SPSS Data Analysis Using Poisson Regression For Count Data

Poisson regression is most often used to model count variables, this is what we will be doing in this practical. Please also utilise the lecture slides for guidance. In this session you will use a simulated dataset to complete Poisson Regression in SPSS. This practical ***will not*** go over all the aspects of good research such as data cleaning, verification of model assumptions, model diagnostics or potential follow up analyses.

Opening the data

Save the "PoissonPractical" data file to your machine and open in SPSS. Using either the drop down menus

->File -> Open -> Data select file in windows directory box and click open (remember clicking past when using menus will paste syntax into a syntax editor)

Or, the following script amended where emboldened and italicised to indicate where you have the file saved.

```
GET FILE='\\C$\Users\fpearson\Documents\PoissonPractical.sav'.
DATASET NAME PoissonPractical WINDOW=FRONT.
```

Task for this afternoon

The data set has 7 variables, '**daysInHosp**' is the outcome variable indicating the days each study participant spent in hospital in the last year. In this session you will analyse variable '**daysInHosp** to see if there is an effect due to '**gender'** and **'FirstTestResult'** or **'SecondTestResult'** using poisson regression for count data.

Below is a list of other analysis methods discussed earlier which you could use to analyse this dataset. Some of the methods listed may give a better model fit than the Poisson model.

- Negative binomial regression - Negative binomial regression can be used for over-dispersed count data, that is when the conditional variance exceeds the conditional mean. It can be considered as a generalization of Poisson regression since it has the same mean structure as Poisson regression and it has an extra parameter to model the over-dispersion. If the conditional distribution of the outcome variable is over-dispersed, the confidence intervals for Negative binomial regression are likely to be narrower as compared to those from a Poisson regession.
- Zero-inflated regression model - Zero-inflated models attempt to account for excess zeros. In other words, two kinds of zeros are thought to exist in the data, "true zeros" and "excess zeros". Zero-inflated models estimate two equations simultaneously, one for the count model and one for the excess zeros.

Exploring the data further using descriptive statistics and graphs

Now you have the data open the first thing to do is explore the variables using descriptive statistics, graphs and where appropriate cross-tabulations.
Again, use either drop down menus, or, use script in the syntax window – remember using the menus and then pressing ***paste*** rather than ***ok*** will show you the syntax you need for each function.

Output you may wish to generate:

A histogram of the outcome variable. In the Poisson lecture you were warned against using OLS regression for count data as it often breaks the assumptions of OLS. A simple histogram can show us if this is a good recommendation.

1. What does the histogram show you?

Descriptive statistics of variables including the mean and variance.

2. What do you learn from the descriptive statistics of the included variables?

In the lecture we discussed the fact that for a poisson distribution the mean and variance are the same.

Before using any further analysis methods, let's run a poisson regression, even though you may believe that the poisson distribution will not give the best model fit. Poisson regression can be followed up with further commands which test the model fit.

Run poisson regression for count data

Again, you can use either the drop down menus (which are shown below), or, script in the syntax window – remember using the menus and then pressing ***paste*** rather than ***ok*** will show you the syntax you need for each function.

->Analyze -> Generalized Linear Models -> Generalized Linear Models

Further info: The 'General Linear Models' option is used for OLS regression and the 'General Estimating Equations' option is used for multi-level modelling of clustered data

Check options Poisson loglinear in type of model tab → go to the response tab → identify dependent variable (see outlined task) → go to the predictors tab → identify any factors and covariates (see outlined task, think about which variables are categorical and which are numerical) → go to the model tab → identify the model (see outlined task) → go to the statistics tab → include exponential parameter estimates (check the box) → hit OK (Or PASTE!)

3. What are your findings (think about goodness of fit, effect measures and inferences that can be drawn)?

Further Tasks

If you have time re-run the analysis including ethnicity '***ethnic***' and hospital type '***hospital***' in the model

4. Does this model produce a better fit? What other things might you now check for to improve model fit (HINT: Already identified during descriptive analysis)?

If you have time, re-run the analysis in question 3, but this time using negative binomial regression. Negative binomial regression is specified in almost exactly the same way as a Poisson model the only difference is on the Type of Model tab the Negative binomial with log link option is selected. The output are interpreted in the same way as for Poisson regression.

5. Does this model produce a better fit? Undoubtedly the effect estimates for the model have changed, but have the main inferences drawn from the model changed too?