York SPIDA
Spring 2010
J. Fox

Problems: Generalized Linear Models

1. Long (1990, 1997) investigates factors affecting the research productivity of doctoral students in biochemistry. Long's data on 915 biochemists are in the file
< http://socserv.mcmaster.ca/jfox/Books/Applied-Regression-2E/datasets/Long.txt>
and a codebook for the data set located is at
< http://socserv.mcmaster.ca/jfox/Books/Applied-Regression-2E/datasets/Long.pdf >
You can read the data into R with the command

```
Long <- read.table(
    "http://socserv.mcmaster.ca/jfox/Books/Applied-Regression-2E/datasets/Long.txt",
    header=TRUE)
```

The response variable in this investigation, `art`, is the number of articles published by the student during the last three years of his or her PhD programme. The explanatory variables include `fem` (gender), `mar` (marital status), `kid5` (children five years old or younger), phd (prestige of PhD department), and `ment` (articles published by the student's mentor).

Examine the distribution of the response variable. Based on this distribution, does it appear promising to model these data by linear least-squares regression, perhaps after transforming the response? Then, following Long, perform a Poisson regression of `art` on the explanatory variables. Refit Long's model allowing for overdispersion (using a quasi-Poisson model). Does this make a difference to the results? Finally, apply "regression diagnostics" to the Poisson or quasi-Poisson regression model. If you identify any problems, try to deal with them.

2. Employ "regression diagnostics" to check the binary logistic-regression model that you fit to the Chilean plebiscite data. If you find any problems, try to deal with them. Two of the quantitative explanatory variables in the model, `population` and `income`, are discrete, with 10 and 8 unique values, respectively. Test for linearity on the logit scale by treating these variables as factors and performing appropriate likelihood-ratio tests.