Exercises

A population of women who were at least 21 years old and of Pima Indian heritage was tested for diabetes. The data consist of 532 complete records.

These are the variables in the **sasuser.pima** data set:

diabetes	diabetes status (1=yes, 0=no)
number_preg	number of pregnancies
glucose	plasma glucose concentration in an oral glucose tolerance test
diastolic_bp	diastolic blood pressure (mm Hg)
skin_thickness	triceps skin fold thickness (mm)
body_mass	body mass index (weight in kg/(height in meters)^2)
pedigree	diabetes pedigree function
age	in years.

Note: This data set was obtained from Ripley (1996).

1. Bayesian Analysis with a Noninformative Prior

a. Submit the program **stbay00d01.sas**. Generate a Bayesian analysis of diabetes in the Pima Indian data set. Fit a logistic regression model and specify **diabetes** as the response variable (use the DESC option in the PROC GENMOD statement to model the probability of diabetes status) and specify all of the predictor variables. In the MODEL statement, use the DIST= binomial option to specify a binomial distribution and a LINK=LOGIT option to specify the logit link function. Use the BAYES statement and specify a seed of 27513, request all the diagnostic statistics, display a fitted penalized B-spline curve for each trace plot, specify the Gamerman sampling algorithm, specify 150000 iterations after the burn-in, specify a thinning rate of 5, and create an output data set with the generated posterior samples.

1) Do any of the plots show problems with convergence?

2) Which parameters show evidence that their slopes are different from 0 based on the highest posterior density intervals?

b. Compute posterior probabilities for the regression coefficient parameters. First, print out the first 10 observations from the output data set of the generated posterior samples. Then, compute the probability that the parameter estimates are greater than 0.

1) What is the probability that the parameter estimates for number of pregnancies are greater than 0?

2. Bayesian Analysis with an Informative Prior

a. Create a data set called **prior** that specifies the means and variances of the prior distributions of the parameters. For the parameter of glucose, specify a mean of 0.0953 and a variance of 0.000536. For the other parameters, specify 0 for the mean and 1e6 for the variance. Fit the same logistic regression model as before and use the BAYES statement to specify a SAS data set containing the mean and covariance information of the prior distribution, create an output data set with the posterior samples, display a fitted penalized B-spline curve for each trace plot, specify the Gamerman sampling algorithm, specify 150000 iterations after the burn-in, specify a thinning rate of 4, and request all the diagnostic convergence statistics.

1) Do any of the diagnostic tests indicate problems with convergence?

2) Did the informative prior change the final parameter estimate for glucose?

b. Create side-by-side histograms of the posterior density distribution of glucose where one histogram is based on the noninformative prior and the other histogram is based on the informative prior.

1) What differences do you see in the two distributions?

End of Exercises