MAST90104: Introduction to Statistical Learning

Week 9 Lab and Workshop

1. The dataset `discoveries` lists the number of great scientific discoveries for the years 1860 to 1959, as chosen by "The World Almanac and Book of Facts", 1975 Edition. Has the discovery rate remained constant over time?

   To answer this question, fit a Poisson regression model with a log link, and use the deviance to compare a null model with models including the year and year squared as predictors.

2. The `ships` dataset from the `MASS` package gives the number of damage incidents and aggregate months of service for different types of ships broken down by year of construction and period of operation. Load the dataset using the commands `library(MASS)` then `data(ships)`.

   Develop a model for the rate of incidents (i.e. a Poisson regression model with log link), describing the effect of the important predictors.

3. The `infert` dataset from the `survival` package presents data from a study of infertility after spontaneous and induced abortion. Using a logistic regression model, analyse and report on the factors related to infertility based on this data. (Don't use the factor stratum, as it is confounded with the other predictors.)

4. The dataset `africa` from the `faraway` package gives information about the number of military coups in sub-saharan Africa and various political and geographical information.

   Use the AIC to choose a parsimonious generalised linear model for the number of coups. Give an interpretation of the effect on the response of the variables you include in your model.

5. The `cornnit` dataset in the `faraway` package contains data on the effect of nitrogen on the yield of corn. Fit a gamma regression to this data, using the `glm` command. You will need to pay attention to the choice of link function (inverse, identity or log), and consider transforming the predictor variable (your first step should be to plot the data).

   (a) Extract the Pearson residuals from the fitted model using the `residuals` function, then use them to estimate the dispersion parameter. Check that your answer agrees with the summary output from your model.

   (b) Suppose your fitted model is `gmod`, then the command `anova(gmod, test="F")` will compare your model against the null model, using an F test. Using the deviances and dispersion estimates reported by `summary(gmod)`, check that the F statistic reported by the `anova` function is correct.

   (c) Now do some diagnostic plots. Can you identify a potential outlier?

   (d) Fit a linear model to the `cornnit` data.
   Which do you prefer, the linear model or the gamma model, and why?

6. The `dvisits` data in the `faraway` package comes from the Australian Health Survey of 1977–78 and consist of 5190 observations on single adults, where young and old have been oversampled.

   (a) Build a Poisson regression model with `doctorco` as the response and `sex, age, agesq, income, levyplus, freepoor, freerepa, illness, actdays, hscore, chcond1` and `chcond2` as possible predictor variables. Considering the deviance of this model, does this model fit the data?

   (b) Plot the response residuals against the fitted values. Why are there lines of observations on the plot?

   (c) Use backward elimination with a critical p-value of 5% to reduce the model as much as possible.

   (d) What sort of person would be predicted to visit the doctor the most under your selected model?

(e) For the last person in the dataset, compute the predicted probability distribution for their visits to the doctor, i.e., give the probability they visit 0,1,2, etc. times.

(f) Fit a comparable (Gaussian) linear model and graphically compare the fits. Describe how the Gaussian and Poisson models differ.

7. 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

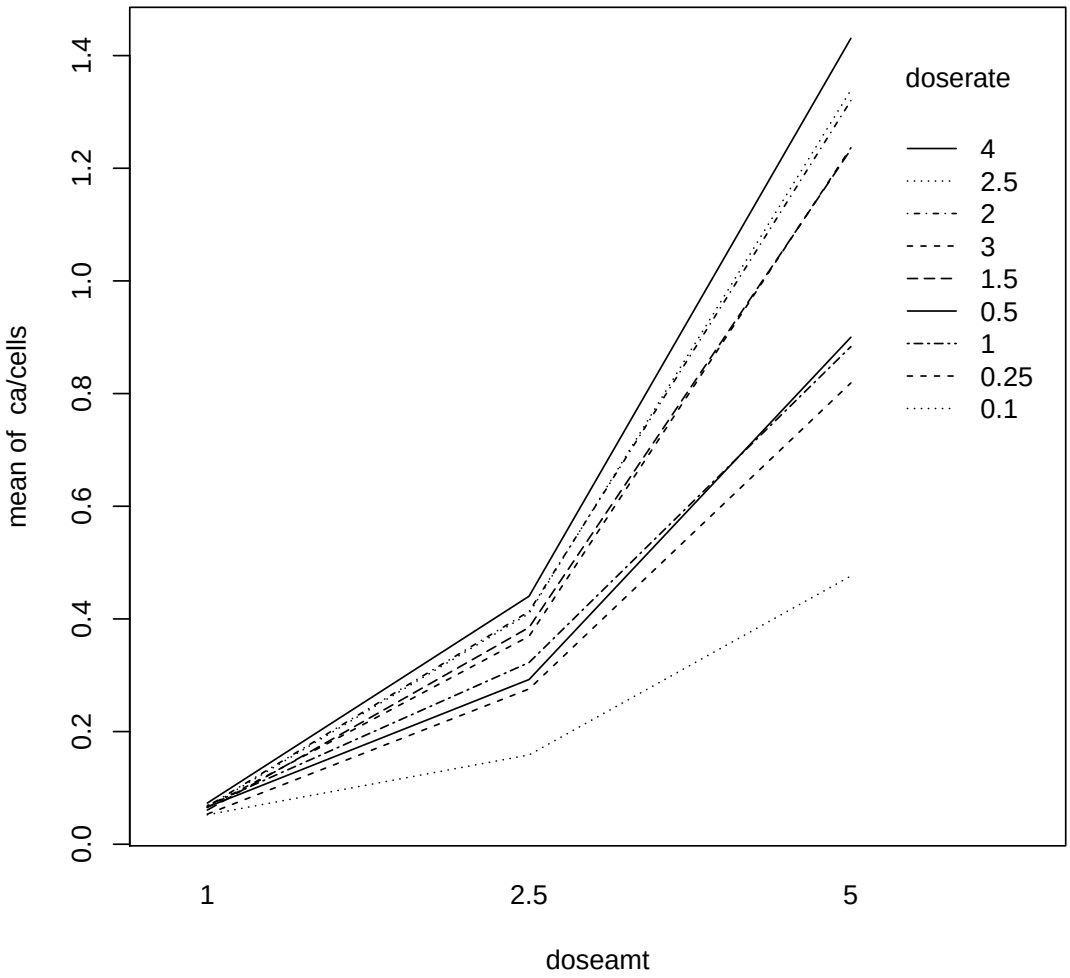
   $$\log(\lambda_i) = \log t_i + x_i^T\beta.$$

   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. Use it to predict the rate of abnormalities when you have 200 cells, `doserate` 3.5 and `doseamt` 5.

8. Verify that for the binomial regression model with logistic link

$$\begin{aligned} \mathbb{E}\frac{\partial l(\boldsymbol{\theta};\mathbf{Y})}{\partial \theta_i} &= 0 \\ -\mathbb{E}\frac{\partial^2 l(\boldsymbol{\theta};\mathbf{Y})}{\partial \theta_i \partial \theta_j} &= \mathbb{E}\left(\frac{\partial l(\boldsymbol{\theta};\mathbf{Y})}{\partial \theta_i}\frac{\partial l(\boldsymbol{\theta};\mathbf{Y})}{\partial \theta_j}\right) \end{aligned}$$

9. Suppose that $\mathbf{Y}$ has pdf $f(\mathbf{y};\boldsymbol{\theta})$ with parameter $\boldsymbol{\theta}$ in some fixed open set of $\mathbb{R}^k$ for some $k = 1, 2, \cdots$. Show that:

$$\begin{aligned} \mathbb{E}\frac{\partial l(\boldsymbol{\theta};\mathbf{Y})}{\partial \theta_i} &= 0 \\ -\mathbb{E}\frac{\partial^2 l(\boldsymbol{\theta};\mathbf{Y})}{\partial \theta_i \partial \theta_j} &= \mathbb{E}\left(\frac{\partial l(\boldsymbol{\theta};\mathbf{Y})}{\partial \theta_i}\frac{\partial l(\boldsymbol{\theta};\mathbf{Y})}{\partial \theta_j}\right) \end{aligned}$$

(Hint: You may assume that f is sufficiently regularly that you may interchange integration and differentiation in computing the expectations.)

10. Suppose that students answer questions on a test and that a specific student has an aptitude T. A particular question might have difficulty d_i and the student will get the answer correct only if $T > d_i$. Consider d_i fixed and $T \sim N(\mu, \sigma^2)$, then the probability that a randomly selected student will get the answer wrong is $p_i = \mathbb{P}(T < d_i)$.

    Show how you might model this situation using a probit regression model.

11. Show that the Gamma density, f, in the form

$$f(y; \lambda, \alpha) = \frac{1}{\alpha}\lambda^{\alpha} y^{\alpha-1} e^{-\lambda y}$$

    is an exponential family with $\theta = -\frac{\lambda}{\alpha}, \phi = \frac{1}{\alpha}$. Identify the functions a, b, c and find the mean and variance functions as functions of θ.

12. Show that the inverse Gaussian density, f, in the form

$$f(y; \mu, \lambda) = \frac{\lambda}{\sqrt{2\pi y^3}} e^{-\frac{\lambda(y-\mu)^2}{2\mu^2 y}}$$

    is an exponential family with $\theta = \frac{1}{\mu^2}, \phi = \frac{1}{\lambda}$. Identify the functions a, b, c and find the mean and variance functions as functions of μ, λ.