The Byssinosis Data

Arden Miller

Executive Summary

A logistic regression model was used to investigate how the odds of byssinosis for workers in the US cotton industry are affected by race, sex, smoking status, length of employment, and the amount of dust in the workplace. It was found that all of these factors except race had an impact on the odds of byssinosis. The odds of byssinosis are higher for smokers than non-smokers and increase as length of employment increases. The effect of the amount of dust in the workplace on the odds of byssinosis depends on the sex of the worker. Although, the odds of byssinosis are higher in a high dust workplace than in a medium dust or low dust workplace for both men and women, this effect is considerably larger for men than for women.

The Data

The data comes from a survey of workers in the US cotton industry and records whether the workers were suffering from a lung disease called byssinosis. Also recorded were the values for five categorical explanatory variables: the race, sex and smoking status of the worker, the length of employment and the amount of dust in the workplace. The amount of dust in the workplace is thought to be a major factor in the occurrence of byssinosis but the other measured variables may also have an impact.

dust	amount of dust in workplace: 1-high, 2-medium, 3-low.
race	ethnic origin of worker: 1-white, 2-other.
sex	sex of worker: 1-male, 2-female.
smoking	smoking status: 1-smoker, 2-nonsmoker.
employ	years of employment: 1-less than 10, 2-10 to 20, 3-more than 20.

Modelling Byssinosis

This data was analysed using logistic regression to investigate how the explanatory factors are related to the occurrence of byssinosis. It was found that dust, sex, smoking and employ all had an impact on the occurrence of byssinosis but there is no evidence that race has an impact. In addition it was found that dust and sex interact. In simple terms this means that the impact dust has on the occurrence of byssinosis depends on the level of sex and vice versa.

It is easier to interpret the logistic regression model if the odds of byssinosis are discussed rather than the probability of byssinosis. There is a simple connection between the odds and the probability of an event: the odds are defined as the probability an event occurs divided by the probability it doesn't occur. Thus if the odds of byssinosis is 1, this means it is probability it occurs is equal to the probability it does not occur.

The following model can be used to predict the odds of byssinosis for different levels of the explanatory variables:

$$\widehat{\text{odds}} = \exp(-1.753 - .658 \times I_{sm} + .464 \times I_{e2} + .637 \times I_{e3} - .999 \times I_{se} - 3.277 \times I_{d2} - 2.878 \times I_{d3} + 2.006 \times I_{se} \times I_{d2} + 1.258 \times I_{se} \times I_{d3})$$

The I's in this equation are indicator variables that are defined as:

$I_{sm} = 1$ for smoking= 2 and $I_{sm} = 0$ for smoking= 1,
$I_{e2} = 1$ for employ= 2 and $I_{e2} = 0$ for employ= 1 or 3,
$I_{e3} = 1$ for employ= 3 and $I_{e3} = 0$ for employ= 1 or 2,
$I_{se} = 1$ for sex= 2 and $I_{se} = 0$ for sex= 1,
$I_{d2} = 1$ for dust= 2 and $I_{d2} = 0$ for dust= 1 or 3,
$I_{d3} = 1$ for dust= 3 and $I_{d3} = 0$ for dust= 1 or 2.

This model indicates the following relationships between the explanatory variables and the odds of byssinosis.

1. The odds are higher for smokers than non-smokers. If the levels of dust, employ and sex are fixed at any specific levels then the odds for a smoker are estimated to be 1.93× those for a non-smoker.

2. The odds increase as length of employment increases. If the levels of dust, smoking and sex are fixed then the odds for someone who has been employed for 10-20 years are estimated to be 1.59× the odds for someone who has been employed for less than 10 years. The odds for someone who has been employed for more than 20 years are estimated to be 1.89× the odds for someone who has been employed for less than 10 years.

3. Since sex and dust interact, the effect of sex must be assessed separately for each level of dust. Suppose that the levels of smoking and employ are fixed. For low dust the odds for men are estimated to be .37× the odds for women, for medium dust the odds for men are estimated to be 2.74× the odds for women, and for high dust the odds for men are estimated to be 1.30× the odds for women. It is surprising that the odds for men should be lower for low dust but higher for medium and high dust levels.

4. Again due to the interaction between sex and dust the difference between dust levels must be assessed separately for men and women. Suppose that he levels of smoking and employ are fixed. Then for women the odds of byssinosis at a medium dust level are estimated to be 1.42× the odds at a low dust level and the the odds at a high dust level are estimated to be 5.06× the odds at a low dust. For men the odds at a medium dust level are estimated to be .67× the odds at a low dust level and the the odds at a high dust level are estimated to be 17.8× the odds at a low dust. Thus for both men and women the odds of byssinosis are much higher at a high level of dust than at a low or a medium level of dust. The results also suggest than men are affected more severely by high levels of dust than women.

According to this model the highest odds occur for male workers who are smokers, work in a high dust workplace, and have been employed for more than 20 years. The odds of byssinosis are estimated to be .328 with a 95% confidence interval of from .256 to .421 (i.e. we can be 95% confident that the true odds are in this interval). The lowest odds are predicted to occur for male workers who are non-smokers, work in a medium dust workplace, and have been employed for less than 10 years. In this case the estimated odds are .0034 with a 95% confidence interval from .0014 to .0081.

It seems odd that the estimated odds should be lowest for a medium dust workplace rather than a low dust workplace. If we consider male workers who are non-smokers, have been employed for less than 10 years but work in a low dust work place the estimated odds are .0050 with a 95% confidence interval of from .0031 to .0082. So although the estimated odds are higher than for similar workers from a medium dust workplace (from the previous paragraph) there is a great deal of overlap between the two confidence intervals. This indicates there may, in fact, be very little difference between low dust and medium dust workplaces.

Statistical Appendix

I used the step.glm function to select a suitable model. It identified the model that contained dust, sex, employ, smoking, and dust:sex. The analysis of deviance table for this model is:

	Df	Deviance	Resid. Df	Resid. Dev	Pr(Chi)
NULL			64	322.5268	
dust	2	252.1082	62	70.4185	0.0000000
sex	1	0.6577	61	69.7608	0.4173570
smoking	1	11.3730	60	58.3878	0.0007452
employ	2	14.8015	58	43.5864	0.0006108
dust:sex	2	7.5676	56	36.0188	0.0227363

The output from dummy.coef for this model is:

```
$"(Intercept)": (Intercept)
```

```
                -1.752926

$dust:    1         2         3
          0 -3.277159 -2.878296

$sex: 1          2
      0 -0.9990074

$smoking: 1          2
          0 -0.6577744

$employ: 1         2         3
         0 0.4639903 0.6366699

$"dust:sex": 11 21 31 12       22       32
              0  0  0  0 2.005788 1.257592
```

The odds and 95% confidence intervals reported on page 3 of the report were created using this model. The intervals were produced by first creating 95% confidence intervals for logit π and then using the exponential transformation to convert these to intervals for odds.

```
> new.df<-data.frame(dust=factor(c(1,2,3)),sex=factor(c(1,1,1)),
   smoking=factor(c(1,2,2)),employ=factor(c(3,1,1)))
> preds<-predict.gam(byss.fit2,new.df,se=T)
> exp(preds$fit)
         1           2           3
 0.3275036 0.003386836 0.005046822
> exp(preds$fit-1.96*preds$se.fit)
         1          2           3
 0.2559091 0.00141443 0.003124704
> exp(preds$fit+1.96*preds$se.fit)
         1           2           3
 0.4191277 0.008109741 0.008151305
```

Diagnostic plots for the fitted model are given in Figure 1. Two points (44 and 46) show up as having unually large values of Cook's Distance. The observations in the dataset corresponding to these points are:

```
> byss.df[c(44,46),]
   dust race sex smoking employ yes total
44    1    1   1       1      3  31   108
46    3    1   1       1      3  12   507
```

These should be checked to make sure they were recorded correctly. If they are dropped and the model is refitted then the estimated coefficients are

```
> dummy.coef(byss.fit3)
$"(Intercept)": (Intercept)
                  -1.807518

$dust: 1         2         3
       0 -3.124386 -3.075423

$sex: 1         2
      0 -0.976482

$smoking: 1          2
          0 -0.4987707

$employ: 1         2         3
         0 0.4922564 0.3547285

$"dust:sex": 11 21 31 12       22       32
              0  0  0  0 1.935123 1.523789
```

The biggest changes occur for smoking, dust:3, employ:3, and dust:sex:32. The impact of these changes would be to increase the odds of byssinosis for smokers, reduce the odds when employ=3, and to reduce the difference in odds between dust=2 and dust=3 for male workers. However, qualitatively the conclusions from the model remain the same (the odds of byssinosis is still higher for smokers than for non-smokers etc.). For this reason I have only included the results for the full data model in the report.

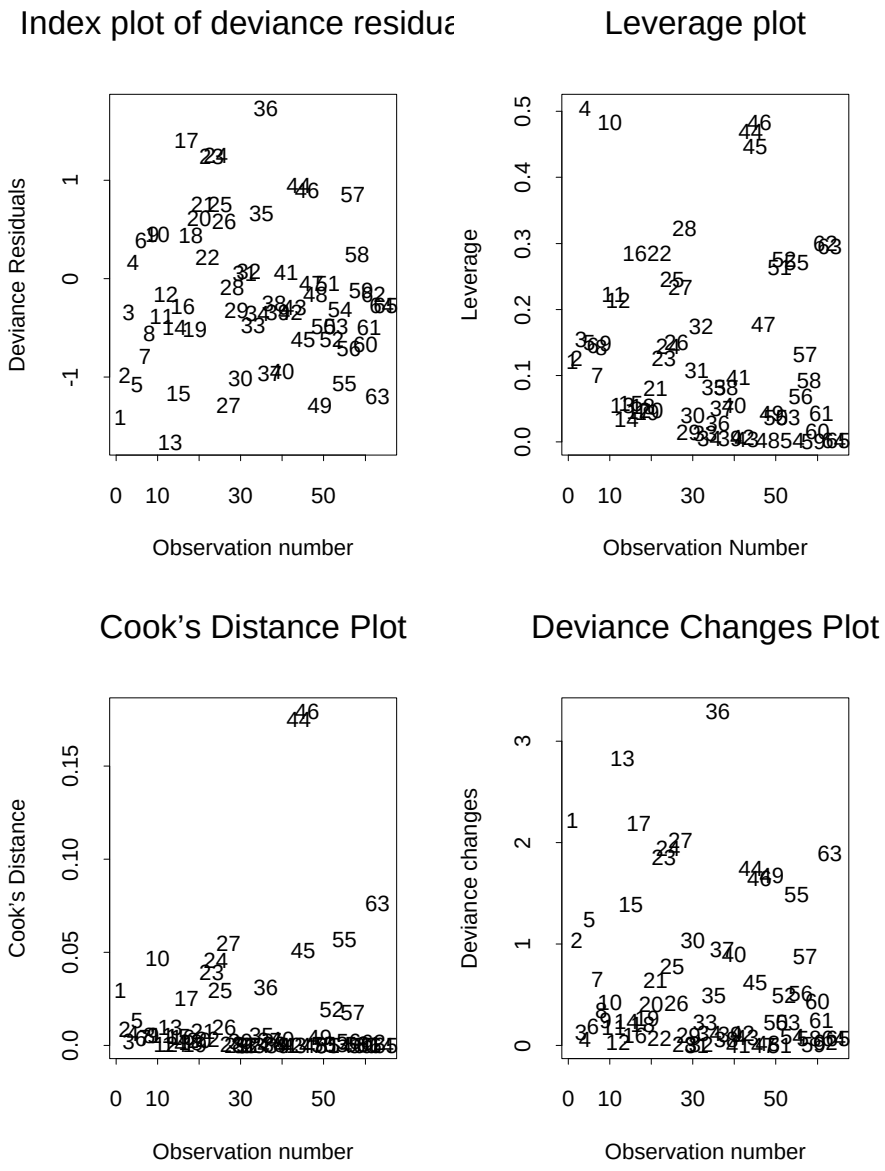

Figure 1: Diagnostic Plots for the Fitted Model

475.330 Assignment 4: Marking Guide

This assignment asks the students to use logistic regression to analyse the Byssinosis data. The assignment is worth a total of 15 marks.

The assignment should consist of two parts: a report that is understandable by a non-statistics major and a statistical appendix that explains their analysis.

The Report

The report should contain the following:

1. An executive summary. This should be a short paragraph that summarises the main findings from their report.
2. A brief description of the data. A plot is not necessary.
3. They should clearly explain how the probability or the odds of byssinosis is related to the explanatory variables. They should discuss how each of the explanatory variables affects π or the odds. When discussing the effects of sex and dust make sure that they take the dust:sex interaction into account.
4. They should give an indication of the range of $\hat{\pi}$ (or the estimated odds) that can occur for different combinations of the explanatory variables.
5. They need to identify points 44 and 46 as being unusual and discuss their impact on the estimated probabilities.

Statistical Appendix

The statistical appendix should contain:

1. A justification of the model they selected. I think the only suitable model is the one identified in the model answers.
2. They need to include diagnostic plots and a discussion of the impact of points 44 and 46.

Allotment of marks

- Give 5 marks for presentation. The writing should be clear and concise. The report and the statistical appendix should include all the parts listed above. In the report, look for explanations that would be understandable to non-statistics students. Graphs should have informative captions. The statistical appendix should give a explanation of their analysis.
- Give 5 marks for content of the report.
  - Give 4 marks for explaining how the factors affect the occurrence of byssinosis. Make sure that they take the interaction into account of the 2 active interactions.
  - Give 1 mark for evaluating the range of probabilities (or odds) that can occur. They should do some confidence intervals for these.
- Give 5 marks for the statistical appendix.
  - Give 2 marks for identifying a suitable model and justifying this model in the statistical appendix.
  - Give 1 mark for explaining how they used that model to get estimates and confidence intervals.
  - Give 1 marks for identifying points 44 and 46 as being influential.
  - Give 1 marks for discussing the influential points. They should note that since these points have large Cook's distance they will have an impact on the fitted model.

Note:

- Include short comments indicating why a student has lost marks.