

UNIVERSITY OF WOLLONGONG
School of Mathematics and Applied Statistics
STAT902. Advanced Data Analysis

ASSIGNMENT 4

Due: 5:00pm Monday 12th April, 2010.

Local students: please put under lecturer's door.

Remote students: please e-mail or post by this date.

1. For each of the density functions and corresponding means listed below; write $\ln[y]$ in canonical form:

$$\ln[y; \eta] = y\eta - b(\eta) + c(y)$$

Identify the *canonical link* transformation g defined by $\eta = g(\mu)$ and the functions b and c .

- (a) $[y; \mu] = \mu^y e^{-\mu}/y!$, $y = 0, 1, \dots$; $\mu > 0$
(b) $[y; \mu] = (1/\mu)e^{-y/\mu}$, $y > 0$; $\mu > 0$.
(c) $[y; \mu] = \sqrt{\frac{1}{2\pi y^3}} e^{-(y-\mu)^2/(2\mu^2 y)}$, $y > 0$; $\mu > 0$.

Hint: Note that the Appendix for this assignment does the required calculations for the Bernoulli distribution.

2. The following set of observed data:

1.448, 4.759, 4.635, 2.972, 6.151, 3.556, 3.256, 1.860, 1.248, 1.454

is modelled to come from the Gamma($\alpha, 1$) distribution for some $\alpha > 0$ to be estimated via maximum likelihood. The log-likelihood for general random sample X_1, \dots, X_n from the Gamma($\alpha, 1$) distribution is

$$\ell(\alpha) = (\alpha - 1) \sum_{i=1}^n \ln(X_i) - n\bar{X} - n \ln \Gamma(\alpha).$$

Note that this cannot be maximised algebraically, so we need to call upon a numerical technique such as Newton-Raphson to find $\hat{\alpha}$, the maximum likelihood estimator of α . The following R script has been set up to solve this problem, but some details have been omitted (see occurrences of `### TASK FOR THIS ASSIGNMENT ###` below). Complete the details, run the script and report the solutions. Note that the asymptotic standard error is obtained as a by-product of Newton-Raphson iteration and that is also illustrated here. Finally, results on differentiation of the function $h(x) = \ln \Gamma(x)$ is given in the notes 'Likelihood Theory and Methods', handed out in Class 2, and the required functions (e.g. digamma) are available in R.

```

x <- c(1.448, 4.759, 4.635, 2.972, 6.151, 3.556, 3.256, 1.860, 1.248, 1.454)
n <- 10
alphaHat <- 5 # Starting guess
for (i in 1:8) # Perform 8 Newton-Raphson iterations.
{
  FirstDrv <- ### TASK FOR THIS ASSIGNMENT ###
  SecondDrv <- ### TASK FOR THIS ASSIGNMENT ###
  alphaHat <- alphaHat - FirstDrv/SecondDrv
  print(alphaHat)
}
cat("\n\n Maximum likelihood estimate of alpha is:\n")
print(alphaHat)
cat("\n\n Asymptotic standard error is:\n")
print(1/sqrt(-SecondDrv))

```

3. This question involves some aspects of Poisson regression, illustrated via data on ragweed pollen counts – corresponding to the paper:

Stark, P. C., Ryan, L. M., McDonald, J. L. and Burge, H. A. (1997). Using meteorologic data to model and predict daily ragweed pollen levels. *Aerobiologia*, **13**, 177–184.

Prediction of pollen counts is important since it can help allergy sufferers plan medication.

- Download the dataset `ragweed1994.txt` from the Computer Code and Data page on course web-site. The advice in Assignment 3 about getting data from the web to your computer in the right format applies.
- Open an R session and, if working in Windows, make sure that you change directories to that where the file `ragweed1994.txt` lives.
- Type `ragweed1994 <- read.table("ragweed1994.txt", header=TRUE)` to make the data available to the current session.
- Type `pairs(ragweed1994)` to visualise the data. This command provides pairwise scatterplots of all variables in the spreadsheet. The response variable is `pollenCount`. The other variables are meteorological, such as `windSpeed`. The variable `transDayNum` is the logarithm of the day number plus one. This transformation was used by Stark *et al.* (1997) since it modelled the pollen counts as a function of time reasonable well.
- Type `attach(ragweed1994)` to make all components of the data-frame `ragweed1994` available to the current session. For example, the variable `windSpeed`, corresponding to the column of `ragweed1994` with the same name, is now a variable available in the current session.
- Type `plot(dayNum, pollenCount, type="l")` to show the time series of ragweed pollen in Kalamazoo, Michican, for the summer of 1994.
- It is now time to fit a regression model. Since the response variable is a count we will treat it as Poisson (the appropriateness of the Poisson model will be postponed for now). Type

```

fit <- glm(pollenCount ~ dayNum + transDayNum + tempResid + rain + windSpeed,
family=poisson)

```

to fit the Poisson GLM:

$$\text{pollenCount}_i \sim \text{Poisson}\{\exp(\beta_0 + \beta_1 \text{dayNum}_i + \beta_2 \text{transDayNum}_i + \beta_3 \text{tempResid}_i + \beta_4 \text{rain}_i + \beta_5 \text{windSpeed}_i)\}, \quad 1 \leq i \leq 75$$

- (h) Type `print(summary(fit))` to get a summary of the regression fit. Notice that all 5 predictor variables are strongly statistically significant.
- (i) We will now discuss what is going on behind the scenes to get the standard errors of the coefficients. Based on theory in class, the Fisher information matrix of $\boldsymbol{\beta} = (\beta_0, \dots, \beta_5)$ is

$$I_{75}(\boldsymbol{\beta}) = \mathbf{X}^T \text{diag}\{\exp(\mathbf{X}\boldsymbol{\beta})\} \mathbf{X}.$$

- (j) Explain why the function $\exp(x)$ appears as the operator on $\mathbf{X}\boldsymbol{\beta}$.
- (k) Issue the following commands:

```
X <- cbind(rep(1, 75), dayNum, transDayNum, tempResid, rain, windSpeed)
print(X)
betaHat <- fit$coef
print(betaHat)
etaHat <- as.vector(X%*%betaHat)
FishInfo <- t(X)%*%diag(exp(etaHat))%*%X
print(FishInfo)
invFishInfo <- solve(FishInfo) # matrix inversion
print(invFishInfo)
print(sqrt(diag(invFishInfo)))
```

Comment on the final set of numbers in relationship to previous output.

Your submitted assignment should include the print-outs from issuing the R commands.

(Warning: Please note that submission of a print-out corresponding to work done by another participant will lead to, possibly severe, penalty.)

Appendix: Canonical Transformation Calculations for the Bernoulli Distribution

$$[y] = [y; \mu] = \mu^y (1 - \mu)^{1-y}, \quad y = 0, 1; 0 < \mu < 1.$$

Hence,

$$\ln[y] = y \ln(\mu) + (1 - y) \ln(1 - \mu) = y \ln\{\mu/(1 - \mu)\} + \ln(1 - \mu).$$

The canonical parameter is therefore

$$\eta = \ln\{\mu/(1 - \mu)\} = \text{logit}(\mu)$$

and the canonical link is $g(x) = \text{logit}(x)$. Note that, via function inversion methods, we obtain

$$\mu = g^{-1}(\eta) = \frac{e^\eta}{1 + e^\eta}.$$

Therefore,

$$1 - \mu = \frac{1}{1 + e^\eta} \implies \ln(1 - \mu) = -\ln(1 + e^\eta)$$

and, in terms of the canonical parameter η ,

$$\ln[y; \eta] = y\eta - \ln(1 + e^\eta) = y\eta - b(\eta) + c(y)$$

where

$$b(x) = \ln(1 + e^x) \quad c(y) = 0.$$