STA 2201S Homework 2. *due Wednesday March 4 2015 at 11.59pm on Blackboard*

When answering questions requiring numerical work, the results are to be reported in a narrative summary, in your own words. Tables and Figures may be included, but must be formatted along with the text. Do not include raw `R` code as part of your answer. An exception can be made if you are using `knitr`: code chunk displays in text are permitted. Whether or not you use `knitr`, a complete set of scripts should be provided as an appendix.

1. Log-linear models

   (a) Suppose we observe independent counts y_{jkl} following a Poisson distribution with means μ_{jkl}, $j = 1, \dots, J; k = 1, \dots, K; l = 1, \dots, L$. These could be arranged in a contingency table with J rows, K columns, and L layers. Show that the conditional distribution of $\underline{y}$, conditional on the row totals in each layer, y_{j+l}, is a *product multinomial* distribution, with parameters θ_{jkl}, and give an expression for θ in terms of μ. (Notation: $y_{j+l} = \Sigma_k y_{jkl}$)

   (b) The data in Table 1 is from a retrospective case-control study. A group of ulcer patients was assembled, and a group of control patients without peptic ulcer, and matched on age and gender, was also assembled. Blood groups were ascertained for all subjects and the study was replicated in three cities. The numbers of cases and controls are fixed for each city, so the product multinomial model from (a) seems appropriate. We will write

   $$\log \mu_{jkl} = \gamma + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \eta_l + (\alpha\eta)_{jl} + (\beta\eta)_{kl}.$$

   Identify the constraints on this model corresponding to the following three hypotheses:

      i. Is the distribution of blood groups the same in all cities?

      ii. In each city is there an association between blood group and peptic ulcer?

      iii. Is any such association the same in all cities?

   (c) Analyse the data and summarize the results with reference to these three questions of interest.

		Blood Groups		
		A	O	Total
London	Cases	579	911	1490
	Controls	4219	4578	8797
Manchester	Cases	246	361	607
	Controls	3775	4532	8307
Newcastle	Cases	291	396	687
Upon Tyne	Controls	5261	6598	11859

Table 1: Data from a study of peptic ulcer incidence. *Source:*

2. ELM, 6.5 The Conway-Maxwell-Poisson distribution has probability function

$$\Pr(Y = y) = \frac{\lambda^y}{(y!)^\nu} \frac{1}{Z(\lambda, \nu)}, \quad y = 0, 1, 2, \ldots,$$

where

$$Z(\lambda, \nu) = \sum_{i=0}^{\infty} \frac{\lambda^i}{(i!)^\nu}.$$

(a) Show that this is an exponential family, and identify the mean and variance functions, the canonical parameter θ and the function $b(\theta)$.

(b) When might this distribution prove useful? Show that it allows for over/under dispersion relative to the Poisson distribution.

(c) Assuming a sample $y_1, \ldots, y_n$ where $y_i \sim \mathrm{CMP}(\lambda_i)$, and a regression model based on the log-link; i.e. $\log \lambda_i = x_i^{\mathrm{T}} \beta$, find an expression for the (scaled) deviance:

$$D = 2\{\ell(\tilde{\lambda}; y) - \ell(\hat{\lambda}; y)\},$$

where $\tilde{\lambda}_i$ is the estimate of λ_i under the saturated model, and $\hat{\lambda}_i = \lambda_i(\hat{\beta})$ is the estimate from the regression model.

(d) Fit this model[1] to the Galapagos Islands data analysed in ELM, §3.1, which did show evidence of over-dispersion. Which model do you prefer, this one or the quasi-Poisson? Why? Correction! The CMP model cannot be estimated for the Galapagos data; the factorials involved in the log-likelihood are too large. I tried using a subset of the data (counts less than 100), and while the model could be fit, the over dispersion estimate was on the order of 170, which doesn't quite make sense. This may explain why Faraway doesn't investigate this data any further in Chapter 3. Instead of fitting a CMP model, try fitting a negative binomial model, and see if this offers any improvement over the quasi-Poisson.

3. The article "An estimate of the science-wise false discovery rate and application to the top medical literature" by Jager & Leek (*Biostatistics*, 2014), is posted on the course web page and available via the link in (c). In this paper they attempted to estimate the rate of false discoveries in papers published in leading medical journals.

(a) Construct a 2×2 table with "Null hypothesis true/false" as the two column headings, and "Discovery/No Discovery" as the two row headings. Give a definition (algebraic) of the false discovery rate as a function of the entries in your table.

(b) What model did Jager & Leek use for the distribution of p-values?

(c) Their conclusion was that the rate of false discoveries among published results was 14% with an estimated standard error of 1%. How was the standard error estimated?

(d) There were several discussants of this paper, and all the discussions can be found at `biostatistics.oxfordjournals.org/content/15/1.toc`. Choose one discussion and summarize in a paragraph the main point of the discussant. Comment briefly on this point, and on the reply by Jager & Leek.

[1] There is an `R` package, `COMPoissonReg` available on CRAN.

4. *Data Set for Final Project*: Please provide the following information about the data set you will analyze for your final project. If you have submitted some of this information with HW1, please copy and paste here again.

   (a) The data source

   (b) The size of the data – number of observations and number of covariates

   (c) the response variable(s)

   (d) a description of the potential covariates

   (e) the scientific questions of interest

   When you submit your final project, it will consist of (at least) the following parts:

   (a) a description of the scientific problem of interest

   (b) how (and why) the data being analyzed was collected

   (c) preliminary description of the data (plots and tables)

   (d) models and analysis

   (e) summary for a statistician of the analysis and conclusions

   (f) non-technical summary for a non-statistician of the analysis and conclusions