AMS274 - Fall 2017
Homework 1

1 Problem 1

Consider a logistic regression model of the form

$$y_i \mid \theta_i \sim \mathsf{Ber}(\theta_i) \qquad\qquad \operatorname{logit}(\theta_i) = \mu + \mathbf{x}_i^T \boldsymbol{\beta}$$

where $\mathbf{x}_i^T = (x_{i,1}, \ldots, x_{i,p})$ is a vector of regressors associated with the i-th response, which have been standardized (i.e., you have substracted the means and divided by the variances of each of them). Note that his parameterization is slightly different from the one we have been using until this point because I have pulled the intercept out of the vector of regression coefficients. In this problem we will explore three different ways to do variable selection for this model in the context of the Pima Indians training dataset (`Pima.tr`) we used in the previous homework.

1. First, use backwards regression to identify a subset of variables that significantly affect the probability of Diabites in this population. Recall that in backward regression you start by fitting the model with the all variables and perform independent checks of significance on each regression coefficient (using t-tests in the Gaussian case, or z-tests in the case of a GLM). If any of them indicates that the regressor is not significant, the variable is dropped and the model refit. The process stops when all remaining variables appear to be significant. For example, in the case of `Pima.tr`

```
> library(MASS)
> data(Pima.tr)
> mod1 = glm(type˜npreg+glu+bp+skin+bmi+ped+age,
+            family=binomial(link="logit"), data=Pima.tr)
> summary(mod1)



Call:
glm(formula = type ˜ npreg + glu + bp + skin + bmi + ped + age,
    family = binomial(link = "logit"), data = Pima.tr)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9830  -0.6773  -0.3681   0.6439   2.3154

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -9.773062   1.770386  -5.520 3.38e-08 ***
npreg        0.103183   0.064694   1.595  0.11073
```

```
glu         0.032117   0.006787   4.732 2.22e-06 ***
bp         -0.004768   0.018541  -0.257  0.79707    
skin       -0.001917   0.022500  -0.085  0.93211    
bmi         0.083624   0.042827   1.953  0.05087 .  
ped         1.820410   0.665514   2.735  0.00623 ** 
age         0.041184   0.022091   1.864  0.06228 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 256.41  on 199  degrees of freedom
Residual deviance: 178.39  on 192  degrees of freedom
AIC: 194.39

Number of Fisher Scoring iterations: 5
```

The table of coefficients suggest dropping first the variable `skin`, which hgas the highest p-value above 0.05.

2. Implement an MCMC algorithm based on the Pòlya-Gamma augmentation to perform a Bayesian variable selection approach for this problem. Note that since there are 7 variables you will need to compare $2^7 = 128$ models (we assume that all models include at least an intercept). Use the following priors in your analysis:

   (a) For the prior on the models, assume that each variable is included independently with probability θ, which is assigned a uniform hyperprior on the $[0, 1]$ interval. If you encode your model by introducing variables $\gamma_1, \ldots, \gamma_7$ such that $\gamma_k = 1$ if and only if variable k is included in the regression, this choice implies that

   $$p(\gamma_1, \ldots, \gamma_7) = \int \theta^{\sum_{i=1}^{7} \gamma_i} (1-\theta)^{7-\sum_{i=1}^{7} \gamma_i} d\theta = \frac{\Gamma\left(\sum_{i=1}^{7} \gamma_i\right) \Gamma\left(7 - \sum_{i=1}^{7} \gamma_i\right)}{\Gamma(7)}$$

   (b) For the prior on the coefficients given the model, assume that significant coefficients follow independent standard normal distribution (coefficients not included by the model will naturally be assigned independent point masses at zero).

   Use your MCMC to analyze the `Pima.tr` dataset. Note that you have two ways to summarize your posterior:

   (a) Reporting the posterior probability of each of the 128 models, which corresponds to the joint $\mathsf{P}(\gamma_1, \ldots, \gamma_7 \mid \text{data})$, and selecting as your optimal model the one with the highest posterior probability (highest frequency in your posterior sample).

(b) Reporting the marginal probability associated with each of the 7 variables, which corresponds to the marginals $\mathsf{P}(\gamma_k \mid \text{data})$ for $k = 1, \ldots, 7$, and choosing to retain variables where this probability is greater than 1/2.

Report both summaries (note that these two summaries do not necessarily choose the same model). Compare these models against the one you chose using backwards regression.

3. Penalized likelihood methods are a very popular method for doing variable selection in generalized linear models. The estimators can be interpreted as posterior modes from a Bayesian model with a suitably chosen prior for the coefficients of the model (in the case of L1 penalized regression, also known as LASSO, these are independent double exponential distributions). In the case of logistic regression,

$$\hat{\boldsymbol{\beta}}_{L1} = \arg\max_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} y_i \left(\mu + \mathbf{x}_i^T \boldsymbol{\beta}\right) - \left(1 + \exp\left\{\mu + \mathbf{x}_i^T \boldsymbol{\beta}\right\}\right) \right\} - \lambda \sum_{k=1}^{p} |\beta_k|$$

where λ is a tuning parameter. Note that if $\lambda = 0$ then $\hat{\boldsymbol{\beta}}_{L1} = \hat{\boldsymbol{\beta}}_{MLE}$. For $\lambda > 0$ the solution to the optimization problem might include exact zeros for any of the β_k because the absolute value makes the penalty not differentiable at zero (in fact, $\lambda \to \infty$ implies that $\beta_k \to 0$ for all k). Hence, L1 penalized regression can be used for joint estimation/model selection.

Use L1 penalized regression to identify variables that are associated with Diabetes in the Pima Indians dataset. I do not intent you to program your own optimization algorithm. Instead, note that the `R` package `glmnet` implements penalized regression in the context of various generalized linear models, including logistic regression:

```
> install.packages("glmnet")
> library(glmnet)
> x <- matrix(rnorm(100*20),100,20)
> y <- sample(1:2,100,replace=TRUE)
> mod2 <- glmnet(x, y, family="binomial", alpha=1) # alpha=1
>                                                  # for LASSO
```

You can find more information about the package at `https://web.stanford.edu/~hastie/glmnet/glmnet_alpha.html`.

2 Problem 2

This problem refers to data from a study of nesting horseshoe crabs (J. Brockmann, Ethology 1996). Each female horseshoe crab in the study had a male crab attached to her in her nest. The study investigated factors that affect whether the female crab had any other males, called satellites, residing near her. Explanatory variables that are thought to affect this included the female crabs color (C), spine condition (Sc), weight (Wt), and carapace width (W). The response outcome for each female crab is her number of satellites (S). There are 173 females in this study. The data is available in the `crab.txt` file.

We want to fit a Bayesian loglinear model of the form

$$S_i \mid \lambda_i \sim \mathsf{Poi}(\lambda_i), \qquad \qquad \log \lambda_i = \mathbf{x}_i \boldsymbol{\beta},$$

where the $\mathbf{x}_i$ contains the value of the regressors associated with the various available covariates (note that color is a categorical variable, so you will need to encode it using some dummy variables). We will use independent Gaussian priors with mean 0 and variance 100 for the coefficients.

Construct the two MCMC samplers for this model that were discussed in class:

1. A random-walk Metropolis Hastings algorithms with joint Gaussian proposals for $\boldsymbol{\beta}$ centered at the current value of the paramters.

2. A slice sampler along the lines described in

   Damlen, P., Wakefield, J., & Walker, S. (1999). Gibbs sampling for Bayesian nonconjugate and hierarchical models by using auxiliary variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 61(2)*, 331-344.

In addition to reporting posterior means and posterior credible intervals for all the regression coefficients, compute the effective sample sizes associated with 100,000 iterations (obtained after burn-in) for each coefficient under each of the two sampling methods (use a table to report these results). Remember that you can compute effective sample sizes using the `coda` package

```
> install.packages("coda")
> library(coda)
> effectiveSize(rnorm(1000))
```