BASKIN SCHOOL OF ENGINEERING

Department of Applied Mathematics and Statistics

2015 First Year Exam Retake, Take Home Question (Statistics)

Due by 1PM, Friday September 18, 2015

Instructions:

Please work individually on this problem. You are allowed to consult any material you wish, but do not share with any other individual any information or comments about your findings or the models and methods you use. You are required to write a report using a word processing software (i.e., LaTeX or Microsoft Word). You are required to email your report as **one pdf file** to the graduate director at `thanos@soe.ucsc.edu`

by 1PM, Friday September 18, 2015

Please organize and present the material in the best possible way. Be informative but concise. You should include a summary of your work at the beginning of the report, include and annotate all relevant figures and tables in the body of the report, write your conclusions in a separate section, and list your references (if any). Your report should consist of no more than 8 letter-size pages (typeset with 11pt or larger font and margins on all four sides of at least 1 inch), including all figures, tables, and appendices (but excluding the numerical codes); answers longer than 8 pages will lose credit for excess length. For those of you that will type your report in LaTeX, it is suggested (but not required) to use the template from `https://courses.soe.ucsc.edu/courses/ams207/Spring15/01`

For the implementation of the models included in this problem, you can use any language you feel comfortable with, but you are **not** allowed to use a pre-programmed sampler, such as the ones implemented in BUGS or STAN. You must include your MCMC codes (in R or other programming languages) at the end of your report; the codes do not count toward the 8-page limit.

Exam Problem:

The concepts of specificity and sensitivity are key in the evaluation of medical tests. The sensitivity (also called the true positive rate, TPR) of a test measures the proportion of positives (diseased individuals) that are correctly identified as such, while the specificity (also called the true negative rate, TNR) measures the proportion of negatives (healthy individuals) that are correctly identified. In the case of tests derived from a (univariate) continuous biomarker, different cutoff values for the classification leads to different values of the TPR and TNR, and the resulting tradeoffs are often summarized using the so-called receiver operating characteristic curve (ROC curve). The ROC curve is defined as the set of points $\{(1 - TNR(u), TPR(u) : u \in (-\infty, \infty)\}$, where $TNR(u)$ ($TPR(u)$) is the TNR (TPR) rate that results from using a threshold u to decide on who is positive or negative.

For this project you will use the data from a study by Etzioni et. al. (1999) [1] to evaluate the performance of total prostate specific antigen (PSA) as a biomarker that can be used to construct a test for prostate cancer. The data set is available from:

`https://users.soe.ucsc.edu/~abel/psa.csv`

and it contains three columns: the first column is the true disease status of the individual (0 = healthy, 1 = cancer), the second column is the total PSA, and the third column is the age of the individual.

In the sequel, let $y_{i,j}$ and $x_{i,j}$ be, respectively, the level of total PSA and the age for individual $j = 1, \ldots, n_j$ in population $i \in \{0, 1\}$.

1. First, ignore the information about the age of the individual. Perform a descriptive analysis of the data. Then, construct an estimate of the ROC for PSA using a Bayesian binormal model. The binormal model assumes that, after potentially transforming the data, the levels of total PSA on each group follow independent normal distributions with unknown means and variances, i.e.,

$$y^*_{i,j} \mid \theta_i, \sigma_i^2 \sim \mathsf{N}\left(\theta_i, \sigma_i^2\right), \qquad y^*_{i,j} = g\left(y_{i,j}\right), \quad \text{for } i \in \{0, 1\}$$

where g is a one-to-one transformation of the data (note that $g(x) = x$, i.e., not transforming the data, is a valid option). Justify your choice of priors and transforma-

[1] Etzioni, R., Pepe, M., Longton, G., Hu, C. & Goodman, G. (1999). Incorporating the time dimension in receiver operating characteristic curves: a case study of prostate cancer. *Med Decis Making*, 19, 242-251.

tion and remember to provide not only a point estimate of the ROC curve, but also (pointwise) uncertainty bands for it.

2. The area under the ROC curve (AUC) is often used as a summary of the quality of a test. An AUC of 1 corresponds to a perfect biomarker, while an area of 0.5 corresponds to a random test. (A biomarker with an ROC of less than 0.5 can always be improved by reversing the test.) For the binormal model, the AUC can be computed as

$$\Phi\left(\frac{\theta_0 - \theta_1}{\sqrt{\sigma_0^2 + \sigma_1^2}}\right),$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution. Test the hypotheses $H_0 : AUC \geq 0.75$ vs. $H_a : AUC < 0.75$. What are the prior odds and the posterior odds associated with these hypotheses?

3. Consider now including the information about the age of the individual. Since the TPR and TNR can now potentially change with age, evaluating the biomarker in this case means constructing age-dependent ROC curves. Perform a descriptive analysis of the data. Then, fit a Bayesian (hierarchical) homocedastic linear regression model for each group

$$y_{i,j} \mid \alpha_i, \beta_i, \sigma_i^2 \sim \mathsf{N}\left(\alpha_i + \beta_i x_{i,j}, \sigma_i^2\right),$$

where

$$\begin{pmatrix}\alpha_i \\ \beta_i\end{pmatrix} \mid \boldsymbol{\mu}, \boldsymbol{\Omega} \sim \mathsf{N}\left(\boldsymbol{\mu}, \boldsymbol{\Omega}\right), \qquad \sigma_i^2 \mid \tau \sim \mathsf{IGam}(e, \tau),$$

and

$$\boldsymbol{\mu} \sim \mathsf{N}\left(\mathbf{c}, \mathbf{C}\right), \qquad \boldsymbol{\Omega} \sim \mathsf{IWis}\left(\mathbf{d}, \mathbf{D}\right), \qquad \tau \sim \Gamma(a, b)$$

with e, $\mathbf{c}$, $\mathbf{C}$, $\mathbf{d}$, $\mathbf{D}$, a and b fixed prior parameters, which you need to specify. Then, use this regression model to estimate the ROC curve at ages 50, 55, 60, 65, 70 and 75. As before, provide both point and interval estimates for the curve, justify your choice of priors, perform model checking and discuss model adequacy.