>%
  add_footnote("Accuracy: 0.9658672") ## adding accuracy

# Workspace + Smoking
prediction <- predict(finalboss_bic, test, type = "response")
```

```r
prediction2 <- ifelse(prediction > 0.5, 1, 0)
## using confusionMatrix to compute matrix
cm2 = confusionMatrix(factor(prediction2), factor(test$Byssinosis))
## using kbl to style the table
kbl(cm2$table, caption = "BIC Confusion Table", booktabs = T) %>%
  kable_styling(latex_options = c("striped", "hold_position"),
                position = 'center') %>%
  column_spec(1, width = "1cm") %>%
  column_spec(2, width = "1cm")  %>%
  column_spec(3, width = "1cm")  %>%
  add_footnote("Accuracy: 0.9658672")## adding accuracy
```