**BASKIN SCHOOL OF ENGINEERING**
**Department of Applied Mathematics and Statistics**
Graduate Program in Statistics and Applied Mathematics

Student number: ____________________________________

You are to work individually on this problem. Do not share with anyone any information or comments about your findings or the models and methods you use. You must turn in your solution by 5pm on Wednesday 17 June 2009. You can either print it and take it to Nic Brummell's office (BE125), or send a PDF file to `brummell@ams.ucsc.edu` . Please take care to organize and present the material in the best possible way; be informative but concise. Your paper should consist of no more than 12 letter-size pages (with 11pt or larger type and margins on all four sides of at least 1 inch), including tables and figures. Provide information about the methods and the software you use to fit the models. The amount of space devoted to each of the four parts of the problem below in your write-up should be roughly proportional to the point values of the problem parts.

**Problem.** The percentage of body fat (PBF) in humans is an important indicator of both fitness and potential health problems. The most accurate ways to measure PBF involve complicated techniques such as skinfold calipers, bioelectric impedance and underwater weighing; simpler methods are needed. One approach is based on demographics (including age and gender) and basic body measurements including height, weight and the circumferences of various body parts (such as the person's neck and chest). The website

http://users.soe.ucsc.edu/˜brummell/Work/UCSC-AMS/FYE/fye.html

(very bottom of page, "Stuff you might need" , right click on "Spring09 statistics takehome data") contains a downloadable text file with PBF and 13 predictor variables, measured on a representative sample of $n = 252$ American men in the 1980s whose ages ranged from 22 to 81; the first line in the file gives the variable names. In this data set PBF is in percentage points, age is in years, weight is in pounds, height is in inches, and all of the circumference measurements (neck, ..., wrist) are in centimeters.

(1) Perform an exploratory analysis on this data set, identifying any outliers that may be present (and choosing what to do about them in subsequent analyses) and deciding on the form of an appropriate model for predicting PBF from the other variables. Explain briefly why it is reasonable to model the outcome variable as a Gaussian directly on the percentage points scale. *[10 points]*

(2) Fit an appropriate Bayesian regression model (given your exploratory analysis in part (1)) to this data set with the predictive goal mentioned above (using all the

Table 1: *Posterior probabilities (%) of variable inclusion in the body fat data, from (complicated) Bayesian Method 1.*

| Method 1 | $x_6$ | $x_2$ | $x_{13}$ | $x_{12}$ | $x_4$ | $x_{11}$ | $x_8$ | $x_7$ | $x_5$ | $x_1$ | $x_9$ | $x_3$ | $x_{10}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $p(\beta_j \neq 0 \,\vert\, D)$ | 100 | 97 | 62 | 35 | 19 | 17 | 15 | 13 | 6 | 5 | 5 | 4 | 4 |

predictors, and incorporating no interaction terms), assuming that little is known about the regression relationship external to the data set. Summarize both the overall fit and the predictive distribution for a man who is one standard deviation above the mean on all predictors. Does your model make accurate enough predictions of percent body fat for careful scientific work? Explain briefly. (**Hint:** keep it simple.) *[35 points]*

(3) Variable selection (choosing a subset of the available predictors that may in some sense be better than using all of them) is a topic that has been considered from many angles in the Bayesian (and non-Bayesian) literature. Table 1 gives posterior probabilities of variable inclusion using a complicated Bayesian approach referred to here as Method 1; the variables are numbered in the order given in the data set $D$, and $\beta_j$ is the coefficient of variable $j$ in the regression model. Method 1 takes a long time to program and has the additional drawback that it treats uncertainty about the regression coefficients in a non-continuous manner (by focusing on $p(\beta_j \neq 0 \,|\, D)$), which is not necessarily scientifically reasonable. Here's a considerably simpler Method 2, which treats uncertainty about the $\beta_j$ continuously:

(a) Convert the outcome $y$ and all of the predictors $x_j$ to standard units, by subtracting off their means and dividing by their standard deviations, obtaining $y^*$ and $x_j^*$; this goes some distance toward putting the predictors on a common scale.

(b) Fit a Bayesian regression model (with a choice of priors that's appropriate to the scientific setting) for predicting $y^*$ from the $x_j^*$ (similar to what you did in part (2) above).

(c) People who claim that a (standardized) regression coefficient $\beta_j$ is zero can't really mean that; no continuous quantity is ever precisely zero. What they presumably mean is that $\beta_j$ is close enough to zero that the effect of $x_j$ on $y$ is close to negligible from a practical significance point of view. Therefore, introduce the idea of a *practical significance threshold* $c$ and assert that $\beta_j$ is *practically* nonzero if $p(|\beta_j| \geq c \,|D)$ is large.

Find a value of $c$ that produces $p(|\beta_j| \geq c \,|D)$ values similar to the $p(\beta_j \neq 0 \,|\, D)$ values from Method 1, and interpret this value of $c$ (i.e., a predictor $x$ has a practically-significant effect if changing it by ...). *[35 points]*

(4) Variable selection forces each predictor to "either be all the way in or all the way out of the model." An alternative that may be better is to fit a hierarchical model of the form

$$\begin{aligned} (\gamma, \tau^2, \sigma^2) & \sim p(\gamma, \tau^2, \sigma^2) \quad \text{(diffuse)} \\ (\beta \,|\, Z, \gamma, \tau^2) & \sim N_p(Z\gamma, \tau^2 I_p) \\ (y \,|\, X, \beta, \sigma^2) & \sim N_n(X\beta, \sigma^2 I_n), \end{aligned} \tag{1}$$

where $y$ is the $n$-vector of outcomes (in standard units), $X$ is the $n \times p$ design matrix from the "full model" with all of the predictors standardized, $\beta$ is the $p$-vector of regression coefficients (no intercept is needed after standardization of both $y$ and the $x_j$), $Z$ is a vector or matrix quantifying prior information about the signs and relative magnitude of the "effects of the $x_j$ on $y$," and $I_k$ is the $k \times k$ identity matrix. Consultation with medical and physiological experts has suggested the $Z$ vector $(+1, -4, 0, -2, 0, +10, -2, +2, 0, 0, +2, +1, -1)$ in the setting of this problem. Fit model (1) to the body fat data and contrast its conclusions with those you obtained in parts (2) and (3). *[20 points]*