STA 138 Final Project

STA 138 | Andrew Muench, Michelle Tsang, Connor Young | December 2023

Introduction

In this study, we will be analyzing data collected in 1973 from a large cotton textile company in North Carolina. This data was collected to investigate the presence of a disease called byssinosis which workers exposed to cotton dust may contract. The purpose of our study is to investigate the relationship between this disease and the level of dustiness of the workplace, smoking status, sex, race, and length of employment.

These relationships are important to investigate because if there is a significant relationship between these factors and the chance of contracting byssinosis, then the company may be liable and need to implement changes to protect employees. To do so, we will fit and interpret several logistic regression models.

Data Description

The data was collected on 5,419 workers and consists of six variables. The variables are there possible values are shown below.

- Type of work place [1 (most dusty), 2 (less dusty), 3 (least duty)]
- Employment, years [<10, 10-19, >=20]
- Smoking [Yes, No (not in the last 5 years)]
- Sex [M (Male), F (Female)]
- Race [W (White), O (Other)]
- Byssinosis [Yes, No]

Results

Full Additive Model

We will begin our analysis by fitting the data to an additive logistic model which uses all of the variables and no interaction terms. The estimated coefficients produced by this model, their associated Wald test p-values, and model's BIC are shown below.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-5.103	0.264	-19.308	0.000
Employment≥20	0.744	0.217	3.432	0.001
Employment10-19	0.568	0.262	2.168	0.030
SmokingYes	0.671	0.196	3.421	0.001
SexM	-0.150	0.230	-0.655	0.513
RaceW	-0.118	0.208	-0.566	0.571
Workspace2	0.171	0.287	0.597	0.551
Workspace1	2.762	0.217	12.713	0.000
BIC	182.377			

The estimated coefficients for each parameter represent the log-odds ratio compared to the base case used by the model. The base case is Employment: < 10 years, Smoking: No, Sex: Female, Race: Other, and Workspace: 3 (least dusty). A positive log-odds ratio for a parameter indicates that, all other things equal, an individual will have a higher log-odds (chance) of contracting byssinosis compared to the base case. A negative log-odds ratio indicates

under the same conditions represents a lower log-odds (chance) of contracting byssinosis compared to the base case. Thus, the results from this model indicate that longer employment, smoking, and a dustier workplace increase the chance of byssinosis and being male and white reduce the chance of byssinosis.

However, when analyzing the Wald test p-values for each parameter with a Bonferroni Correction, we find that several parameters are not significant to the model's performance. We will be using a significance level of $\alpha = 0.05$, so with 8 parameters, our Bonferroni corrected significance level for each parameter is 0.00625. Since the Wald test uses a null hypothesis of the parameter coefficient being equal to 0, using this significance level, we find we fail to reject this null hypothesis for Employment 10-19, Sex M, Race W, and Workspace 2. Thus, the coefficients for these parameters are not significantly different from 0 and are not related to the contraction of byssinosis.

Additionally, the BIC of this model is 182.377. We will compare our following models to this value to measure potential improvement. For our next model. we will use Forward Stepwise Regression with BIC as the step criteria. The results of fitting this model are shown below.

Forward Stepwise Regression

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-5.208	0.237	-21.931	0.000
Workspace2	0.192	0.285	0.675	0.500
Workspace1	2.739	0.191	14.315	0.000
SmokingYes	0.646	0.193	3.358	0.001
Employment≥20	0.660	0.182	3.625	0.000
Employment10-19	0.505	0.249	2.029	0.043
BIC	174.779			

From the above table, we see that the process of stepwise regression has concluded that Sex and Race are not sufficiently significant to be included. This agrees with the results of the Wald tests performed on our additive model, so this is not surprising. Similarly to our additive model, this model also indicates positive log-odds ratios for more dustiness, smoking, and longer employment. This model also has a BIC of 174.779 which is lower than the BIC of our first model, so this model has a better fit.

We can compare the equivalency of our two models using a likelihood ratio test with a null hypothesis of the models being equivalent, so the simpler model is preferred. The resulting p-value of this test is 0.687. Using a significance level of $\alpha = 0.05$, we fail to reject the null hypothesis. Thus, we conclude that the full additive and forward stepwise models are equivalent, so we prefer the stepwise model because it is simpler.

When analyzing the Wald test p-values for each parameter with a Bonferroni corrected significance level of 0.00833, we find the log-odds ratios of Workspace 2 and Employment 10-19 are not significantly different than 0. To investigate whether we can adjust our data to improve the model, we will modify these variables.

Forward Stepwise Regression: Workspace Modification

We will first modify the Workspace variable by combining Workspace 2 and 3 into a single category. Thus, the Workspace variable will now have the values [1 (more dusty), 3 (less dusty)]. The result of fitting a Forward Stepwise Regression using BIC and the modified data is shown below.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-5.150	0.219	-23.516	0.000
Workspace1	2.684	0.170	15.749	0.000
SmokingYes	0.645	0.192	3.349	0.001
Employment≥20	0.657	0.182	3.612	0.000
Employment10-19	0.501	0.249	2.011	0.044
BIC	171.0489			

From the results above, we see that this stepwise regression resulted in the same final model as our previous stepwise regression, but without Workspace 2 since that has been combined to the base case. We also see the same positive log-odds ratios for each parameter, so this model also concludes that a dustier workspace, smoking, and longer

employment are related to a worker's chance of contracting byssinosis. This model also has a BIC of 171.0489 which is lower than our previous stepwise regression, which would indicate that this model fits our data better.

When computing a likelihood ratio test to compare the equivalency between this new stepwise model and our previous stepwise model, we get a p-value of 0.505. Using a significance level of $\alpha = 0.05$, we fail to reject the null hypothesis. Thus, we conclude that our new stepwise and previous stepwise models are equivalent, so we prefer the newer stepwise model because it is simpler.

When analyzing the Wald test p-values for each parameter with a Bonferroni corrected significance level of 0.01, we find the log-odds ratio of Employment 10-19 is not significantly different than 0. Thus, we will modify this variable and create our final stepwise model.

Forward Stepwise Regression: Employment Modification

Using Bonferroni Correction suggests that Employment 10-19 is not significantly different compared to Employment < 10. However, since the p-value is still relatively low (<0.05), we will combine Employment 10-19 and Employment ≥ 20. Thus, the employment variable will now have the values [<10, ≥ 10].

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-5.143	0.218	-23.538	0.000
Workspace1	2.683	0.170	15.746	0.000
Employment≥10	0.615	0.170	3.607	0.000
SmokingYes	0.636	0.192	3.314	0.001
BIC	167.289			

From the results above, we see that this stepwise regression resulted in the same final model as our previous stepwise regressions, but with the effects of our variable modifications. We also see the same positive log-odds ratios for each parameter, so this model also concludes that a dustier workspace, smoking, and longer employment are related to a worker's chance of contracting byssinosis. This model also has a BIC of 167.289 which is lower than our previous stepwise regressions, which indicates that this model fits our data best.

When computing a likelihood ratio test to compare the equivalency between this new stepwise model and our first stepwise model, we get a p-value of 0.651. Using a significance level of $\alpha = 0.05$, we fail to reject the null hypothesis. Thus, we conclude that our new stepwise and our first stepwise models are equivalent, so we prefer this newer stepwise model because it is simpler.

We can also compute a likelihood ratio test to compare the equivalency between this new stepwise model and our modified previous stepwise model, and we get a p-value of 0.52. Using a significance level of $\alpha = 0.05$, we fail to reject the null hypothesis. Thus, we conclude that our new stepwise and previous stepwise models are equivalent, so we prefer this newer stepwise model because it is simpler.

When analyzing the Wald test p-values for each parameter with a Bonferroni corrected significance level of 0.0125, we find that all of the parameters are significantly different than 0, so all of the parameters are significant and related to byssinosis.

Logistic Regression Diagnostics

To determine how well our model fits the data, we will look at its deviance residuals. A residual is a measure of the difference between an observed value and a predicted value. The most basic residual for our models is the raw difference between the observed value (Byssinosis status: $y_i \in \{0, 1\}$) and the predicted probability ($\hat{\pi}_i \in [0, 1]$). This residual will be denoted as e_i and has the form

$$e_i = y_i - \hat{\pi}_i \in [-1, 1].$$

The deviance residual, d_i, is based on deviance which is derived from a likelihood ratio test comparing the fitted logistic model and a *saturated* model which fits each data point perfectly. The formula to calculate the deviance residuals is

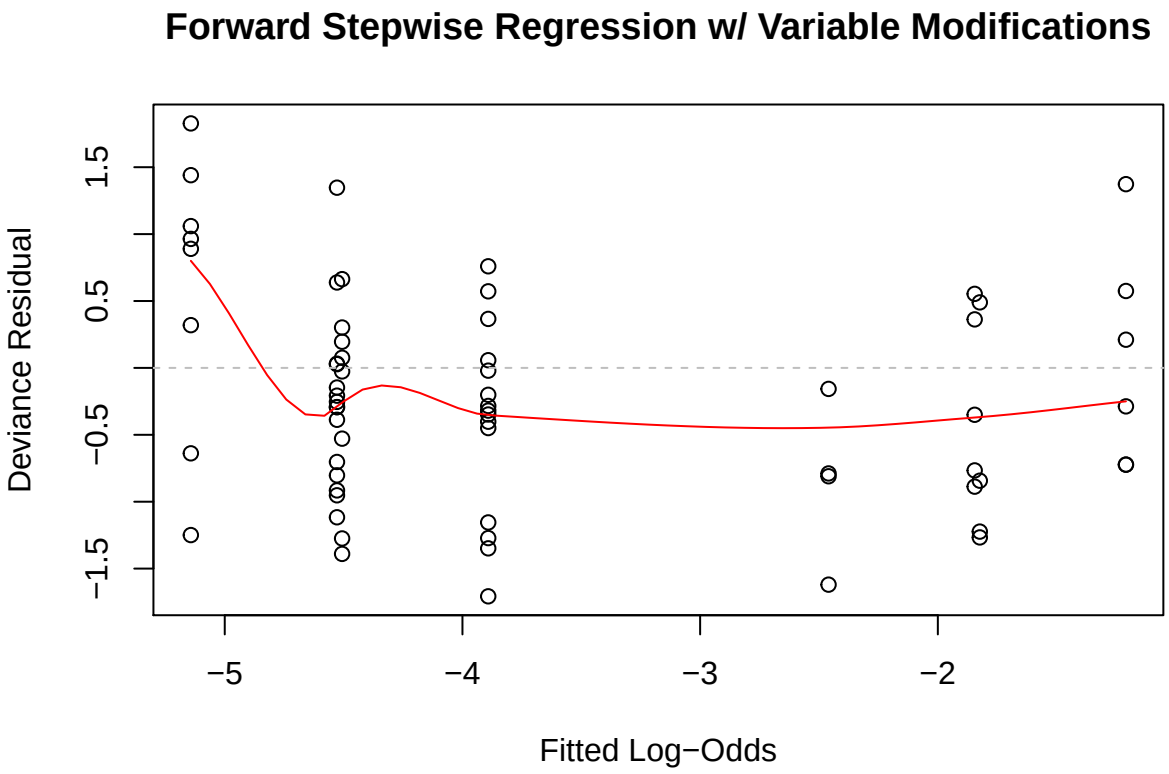
$$\begin{aligned} d_i &= \text{sign}(e_i)\sqrt{\text{Deviance}} \\ &= \text{sign}(e_i)\sqrt{-2\left[y_i \log \hat{\pi}_i + (1 - y_i)\log(1 - \hat{\pi}_i)\right]}. \end{aligned}$$

The values of these deviance residuals measure how much the predicted probabilities from our model differ from the observed data and whether our model is under- or over-fitting for certain predictor variable values.

Large values for deviance residuals indicate poor fit and small values indicate good fit. A positive deviance residual indicates our model is under-predicting for a certain set of predictor variable values, i.e., the observed value is 1 and the predicted probability is less than 1. A negative deviance residual indicates our model is over-predicting for a certain set of predictor variable values, i.e., the observed value is 0 and the predicted probability is greater than 0.

A good model fit is indicated by deviance residuals which are evenly distributed and relatively close to 0. Thus, if a smooth line is drawn to "average" the deviance residuals, we would like this line to be relatively flat and close to the line at $d_i = 0$.

The plot and interpretation of the deviance residuals for our logistic regression model is shown below.



From the plot above, we see that the fitted smooth line is relatively straight (after log-odds of -4) and close to 0, with a peak at around $(-5, 0.8)$ and the rest of the line at $d_i \approx -0.35$. It appears that at low log-odds, this model is somewhat over-predicting, but at log-odds greater than around -4.75, this model is somewhat under-predicting. Overall, the deviance residuals appear to be relatively equally distributed above and below 0 (~37% and ~63% respectively), and ~95% of the residuals have an absolute value less than 1.5. Thus, we consider this model to be a relatively good fit.

Discussion

In this study, we fit four logistic regression models:

- Model 1: All additive terms with no interaction terms
- Model 2: Forward Stepwsie Regression
- Model 3: Forward Stepwise Regression with Workspace modifications
- Model 4: Forward Stepwise Regression with Workspace and Employment modifications

Of these models, Model 4 had the smallest BIC and all of its parameters were significant based on Wald tests at Bonferroni corrected significance levels. BIC is a measure of goodness of fit which accounts for model complexity through the addition of a penalty term that scales on the number of parameters in the model. Having a low BIC indicates that the model is a better fit compared to models with higher BICs. Thus, BIC suggests that Model 4 has the best fit for our data. We also conducted likelihood ratio tests to compare the equivalency of our models as parameters were dropped and variables were modified. In doing so, we found that our simpler models were still equivalent to full models, so the simplest model, Model 4, is still preferred. Lastly, from the diagnostic plot of Model 4, we can see that its deviance residuals are relatively equally distributed and close to 0, which suggests the model is a relatively good fit. Taking everything into account, we conclude that Model 4 is the best fit.

Conclusion

Using Model 4 as reference since it is the best fit, we find that Workspace, Employment, and Smoking all have positive, statistically significant effects on the chance of contracting byssinosis. Since the estimated coefficients and log-odds ratios of these parameters compared to the base case are positive, we conclude that a dustier Workspace, longer Employment ($\geq$ 10 years), and Smoking all contribute to an increased chance of a worker developing byssinosis. Conversely, we found that Sex and Race do not have a statistically significant effect on the chance of a worker developing byssinosis.

Appendix

```
knitr::opts_chunk$set(echo = FALSE)
library(xtable)
# load data
byss = read.csv("./Dataset/Byssinosis.csv")
# change Workspace to factor
byss$Workspace = factor(byss$Workspace, levels = c("3", "2", "1"))
# add Total column equal to Byssinosis + Non.Byssinosis
byss$Total = byss$Byssinosis + byss$Non.Byssinosis
# remove rows which have no total
byss = byss[byss$Total!=0,]
glm1 = glm(cbind(Byssinosis, Non.Byssinosis)~Employment + Smoking + Sex + Race + Workspace,
           family=binomial(),
           data=byss)
glm1_table = round(summary(glm1)$coefficients, digits=3)

glm1_bic = BIC(glm1)
result1 = step(glm(cbind(Byssinosis, Non.Byssinosis)~1, binomial(), byss),
                  scope = ~Employment*Smoking*Sex*Race*Workspace,
                  k=log(72),
                  trace=0,
                  direction="forward")
result1_table = round(summary(result1)$coefficients, digits=3)
result1_bic = BIC(result1)
# H_0: models are equivalent -> use simpler model
# H_1: models are not equivalent -> use more complex model
anv = anova(result1, glm1)
result1_pvalue = 1-pchisq(anv[2, 4], anv[2,3])
# changing Workplace var
# 2 & 3 -> 3 (less dusty)
# 1 -> 1 (more dusty)
byss2 = byss
byss2$Workspace[byss2$Workspace == 2] = 3
# new stepwise w/ Workplace modification
new1 = step(glm(cbind(Byssinosis, Non.Byssinosis)~1, binomial(), byss2),
                  scope = ~Employment*Smoking*Sex*Race*Workspace,
                  k=log(72),
                  trace=0,
                  direction="forward")
new1_table = round(summary(new1)$coefficients, digits=3)
new1_bic = BIC(new1)
new1_pvalue = 1-pchisq(anova(new1, result1)[2, 4], anova(new1, result1)[2,3]) # compare new1 & result1 w/ L
# changing Employment var
# 10-19 & >=20 -> >=10
byss2$Employment[byss2$Employment == "10-19" | byss2$Employment == ">=20"] = ">=10"
# new stepwise w/ Employment modification
new2 = step(glm(cbind(Byssinosis, Non.Byssinosis)~1, binomial(), byss2),
                  scope = ~Employment*Smoking*Sex*Race*Workspace,
                  k=log(72),
                  trace=0,
                  direction="forward")
new2_table = round(summary(new2)$coefficients, digits=3)
new2_bic = BIC(new2)
new2_pvalue1 = 1-pchisq(anova(new2, result1)[2, 4], anova(new2, result1)[2,3]) # compare new2 & result1 w/
new2_pvalue2 = 1-pchisq(anova(new2, new1)[2, 4], anova(new2, new1)[2,3]) # compare new2 & new1 w/ LRT
# new2
```

```r
ry = residuals(new2, type="deviance")
rx = qlogis(fitted.values(new2))

plot(rx, ry,
     xlab = "Fitted Log-Odds",
     ylab = "Deviance Residual",
     main = "Forward Stepwise Regression w/ Variable Modifications",
     col = "black")
lines(loess.smooth(rx, ry), col="red")
abline(h=0, lty=2, col="grey")
```