Modeling Byssinosis Data

Christina Chang

12/5/2019

This project investigates the relationship between the Byssinosis disease and smoking status, sex, race, length of employment, smoking, and dustiness of the workplace.

Fit the Logistic Regression Model

```
##
## Call:
## glm(formula = cbind(Byssinosis, Non.Byssinosis) ~ ., family = binomial(),
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.3356  -0.7653  -0.2712   0.2501   2.1264
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -1.3942     0.3175  -4.392 1.12e-05 ***
## Employment>=20    0.7834     0.2160   3.627 0.000287 ***
## Employment10-19   0.5910     0.2602   2.271 0.023147 *
## SmokingYes        0.6208     0.1934   3.210 0.001326 **
## SexM              0.2457     0.2124   1.157 0.247452
## RaceW            -0.2587     0.2062  -1.255 0.209596
## Workspace        -1.3758     0.1157 -11.886  < 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:  68.779  on 58  degrees of freedom
## AIC: 189.46
##
## Number of Fisher Scoring iterations: 5
```

From the results of the logistic regression model with no interactions, sex and race are not statistically significant. As for the statistically significant variables, workspace has the smallest p-value, therefore, there is a strong association between the type of work place and the disease. Since the coefficient for workspace is negative, this means that a less dusty environment decreases the odds of having the disease, given that all other variables are held constant. More specifically, a one unit increase in the workspace variable reduces the log odds by -1.376, or reduces the odds by 0.253.

I also introduced interactions into the model. I fitted the logistic regression model with all possible interactions. I found that the interaction term between male and workspace had the strongest association with the disease. The interaction term between male and workspace had the smallest p-value of 0.0087 and the coefficient was -0.907.

Model Selection

Perform stepwise variable selection

To perform model selection, I examined the results from forward subset, backward subset, and bidirectional stepwise regression with AIC.

```
##
## Call:  glm(formula = cbind(Byssinosis, Non.Byssinosis) ~ Workspace +
##     Smoking + Employment, family = binomial(), data = df)
##
## Coefficients:
##     (Intercept)        Workspace       SmokingYes   Employment>=20
##         -1.1858          -1.4663           0.6670           0.6699
## Employment10-19
##          0.5328
##
## Degrees of Freedom: 64 Total (i.e. Null);  60 Residual
## Null Deviance:       322.5
## Residual Deviance: 71.83     AIC: 188.5
```

These three methods of model selection using AIC yielded ths same model. The best model is:

$$log\left(\frac{\pi}{1-\pi}\right) = -1.1858 - 1.4663x_{\text{Workspace}} + 0.6670x_{\text{SmokingYes}} + 0.6699x_{\text{Employment>=20}} + 0.5328x_{\text{Employment10-19}}$$

This suggests that being in a less dusty workspace decreases the log odds of having the disease. Smoking, employment between 10 and 19 years, and employment greater than or equal to 20 years all increase the log odds of having the disease. However, the type of workspace seems to have the largest effect on the odds of having the disease.

Also, notice that model selection removed the terms for sex and race. This implies that these two variables are not very important for determining if a person has the disease.

LR test for interactions

```
## Likelihood ratio test
##
## Model 1: cbind(Byssinosis, Non.Byssinosis) ~ (Employment + Smoking + Workspace)^2
## Model 2: cbind(Byssinosis, Non.Byssinosis) ~ Employment + Smoking + Workspace
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1  10 -85.212
## 2   5 -89.253 -5 8.0836     0.1517
```

I used the likelihood ratio test to test for interactions between all variables. I compared the model with all interactions to the model with employment, smoking, and workspace. So, the null hypothesis is that all the coefficients of the interaction terms equal zero. The alternative hypothesis is that at least one coefficient of the interaction terms is not zero.

The LR test statistic is 8.0836 and the p-value is 0.152. Hence, there is not enough statistical evidence to reject the null hypothesis when $\alpha = 0.01$. We fail to reject the null hypothesis and leave out the interaction terms from the model.

Model Diagnostics

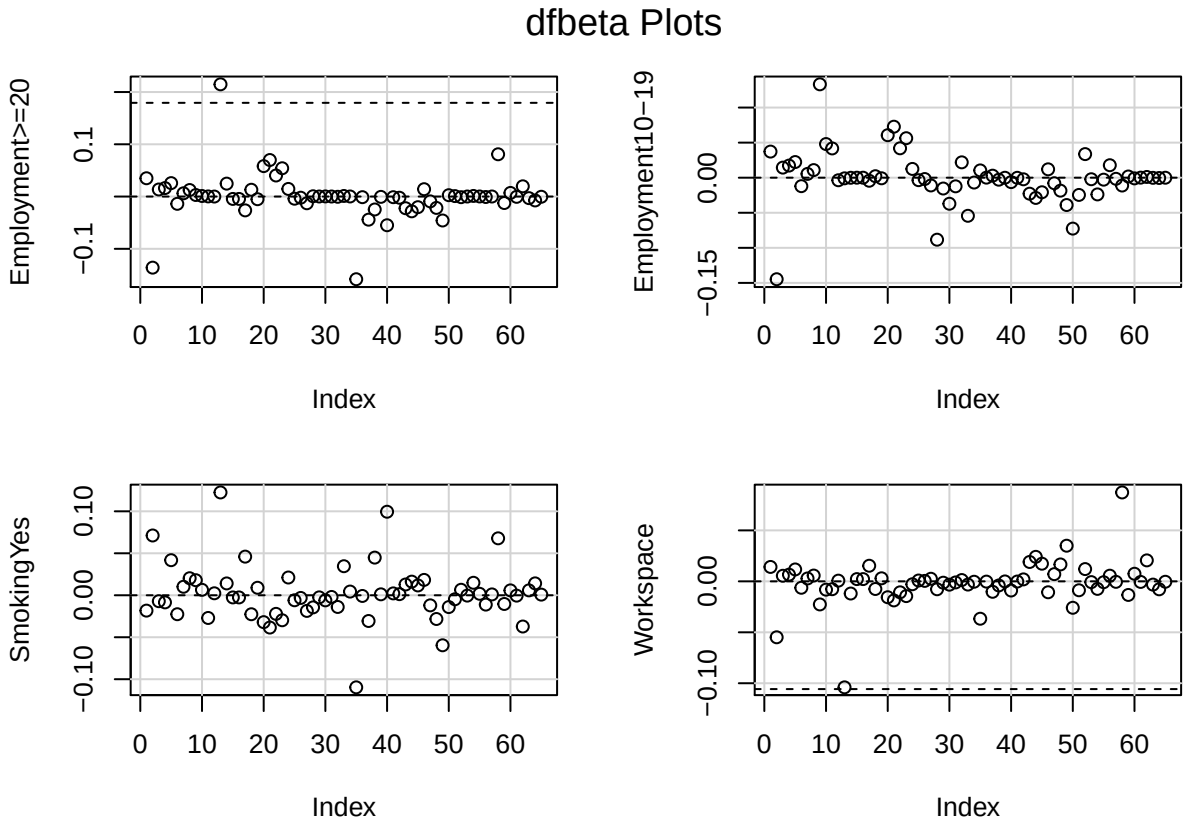
I performed model diagostics on the model that only includes the main effects of employment, smoking, and workspace. It does not include any interactions.

Pearson's residuals plot

Below is a histogram of Pearson's residuals. All of the residuals are within -3 and 3. This suggests that the model fits the data fairly well.

DFbeta plots

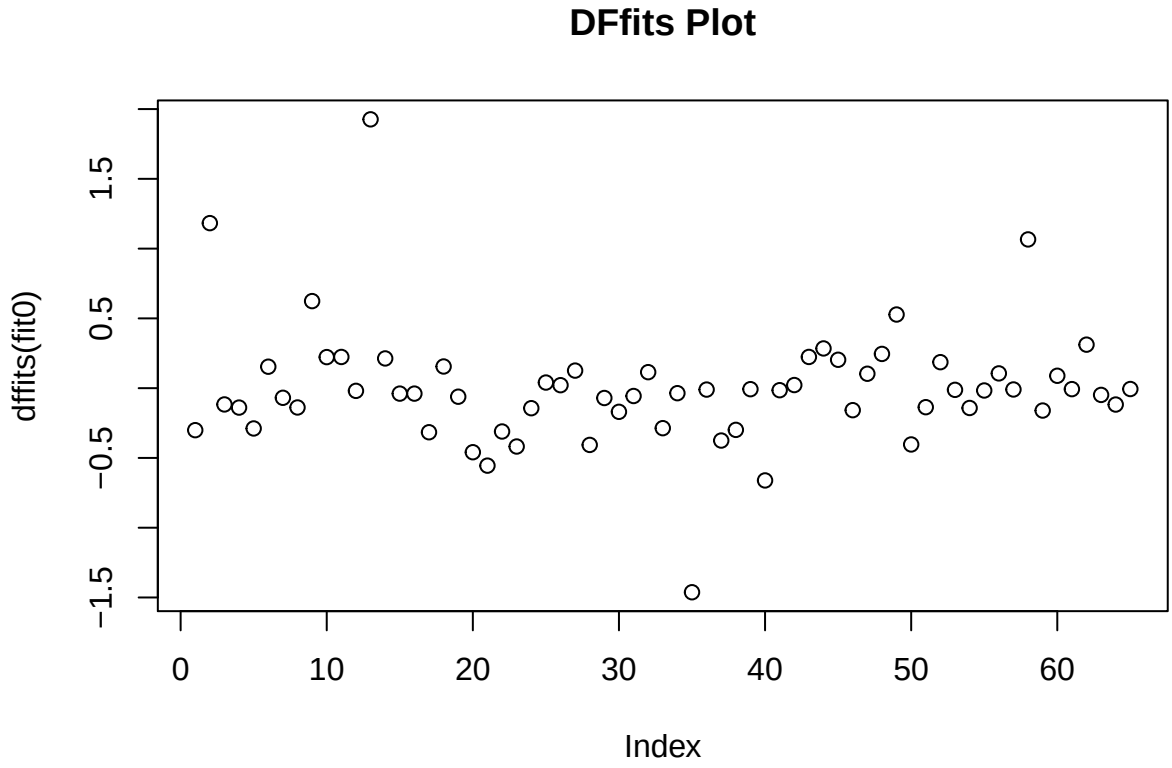
The following are index plots of DFbeta. These plots show the effect on coefficients after deleting an observation. Each point has a DFbeta. A large DFbeta indicates that the observation has a large influence on the parameter estimate.



A data point is suspicious if the DFbeta is greater that 1. There are no DFbetas that are greater than 1 in the dataset. There are some points have a absolute DFbeta value around 0.1 or 0.2 in all plots.

DFfits plot

The following is a plot of DFfits. DFfits measures how the regression function changes when an observation is



deleted.

There are four points that have DFfits magnitudes greater than 1:

```
##    Employment Smoking Sex Race Workspace Byssinosis Non.Byssinosis
## 2         <10     Yes   M    O         1         25            139
## 13      10-19      No   M    W         1          2              8
## 35      10-19     Yes   F    W         2          1             33
## 58      10-19     Yes   M    O         3          0             33
```

These observations have a high influence on the model.

Conclusions

From fitting the logistic regression model and exploring model selection and diagnostics, I found that workspace seems to have the most influence on having the Byssinosis disease. People who work in dustier environments have a higher odds of having the disease when compared to people who work in less dusty environments. Smoking, employment between 10 and 19 years, and employment greater than 20 years increases the odds of having the disease.

After model selection, the variables for sex and race were removed, hence, they were not as important for determining if a person has the Byssinosis disease.

The Pearson's residual plots showed that all residuals were within -3 and 3, meaning that the model fits the data well. The DFbetas plot did not reveal any observations that have a particularly large influece on the coefficients. However, the DFfits plot showed that four observations have a large infuence on the regression function.

Appendix

```
knitr::opts_chunk$set(echo = FALSE, include = TRUE)
# Import libraries.
library(bestglm)
```

```r
library(lmtest)
library(car)

# Read the data.
df = read.csv("byssinosis.csv")
# Model fitting
model = glm(formula = cbind(Byssinosis,Non.Byssinosis) ~ .,
            family = binomial(), data = df)
summary(model)
# Consider interaction terms.
model = glm(formula = cbind(Byssinosis,Non.Byssinosis) ~ (.)^2,
            family = binomial(), data = df)
summary(model)
# Model selection
fullmodel = glm(formula = cbind(Byssinosis,Non.Byssinosis) ~ .,
            family = binomial(), data = df)
nullmodel = glm(formula = cbind(Byssinosis,Non.Byssinosis) ~ 1,
            family = binomial(), data = df)
# Forward AIC
forwardAIC = step(nullmodel,
                  scope = list(lower = nullmodel,
                               upper = fullmodel),
                  direction = "forward")
# Backward AIC
backwardAIC = step(fullmodel,
                  scope = list(lower = nullmodel,
                               upper = fullmodel),
                  direction = "backward")

# Bidirectional AIC
bidirectAIC = step(fullmodel,
                  scope = list(lower = nullmodel,
                               upper = fullmodel),
                  direction = "both")
forwardAIC
# LR test for interactions
fitf = glm(formula = cbind(Byssinosis,Non.Byssinosis) ~
             (Employment + Smoking + Workspace)^2,
           family = binomial(), data = df)
fit0 = glm(formula = cbind(Byssinosis,Non.Byssinosis) ~
             Employment + Smoking + Workspace,
           family = binomial(), data = df)

lrtest(fitf,fit0)
# Pearson's residuals
res = residuals(fit0, "pearson")
hist(res, main = "Plot of Pearson's Residuals", xlab = "Residual")
# DFbeta plots
dfbetaPlots(fit0)
# DFfits plot
plot(dffits(fit0), main = "DFfits Plot")
df[which(abs(dffits(fit0)) > 1),]
```