Francesco Caporali, Isabel Muzio
15 November 2022

Bayesian Statistics

Assignment 2

Question 1: Probit regression (Hoff 6.3)

A panel study followed $n = 25$ 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.RData`. We will model these data with probit regression, in which a binary variable Y_i is described in terms of an explanatory variable x_i via the following latent variable model:

$$\begin{aligned} Z_i &= \beta x_i + \varepsilon_i \\ Y_i &= \mathbb{1}_{(c,+\infty)}(Z_i), \end{aligned}$$

where β and c are unknown coefficients, $\varepsilon_1, \ldots, \varepsilon_n \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ and $\mathbb{1}_{(c,+\infty)}(z) = 1$ if $z > c$ and equals zero otherwise. In the following, since the covariates x_i are known, they will be treated as constants and so not explicitly written in the conditioning part.

Point a.

Assuming $\beta \sim \mathcal{N}\left(0, \sigma_\beta^2\right)$, obtain the full conditional distribution $p(\beta \mid y_{1:n}, z_{1:n}, c)$.

First of all let us write explicitly the conditional distributions which we can deduce from the text:

- $\forall i = 1, \ldots, n$ we know $p(z_i \mid \beta)$:

$$Z_i(\omega) \mid \beta = \beta x_i + \varepsilon_i(\omega) \sim \beta x_i + \mathcal{N}(0,1) \sim \mathcal{N}(\beta x_i, 1) \implies Z_i \mid \beta \sim \mathcal{N}(\beta x_i, 1)$$

$$\Downarrow$$

$$p(z_i \mid \beta) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z_i - x_i\beta)^2};$$

- $\forall i = 1, \ldots, n$ we know $p(y_i \mid c, z_i)$:

$$Y_i(\omega) = \mathbb{1}_{(c,+\infty)}(Z_i) = \begin{cases} 1 & \text{if } Z_i > c \\ 0 & \text{otherwise} \end{cases}$$

$$\Downarrow$$

$$\begin{aligned} p(y_i) &= \mathbb{P}(Y_i = y_i) = \mathbb{P}\left(\mathbb{1}_{(c,+\infty)}(Z_i) = y_i\right) = \\ &= \begin{cases} \mathbb{P}\left(\mathbb{1}_{(c,+\infty)}(Z_i) = 1\right) & \text{if } y_i = 1 \\ \mathbb{P}\left(\mathbb{1}_{(c,+\infty)}(Z_i) = 0\right) & \text{if } y_i = 0 \\ 0 & \text{otherwise} \end{cases} = \\ &= \begin{cases} \mathbb{P}(\{Z_i > c\}) & \text{if } y_i = 1 \\ \mathbb{P}\left(\{Z_i > c\}^C\right) & \text{if } y_i = 0 \\ 0 & \text{otherwise} \end{cases} = \\ &= \left(y_i \mathbb{P}(\{Z_i > c\}) + (1 - y_i)\mathbb{P}\left(\{Z_i > c\}^C\right)\right) \mathbb{1}_{\{0,1\}}(y_i), \end{aligned}$$

hence $Y_i \sim \text{Bernoulli}\left(\mathbb{P}\left(Z_i > c\right)\right)$.

It follows that, conditionally on Z_i, c, the r.v. Y_i is no more *random* and it holds[1]

$$p(y_i \,|\, c, z_i) = \left(y_i \mathbb{1}_{(-\infty, z_i)}(c) + (1 - y_i)\mathbb{1}_{(-\infty, z_i)^C}(c)\right) \mathbb{1}_{\{0,1\}}(y_i).$$

The full conditional distribution $p(\beta \,|\, y_{1:n}, z_{1:n}, c)$ can be obtained just from $p(z_i \,|\, \beta)$, indeed

$$\begin{aligned}
p(\beta \,|\, y_{1:n}, z_{1:n}, c) &= \frac{p(\beta, y_{1:n}, z_{1:n}, c)}{p(y_{1:n}, z_{1:n}, c)} \frac{p(\beta, z_{1:n}, c)}{p(\beta, z_{1:n}, c)} \frac{p(\beta, c)}{p(\beta, c)} \frac{p(c)}{p(c)} \propto \\
&\propto \frac{p(\beta, y_{1:n}, z_{1:n}, c)}{p(\beta, z_{1:n}, c)} \frac{p(\beta, z_{1:n}, c)}{p(\beta, c)} \frac{p(\beta, c)}{p(c)} = \\
&= p(y_{1:n} \,|\, \cancel{\beta}, c, z_{1:n}) p(z_{1:n} \,|\, \beta, \cancel{c}) p(\beta \,|\, \cancel{c}) \propto \\
&\propto p(z_{1:n} \,|\, \beta) p(\beta).
\end{aligned}$$

So we can write explicitly

$$\begin{aligned}
p(\beta \,|\, y_{1:n}, z_{1:n}, c) &\propto p(z_{1:n} \,|\, \beta) p(\beta) = \\
&= \prod_{i=1}^{n} p(z_i \,|\, \beta) p(\beta) \propto \\
&\propto \exp\left(-\frac{1}{2} \sum_{i=1}^{n} (z_i - x_i \beta)^2\right) \exp\left(-\frac{1}{2} \frac{1}{\sigma_\beta^2} \beta^2\right) = \\
&= \exp\left(-\frac{1}{2}\left(\beta \sum_{i=1}^{n} x_i^2 + \cancel{\sum_{i=1}^{n} z_i^2} - 2\beta \sum_{i=1}^{n} x_i z_i + \beta^2 \frac{1}{\sigma_\beta^2}\right)\right) = \\
&= \exp\left(-\underbrace{\left(\sum_{i=1}^{n} x_i^2 + \frac{1}{\sigma_\beta^2}\right)}_{\stackrel{\text{def}}{=} (\sigma_{\beta,n}^2)^{-1}} \frac{\beta^2}{2} + \underbrace{\left(\sum_{i=1}^{n} x_i z_i\right)}_{\stackrel{\text{def}}{=} \frac{\mu_{\beta,n}}{\sigma_{\beta,n}^2}} \beta\right),
\end{aligned}$$

where from the 1[st] to the 2[nd] line we used $(Z_i \,|\, \beta)_{i=1}^n$ independent, identically distributed r.v.'s.
So we can conclude that

$$\beta \,|\, y_{1:n}, z_{1:n}, c \sim \mathcal{N}\left(\mu_{\beta,n}, \sigma_{\beta,n}^2\right) \text{ with } \begin{cases} \sigma_{\beta,n}^2 = \left(\sum_{i=1}^n x_i^2 + \frac{1}{\sigma_\beta^2}\right)^{-1} \\ \mu_{\beta,n} = \sigma_{\beta,n}^2 \left(\sum_{i=1}^n x_i z_i\right) \end{cases}$$

$$\Downarrow$$

$$p(\beta \,|\, y_{1:n}, z_{1:n}, c) = \frac{1}{\sqrt{2\pi\sigma_{\beta,n}^2}} \exp\left(-\frac{1}{2\sigma_{\beta,n}^2}(\beta - \mu_{\beta,n})^2\right).$$

□

Point b.

Assuming $c \sim \mathcal{N}\left(0, \sigma_c^2\right)$, show that $p(c \,|\, y_{1:n}, z_{1:n}, \beta)$ is a constrained normal density, i.e. proportional to a normal density but constrained to lie in an interval. Similarly, show that $p(z_i | y_{1:n}, z_{-i}, \beta, c)$ is proportional to a normal density but constrained to be either above c or below c, depending on y_i.

[1]We replace $\mathbb{P}\left(\{z_i > c\}\right)$ with $\mathbb{1}_{(-\infty, z_i)}(c)$ because we will use this characterization afterwards.