Lab 10: Probit and Logitistic Models

The attached data set contains information on 35 young automobile drivers. Recorded for each individual is the response, *y*, which is 1 if the individual has been involved in 1 or more automobile accidents in the last 6 months, 0 otherwise; the driver's age in years; whether or not the driver has taken an extensive driver's education course (1 for yes); and how the driver responded when asked if s/he enjoys the discipline of statistics (1 for yes). Specific questions to be answered are in bold.

Prove that the latent variable construction where $\boldsymbol{y_i^* = \ x_i'\beta, +\epsilon_i}$ and $\boldsymbol{y_i = I(y_i^* > 0)}$, for $\boldsymbol{\epsilon_i \sim N(0, 1)}$, is equivalent to probit regression: $\boldsymbol{Pr(y_i = 1|x_i) = \Phi(x_i'\beta)}$, where $\boldsymbol{I(\,.\,)}$ denotes the indicator function and $\boldsymbol{\Phi(.)}$ denotes the standard normal CDF. Prove that the same latent variable construction with $\boldsymbol{\epsilon_i \sim Logistic(0, 1)}$ is equivalent to logistic regression: $\boldsymbol{Pr(y_i = 1|x_i) = \mathrm{logit}^{-1}(x_i'\beta)}$, where $\boldsymbol{\mathrm{logit}^{-1}(a)}$ is the standard logistic CDF evaluated at $\boldsymbol{a}$, or the inverse of $\boldsymbol{\mathrm{logit}(a) = \log\left\{\frac{a}{1-a}\right\}}$.

Using independent diffuse (improper) normal priors, perform Bayesian probit regression on these data. You should use the latent variable method and Gibbs sampling to do this. In particular, for $x_i' = (1, age_i, course_i, likeStats_i)$, let

$$y_i^*|x_i, \beta \sim N(x_i'\beta, 1), \text{ and}$$

$$y_i = I(y_i^* > 0).$$

Derive full conditionals for the vector $\boldsymbol{\beta}$ and each $\boldsymbol{y_i^*}$ (form is the same for all $\boldsymbol{i}$) assuming a generic multivariate normal prior with mean b0 and variance B0. Show work, but feel free to use shortcuts learned from other labs. Explain why Gibbs sampling in the logistic regression case is more difficult.

Using 10,000 post-burn-in samples calculate the posterior predictive distribution for your 26-year-old TA who has never had a driver's course. **Display the posterior predictive on the space of probability of an accident using a histogram with at least 20 bins (breaks=20). What is the probability that your TA was involved in an accident in the last six months?**

Given only the data from this lab, whom would you choose to park your new Mercedes-Benz: a 17-year-old who has done a driver's education course and hates statistics, or an 18-year old who has done a driving course and loves statistics? Use the probit output. Assume that the probability of crashing while valet parking is proportional to the probability of having crashed in the past 6 months.