M.Sc. Practical MT2005 week 6

Logistic and Log-linear Models

1 Low Birth Weights

Hosmer & Lemeshow (1989) give a dataset on 189 births at a US hospital, with the main interest being in low birth weight. The following variables are available in data frame `birthwt` in library `MASS`

`low`	birth weight less than 2.5 kg (0/1),
`age`	age of mother in years,
`lwt`	weight of mother (lbs) at last menstrual period,
`race`	white / black / other,
`smoke`	smoking status during pregnancy (0/1),
`ptl`	number of previous premature labours,
`ht`	history of hypertension (0/1),
`ui`	has uterine irritability (0/1),
`ftv`	number of physician visits in the first trimester,
`bwt`	actual birth weight (grams).

Although the actual birth weights are available, we concentrate here on predicting if the birth weight is low from the remaining variables. The dataset contains a small number of pairs of rows that are identical apart from the ID; it is possible that these refer to twins but identical birth weights seem unlikely.

We use a logistic regression with a binomial (in fact 0/1) response. It is worth considering carefully how to use the variables. It is unreasonable to expect a linear response with `ptl`. Since the numbers with values greater than one are so small we reduce it to an indicator of past history. Similarly, `ftv` can be reduced to three levels. With non-Gaussian GLMs it is usual to use treatment contrasts.

```
library(MASS)
?birthwt
options(contrasts = c("contr.treatment", "contr.poly"))
attach(birthwt)
race <- factor(race, labels = c("white", "black", "other"))
table(ptl)
ptd <- factor(ptl > 0)
ftv <- factor(ftv)
table(ftv)
levels(ftv)[-(1:2)] <- "2+"
table(ftv)  # as a check
bwt <- data.frame(low = factor(low), age, lwt, race,
    smoke = (smoke > 0), ptd, ht = (ht > 0), ui = (ui > 0), ftv)
detach(); rm(race, ptd, ftv)
```

We can then fit a full logistic regression.

```
birthwt.glm <- glm(low ~ ., family = binomial, data = bwt)
summary(birthwt.glm, cor = F)
```

Since the responses are binary, even if the model is correct there is no guarantee that the deviance will have even an approximately chi-squared distribution, but since the value is about in line with its degrees of freedom there seems no serious reason to question the fit. We use stepwise selection by AIC:

```
birthwt.step <- stepAIC(birthwt.glm)
birthwt.step$anova
birthwt.step2 <- stepAIC(birthwt.glm, ~ .^2 + I(scale(age)^2)
                         + I(scale(lwt)^2), trace = F)
birthwt.step2$anova
summary(birthwt.step2, cor = F)$coef
table(bwt$low, predict(birthwt.step2) > 0)
plot(birthwt.step2)
```

Note that although both `age` and `ftv` were previously dropped, their interaction is now included, the slopes on `age` differing considerably within the three `ftv` groups.

Try also three-way interactions.

An alternative approach is to predict the actual live birth weight and later threshold at 2.5 kilograms. Try this: it produces somewhat worse predictions with around 52 errors.

2 Speed Limits in Sweden

An experiment was performed in Sweden in 1961–2 to assess the effect of speed limits on the motorway accident rate. The experiment was conducted on 92 days in each year, matched so that day j in 1962 was comparable to day j in 1961. On some days the speed limit was in effect and enforced, while on other days there was no speed limit and cars tended to be driven faster. The speed limit days tended to be in contiguous blocks.

The data set is given in the data frame `Traffic` with factors `year`, `day` and `limit` and the response is the daily traffic accident count, `y`.

Fit Poisson log-linear models and summarize what you discover. Some first steps in the analysis would be

```
options(contrasts = c("contr.treatment", "contr.poly"))
try(rm(y)) # to be careful
attach(Traffic)
tr <- data.frame(y, day = factor(day), year = factor(year), limit)
detach()
fm <- glm(y ~ day + year + limit, data = tr, family = poisson)
fm2 <- stepAIC(fm, scope = list(upper = . ~ . + limit:year))
plot(fm2)
exp(coef(fm2)["limit"])
```

3 Cancer Deaths of Atomic Bomb Survivors

This example was set as an assessed practical in 2003.

> The table shows data on cancer deaths amongst survivors of the atomic bombs dropped on Japan in WWII, categorized by the time (in years) after the bomb that death occurred and the amount of radiation exposure received (in rads).

exposure	0–7	8–11	12–15	16–19	20–23	24–27	28–31
0	10/262	12/243	19/240	31/237	35/253	48/227	73/220
25	17/313	17/290	17/285	47/280	50/275	65/269	71/262
75	0/38	2/36	1/35	5/34	8/34	7/33	12/32
150	1/28	0/26	4/25	1/25	6/24	12/24	11/23
250	1/13	1/12	0/12	4/12	3/11	7/11	13/10
400	0/15	2/14	5/14	3/14	2/13	3/13	5/13

> The entries in the table are
>
> > number of deaths / person-years at risk
>
> where the 'denominator' is computed by summing the number of years (includng fractional years) alive in the time interval over all the people who had that amount of radiation and who were alive at the beginning of the interval.
>
> Find a suitably parsimonious model for the rate of cancer deaths from this dataset, and express its results graphically. Your conclusions should include a statement on the effect of radiation on the cancer death rate, including quantifying your uncertainty.

It should be clear that we want to use a Poisson log-linear regression model for the numbers of death with offset log of the person-years at risk. I set up a 42-row data frame[1] `Atomic` with the data by row, response `Deaths`, offset variable `AtRisk` and factors `dose` and `tyears`. The basic model should be something like

```
options(contrasts=c("contr.treatment", "contr.poly"), digits=5)
fm <- glm(Deaths ~ tyears + dose + offset(log(AtRisk)), family = poisson,
          data = Atomic)
summary(fm, cor=F)
```

Note that the fit is fairly good, although the residual deviance is a little large. Given how small some of the counts are, we should not take this too seriously.

Now we can explore if either factor could be dropped (*a priori* unlikely)

```
dropterm(fm, test="Chisq")
```

Can we simplify this? It looks as if a linear effect in tyears would be OK (and as the intervals are of different length, I used the midpoints).

[1] which you can load from `Atomic.sdd` on the course website – just double-click on this to load into S-PLUS.

```
Atomic$ot <- c(4,10,14,18,22,26,30)[Atomic$tyears]
fm2 <- glm(Deaths ~ ot + dose + offset(log(AtRisk)), family = poisson,
        data = Atomic)
anova(fm, fm2, test="Chisq")
```

shows this is a reasonable simplification.

We can use prediction to get the fitted rates and hence mean survival times.

```
preddata <- Atomic
preddata$AtRisk <- rep(1,42)
Atomic$lrates <- predict(fm2, preddata)
tab <- matrix(exp(-Atomic$lrates),6,7, byrow=T)
dimnames(tab) <- list(levels(Atomic$dose), levels(Atomic$tyears))
round(tab,1)
```

Note that the Poisson model corresponds to exponential survival times, so this is a reasonable interpretation. We could show this graphically by

```
trellis.device()
xyplot(exp(-lrates) ~ ot | dose, data=Atomic, type="b")
```

or using multiple lines on one graph. They should include some confidence intervals, too. For example

```
Atomic$selrates <- predict(fm2, preddata, se=T)$se
Atomic$up <- Atomic$lrates + 2*Atomic$selrates
Atomic$low <- Atomic$lrates - 2*Atomic$selrates

xyplot(exp(lrates) ~ ot | dose, data=Atomic, type="b", subscripts=T,
  panel = function(x,y,subscripts, ...) {
     panel.xyplot(x,y, ...)
     error.bar(x, y, exp(Atomic$low[subscripts]),
               exp(Atomic$up[subscripts]), add=T, incr=F)
  }
)

xyplot(lrates ~ as.numeric(as.character(dose)) | tyears, data=Atomic,
       type="b", subscripts=T,
  panel = function(x,y,subscripts, ...) {
     panel.xyplot(x,y, ...)
     error.bar(x, y, Atomic$low[subscripts],
               Atomic$up[subscripts], add=T, incr=F)
  }
)
```

although for a report make sure you use better labels (by creating new variables). Note that any of the panels of the second trellis could be regarded as answering the question, as could a plot of the dose effects with standard errors superimposed, on rate or survival time scale (possibly plotted on log scale but labelled on one of those scales).