DEPARTMENT OF STATISTICS

Course STATS 330: Advanced Statistical Modelling

Tutorial Sheet 8: October 2, 2008

This tutorial is designed to give you practice in fitting logistic models with categorical explanatory variables (binary anova).

In this tutorial you will investigate data from a survey of workers in the US cotton industry. The data 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.

The data has been put into a file called **`byssinosis.txt`** which can accessed from the webpage. This data set contains 7 variables: **`dust, race, sex, smoking, employ, yes`** and **`total.`**

The first five of these are the categorical explanatory variables which have the following codings for their levels:

`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.

For each combination of the explanatory variables, **`total`** represents the total number of workers and **`yes`** represent the number that suffer from byssinosis. This is grouped data.

We will use binary ANOVA to investigate the relationship between the explanatory variables and the occurrence of byssinosis. Primary interest is in how the incidence of byssinosis is affected by the amount of dust in the workplace but the other variables need to be taken into account as well.

You need to fit a model that explains the connection between the probability of contracting byssinosis and the other variables. You should run suitable diagnostic checks for your fitted model.

Task 1: Read in the data and create a suitable data frame for the analysis

This should be routine by now. The data is found in "Data for past assignments"; it was assignment 4 in 2001.

Caution: since the factors are coded as integers, you will need to make sure that the explanatory variables are converted to factors. See tutorial sheet 6, and below.

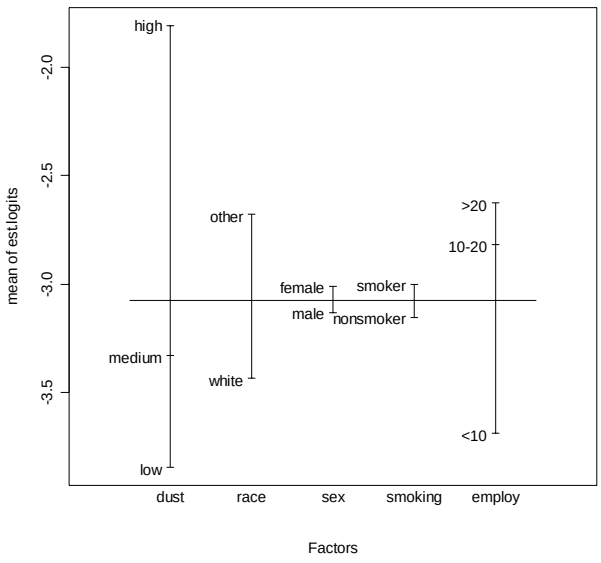
Task 2: Plot the data

A good plot is that drawn by the **`plot.design`** function. (See lecture 24). You will need to use the logits (log-odds) of the different factor level combinations. It is helpful to create a new data frame, which includes the logits, and also recodes the categorical variables with more meaningful labels than 1, 2 or 3. This can be done with the **`data.frame`** function. Note the use of **`labels`** to recode the factors. The following assumes we have attached the original data frame.

```
tut8.df<-
data.frame(dust=factor(dust,labels=c("high","medium","low")),
race=factor(race,labels=c("white","other")),
sex=factor(sex,labels=c("male","female")),
smoking=factor(smoking,labels=c("smoker","nonsmoker")),
employ=factor(employ,labels=c("<10","10-20",">20")),
yes, total, est.logits=log((yes+0.5)/(total-yes+0.5)))
```

We can get the desired plot from the code

```
plot.design(est.logits ~ dust*race*sex*smoking*employ,
data=tut8.df)
```



Which factors have the biggest impact on byssinosis?

Task 3: Fit the model

The data are in grouped form. To fit a separate probability to each combination of factor levels, we type

```
byss.glm<-glm(cbind(yes,total-yes)~dust*race*sex*smoking*employ,
family=binomial,data=tut8.df)
```

This is in fact the maximal model. The warnings are due to the fact that there are a lot of zero counts. You can ignore them. Let's use **anova** to see if we can drop some terms.

```
anova(byss.glm)
```

This indicates that the terms dust, smoking and employ are important (why?). Also important is the sex:dust interaction, so we put sex in as well. (Remember if we put in an A:B interaction we must put in A and B as well.) This leads to the model

```
cbind(yes,total-yes)~dust+smoking+employ + sex + sex:dust
```

Let's fit this and look at the results:

```
> model2<-glm(cbind(yes,total-yes)~dust+smoking+employ + sex + sex:dust,
family=binomial,data=tut8.df)
> summary(model2)

Call:
glm(formula = cbind(yes, total - yes) ~ dust + smoking + employ +
    sex + sex:dust, family = binomial, data = tut8.df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6584  -0.6089  -0.2756   0.2555   1.7389

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)           -1.7529     0.1564 -11.207  < 2e-16 ***
dustmedium            -3.2770     0.5101  -6.424 1.33e-10 ***
dustlow               -2.8783     0.2422 -11.883  < 2e-16 ***
smokingnonsmoker      -0.6578     0.1945  -3.382 0.000719 ***
employ10-20            0.4640     0.2512   1.847 0.064704 .
employ>20              0.6367     0.1836   3.467 0.000526 ***
sexfemale             -0.9990     0.6106  -1.636 0.101792
dustmedium:sexfemale   2.0056     0.8309   2.414 0.015791 *
dustlow:sexfemale      1.2576     0.6823   1.843 0.065286 .
---
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:  36.019  on 56  degrees of freedom
AIC: 160.69
```

Task 4. Interpret the coefficients

Take careful note of which factor levels are the baseline. For example, the baseline for **`dust`** is "high". The coefficients for dustmedium and dustlow are negative, so high dust is clearly more harmful than low or medium dust. Also, the probability increases as the length of exposure increases. How would you interpret the dust:sex interaction?

Task 5: check the model

Run the diagnostics. Is the deviance OK? Any strange points?

Refer to Arden's model answers to assignment 4, 2001 for more details.