

Statistics 580

Assignment No.4 (40 points)

1. Use function `nlm()` in R to obtain the maximum likelihood estimates of α and β for Poisson regression model using the Australian AIDS deaths data. Set-up the objective, gradient, and the Hessian using the `deriv3()` function in R as in the logistic regression example. Use starting values you used in Problem #3 (Assignment #3).
2. Write a C program to obtain the maximum likelihood estimates of α and β for Poisson regression model using the Australian AIDS deaths data using the GSL function `gsl_multimin_fdfminimizer_vector_bfgs`. Follow the C programs supplied for using this function for the logistic problem. Recall that for using the above function we need only provide the function and its first derivative. Use appropriate values for the *step size* and *tolerance* parameters. Use the same starting values as above. Remember that we need to specify the *negative* of the likelihood function and its derivatives .
3. Use the Metropolis algorithm (given in the notes) to sample 4000 values from the bivariate normal distribution $\mathbf{x} \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\mu} = \begin{pmatrix} 1 \\ 2 \end{pmatrix} \text{ and } \boldsymbol{\Sigma} = \begin{pmatrix} 1 & .9 \\ .9 & 1 \end{pmatrix}$$

with the pdf of $\mathbf{y} \sim N_2(\mathbf{x}, \mathbf{D})$ where $\mathbf{D} = \begin{pmatrix} .6 & .0 \\ .0 & .4 \end{pmatrix}$ as the candidate generating density. Obtain a scatterplot of x_1 vs. x_2 , similar to those on Figure 3 of Chib and Greenberg(1995). For comparison, obtain a similar scatterplot of 4000 values generated from the same bivariate normal distribution as above, using the Cholesky factor of $\boldsymbol{\Sigma}$.

4. Use the R function `MH.normal.variance()` supplied, along with R functions you will write, to simulate from the posterior of the Gaussian variance, σ^2 , using the Metropolis-Hastings sampler, where

$$\text{Prior: } \sigma^2 \sim \text{Scaled Inverse-}\chi^2(\nu_p, \sigma_p^2)$$

$$\text{Data Model: } y_i | \sigma^2 \stackrel{i.i.d}{\sim} N(\mu, \sigma^2), \quad i = 1, \dots, n$$

Use the NB10 data, and pretend that μ for the data distribution is known and is equal to the the sample mean 404.59. Use a burn-in of 1000 iterations and a total of 6000 iterations with a thinning parameter of 1. Details of other parameter values to use are provided along with the function. Provide the following as your results using the saved output (after burn-in): the time series trace, the density trace overlaid with the actual posterior density, and plots of autocorrelations and partial autocorrelations obtained from standard R functions `acf()` and `acf.plot()`.

Note: All information on the functions required are in your MCMC notes. The NB10 data and `MH.normal.variance()` are to be downloaded from the webpages.

5. Consider the regression $y_j = \beta_0 + \beta_1 x_j + \epsilon_j$ $j = 1, \dots, n$ where $\epsilon_j \sim \text{iid } N(0, \sigma^2)$, i.e., $y_j | x_j, j = 1, \dots, n$ are independently distributed as $N(\beta_0 + \beta_1 x_j, \sigma^2)$. Suppose that y_j is missing (at random) for two cases. Label these to be the observations $(y_{n-1}, x_{n-1}), (y_n, x_n)$ i.e., y_{n-1} and y_n have not been observed. We are interested in obtaining m.l.e.'s of $\theta = (\beta, \sigma^2)$. The complete-data vector is $(y_j, x_j), j = 1, \dots, n$. Use the EM algorithm to obtain m.l.e.'s of β_0, β_1 and σ^2 based on the observed data $\{(y_j, x_j), j = 1, \dots, n-2, \text{ and } x_{n-1}, x_n\}$ as described below:

- Show that the sufficient statistics for $(\beta_0, \beta_1, \sigma^2)$ are $(\sum_{j=1}^n y_j, \sum_{j=1}^n x_j y_j, \sum_{j=1}^n y_j^2)$.
- Formulate the E-step computations by obtaining expectations of the sufficient statistics conditional on the observed data.
- Show that (no need to do this part as you know these) the complete-data m.l.e.'s for β_0, β_1 and σ^2 are

$$\begin{aligned}\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 &= \frac{\sum_{j=1}^n (x_j - \bar{x})(y_j - \bar{y})}{\sum_{j=1}^n (x_j - \bar{x})^2} \\ \hat{\sigma}^2 &= \frac{\sum_{j=1}^n (y_j - \hat{\beta} x_j)^2}{n}\end{aligned}$$

- Formulate the M-step computations by substituting conditional expectations from part (b) in the expressions for complete-data m.l.e.'s in part (c).
- Write an R function to implement this algorithm for the data (`house.data` with two missing y values) available from the Stat580 webpage. Print the MLE's, SSE, and the estimates of the missing y -values resulting from the EM algorithm. Substitute \bar{y} for the two missing values, estimate (β_0, β_1) using `lm()` and use them as starting values $(\beta_0^{(0)}, \beta_1^{(0)})$. Use both the Euclidean length of the successive differences and the change in likelihood (or SSE) in your stopping rule. Print iteration info.

Due Tuesday, April 29, 2008
