Methods in Epidemiologic Research

Sample Problems

Chapter 18 - Modelling Count Data

Preparation

As we indicated in the sample problems for Chapter 15, we are going to carry on with the -mi- dataset but now we will use Poisson and negative binomial models to evaluate how various factors influence the length of hospital stay (-los-).

The variables we will use in this exercise are listed below. The outcome will be -los-. Most of the variables listed have already been defined. -age_inv- and -bmi_ct- are transformed versions of -age- and -bmi-, respectively (see below).

```
Contains data from C:\mer\data\mi.dta
  obs:         2,965
 vars:            13                          28 Feb 2012 10:53
 size:       112,670
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
id              float   %9.0g                 patient id
los             int     %8.0g                 length of hospital stay
sex             byte    %8.0g                 gender
age             float   %9.0g                 age at admission
age_inv         float   %9.0g                 transformed and centred age (1/age)-0.015
white           float   %9.0g                 race=White
mar_c2          float   %9.0g                 married Y/N
bmi             float   %9.0g                 body mass index
bmi_ct          float   %9.0g                 centred bmi (bmi - 28)
prmi            byte    %8.0g                 previous MI
card            byte    %8.0g                 cardiac arrest during hospitalization
cabg            byte    %8.0g                 coronary artery bypass surgery
died_hosp       float   %9.0g                 died in hospital
-------------------------------------------------------------------------------
Sorted by:
```

Note: It is important to note that, for the purpose of this exercise, we are ignoring the possible clustering of lengths of stay within hospital (*ie* some hospitals may have, on average, longer stays than others). We will evaluate the impact of this in the exercises for Chapter 22.

Questions

Your primary interest is how marital status (married vs no-married) (-mar_c2-) and body mass index (-bmi-) influence the length of stay. However, we also investigate the role of other factors.

1. Draw a causal diagram incorporating all of the predictors listed above.
2. Is there any evidence that the effect of age is not linear? (You should revisit what was done in Chapter 15 to address this question).
   (a) create quadratic terms and add them to the model
3. Poisson model

(a) Rather than going through a full model building exercise, we will start with the final model from the sample problems for Chapter 15. (-los- as outcome and the following as predictors: -sex-, -age_inv-, -white-, -mar_c2-, -bmi_ct-, -prmi-)

(b) Compute the expected number of days in hospital for: a "baseline" individual (female, age=67, non-white, not married, bmi=28, no previous mi); the same except married; a "long stay" individual (age~=85 (age_inv=-0.003), bmi=48 (bmi_ct=20) and had a previous mi)

(c) Fit the same Poisson regression model in the GLM framework

(d) Express the same model .... but in terms of count ratios (called incidence rate ratios (IRR) in the statistical output, because the data are quite often incidence data)

4. Poisson model diagnostics - overall model fit

(a) Compare the observed and predicted counts of days in hospital

(b) Compute both deviance and Pearson residual based goodness-of-fit statistics (dispersion tests)

(c) Given that there is clear evidence of overdispersion, refit the model but compute scaled SE (scaled by the Pearson dispersion parameter)

5. Detailed diagnostics

(a) Compute residuals (deviance, Pearson and Anscombe) as well as Cook's distance values for each observation. As a first step, determine if there are many Pearson residuals which you would consider extreme? The determine if any individuals had particularly large Cook's D.

(b) Diagnostic plots - generate plots of Anscombe residuals against predicted values and Cook's distances.

(c) Refit the model without obs. # 366.

6. Negative binomial regression

(a) Fit the same model as developed above using negative binomial regression. Is there evidence that an NB model would be preferred to a Poisson model?

(b) Are the estimates of the fixed effects similar from a Poisson and NB model?

(c) Are NB models fit by maximum likelihood and by GLM comparable?

7. NB diagnostics

(a) Obtain both deviance and Pearson χ^2 statistics. Do they provide evidence of lack of fit?

(b) Compute Pearson residuals for all observations. Are there an excess of observations <-3 or >3?

(c) Compute Cook's D for all observations. Do any observations stand out as having very large values? If so, refit the model with this/these observation(s).

8. Zero-inflated models

(a) Fit a zero-inflated NB model. Is there any evidence that there are more values of zero (*ie* patients discharged on the same day as admission) than would be expected?