MAST30027: Modern Applied Statistics

Week 5 Lab

1. In a binomial model we assume that any given observation is from a $\text{bin}(m, p)$ distribution, for some m and p. That is, we count the number of successes from m i.i.d. bernoulli(p) trials. There are two main ways that overdispersion can arise: the trials are not identically distributed, or the trials are not independent.

   A common way in which we can have heterogeneous trials is via clustering. Suppose that we have m trials split into h clusters of size $k = m/h$, and that the probability of success for a trial in the i-th cluster is p_i. Now suppose that p_i is random, with $\mathbb{E}p_i = p$ and $\text{Var}\, p_i = \tau^2 p(1-p)$. Let the number of successes from cluster i be Z_i and let the total number of successes be $Y = Z_1 + \cdots + Z_h$.

   Show that

   (a)
   $$\mathbb{E}Y = mp$$

   (b)
   $$\text{Var}\, Y = (1 + (k-1)\tau^2)mp(1-p)$$

   Hint: $\text{Var}\, Y = \mathbb{E}\text{Var}\,(Y|X) + \text{Var}\,\mathbb{E}(Y|X)$.

   Thus Y is overdispersed, relative to a binomial.

2. Suppose that $Y_i \sim \text{Poisson}(\lambda_i)$, where $\lambda_i \propto t_i$. For example, if we record the number of burglaries reported in different cities, the observed number will depend on the number of households in these cities. In other cases, the size variable t may be time. For example, if we record the number of customers served by sales people, we must take account of the differing amounts of time worked.

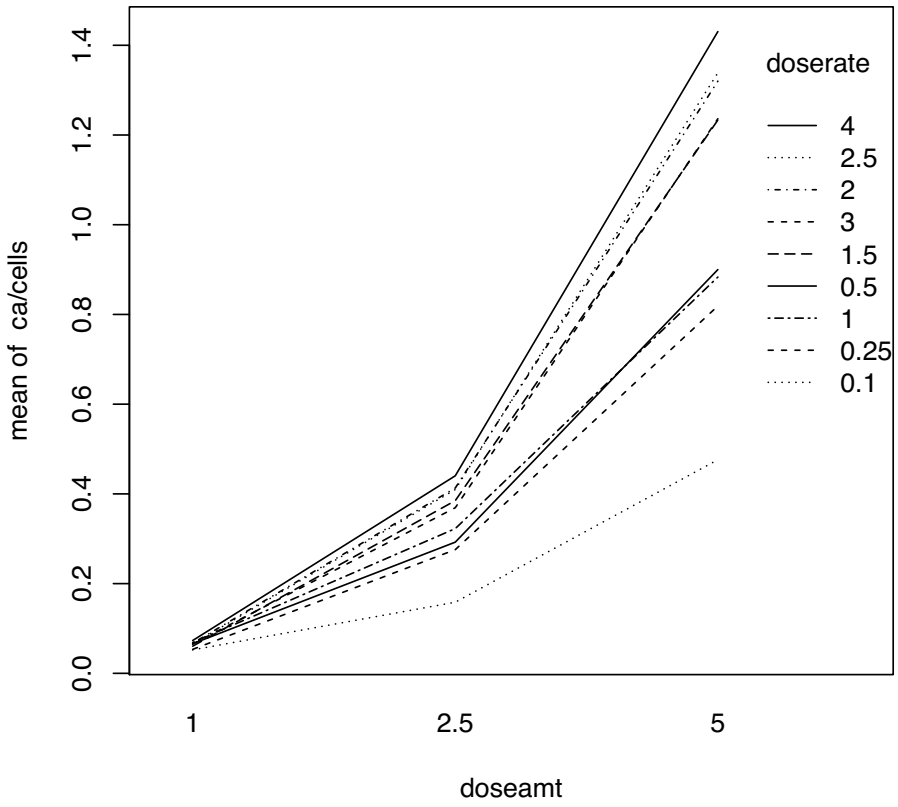
   We can model the rate *per unit time* using a log link via
   $$\log(\lambda_i/t_i) = x_i^T \beta$$
   where x_i are known predictors and β unknown parameters. That is
   $$\log(\lambda_i) = \log t_i + x_i^T \beta.$$

   This is of the form of a Poisson glm with log link, but where the coefficient of $\log t_i$ has been constrained to be 1. This is called a *rate model*.

   In an R model description we can fix the coefficient of a variable to 1 by enclosing it in the `offset` function, viz `y ~ offset(log(t)) + x1 + x2 + ...`.

   In Purott and Reeder (1976), some data is presented from an experiment conducted to determine the effect of gamma radiation on the numbers of chromosomal abnormalities (ca) observed. The number (cells), in hundreds of cells exposed in each run, differs. The dose amount (doseamt) and the rate (doserate) at which the dose is applied are the predictors of interest. We can plot the data as follows

```
> library(faraway)
> data(dicentric)
> with(dicentric, interaction.plot(doseamt, doserate, ca/cells))
```



Fit a rate model to this data and perform diagnostics.