GPH-GU2372/3372
Applied Bayesian Analysis in Public Health

Lesson 11: Probit regression

Hai Shu, PhD

12/05/2022

Topics

- Logistic regression
- Probit regression for binary data
- Probit regression for ordinal data
- R implementation

Binary data

- We have seen (Lecture 9) that Bayesian inference for generalized linear model requires using an MCMC algorithm, and in particular a Metropolis or Metropolis-Hastings algorithm.
- In Lecture 9 we have seen the case of a Poisson regression model.
- Here we now consider the case of binary data.
  Example: 54 elderly people completed a subtest of the Wechsler Adult Intelligence Score (WAIS) resulting in a discrete score with a range from 0 to 20. We want to identify people with senility symptoms (binary variable) using the WAIS score. We also want to determine WAIS scores that correspond to increased probability of senility symptoms (i.e. scores with $\pi > 0.5$).

Logistic regression

- As the response variable, presence or not of senility symptoms, is a binary variable, we study the association between WAIS score and senility by fitting a logistic regression.
  Thus, we assume that

$$Y_i|\pi_i \overset{ind}{\sim} \text{Bernouilli}(\pi_i) \qquad i = 1, \ldots, n = 54$$

  with

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_1 + \beta_2 x_i = \mathbf{X}_i'\beta \qquad i = 1, \ldots, n$$

  where x_i is the WAIS score for individual i and $\mathbf{X}$ is the matrix with i-row the 1×2 vector $\mathbf{X}_i' = (1\ x_i)$.
- We complete the specification of the model by placing a prior on the logistic regression coefficients β_1, β_2.
- The usual choice for the prior on the regression coefficients is to assume $\beta = (\beta_1\ \beta_2)' \sim N_2(\beta_0, \mathbf{\Sigma}_0)$.
- Note that if $\mathbf{\Sigma}_0$ is a diagonal matrix, then the prior $\beta \sim N_2(\beta_0, \mathbf{\Sigma}_0)$ is equivalent to $\beta_j \sim N(\beta_{0j}, \mathbf{\Sigma}_{0,jj})$ for $j = 1, 2$

Logistic regression

- In a logistic regression model, the regression coefficient β_2 represents the relative change in log-odds, that is in $\log \frac{\pi}{1-\pi}$ for a one unit increase in the covariate X.
- A similar interpretation holds if the covariate X is a categorical variable with K categories.
- Since

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_1 + \beta_2 x_i \qquad i = 1, \ldots, n$$

setting $x_i = -\frac{\beta_1}{\beta_2}$ will yield $\text{logit}(\pi_i) = 0$, that is $\pi_i = 0.5$. Hence, $x_i = -\frac{\beta_1}{\beta_2}$ gives the value of X for which both probabilities, of success and failure, are equal to 0.5.

Logistic regression

- To make posterior inference on β_1, β_2 in a logistic regression model, we need to derive the posterior distribution of β_1, β_2 **given** the data, $y_1, \ldots, y_n, x_1, \ldots, x_n$.
- The likelihood implied by the logistic regression model is:

$$
\begin{aligned}
p(y_1, \ldots, y_n | \beta_1, \beta_2, x_1, \ldots, x_n) &= \prod_{i=1}^{n} \left[\pi_i^{y_1} \cdot (1 - \pi_i)^{1 - y_i} \right] \\
&= \prod_{i=1}^{n} \left\{ \left(\frac{\exp(\beta_1 + \beta_2 x_i)}{1 + \exp(\beta_1 + \beta_2 x_i)} \right)^{y_i} \right. \\
&\quad \left. \cdot \left(\frac{1}{1 + \exp(\beta_1 + \beta_2 x_i)} \right)^{1 - y_i} \right\} \\
&= \prod_{i=1}^{n} \left\{ \left(\frac{\exp(\mathbf{X}_i' \beta)}{1 + \exp(\mathbf{X}_i' \beta)} \right)^{y_i} \cdot \left(\frac{1}{1 + \exp(\mathbf{X}_i' \beta)} \right)^{1 - y_i} \right\}
\end{aligned}
$$

Logistic regression

- Hence, the posterior distribution of β_1, β_2 is:

$$
\begin{aligned}
p(\beta_1, \beta_2 | y_1, \ldots, y_n, x_1, \ldots, x_n) &\propto p(y_1, \ldots, y_n | \beta_1, \beta_2, x_1, \ldots, x_n) \cdot p(\beta_1, \beta_2) \\
&= \left[\prod_{i=1}^{n} \left(\frac{\exp(\mathbf{X}_i' \beta)}{1+\exp(\mathbf{X}_i' \beta)} \right)^{y_i} \cdot \left(\frac{1}{1+\exp(\mathbf{X}_i' \beta)} \right)^{1-y_i} \right] \\
&\cdot \frac{1}{(2\pi)|\mathbf{\Sigma}_0|^{\frac{1}{2}}} \exp\left(-\tfrac{1}{2}(\beta - \beta_0)' \mathbf{\Sigma}_0^{-1} (\beta - \beta_0)\right)
\end{aligned}
$$

- This is clearly not a distribution that we know. To approximate the posterior distribution of $\beta = (\beta_1 \ \beta_2)'$ using simulation methods, we will need to use the Metropolis/Metropolis-Hastings algorithm.

Probit regression

- There is a different way to parameterize a regression model for binary data that renders Bayesian computation more amenable.
- Suppose that there is a latent random variable Z_i such that

$$z_i = \beta_1 + \beta_2 x_i + \varepsilon_i \qquad \varepsilon_i \sim N(0,1) \qquad i = 1, \ldots, n$$

and

$$Y_i = \begin{cases} 1 & \text{if } Z_i > 0 \\ 0 & \text{if } Z_i \leq 0 \end{cases}$$

- Then, since $Z_i|\beta = (\beta_1 \ \beta_2)' \sim N(\beta_1 + \beta_2 x_i, 1) = N(\mathbf{X}_i'\beta, 1)$ it follows that

$$\begin{aligned} \pi_i &= P(Y_i = 1) = P(Z_i > 0) = 1 - P(Z_i \leq 0) \\ &= 1 - P(Z_i - \mathbf{X}_i'\beta \leq -\mathbf{X}_i'\beta) \\ &= 1 - \Phi(-\mathbf{X}_i'\beta) = \Phi(\mathbf{X}_i'\beta) = \Phi(\beta_1 + \beta_2 x_i) \end{aligned}$$

Probit regression

- Hence, in this model we have

$$\begin{aligned} Y_i|\pi_i &\overset{ind}{\sim} \text{Bernouilli}(\pi_i) \qquad i = 1, \ldots, n \\ \pi_i &= \Phi(\beta_1 + \beta_2 x_i) \qquad i = 1, \ldots, n \end{aligned}$$

or equivalently

$$Y_i|Z_i = \begin{cases} 1 & \text{if } Z_i > 0 \\ 0 & \text{if } Z_i \leq 0 \end{cases} \qquad i = 1, \ldots, n$$

$$Z_i|\beta = (\beta_1 \ \beta_2)' \overset{ind}{\sim} N(\mathbf{X}_i'\beta, 1) \qquad i = 1, \ldots, n$$

- Such a model, where the probability of success π_i is linked to the regression coefficients β_1, β_2 via the cumulative distribution function of a standard normal random variable is called a probit regression model.
- In a probit regression model, the regression coefficient β_2 represents the expected change in the latent variable Z_i for a one-unit increase in the explanatory variable X_i.

Probit regression

- Interpreting the regression coefficients in a probit regression model in terms of changes in the probability of success is harder.
- Consider the difference in the probability of success $\pi(X = x+1)$ versus the probability of sucess $\pi(X = x)$ under the probit regression model. We have

$$\pi(X = x+1) = \Phi(\beta_1 + \beta_2(x+1)) = \Phi(\beta_1 + \beta_2 x + \beta_2) = \Phi(\Phi^{-1}(\pi(x)) + \beta_2)$$

- A better interpretation of the regression coefficient β_2 can be obtained if we consider the value of $X = x_c = -\frac{\beta_1}{\beta_2}$. In that case:

$$\pi(x_c + 1) = \Phi\left(\beta_1 + \beta_2\left(-\frac{\beta_1}{\beta_2} + 1\right)\right) = \Phi(\beta_2)$$

- If for example $\beta_2 = 0.5, 1, 2, 3$ then the probability that $Y = 1$ is respectively, $0.69, 0.84, 0.977$ and 0.99865.

Probit regression

- Since $\pi(x_c) = \Phi(\beta_1 + \beta_2 \cdot \left(-\frac{\beta_1}{\beta_2}\right)) = \Phi(0) = 0.5$, this implies that if $\beta_2 = 0.5, 1, 2, 3$ a one-unit increase in X from x_c to $x_c + 1$ corresponds to an increase in the probability of success by $\frac{\pi(x_c+1)}{\pi(x_c)}$, that is, by $38\%, 68\%, 95\%$ and 100%, respectively.
- The loss of interpretability in probit regression models is accompanied by an ease in Bayesian computation.In fact, using the common choice of a $N_2(\beta_0, \mathbf{\Sigma}_0)$ prior for β, we have:

$$
\begin{aligned}
Y_i|Z_i &= \begin{cases} 1 & \text{if } Z_i > 0 \\ 0 & \text{if } Z_i \leq 0 \end{cases} \qquad i = 1, \ldots, n \\
Z_i|\beta = (\beta_1\ \beta_2)' &\overset{ind}{\sim} N(\mathbf{X}_i'\beta, 1) \qquad i = 1, \ldots, n \\
\beta &\sim N_2(\beta_0, \mathbf{\Sigma}_0)
\end{aligned}
$$

Probit regression

- The parameters in this model are $\beta, z_1, \ldots, z_n$. Thus, we want to derive the joint posterior distribution $p(\beta, z_1, \ldots, z_n | y_1, \ldots, y_n, x_1, \ldots, x_n)$.
- We have:

$$\begin{aligned} p(\beta, z_1, \ldots, z_n | y_1, \ldots, y_n, x_1, \ldots, x_n) &\propto \{\prod_{i=1}^n p(y_i | z_i)\} \\ &\cdot \{\prod_{i=1}^n p(z_i | \beta)\} \cdot p(\beta) \end{aligned}$$

- We approximate this joint posterior distribution by devising an MCMC algorithm that generates a Markov chain that has as stationary distribution the joint posterior distribution. We use the Gibbs sampling algorithm as MCMC algorithm.
- Thus, we will sample sequentially all the parameters from their respective full conditional distributions.
- We need to derive the full conditional of β and $z_1, \ldots, z_n$.

Probit regression

- The full conditional distribution of β is:

$$\begin{aligned} p(\beta|z_1,\ldots,z_n,y_1,\ldots,y_n,x_1,\ldots,x_n) &\propto \{\prod_{i=1}^n p(z_i|\beta)\}\cdot p(\beta) \\ &= \{\prod_{i=1}^n N(z_i;\mathbf{X}_i'\beta,1)\}\cdot N_2(\beta;\beta_0,\mathbf{\Sigma}_0) \\ &= N_n(\mathbf{z};\mathbf{X}\beta,\mathbf{I}_n)\cdot N_2(\beta;\beta_0,\mathbf{\Sigma}_0) \end{aligned}$$

- The full conditional of β is $N_2(\beta_n,\mathbf{\Sigma}_n)$ where

$$\begin{aligned} \mathbf{\Sigma}_n &= \left(\mathbf{X}'\mathbf{X}+\mathbf{\Sigma}_0^{-1}\right)^{-1} \\ \beta_n &= \mathbf{\Sigma}_n\cdot(\mathbf{X}'\mathbf{z}+\mathbf{\Sigma}_0^{-1}\beta_0) \end{aligned}$$

Probit regression

- The full conditional distribution of z_i is:

$$p(z_i|\beta, \mathbf{z}_{-i}, y_1, \ldots, y_n, x_1, \ldots, x_n) \propto p(y_i|z_i) \cdot p(z_i|\beta)$$

$$= \delta_{z_i}(R_{y_i}) \cdot N(z_i; \mathbf{X}_i'\beta, 1)$$

  where

  - R_{y_i} is the subset where z_i should lie based on the value of y_i. That is, if $y_i = 1$, then $R_{y_i} = (0, +\infty)$ while if $y_i = 0$, then $R_{y_i} = (-\infty, 0]$
  - $\delta_{z_i}(R_{y_i})$ is the indicator function for z_i that is equal to 1 if $z_i \in R_{y_i}$ and 0 otherwise.

- The full conditional of z_i is a truncated normal distribution, truncated to the interval $(0, +\infty)$ if $y_i = 1$, truncated to $(-\infty, 0]$ otherwise.

Probit regression

- A Gibbs sampling algorithm for the probit regression model works as follows:
  - Choose a number of iterations S.
  - Determine initial values $\beta^{(0)}$ and $z_1^{(0)}, \ldots, z_n^{(0)}$ for β and $z_1, \ldots, z_n$.
  - For iterations $k = 1, \ldots, S$, repeat the following two steps
    1. sample $\beta^{(k)}$ from the full conditional distribution $p(\beta | z_1^{(k-1)}, \ldots, z_n^{(k-1)}, \mathbf{X})$
    2. for $j = 1, \ldots, n$ sample $z_j^{(k)}$ from the full conditional distribution $p(z_j | y_j, \beta^{(k)})$.

Probit regression

- In R, fitting a probit regression model is easy:
  - To simulate from the full conditional of β we simply need to sample from a multivariate normal. In R, we can do that using the function `mvrnorm`.
  - To simulate from the full conditional of z_i, that is, a truncated normal distribution, we could either use the `msm` package in R that contains the `rtnorm` function to simulate truncated normal random variables with a given mean μ, given standard deviation σ and truncated to be in the interval (a, b).

Probit regression

- For the data on the 54 elderly people that took a WAIS test and for which we recorded the outcome variable, presence or absence of senility symptoms, we have the following posterior summaries for the two choices of link function

Variable	Logit link Mean (SD)	Probit link Mean (SD)
β_1	2.507 (1.229)	1.402 (0.661)
β_2	-0.339 (0.119)	-0.191 (0.061)
WAIS ($\pi = 0.5$)	6.975 (2.104)	6.677 (3.195)
DIC	55.105	54.997

- Both models indicate a negative association between the WAIS score and the presence of senility symptoms.

Probit regression

Variable	Logit link Mean (SD)	Probit link Mean (SD)
β_1	2.507 (1.229)	1.402 (0.661)
β_2	-0.339 (0.119)	-0.191 (0.061)
WAIS ($\pi = 0.5$)	6.975 (2.104)	6.677 (3.195)
DIC	55.105	54.997

- From the logistic regression model, the posterior odds of senility symptoms for individuals scoring zero in WAIS are expected to be equal to $\exp(\beta_1) = 12.27$. For each additional point of the WAIS score, we expect a decrease of 28.8% in the odds of senility symptoms.
- For the probit model, the corresponding approximate posterior odds decrease of 26.3% for an additional point to the WAIS score=6.677.
- For the threshold disease value, both models indicate that we classify somebody as a case whenever WAIS ≤ 6.

Probit regression for ordinal variables

- The DIC value for the probit regression model is the lowest, therefore the probit regression model provides the best fit. However, the difference between the two models is quite small indicating minor differences between the two models.
- A probit regression model can also be used if we are interested in studying the association between an ordinal variable and other variables.
- An ordinal variable is a random variable for which there is a logical ordering of the sample space. If the variable has a meaningful numeric scale, then we call it an ordinal numeric variable (example: number of children an individual has).
  A variable is said continuos if it is numeric and it can take any value in an interval (example: height, weight; they are both ordinal, numeric, continuous variable.

Probit regression for ordinal variables

- Example: Suppose that we are interested in describing the relationship between the educational attainment and the number of children for individuals in a population. Additionally, we believe that an individual educational attainment is influenced by their parent's education level.
- The 1994 General Social Survey provides data on DEG_i, CHILD_i and PDEG_i, respectively, the highest degree obtained by individual i, his/her number of children and an indicator on whether or not either parent of individual i have obtained a college degree.
- Then:
  - DEG_i is an ordinal but not numeric variable
  - CHILD_i is a an ordinal, numeric and discrete variable
  - PDEG_i is an ordinal, numeric and discrete variable
- To relate DEG_i and CHILD_i, PDEG_i and the interaction $\text{CHILD}_i \times \text{PDEG}_i$ we cannot use a linear regression model but we can use an ordered probit regression model.

Ordered probit regression

- Therefore, we assume that there exists a latent underlying numeric process Z that drives the observed variable DEG.
- As in the case of probit regression, we assume that the variable Z is related to the covariates $X_1, \ldots, X_p$ via a linear regression model, that is: for $i = 1, \ldots, n$

$$\begin{aligned} z_i &= \beta_1 x_{1,i} + \beta_2 x_{2,i} + \ldots \beta_p x_{p,i} + \varepsilon_i \\ &= \mathbf{X}_i' \beta + \varepsilon_i \qquad \varepsilon_i \sim N(0,1) \qquad i = 1, \ldots, n \end{aligned}$$

implying that $Z_i | \beta \overset{ind}{\sim} N(\mathbf{X}_i' \beta, 1)$and

$$Y_i = g(Z_i)$$

where g is a non-decreasing function that relates the value of Z_i to the observed value of Y_i.

Ordered probit regression

- Note that in the probit regression model we have taken the variance of $\varepsilon_1, \ldots, \varepsilon_n$ to be equal to 1. This is because the scale of the distribution of Y can be incorporated in the g function.
- Similarly, the g function also includes information on the location of the distribution of Y, therefore, there is no need to include an intercept term among the covariates $X_1, \ldots, X_p$.
- If Y can take on K values, say $\{1, 2, \ldots, K\}$, the function g can be described with $K-1$ parameters, $g_1, \ldots, g_{K-1}$ and we set

$$Y = \begin{cases} 1 & \text{if} \quad -\infty = g_0 < Z \leq g_1 \\ 2 & \text{if} \quad g_1 < Z \leq g_2 \\ \vdots & \vdots \quad \vdots \\ K & \text{if} \quad g_{K-1} < Z < g_K = +\infty \end{cases}$$

- The values $g_1, \ldots, g_{K-1}$ can be thought of as thresholds so that when Z_i is above a certain threshold, then Y_i is in the next highest category.

Ordered probit regression

- We complete the specification of the ordered probit regression model by placing a prior on β and on the thresholds $g_1, \ldots, g_{K-1}$ leading to the following model

$$\begin{aligned} Y_i | Z_i &= g(Z_i) \\ Z_i | \beta &\overset{ind}{\sim} N(\mathbf{X}_i' \beta, 1) \\ \beta &\sim N_p(\beta_0, \mathbf{\Sigma}_0) \\ \mathbf{g} = \{g_1, \ldots, g_{K-1}\} &\sim p(\mathbf{g}) \end{aligned}$$

- The parameters in this model are $\beta, \mathbf{g}, z_1, \ldots, z_n$. We infer upon them by deriving the joint posterior distribution.

Ordered probit regression

- The joint posterior distribution $p(\beta, \mathbf{g}, z_1, \ldots, z_n | y_1, \ldots, y_n, \mathbf{X})$ is:

$$\begin{aligned} p(\beta, \mathbf{g}, z_1, \ldots, z_n | y_1, \ldots, y_n, \mathbf{X}) &\propto \{\prod_{i=1}^n p(y_i | z_i)\} \\ &\cdot \{\prod_{i=1}^n p(z_i; \beta)\} \cdot p(\beta) \cdot p(\mathbf{g}) \\ &= \{\prod_{i=1}^n p(y_i | z_i)\} \\ &\cdot p(\mathbf{z}; \beta) \cdot p(\beta) \cdot p(\mathbf{g}) \end{aligned}$$

- We approximate this joint posterior distribution via simulations. We use a Gibbs sampling algorithm: we sequentially sample every parameter from its full conditional and we generate a Markov chain with stationary distribution the joint posterior distribution $p(\beta, \mathbf{g}, z_1, \ldots, z_n | y_1, \ldots, y_n, \mathbf{X})$.

Ordered probit regression

- Since $p(\mathbf{z};\beta) = N(\mathbf{z};\mathbf{X}\beta,\mathbf{I})$ and $p(\beta) = N_p(\beta_0,\mathbf{\Sigma}_0)$, it follows that the full conditional of β is $N_p(\beta_n,\mathbf{\Sigma}_n)$ where

$$\mathbf{\Sigma}_n = \left(\mathbf{X}'\mathbf{X} + \mathbf{\Sigma}_0^{-1}\right)^{-1} \qquad \beta_n = \mathbf{\Sigma}_n\left(\mathbf{X}'\mathbf{z} + \mathbf{\Sigma}_0^{-1}\beta_0\right)$$

- As in the case of probit regression, the full conditional of z_i is given by:

$$\begin{aligned} p(z_i|\mathbf{z}_{-i},\beta,\mathbf{g},y_1,\ldots,y_n,x_1,\ldots,x_n) &\propto p(y_i|z_i)\cdot p(z_i;\beta) \\ &= p(y_i|z_i)\cdot N(z_i;\mathbf{X}_i'\beta,1) \end{aligned}$$

and this is a truncated normal distribution $TN(\mathbf{X}_i'\beta,1)$ constrained to be in the interval (g_{y_i-1},g_{y_i}) where (g_{y_i-1},g_{y_i}) is the interval relative to the value of y_i.

Ordered probit regression

- For the full conditional of $\mathbf{g}$ we proceed as follows: given the data $y_1, \ldots, y_n$ and the latent variables $z_1, \ldots, z_n$, we know that g_k has to be higher than all those z_i for which $y_i = k$ and lower than all those z_i for which $y_i = k+1$.
- We call $a_k = \max\{z_i : y_i = k\}$ and $b_k = \min\{z_i : y_i = k+1\}$.
- Then, the full conditional of $\mathbf{g}$ is proportional to $p(\mathbf{g})$ but subject to the constraints $a_k < g_k < b_k$ for $k = 1, \ldots, K-1$.
- For example, if $p(\mathbf{g}) = \prod_{k=1}^{K-1} N(g_k; \mu_k, \gamma_k^2)$, then the full conditonal for $\mathbf{g}$ is

$$p(\mathbf{g}|z_1, \ldots, z_n, \beta, y_1, \ldots, y_n, \mathbf{X}) \propto \prod_{k=1}^{K-1} N(g_k; \mu_k, \gamma_k^2) \cdot \delta_{g_k}(a_k, b_k)$$

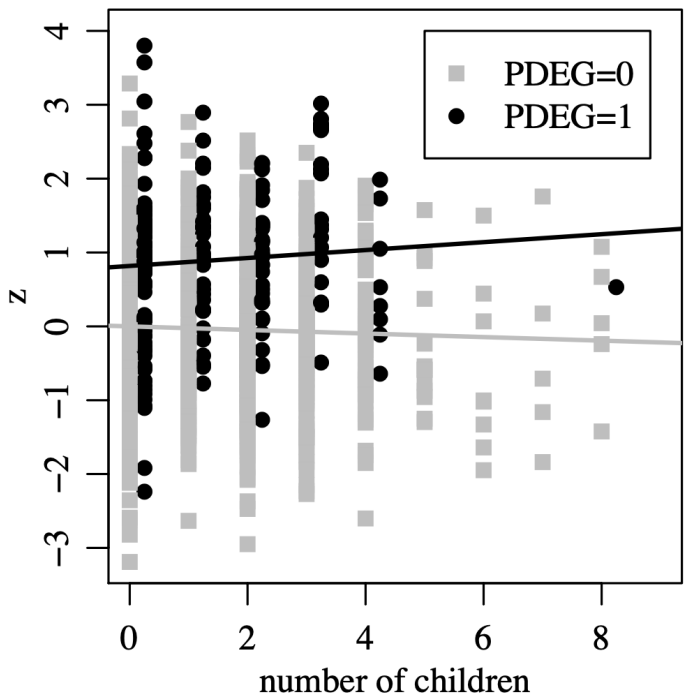
that is, each g_k has a truncated normal distribution with mean μ_k, variance γ_k^2 and constrained to be in the interval (a_k, b_k).

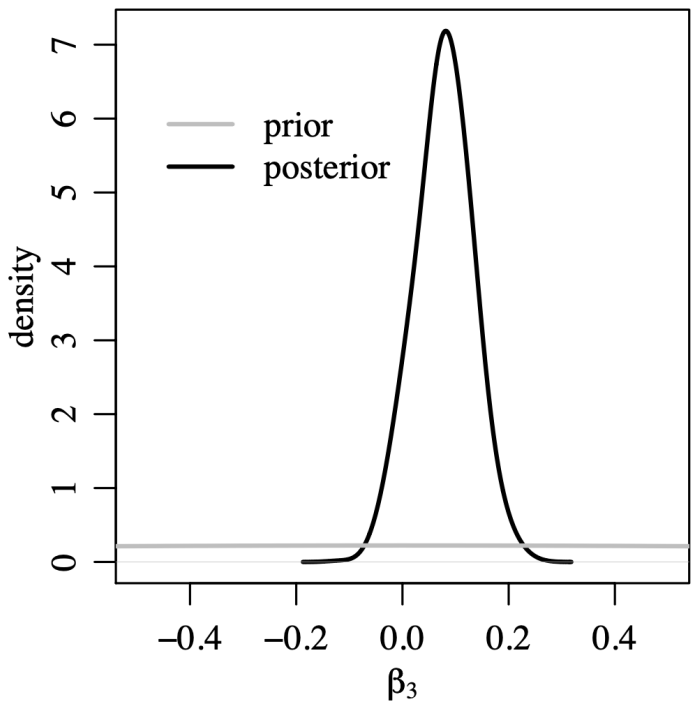
Ordered probit regression

- We return to consider the example on educational attainment.
- In that case, the covariates used were CHILD_i, PDEG_i and $\text{CHILD}_i \times \text{PDEG}_i$.
- As prior on β the book used a $N_3(\mathbf{0}, n(\mathbf{X}'\mathbf{X})^{-1})$ where $\mathbf{X}$ is the $n \times 3$ matrix with the covariates.
- As prior on $\mathbf{g}$ the book used $p(\mathbf{g}) \propto \prod_{k=1}^{K-1} N(g_k; 0, 100)$ with the constraint that $g_1 < g_2 < \ldots < g_{K-1}$.
- A Gibbs sampling algorithm was run for $S = 25,000$ iterations saving the output every 25 iterations. Posterior inference was then based on a sample of 1,000 iterations.

Ordered probit regression

- Plot of the regression line for people with and without a college degree: on the x-axis there is the number of children and on the y-axis the values $z_1^{(S)}, \ldots, z_n^{(S)}$ at the last MCMC iteration.





- Extensions of the ordered probit regression and other examples of models that involve a latent variable are discussed in the book in Section 12.1.2 (rank likelihood) and Section 12.3 (Gaussian copula model).