189.523B Assignment #3 (revised 2001.03.09)
Due Friday 16 March 2001 in class

Notes:

- This Assignment comprises 6 questions on 3 pages.
- 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 are not coded as such**. You will have to modify them appropriately.

1. Let the model M_0 be given by

$$M_0 : \mathbb{E}[\mathbf{Y}] = X\boldsymbol{\beta}$$

where $\mathbf{Y}$ is an $n \times 1$ random column vector, X is a $n \times (n-1)$ real matrix with $\text{rank}(X) = n - 1$ and $\boldsymbol{\beta}$ is an unknown $(n-1) \times 1$ vector of unknown parameters.
Define the i^{th} standardized residual to be

$$\hat{z}_i = \frac{Y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}}}{\sqrt{\hat{\sigma}_0^2(1 - \mathbf{x}_i(X'X)^{-1}\mathbf{x}_i')}}$$

where $\hat{\boldsymbol{\beta}}$ is the least-squares estimator of β under M_0, x_i is a row vector of covariates and $\hat{\sigma}_0^2 = \text{SSE}_0/[n-(n-1)] = \text{SSE}_0$. Show that $\hat{z}_i$ =-1, 0 or 1.

Hint: You may wish to use the following two facts:

- Let $R_{p\times p}$ and $H_{p\times p}$ be two matrices such that
  i. R and H are symmetric and idempotent, i.e. $R = R' = R^2$ and $H = H' = H^2$;
  ii. $RH = 0$;

  then $\text{rank}(R)+\text{rank}(H) = \text{rank}(R+H)$.
- If $R_{p\times p}$ is a symmetric, non-negative definite matrix with rank $q \leq p$, then there exists a matrix $A_{p\times q}$ with rank q such that $R = AA'$.

2. Consider the car data from Assignment 2 and the best model for log(price) you found in question **5. (a)** of that Assignment.

(a) Test the model for outliers at the 5% level.

(b) Produce a scatterplot of the residuals against the fitted values of the models. Identify the outliers on the graph.

(c) Graphically examine the assumption of constant variance by plotting the residuals against the covariates in your model (for expediency, do not plot the residuals against interactions, just the main effects). Use scatterplots for continuous covariates and boxplots for factor (categorical) covariates. What do you conclude?

(d) Produce a p-p plot of the residuals in your model. What do you you conclude?

(e) Compute the Kolmogorov-Smirnov statistic. Comment on the results.

3. **(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.

4. **(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 $Y_1, \ldots, Y_n$.

5. 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.

6. 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.

End of Assignment 3.