DEPARTMENT OF STATISTICS

Course STATS 330: Advanced Statistical Modelling

Tutorial Sheet 7: September 25, 2008

This tutorial is designed to give you practice in fitting simple logistic models.

In this tutorial we will be using the **mice data** which you are required to type in yourselves. It is shown below.

The data for this tutorial comes from a study that investigated the effect of insulin on laboratory mice. The response was whether or not the mice had convulsions when given insulin. We are interested in modelling how the proportion of mice with convulsions varies with the dose applied.

Mice data

Dose (mg)	Number with convulsions	Number of mice
3.4	0	33
5.2	5	32
7.0	11	38
8.5	14	37
10.5	18	40
13.0	21	37
18.0	23	31
21.0	30	37
28.0	27	30

You need to fit a model that explains the connection between the probability of a convulsion and the dose.You should run suitable diagnostic checks for your fitted model.

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

With more complicated data it is best to use an editor to create a text file of data. This can then be read in the usual way. However, this data set is so simple that the data frame can be created directly in R. Use the code

```
dose<-c(3.4,5.2,7.0,8.5,10.5,13.0,18.0,21.0,28.0)
r<-c(0,5,11,14,18,21,23,30,27)
n<-c(33,32,38,37,40,37,31,37,30)
mice.df<-data.frame(dose=dose,r=r,n=n)
```

Note that the data is in "grouped form"

Task 2: Fit an initial model

We will fit a model of the form

$$\text{Probability of a convulsion} = \pi = \exp(\alpha + \beta * \text{DOSE})/\left(1 + \exp(\alpha + \beta * \text{DOSE})\right),$$

or, in log-odds form,

$$\text{log-odds of a convulsion} = \log(\pi/(1-\pi)) = \alpha + \beta * \text{DOSE}.$$

Because the data is in grouped form, we need to use the special way of specifying the response:

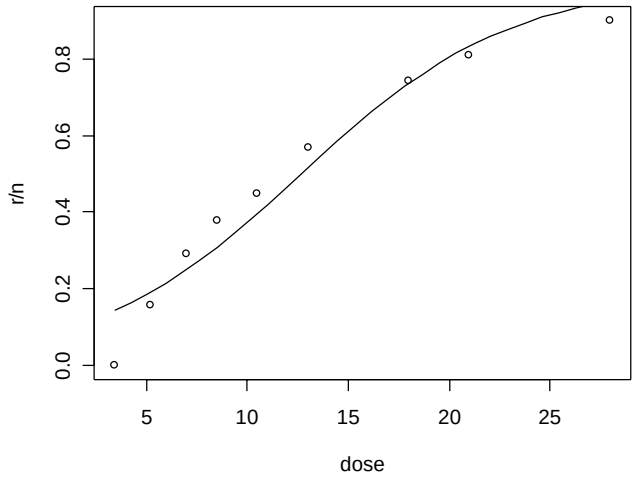
```
> mice.glm<-glm(cbind(r,n-r)~dose,family=binomial,data=mice.df)
```

Task 3: Check the model

Again, since the data are in grouped form, we have an estimate, namely the sample proportion r/n, of the probability of a convulsion at each dose. This estimate doesn't depend on the logistic assumption. We can compare this estimate with the value predicted by the logistic model:

```
> newdose<-seq(3.4,28,length=30)
> est.probs<-predict(mice.glm,newdata=data.frame(dose=newdose),type="response")
> plot(dose,r/n)
> lines(newdose, est.probs)
```

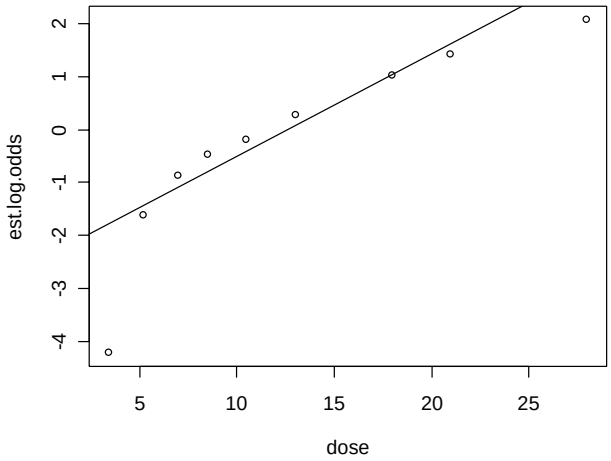
This gives the graph below. The fit is not very good. Another way of plotting is to plot the estimated log-odds against the log-odds predicted by the model:



We estimate the log-odds by log((r+0.5)/(n-r+0.5)). The 0.5 is inserted to handle the case when r is 0 or n (otherwise the estimate would be undefined). If we plot these estimates versus the log-odds predicted by the model, using code

```
> est.log.odds<-log((r+0.5)/(n - r +0.5))
> plot(dose,est.log.odds)
> abline(coef(mice.glm))
```

we get

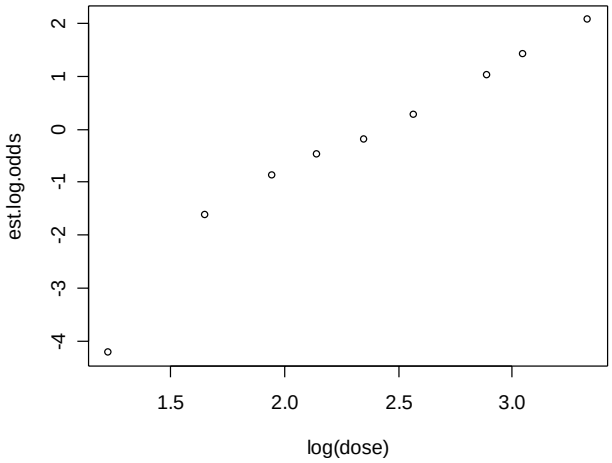


This confirms our feeling that the fit is not very good.

Task 4: Find a better model

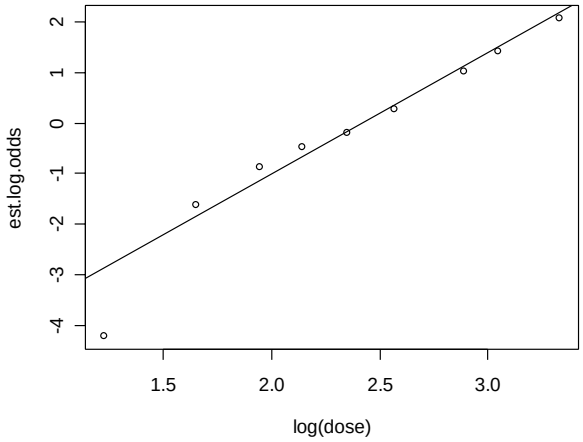
To improve the fit, we can try and "straighten the plot" using a power transformation on dose. After some fiddling about, we try a log:

```
> plot(log(dose),est.log.odds)
```



Apart from the first point, this is much better, and strongly suggests a model where the log-odds is a linear function of the log dose. Let's fit such a model:

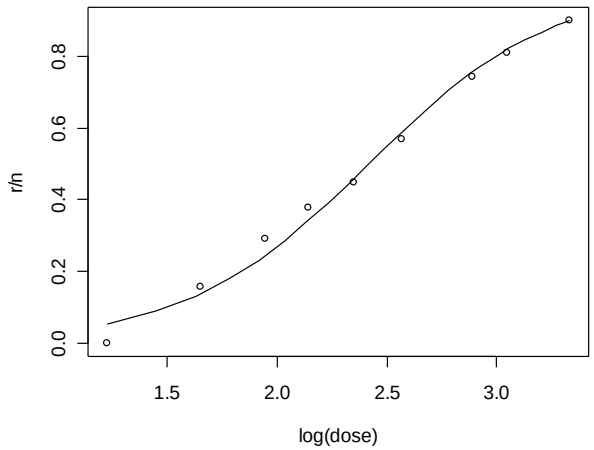
```
> plot(log(dose),est.log.odds)
> logdose=log(dose)
> mice.log.glm<-glm(cbind(r,n-r)~logdose,family=binomial,data=mice.df)
> abline(coef(mice.log.glm))
```



There is a hint the first point is influential. Still, the result is not too bad. On the probability scale, using the code

```
> plot(log(dose),r/n)
> est.probs.logs<-predict(mice.log.glm,newdata=data.frame(logdose=log(newdose)),
type="response")
> lines(log(newdose),est.probs.logs)
```

we get



which is again not too bad. Our final model is (get coefficients from summary)

$$\text{log-odds of a convulsion} = -5.7907 + 2.3964 \times \log(\text{DOSE}).$$