Byssinosis Prevalence Analysis

Brennan Dowling

UC Davis, Statistics 138

**Introduction**

This report details an analysis of byssinosis prevalence data from 1973 for a large North Carolinian textile. Byssinosis is a lung disease that workers in dusty conditions are susceptible to. The data includes the level of dustiness, years worked, sex, race, smoking status, byssinosis count, and a non-byssinosis count.

**Methods**

Each row in the data set had varying numbers of counts for byssinosis and non-byssinosis which summed to 5419, the number of people observed in the trial. Since byssinosis counts were strictly lower than non-byssinosis counts, the byssinosis counts had a small upper bound. This indicates a model using the Poisson distribution would not represent the data well. Over the course of STA 138, I have learned powerful techniques for building and analyzing logistic models; therefore, I manipulated the data to allow logistic regression.

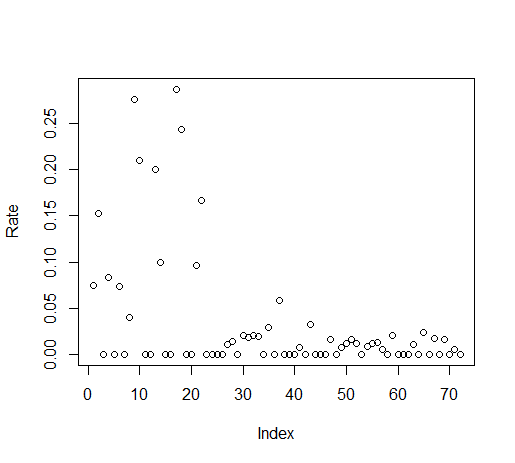
I created byssinosis prevalence rates for each row. These rates are more informative than just a byssinosis count alone. Based upon these rates, I chose a cutoff of 0.01. The model fit to these logic variables will hopefully predict if byssinosis is likely to occur at all in a group. I chose this cutoff somewhat arbitrarily after plotting the rates. The cutoff can be changed depending on what an analyst would like to predict. For example, a larger cutoff could help build a model to predict when rates will be especially large and significant.

I chose against consolidating redundant groups. For example, there were rows that had the same time at company, smoking status, sex, workspace dustiness, and race. Assuming there was some sort of organized grouping for this to be the case, the groups remained as separate unweighted observations in hopes to improve robustness.

With the data prepared for logistic regression, a model was fit. Both backwards step and forwards step algorithms with ‘lowest AIC’ as the criterion were used. After synthesizing the resulting two models, the significance of adding interaction terms was tested using the Likelihood-Ratio Test, Pearson residuals were analyzed, and DFBetas were analyzed.

**Results**

To attain an impression of how the rates were distributed, I plotted the rates studied in the plot below:



Both backwards and forwards step algorithms, found the same optimal variables were employment and race. Fitting a logistic model to these two variables produced the following model coefficients:

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 0.4486 0.5018 0.894 0.3713

RaceW 0.9779 0.5063 1.931 0.0534 .

Employment>=20 -1.1148 0.6273 -1.777 0.0755 .

Employment10-19 -1.2943 0.6313 -2.050 0.0403 \*

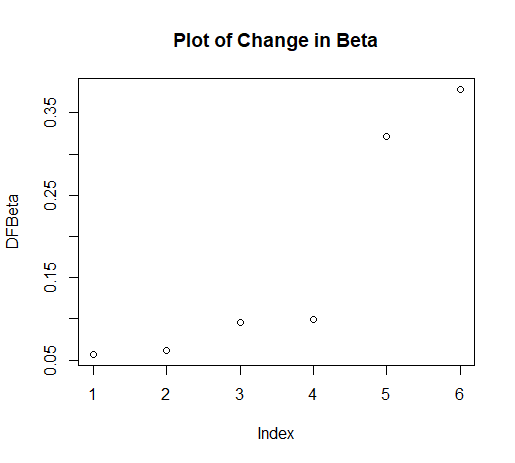
---

Testing the significance of the interaction term for Race and Employment yielded a Likelihood Ration test statistic of 16.89 which had a p-value of 0.000214 with 4 degrees of freedom.

The plotted Pearson Residuals are below:



The plotted DFBetas are below:



**Analysis**

Analyzing the plotted Byssinosis rates, most rates are zero or very close to zero, but there is a significant proportion that have very high byssinosis rates, considering the condition is a rare condition.

Regarding model selection, AIC is used as the sole criterion in leaning towards the frequentists statistical approach. Analyzing the coefficients of the optimized model show the most statistically significant coefficients are if employment has been for over 20 years, if employment has been between 10-20 years, and if race is White. Employment coefficients decreased the logistic regression equation while race ‘white’ increased the logistic regression equation.

The Likelihood Ratio Test indicates the model with interaction terms fits better, given the p-value. However, it was determined not to include the interaction term to prevent overfitting the model to the data.

Analyzing the Pearson residuals, there appears to be a mediocre fit. Since no residuals exceed 3, this model is not too “out-of-fit.”

As for DFBetas: since index 5 and 6 are somewhat high, this indicates outliers may be having too much of an impact on the model.

**Discussion**

While this model is not intended to predict if a single individual will be diagnosed with byssinosis, it can be helpful in predicting if a grouping of workers will have a high rate of byssinosis. Also, if the cutoff is raised, the model can be used for determining especially higher risk groupings.

Based on the chosen logistic model, employment over 10 years reduces likelihood of a significant proportion having the condition by a factor of 30%. On the other hand, white populations, based on the model had an increased likelihood by a factor of 260%. This could indicate a genetic predisposition to contracting the condition. As for years of experience, contracting the condition can be increased by a lack of proper work preparation, so if a worker remains with the job, this indicates they have some proficiency in the job. This proficiency can be the tendency to properly prepare for duties which can reduce the likelihood of contracting the condition.

While Pearson residual indicate that the model fit is mediocre and the DFBetas indicate outliers may be impacting the model, but they are still within acceptable ranges. And since overfitting was avoided, this model may be robust enough to apply to other populations. So management, using these findings, may want to emphasize proper safety and protection for new hires.

**References**

<https://www.albany.edu/faculty/kretheme/PAD705/SupportMat/DFBETA.pdf>

<https://newonlinecourses.science.psu.edu/stat504/node/86/>

**Appendix**