In Chapter 6(a), we studied a Poisson regression model, a type of model for count data.

We now examine the **probit regression model**, which we apply to:

1. Binary (2-category) responses, and
2. Multi-category ordinal responses
Example 1: Consider the response variable $Y \in \{1, 2, 3, 4, 5\}$ that indicates the highest educational degree an individual has obtained.

- The categories for $Y$ correspond to: No degree; High school; Associate’s; Bachelor’s; Graduate degree.

- In a regression model, we consider the explanatory variables:

  $X_1 =$ number of children the individual has

  $X_2 = \begin{cases} 
  1 & \text{if either parent of individual has obtained college degree} \\
  0 & \text{otherwise} 
  \end{cases}$

  $X_3 = X_1 X_2$ (interaction variable)
Example: Ordinal Probit Regression

- Using a normal regression model for $Y$ is inappropriate because:
  1. the normal error assumption will be severely violated
  2. the labels $\{1, 2, 3, 4, 5\}$ imply an “equal spacing” between types of degree that may not exist in reality.

- We assume in **probit regression** that the underlying, say, educational achievement of a person is some unobserved continuous variable $Z$.

- What we observe is the ordinal, categorized version, denoted $Y$. 
Example: Ordinal Probit Regression

Our model is thus:

\[ Y_i = g(Z_i), \quad i = 1, \ldots, n \]
\[ Z_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i \]
\[ \epsilon_1, \ldots, \epsilon_n \text{ iid } \sim N(0, 1) \]

The unknown parameters are: \( \beta = (\beta_0, \beta_1, \beta_2, \beta_3) \) and the nondecreasing function \( g(\cdot) \), which relates the latent variable \( Z \) to the observed variable \( Y \).

Note \( g(\cdot) \) can capture the location and scale of the distribution of the \( Y_i \)'s, so we may let \( \text{var}(\epsilon_i) = 1 \) and let the intercept \( \beta_0 = 0 \).
Since $Y$ takes on $K = 5$ ordered values, define $K - 1$ “thresholds” $g_1, g_2, g_3, g_4$ that cut the range of $Z$ into 5 categories:

$$y = g(z) = \begin{cases} 
1 & \text{if } -\infty < z < g_1 \\
2 & \text{if } g_1 \leq z < g_2 \\
3 & \text{if } g_2 \leq z < g_3 \\
4 & \text{if } g_3 \leq z < g_4 \\
5 & \text{if } g_4 \leq z < \infty 
\end{cases}$$

We will use the Gibbs sampler to approximate the joint posterior of $\{\beta, g_1, g_2, g_3, g_4, Z\}$. 
The full conditional of $\beta$ depends only on $Z$:

$$\pi(\beta|y, z, g) = \pi(\beta|z)$$

If we choose a multivariate normal prior

$$\beta \sim MVN(0, n(X'X)^{-1})$$

then the full conditional is:

$$\beta|z \sim MVN\left[ \frac{n}{n+1}(X'X)^{-1}X'z, \frac{n}{n+1}(X'X)^{-1} \right].$$
We know $Z_i \mid \beta \sim N(\beta' x_i, 1)$.

Given $g$ and $Y_i = y_i$, then $Z_i \in [g_{y_i - 1}, g_{y_i})$. Hence

$$\pi(z_i \mid \beta, y, g) \propto N(\beta' x_i, 1) \times I[a \leq z_i < b]$$

(a constrained normal distribution), where $a = g_{y_i - 1}, b = g_{y_i}$.

This can be sampled from fairly easily in $\mathbb{R}$. 

Given \( y \) and \( z \), we know \( g_k \) must be between
\[
\begin{align*}
a_k &= \max\{z_i : y_i = k\} \quad \text{and} \\b_k &= \min\{z_i : y_i = k + 1\}.
\end{align*}
\]
We can choose constrained normal priors on the \( g_k \)'s so that the full conditional of \( g_k \) is \( N(\mu_k, \sigma_k^2) \) constrained to the interval \( [a_k, b_k) \).
Example 1: Educational achievement data on 959 working males.

Let’s use the priors: \( \beta \sim MVN(0, n(X'X)^{-1}) \) and

\[
p(g) \propto \prod_{k=1}^{4} \text{dnorm}(g_k, 0, 100)
\]

constrained so that \( g_1 < g_2 < g_3 < g_4 \).

R example on course web page: Posterior inference is made on \( \beta_1, \beta_2, \beta_3 \).

See plot of generated \( z_1, \ldots, z_{959} \) against the number of children for individuals 1, \ldots, 959.

Different slopes for \( X_2 = 0 \) and \( X_2 = 1 \).
Note that if $Y$ is binary (two-category), the same model could hold, with $K = 2$.

So we have only one threshold $g_1$ separating the two categories.

**Example 2** (54 elderly patients): Let

$$Y_i = \begin{cases} 
1 & \text{if senility is not present in individual } i \\
2 & \text{if senility is present in individual } i 
\end{cases}$$

Explanatory variable $X =$ score on subset of WAIS intelligence test.

See R example on course web page.