Logistic Regression of Byssinosis Data

Byssinosis, also called *Brown Lung,* or *Brown Lung Disease,* is a type of pneumoconiosis caused by dust from cotton and other fibers. When inhaled, the dust stimulates histamine release, which causes constriction of the air passages, making breathing difficult. Over time the dust accumulates in the lung, producing a typical discoloration that gives the disease its common name. The disease afflicts mill workers and takes several years of exposure before manifestations are noticed. It can progress to chronic bronchitis and emphysema. Between 1979 and 1992 there were 183 deaths due to byssinosis, 110 of which were concentrated in three states (South Carolina 15 cases, North Carolina 76 cases, and Georgia, 19 cases).

The file byssinosis.sas7bdat was compiled from a survey of 5419 workers in the US cotton industry[1]. It contains information on five explanatory variables: race, sex, and smoking status of the worker, his or her level of dust exposure, and the duration of his or her employment in the textile industry.

Table 1 Variable Names and Descriptions

Variable Name	Description
DUST	Dustiness of the workplace (Hi / Med / Low)
RACE	Worker's race (N = non-white / W = white)
SEX	Worker's sex (F = female / M = male)
SMOKE	Worker's smoking status (Smk = Smoker / NSm = non-smoker)
TIME	Length of employment (< 10 / 10-20 / >20)
Byssinosis	Yes/No

Assignment. Analyze the data and write a short report.

Download the data to a convenient directory such as C:\TEMP (***right*** click on the file name). Use SAS PROC Logistic to estimate logistic regression coefficients and adjusted odds ratios for the explanatory variables.

```
LIBNAME MYDATA "C:\TEMP";
PROC LOGISTIC DATA=MYDATA.BYSSINOSIS;
    CLASS DUST(REF="Low") RACE(REF="W") SEX(REF="M")
          SMOKING(REF="NSm") TIME(REF="<10") / PARAM = REF;
    MODEL BYSSINOSIS(EVENT="Yes") = DUST RACE SEX SMOKING TIME;
RUN;
```

Make a table like Table 10.4 in the textbook. Discuss which variables are significant risk factors (or protective factors). Drop the insignificant explanatory variables (race and sex) and re-run the logistic regression. This is called a "reduced model". This time ask for a file of $\hat{p}_x$ values. For the reduced model, make another table like Table 10.4 and one like Table 10.5 in the textbook. Note that there are 18 different categories in the reduced model (3 dust levels by 2 smoking categories by 3 time categories). Your version of Table 10.5 will not mention race and sex, since they are not used as explanatory variables in the reduced model. In your discussion, be sure to specifically address the issue of whether medium dust levels produce a significant increased risk compared to low dust levels.

[1] Higgins, J.E. and Koch, G.G. (1977) Variable selection and generalized chi-square analysis of categorical data applied to a large cross-sectional occupational health survey. *International Statistical Review*, 45, 51-62.

Reduced Model & Predictions.

```
DATA REQUEST;
     INPUT DUST $ SMOKING $ TIME $;
DATALINES;
     Low NSm <10
     Low NSm 10-20
     Low NSm >20
     Med NSm <10
     Med NSm 10-20
     Med NSm >20
     Hi  NSm <10
     Hi  NSm 10-20
     Hi  NSm >20
     Low Smk <10
     Low Smk 10-20
     Low Smk >20
     Med Smk <10
     Med Smk 10-20
     Med Smk >20
     Hi  Smk <10
     Hi  Smk 10-20
     Hi  Smk >20
;;;;

PROC LOGISTIC DATA=MYDATA.BYSSINOSIS;
     CLASS DUST(REF="Low") RACE(REF="W") SEX(REF="M")
          SMOKING(REF="NSm") TIME(REF="<10") / PARAM = REF;
     MODEL BYSSINOSIS(EVENT="YES") = DUST SMOKING TIME;
     SCORE CLM DATA=Request OUT=Predict;
RUN;
PROC PRINT DATA=PREDICT;
RUN;
```