BASKIN SCHOOL OF ENGINEERING

SAM Program, Statistics Track

2018 First Year Exam, Take Home Question

Due by 5PM, Wednesday June 13, 2018

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 email your report as **one pdf file** to the graduate director at `thanos@soe.ucsc.edu`

by 5PM, Wednesday June 13, 2018

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). You are required to write your report in LaTeX, using the template from

`https://ams207-spring18-01.courses.soe.ucsc.edu/textbook-and-grading`

Your report should consist of no more than 10 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 10 pages will lose credit for excess length. You must include your R code for both problems at the end of your report; the codes do not count toward the 10-page limit.

Exam Problems:

1. (30 points) The NHANES data (included in the R package NHANES) consist of survey data collected by the US National Center for Health Statistics (NCHS). NHANES contains 10,000 rows of data (one row corresponds to one individual) that can be thought as a random sample from the American population with 76 variables per individual. Consider the following variables: `Diabetes, BMI, Age, Gender, Education` and `Poverty`.

   (a) (10 points) Using quantitative and graphical tools, summarize the relationships across these variables. Select your tables and/or graphs carefully. Remember there is a page limit restriction for your report.

   (b) (20 points) Consider a logistic regression model with `Diabetes` as the response (you can define a new variable `DiabetesYes` that takes the value 1 if an individual has diabetes and 0 otherwise). Which of the variables listed above are useful to predict diabetes? Justify your answer by fitting the appropriate logistic regression model and interpreting the results of your analysis numerically and/or graphically. You can use the `glm` function in `R` to fit the model and interpret your results. No residual analysis is required.

   (**Note:** You are *not* expected to fit Bayesian models for this question.)

2. (70 points) In the HELP (Health Evaluation and Linkage to Primary Care) study, investigators were interested in determining predictors of severe depressive symptoms (measured by the Center for Epidemiologic Studies–Depression Scale, `cesd`) amongst a cohort enrolled at a substance abuse treatment facility. A number of datasets from this study are available through the `R` package `mosaicData`.

   You will be working with the dataset `HELPrct` which is a subset of the data obtained from the HELP study restricted to 453 subjects. Consider the response variable `cesd` (depression measure at baseline, high scores indicate more depressive symptoms) and the predictor variables `substance` of abuse (alcohol, cocaine, or heroin), `mcs` (a measure of mental well-being), `sex`, and `homeless` (housing status).

   (a) (20 points) Determine which of the variables are statistically significant to predict `cesd` by fitting multiple linear regression models with the `R` function `lm` and then using model comparison tools to select the predictors. Explain your methodology. Once you obtain your final model answer the following questions:

      i. interpret the parameter estimates;

      ii. compute the predicted `cesd` and the associated 95% prediction interval for a female homeless cocaine-involved subject with an `mcs` score of 20;

      iii. perform a residual analysis, what do you conclude?

      iv. what do you conclude about the relationship between the fitted values and the residuals? how about the relationship between `mcs` score and the residuals?

      v. are there any outliers?

   (b) (25 points) Repeat the analysis above using a Bayesian approach with the following model structure. Let

   $$\mathbf{y} = \beta_0 \mathbf{1} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \qquad \epsilon \mid \sigma^2 \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$$

   be the regression model with $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$. Summarize your posterior inference under a prior of the form

   $$p(\beta_0, \boldsymbol{\beta}, \sigma^2) \propto \frac{1}{\sigma^2}.$$

   Explain how you obtained posterior inferences and compare your results to those obtained in part (a) focusing on parts i. and ii.

(c) (25 points) Repeat the analysis above using a Bayesian Lasso model that includes all the predictors listed above (not only the significant ones). More specifically, consider the model $\mathbf{y} \mid \beta_0, \boldsymbol{\beta}, \sigma^2 \sim N(\beta_0\mathbf{1} + \mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$, with a flat prior for β_0, $p(\beta_0) \propto 1$, and the conditional Laplacian prior given by:

$$p(\boldsymbol{\beta} \mid \sigma^2) = \prod_{j=1}^{p} \frac{\lambda}{2\sqrt{\sigma^2}} \exp(-\lambda|\beta_j|/\sqrt{\sigma^2}), \quad p(\sigma^2) \propto 1/\sigma^2,$$

where $\lambda > 0$ is the Bayesian Lasso parameter, which will be fixed for this problem. As shown in the Bayesian Lasso paper by Park and Casella (JASA, 2008), the model can be rewritten as follows to facilitate posterior computation:

$$\begin{aligned} \mathbf{y} \mid \mathbf{X}, \beta_0, \boldsymbol{\beta}, \sigma^2 &\sim N(\beta_0\mathbf{1} + \mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}), \\ \boldsymbol{\beta} \mid \sigma^2, \tau_1^2, \ldots, \tau_p^2 &\sim N(\mathbf{0}, \sigma^2\mathbf{D}), \\ \mathbf{D} &= \mathrm{diag}(\tau_1^2, \ldots, \tau_p^2), \\ p(\tau_1^2, \ldots, \tau_p^2) &= \prod_{j=1}^{p} \frac{\lambda^2}{2} \exp(-\lambda^2\tau_j^2/2) \end{aligned}$$

with $p(\beta_0, \sigma^2) \propto 1/\sigma^2$. A Gibbs sampling algorithm can be used for posterior simulation as described in Park and Casella (2008). Implement this algorithm to obtain posterior samples of the model parameters. Summarize your posterior inference results and compare them with the results you obtained in parts (a) and (b) focusing on i. and ii. For the Bayesian Lasso parameter, you can use $\lambda^2 = 0.1$, but perform a (limited) sensitivity analysis for values of λ^2 in $(0, 0.5)$.

Note that one of the Gibbs sampling steps requires sampling from an inverse-Gaussian distribution. You can use the function `invgauss` in the `statmod` package. There are also packages and functions in `R` that implement the Bayesian lasso (e.g., the function `blasso` from package `monomvn`). You can use one of the available packages/functions or you can write your own code. Either of these options is fine, but you should discuss how you implemented the model. If you use existing software, you should provide the relevant references in your report.