Probit Regression

A panel study followed 25 married couples over a period of five years. One item of interest is the relationship between divorce rates and the various characteristics of the couples. For example, the researchers would like to model the probability of divorce as a function of age differential, recorded as the man's age minus the woman's age. The data can be found in the file *divorce.dat*.
Reference: Problem 6.3, A First Course in Bayesian Statistical Methods 2009 Edition

Our probit regression model:

$$Z_i = \beta x_i + \epsilon_i$$
$$Y_i = \delta_{(c,\infty)}(Z_i)$$

where β and c are unknown coefficients, $\epsilon_1, ..., \epsilon_n \sim$ i.i.d. normal(0,1) and $\delta_{(c,\infty)} = 1$ if $z > c$ and equals zero otherwise.

Parameter estimated would be carried out in the Bayesian paradigm. In particular, we use Gibbs sampling to simulate the sample of $\beta, c, z \mid y, x$.

The sampling scheme is outlined as below:
1.
Setting initial values $\beta^1 = c^1 = 0$, generate $z^1 \sim$ N(0,1), for s=1:10000:
Sample $z_1^{s+1} \sim p(z_1 \mid y, x, z_{-1}^s, \beta^s, c^s)$
Sample $z_2^{s+1} \sim p(z_2 \mid y, x, z_{-1,2}^s, z_1^{s+1}, \beta^s, c^s)$
...
Sample $z_n^{s+1} \sim p(z_n \mid y, x, z_{-n}^{s+1}, \beta^s, c^s)$
2.
Sample $\beta^{s+1} \sim p(\beta \mid y, x, z^{s+1}, c^s)$
3.
Sample $c^{s+1} \sim p(c \mid y, x, z^{s+1}, \beta^{s+1})$

To obtain the full conditional distributions $p(z_i \mid y, x, z_{-i}, \beta, c), p(\beta \mid y, x, z, c)$ and $p(c \mid y, x, z, \beta)$, we made the following assumptions:

i. $\beta \sim N(0, \tau_\beta^2)$
ii. $c \sim N(0, \tau_c^2)$, where $\tau_\beta^2 = \tau_c^2 = 16$

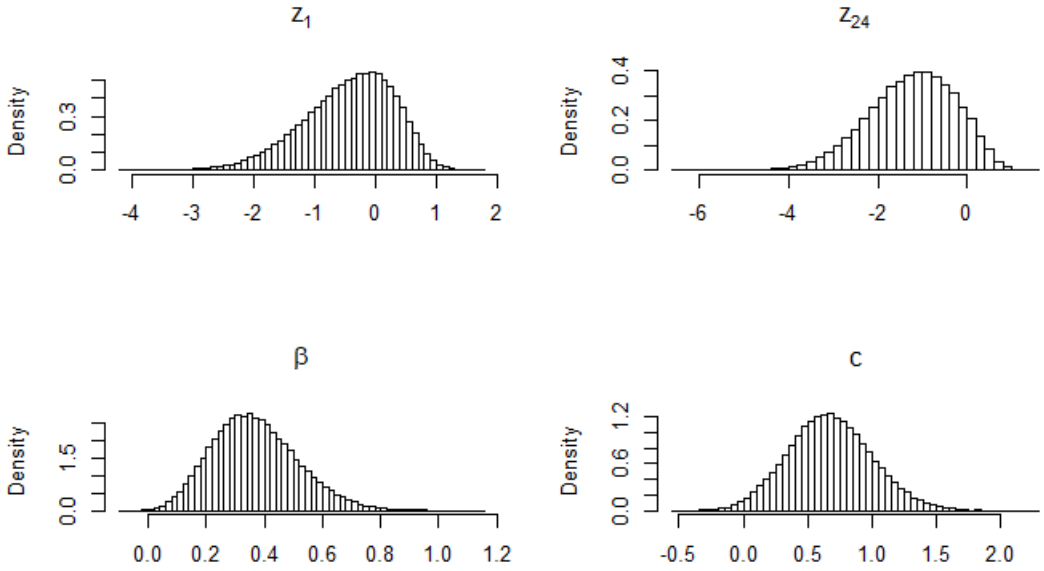
Derivation of full conditions:

$$
\begin{aligned}
p(\beta \mid y, x, z, c) &= p(\beta \mid z, x) \\
&\propto p(\beta) p(z \mid \beta, x) \\
&= \left[\frac{1}{\sqrt{2\pi}}\right]^{n+1} \tau_\beta^{-1} \exp -\frac{1}{2}\left[\frac{\beta^2}{\tau_\beta^2} + \sum_{i=1}^{n}(z_i - \beta x_i)^2\right] \\
&\propto \exp\left\{-\frac{1}{2}\left[\sum_{i=1}^{n} z_i^2 - 2\beta\sum_{i=1}^{n} x_i z_i + \beta^2 \sum_{i=1}^{n} x_i^2 + \tau_\beta^{-2}\beta^2\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2(\sum_{i=1}^{n} x_i^2 + \tau_\beta^{-2})^{-1}}\left[\beta^2 - 2\left(\sum_{i=1}^{n} x_i z_i\right)\left(\sum_{i=1}^{n} x_i^2 + \tau_\beta^{-2}\right)^{-1}\beta\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2(\sum_{i=1}^{n} x_i^2 + \tau_\beta^{-2})^{-1}\left[\beta - (\sum_{i=1}^{n} x_i z_i)\left(\sum_{i=1}^{n} x_i^2 \tau_\beta^{-2}\right)^{-1}\right]^2}\right\}
\end{aligned}
$$

Therefore, $\beta \mid y, x, z, c \sim \mathrm{N}\Big((\sum_{i=1}^{n} x_i z_i)\left(\sum_{i=1}^{n} x_i^2 + \tau_\beta^{-2}\right)^{-1}, \left(\sum_{i=1}^{n} x_i^2 + \tau_\beta^{-2}\right)^{-1}\Big)$

Given y and z, we know c¡z_i if $y_i = 1$ and $c \geq z_i$ if $y_i = 0$. Letting $a = \max\{z_i = y_i = 0\}$ and $b = \min\{z_i = y_i = 1\}$, $p(c \mid y, x, z, \beta)$ is proportional to $p(c)$ but constrained to $\{c : a < c < b\}$. Since $p(c) = \mathrm{N}(0, \tau_c^2)$, hence $c \mid y, x, z, \beta$ follows a constrained normal distribution. Similarly, given c and y_i, we know $z_i > c$ or $\leq c$, depending on $y_i = 1$ or 0 accordingly. Therefore, $p(z_i \mid y, x, z_{-i}, \beta, c) \sim$ a constrained $\mathrm{p}(z_i \mid x_i, \beta)$. Since $z_i \mid x_i, \beta \sim \mathrm{N}(\beta x_i, 1)$, hence $z_i \mid x_i, \beta$ follows a constrained normal distribution as well.

Results of our model:



Inference on β

Based on our samples of β, the 95% posterior c.i. and $\Pr(\beta > 0 \mid y, x)$ are estimated to be $[0.1099, 0.6866]$ and 0.9993 respectively.