MATH 523B Assignment #4
Problem 1 due Thursday 17 April 2002 at 16:00

Notes:

- This Assignment comprises 4 questions on 4 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 may not be coded as such**. You may have to modify them appropriately.

1. Methadone is given to heroin addicts to help them overcome their addiction. Two methadone clinics in Australia collected data on the time to discharge of their patients after the start of a methadone treatment. The data consist in 238 cases (in file *heroin.**). The variables collected are described in the following table:

Variable name	Variable values	Description
Clinic	factor, 1 or 2	clinic ID
Status	indicator, 0 or 1	1=discharged, 0=not yet discharged
Prison	factor, No/Yes	if Yes, patient has a prison record
Time	continuous, >0	time in days to discharge or censoring
Dose	continuous, >0	daily methadone dosage, in mg

SOURCE: Caplehorn, J. (1991). Methadone dosage and retention of patients in maintenance treatment. *Medical Journal of Australia* **154**: 195–199.

The data are assumed to be Exponentially distributed. Possible effects of a previous prison record are ignored.

(a) Is the effect of Methadone dosage on discharge time linear on a log scale?

(b) Is the effect of Methadone dosage on discharge time the same for both clinics?

(c) What is the expected discharge time of a patient in Clinic 1 who receives a dose of 70 mg/day?

(d) What is the probability that a patient from Clinic 1 who receives 70 mg/day of methadone will be discharged before 1 year?

(e) In which clinic is a patient more likely to be observed to be discharged? How does the dosage affect the probability of not yet being discharged? Briefly discuss possible implications on the validity of the model.

2. Consider the motorettes example in the hand-out, page 8.4 (the data are available in file *motors.**.).

(a) Fit a log-linear model, assuming that the time to failure has an Exponential distribution.

(b) Predict the mean time to failure for a motorette run at 130°C, which is the design temperature of interest in the experiment.

(c) What is the probability of a motor running at 130°C failing in the first year?

3. Feigl & Zelen (1965) consider the survival of patients with acute myelogenous leukemia. The patients were divided into two groups, AG+ and AG-, according to the presence or absence respectively of a morphological character in the white blood cell (covariate *AG*). The white blood cell count (covariate *WBC*) was also recorded. Note that for leukemia, a higher blood cell count is usually associated with a poorer prognosis (i.e. a smaller probability of surviving a given length of time).
The data are available in file *wbc.**. AG is coded as a factor in *wbc.dput* with levels "pos" and "neg" for AG+ and AG- respectively. In *wbc.dat*, AG is in the first column, with values of 1 and 2 corresponding to AG+ and AG- respectively. Assume that the survival time (variable *surv*), in weeks, has a Gamma distribution, with a log link between expected survival and the linear models below.

(a) Carry out suitable tests to see if the mean survival time depends on:

(i) the log of the white blood cell count;

(ii) whether the patient is AG+ or AG-.

(b) Is the effect of log(*WBC*) the same for AG+ and AG- patients?

(c) Suppose that the true distribution of the survival times is Exponential. How would you adjust the estimates and standard errors fitted from the larger model in part **(b)** to reflect this fact?

(d) Test to see if the assumption that the data have an Exponential distribution is satisfied.

4. Data concerning the length of stay in Emergency Departments (ED) was collected in several Québec hospitals, and is available in file *edlos.dput*. Some of the variables collected are described in the following table:

Variable	Variable values	Description
los	continuous, > 0	Length of stay in ED in hours
age	continuous, > 0	Age of patient in years
percsev	factor `None`, `Light`, `Medium`, `High`	patient's perception of the severity of his/her condition
ccrank	continuous integer from 1 to 172	external ranking of the severity of the patient's chief complaint
mental	factor, `Yes`, `No`	if `Yes`, history of mental illness
respir	factor, `Yes`, `No`	if `Yes`, history of respiratory disease
mi	factor, `Yes`, `No`	if `Yes`, history of myocardial infarction

This data set is *only* available with the `dget` command, to preserve the level labels.

(a) Plot a histogram of the length of stay on 1) identity scale and 2) log scale. Plot length of stay and log(length of stay) against age. Briefly discuss regression modelling options, with any other plot or fact you need.

(b) Fit a Normal model to the log of the length of stay to test for the significance of *percserv* in the presence of the other variables.

(c) Fit a Gamma regression model to the length of stay data with a log link to test for the significance of *percserv* in the presence of the other variables in the table.

(d) What is the estimated effect of a 10-year increase in age on the expected length of stay in the alternative model (with *percserv*) from parts **(b)** and **(c)**? Discuss the differences in interpretation in both cases.

The rest of this question concerns a Gamma model with log link.

(e) The levels of *percsev* are ordered in a natural way. Create a continuous variable *pscont* which takes on values 0,1,2,3 at the corresponding levels of *percsev*. Test for the linearity of the effect of *percsev* in the presence of the other variables.

(f) Create variables *pscont2* and *pscont3*, respectively the square and cube of variable *pscont*. Fit a model to the length of stay with *pscont*, *pscont2* and *pscont3* (but not *percsev*) in the presence of the other covariates. Explain why you obtain the same fit when you fit these three variables and when you fit factor *percsev* instead.

(g) Consider a quasi-likelihood model with variance function $V(\mu) = \mu^2$ and log link, regressing the length of stay on all other variables. Describe the fit of this model compared to the fit of the alternative model from part **(c)**.

(h) If $Y \sim \Gamma(\alpha, \beta)$, then $\mathbb{E}\left[Y\right] = \alpha\beta$ and $2Y/\beta \sim \chi^2_{2\alpha}$. Use these results to test that the data are normally distributed using the Kolmogorov-Smirnov statistic. (*Hints*: If X is a random variable from a χ^2_d distribution with cdf $F_{\chi^2_d}$, then $F_{\chi^2_d}(X)$ is a Uniform random variable on $(0,1)$. The R command to produce the cdf at `x` of a chi-squared random variable with d degrees of freedom is `pchisq(x,df=d)`.)

(i) The coefficient of variation of a random variable Y is defined as $\mathrm{CV}(Y) = \mathbb{E}[Y]/\sqrt{\mathrm{Var}[Y]}$. Explain why Gamma Exponential family Generalized Linear Models are also called models with constant coefficient of variation. What is the relationship between the dispersion parameter and the coefficient of variation of the data?

End of Assignment 4.