8.7 Practical 8

Dataset required: `medpar.dta` (from Stata website)

Introduction

In this practical we will explore models for a count variable and look at the two approaches described in the lecture for handling overdispersion. We will use publicly available data on length of hospital stay from Arizona in the US.

In our analyses we will focus on the following variables.

Variable	Description
`los`	Length of hospital stay, in days
`age`	Age group (factor variable)
`type1`	Binary variable indicating elective admission
`type2`	Binary variable indicating urgent admission
`type3`	Binary variable indicating emergency admission

Note that admission type is here described using three binary indicator (or dummy) variables, rather than as a 3-level categorical variable.

Aims

1 Understand how to fit models to count data using the glm command (in Stata).

2 Understand how to compare such models (in Stata).

3 Understand the concept of overdispersion, and how to identify it.

Analysis

The dataset is available from the Stata Press website; to load the data type:

```
use http://www.stata-press.com/data/hh3/medpar, clear
```

1 Explore the length of stay variable. How is it distributed? What are the minimum and maximum values in this dataset?

Construct a 95% confidence interval for the mean length of stay using the standard approach taught in the Analytical Techniques course (a Wald 95% confidence interval).

2 Use the `glm` command to fit a Poisson model for length of stay, with no covariates.

What does the constant coefficient in this model represent? Construct a 95% confidence interval for the mean length of stay.

3 Repeat question 2 using the `robust` option with the `glm` command.

Discuss: Compare and contrast the 95% confidence intervals in questions 1, 2 and 3. Which is most appropriate? Which is least appropriate?

4 We will now explore which (if any) factors are related to length of hospital stay. First fit a Poisson model (without robust standard errors) to assess whether length of hospital stay is related to the type of admission. Interpret each of the parameter estimates in your model.

5 Add age as factor variable to your model in question 4 and perform a likelihood ratio test of whether age (treated as categorical) is a predictor of length of hospital stay, adjusting for type of admission. Is this test an appropriate one?

6 The likelihood ratio test cannot be used with robust standard errors. However, we can still use a (multivariate) Wald test, which is based on the robust variance-covariance matrix, to assess if age group has an effect on length of stay, adjusting for admission type. Refit the model in question 5 with robust standard errors, and then perform this Wald test using the following command.

```
testparm i.age
```

Discuss: Compare and contrast the results of the tests in questions 5 and 6. Which is the appropriate test?

Working together with one or more colleagues (in your Breakout Room if online), write a short paragraph (suitable for a medical journal) to summarise your findings concerning the effects of the type of admission on length of stay in hospital for this model. If online, one of you should post your group's paragraph in the Zoom chat.

7 Use `nbreg` to fit a negative binomial regression model to the length of stay variable, with no covariates. Interpret the likelihood ratio test reported at the bottom of the output, and relate it to your findings from earlier questions.

8 Fit a negative binomial regression model with indicator variables for type of admission as the only predictor variables. Use the estimated parameters to calculate the predicted mean and variance of lengths of stay for each of the three types of admission. To do this you will to use the negative binomial regression model result that when the expectation of the outcome is μ then its variance is $\mu(1 + \alpha\mu)$ (see section 8.5.2 of the notes). Compare the predicted variances with the observed variances for each admission type.

Discuss: What do you conclude about the estimates and variances predicted from the negative binomial model? What are your conclusions about the best way to model these data?