STATS 3001 / STATS 4104 / STATS 7054
Statistical Modelling III
Assignment 3
2022

DEADLINE:

- Friday 20th May 2021 5pm (Week 10)

QUESTIONS:

1. Consider the following contingency table

Treatment	Controls	Cases	
A	Y_{11}	Y_{12}	N_1
B	Y_{21}	Y_{22}	N_2

In this question, we show how you could use Poisson regression to test for an effect of treatment on the proportion of cases and controls.

Normally we could use logisitic regression, and so we will show that these are equivalent.

Poisson model

Let

$$Y_{ij} \sim Po(\lambda_{ij})$$

and

$$\log(\lambda_{ij}) = \gamma_0 + \gamma_1 x_i + \gamma_2 x_j + \gamma_3 x_i x_j,$$

where

$$x_i = \begin{cases} 0 & \text{if treatment is } A, \\ 1 & \text{if treatment is } B \end{cases}$$

and

$$x_j = \begin{cases} 0 & \text{if controls,} \\ 1 & \text{if cases} \end{cases}$$

(a) Write in terms of the γs, equations for
   i. $\log(\lambda_{11})$,
   ii. $\log(\lambda_{12})$,
   iii. $\log(\lambda_{21})$, and
   iv. $\log(\lambda_{22})$.

(b) Show that the conditional distribution of

$$Y_{i2}|Y_{i1} + Y_{i2} = n_i$$

is binomial $B(n_i, \pi_i)$ and find an expression for π_i in terms of λ_{i1} and λ_{i2}.

(c) Let

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_i$$

where

$$x_i = \begin{cases} 0 & \text{if treatment is } A, \\ 1 & \text{if treatment is } B \end{cases}$$

Hence, or otherwise, write an expression for β_1 in terms of γs.

(d) Hence, or otherwise, show that testing that treatment has no effect on the probability of being a cases, is equivalent for test for no interaction in the Poisson model.

2. The data set `lung_cancer`, collected by Anderson (1977)[1], contains the population size and number of cases of lung cancer in four Danish cities, stratified by age.

   The purpose of the analysis is to fit a Poisson regression model to predict the number of cases of lung cancer.

   (a) Read the data `lung_cancer.csv` into R.
   (b) Perform an EDA of the data. In particular produce appropriate plots to look at the relationship between the number of cases and age and city. Also look at the relationship between the proportion of the population that was a cases and age and city.
   (c) Fit a Poisson rate regression (denoted `M1`) with cases as the response variable, log(pop) as an offset, and no predictors.
   (d) Fit a Poisson rate regression (denoted `M2`) with cases as the response variable, log(pop) as an offset, and age and city as the predictors.

[1] E.B. Andersen (1977), Multiplicative Poisson models with unequal cell rates, Scandinavian Journal of Statistics, 4:153-158.

(e) Fit a Poisson rate regression (denoted `M3`) with cases as the response variable, log(pop) as an offset, and age, city and log(pop) as the predictors.
(f) Use ANOVA to compare `M1` and `M2`. Use this to decide if age and city are significant predictors in predicting the number of cases.
(g) Find the AIC for all three models.
(h) Produce a summary of the coefficients for `M2`.
(i) Obtain the Pearson residuals for `M2`.
(j) Plot the residuals versus
    i. fitted values,
    ii. age, and
    iii. city.
(k) A health campaign was introduced in Fredericia. In 1980, the number of cases of lung cancer in the `40-54` age group (population now 4000) was 5. By calculating the probability of have five cases or fewer assuming the same rate as given by M2, decide if this is a significant decreases in the rate of lung cases.
(l) Go to myUni and complete quiz.