STA410/2102 (Statistical Computation), Fall 2007

Homework #3

Due: In class by 6:10 p.m. **sharp** on Tuesday November 13. (If you prefer, you may bring your assignment to the instructor's office, Sidney Smith Hall room 6024, any time before it is due; slide it under the door if he is not in.)

Warning: Late homeworks, even by one minute, will be penalised!

Announcement: The final exam will be 7–10 p.m. on Wednesday December 12.

Note: Homework assignments are to be solved by each student individually. You may discuss assignments in general terms with other students, but you must solve it on your own, including doing all of your own computing and writing.

Include at the TOP of the first page: Your name and student number.

Special Note: When writing programs in R for homework, you should include some comment lines to explain what you are doing. Also, you should hand in both the program itself, and the program's output.

Comment: The questions in this assignment are more "open ended" than in previous assignments. You will have to try various different approaches and "see how it goes". Be sure to lease leave yourself enough time to do so! (Also, question #1 is the most involved, so you may wish to leave it until last.)

The assignment:

1. [40 points] Consider the 1000 data pairs $\{(x_i, y_i)\}$ in the file "Rhw3Q1data", available on the course web page. Consider three models for how Y depends on X: (i) $Y = \beta_1 + X^{\beta_2} + \text{error}$, (ii) $Y = \beta_1 + (\cos(X))^{\beta_2} + \text{error}$, (iii) $Y = \beta_1 + (\sin(X))^{\beta_2} + \text{error}$. Using (a) training/testing data, and (b) cross-validation, try to determine which model best describes the relationship, and what is the best fit of the data. Explain all of your steps and choices and reasoning, and include all your R code and output. [Hints: this question is somewhat open-ended. The best solutions will try a number of different tests, and will clearly explain the approaches taken and why. For simplicity, if you wish, you may use the built-in R optimisation functions "optimise" and/or "nlm" [as in the file "Rcross"], but do not use the built-in R routines for cross-validation itself. Note: this data

was generated from an actual statistical model, and the ideal solution will find that actual model, so think of this question as a "treasure hunt"!]

2. [20 points] Consider the 1000 data values "xdata" in the file "Rhw3Q2data", available on the course web page. Attempt to estimate the density from which they were generated. Try each of the methods discussed in class, with various different choices of bandwidth, before settling on a final density estimate. (Do not use R's built-in density estimation routines.) Explain all your steps, and include your R code and output, including plots of the various estimates. [Note: this data was generated from an actual density, so this question is again a "treasure hunt", to find the density from which it was generated.]

3. [20 points] Consider the function

$$g(x,y) \;=\; x^2\; y^3\; \sin(xy)\; \cos(\;\sqrt{xy}\;)\; \exp(x^2+y)\,.$$

Suppose X and Y are two random variables with joint density given by $f_{X,Y}(x,y) = C\,g(x,y)$ for $0 \le x, y \le 1$ (with $f_{X,Y}(x,y) = 0$ for other x, y), for appropriate constant C. Write an R program to use numerical integration to compute the expected values $\mathbf{E}(X)$, $\mathbf{E}(Y)$, and $\mathbf{E}(XY)$. Try at least two different methods of numerical integration (do not use R's built-in numerical integration routines), with various grid sizes. Also, check whether or not $\mathbf{E}(XY) = \mathbf{E}(X)\,\mathbf{E}(Y)$. Explain your choices of method, grid size, etc., as well as your choice of final estimates. Include all your R code and output.