

M5MS04 – Computational Statistics

Homework 1—Autumn 2013

Due: 5:00pm, Tuesday, 29 October 2013

Please submit via Blackboard in pdf format; late submissions will not be accepted.

Hand-in no more than six sides, and append your annotated R code.

1. The *logistic* distribution is a location family probability model for real-valued random variables with density

$$f_Y(y|\mu) = \frac{\exp\{-(y - \mu)\}}{(1 + \exp\{-(y - \mu)\})^2} \quad y \in \mathbb{R}$$

for location parameter μ . This density is symmetric about μ , and yields variance $\text{Var}[Y] = \pi^2/3$. The density is heavy-tailed compared to the Normal density, and is thus useful for carrying out robust estimation, that is, estimation that is less influenced by outlying observations.

The logistic distribution model can be used in linear regression as a model for the residual errors. Specifically, suppose that $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})$ is a (row) vector of p predictors available for observed response $y_i, i = 1, \dots, n$. Suppose that the linear model takes the form

$$Y_i = \mathbf{x}_i\beta + \epsilon_i \quad i = 1, \dots, n \quad (1)$$

where β is a $(p \times 1)$ parameter vector, and $\epsilon_i, i = 1, \dots, n$ are independent and identically distributed residual errors having a logistic distribution with parameter $\mu = 0$. We may write

$$\mathbb{E}[Y_i|\mathbf{x}_i] = \mathbf{x}_i\beta = \mu_i.$$

- (a) Write down the log-likelihood, ℓ , for the regression model in (1), based on observed data $(y_i, \mathbf{x}_i), i = 1, \dots, n$, and hence form the score equations for β .

5 MARKS

- (b) Write code to carry out maximum likelihood (ML) estimation for the regression model using **Newton's method**, and then use it to find the ML estimate, $\hat{\beta}$, based on the data in the file

`A1Data.csv`

that can be found on the Blackboard course site. Report also the second derivative matrix evaluated at $\hat{\beta}$, $-\ddot{\ell}(\hat{\beta})$.

10 MARKS

2. Write code to carry out **iteratively reweighted least squares** estimation of the parameters in a *binary regression* problem using a logistic link

$$l(t) = \log\left(\frac{t}{1-t}\right)$$

and use the code to estimate the parameters $(\beta_0, \beta_1)^\top$ in a model that uses linear predictor

$$\eta_i = \beta_0 + \beta_1 x_i$$

for the data in the R data frame `birthwt` from the `MASS` library. This data set records the incidence of low birthweight births in a data set of 189 births recorded at a US Medical Centre: the binary response variable y records the incidences of low birthweight (1 - low birthweight, 0 - otherwise), and the single predictor x is the mother's weight in pounds at the start of the pregnancy.

The data set can be obtained using the commands

```
>library(MASS)
>y<-birthwt$low
>x<-birthwt$lwt
```

10 MARKS