MATH 523B Assignment #3 (revised 2003.03.23)
Due Thursday 27 March 2003 afternoon

Notes:

- This Assignment comprises 6 questions on 4 pages.
- Problem **3. (a)** corrected 2003.03.23.
- The data are available in R format with the extension `.dput` or in text format with the extension `.dat` with the first row containing the variable names.
- Data files in R format can be read in using the `dget` command and `attach`'ed as in Assignment 1.
- If you use the R format, be aware that the **factors may not be coded as such**. You will have to modify them appropriately.

1. Consider the following data indicating the number of horses which have won their race according to their lane number at the start of the race.

Lane number	1	2	3	4	5	6	7	8
Number of winners	32	21	19	20	16	11	14	11

(a) Plot the number of winners against the lane number. Determine an appropriate generalized linear model for the data (i.e. Exponential family, link function and linear model).

(b) Using the model determined in part **(a)**, test for an effect of lane number on winning.

(c) Test the model you fitted in part **(b)** for goodness of fit.

2. Consider the following contingency table displaying numbers of people surveyed with the indicated self-perception and opinion of small cars.

Op. of small cars	Self-perception		
	Bad	Fair	Good
Bad	79	58	49
Fair	10	8	9
Good	10	34	42

In the following, justify any assumption you make in performing the tests.

(a) Test for association between self-perception and opinion of small cars. Justify any assumption you make in performing your test.

(b) Using a suitable, non-saturated model, test for interaction between self-perception and opinion of small cars.

(c) Test the model you fitted in part **(b)** for goodness of fit.

3. Data on the passenger class/crew status, age, gender and survival status of those aboard the ocean liner Titanic, when it sank off the coast of Newfoundland in 1912, were assembled from various sources to yield the following table:

	Child				Adult			
	Female		Male		Female		Male	
Class	Survived	Died	Survived	Died	Survived	Died	Survived	Died
Crew	0	0	0	0	20	3	192	670
I	1	0	5	0	140	4	57	118
II	13	0	11	0	80	13	14	154
III	14	17	13	35	76	89	75	387

The Titanic thus was carrying 2201 passengers and her sinking led to 1490 deaths, according to these sources.

(a) Explain why a null model log-linear model to fit these data has 27 rather than 31 degrees of freedom. Without using software, compute the intercept of such a log-linear model.

The rest of this question concerns a Binomial regression model.

(b) Under a model involving only the main effects of Age, Class and Gender, what are the log-odds of survival for a Female Crew member?

(c) In the presence of the three main effects above

i. Is the effect of Age the same for every Class?

ii. Is the effect of Gender the same for every Class?

(d) Briefly discuss the interaction estimates of Age and Class that you found in part **(c) i.** and link them with the data.

(e) Is it necessary to account for different effects from both Sex and Age in different Classes simultaneously to model the data appropriately? Can you test the goodness of fit of such a model?

(f) Fit a log-linear model to the data equivalent to **(c) ii.**.

4. **(a)** Show that the MLE of the mean $\hat{\mu}$ of a random sample $Y_1, \ldots, Y_n$ from an Exponential family is given by $\hat{\mu} = \overline{Y}$. (Recall that for a 1-1 function g and parameter θ, the MLE of $g(\theta)$ is $g(\hat{\theta})$.)

(b) Let $Y_1, \ldots, Y_n$ be a random sample from Exponential family f with mean μ and scale parameter ϕ. Show that $\overline{Y}$ is from Exponential family f with mean μ and scale parameter ϕ/n.

5. **(a)** Consider the Inverse Gaussian distribution

$$f(y) = \frac{\exp\sqrt{\chi\psi}}{\sqrt{2\pi y^3}}\sqrt{\chi}\exp\left(-\frac{1}{2}\left[\psi y + \frac{\chi}{y}\right]\right)$$

Show that the Inverse Gaussian is an Exponential family by expressing it in the form

$$f(y) = \exp\left[\frac{a(\mu)y - b(\mu)}{\phi} + C(\phi, y)\right]$$

where $\mu = \mathbb{E}[Y]$. Provide expressions for the mean μ and the scale parameter ϕ in terms of χ and ψ.

(b) Show that the variance function of the Inverse Gaussian is given by $V(\mu) = \mu^3$.

(c) Let Y_i be an Inverse Gaussian random variable with mean μ_i and scale parameter ϕ, $i = 1, \ldots, n$. Consider the model

$$\frac{1}{\mu_i} = \mathbf{x}_i\boldsymbol{\beta}, \qquad i = 1, \ldots, n$$

for some known $1 \times p$ row vectors $\mathbf{x}_i$ and unknown column vector $\boldsymbol{\beta}_{p\times 1}$. Assuming that $X = [\mathbf{x}_1', \ldots, \mathbf{x}_n']'$ has rank p, show that the MLE of $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}} = (X'WX)^{-1}X'\mathbf{e}$$

where $\mathbf{e} = [1, 1, \ldots, 1]'$ and $W = \text{diag}(Y_1, \ldots, Y_n)$, the diagonal matrix with diagonal $Y_1, \ldots, Y_n$.

6. The 100 United States senators were asked to vote on two articles of impeachment regarding then president Bill Clinton, the first concerning perjury, the second concerning obstruction of justice (file *impeach.**). For each senator, several data were collected, including the number of votes against Clinton (variable TOT, taking on values 0, 1 or 2) and an external measure of conservatism for the senator (variable CONSERV, on a scale of 1 to 100).

(a) Fit a Binomial (logistic) regression of TOT on CONSERV (and an intercept). Is the measure of conservatism CONSERV significant?

(b) From your model in part **(a)**, form the new outcomes and weights

$$Y_i = \log\frac{\hat{p}_i}{1-\hat{p}_i} \text{ and } w_i = 2\hat{p}_i(1-\hat{p}_i).$$

Fit a (Normal) linear model $\mathbb{E}[Y_i] = \alpha + \beta\text{CONSERV}_i$ with weights w_i.

i. Explain the relationship between the parameter estimates of the linear and the logistic models.

ii. Explain the value of the estimated dispersion parameter in the linear model.

(c) Now form

$$Y_i^* = Y_i + \frac{tot_i - 2\hat{p}_i}{w_i}.$$

Show that fitting a (Normal) linear model $\mathbb{E}[Y_i^*] = \alpha + \beta \text{CONSERV}_i$ with weights w_i is equivalent to executing one more step in the Gauss-Newton algorithm to fit the original Binomial model. How must the estimated standard errors in the linear model be adjusted to reflect the dispersion parameter for the Binomial family?

(d) Use the above models to perform an approximate test for outliers in the model of part **(a)**.

End of Assignment 3.