COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

STAT 4224/5224

Bayesian Statistics

Dobrin Marchev

Ordinal data

- Many datasets include variables whose distributions cannot be represented by the normal, binomial or Poisson distributions.
- For example, distributions of common survey variables such as age, education level and income generally cannot be accurately described by any of the above- mentioned sampling models. Such variables are often binned into ordered categories, the number of which may vary from survey to survey.
- In such situations, interest often lies not in the scale of each individual variable, but rather in the associations between the variables: Is the relationship be- tween two variables positive, negative or zero?

Example 1 (p. 209)

Suppose we are interested in describing the relationship between the edu- cational attainment and number of children of individuals in a population. Additionally, we might suspect that an individual's educational attainment may be influenced by their parent's education level. The 1994 General Social Survey provides data on variables DEG, CHILD and PDEG for a sample of individuals in the United States, where DEG_i indicates the highest degree obtained by individual i, $CHILD_i$ is their number of children and $PDEG_i$ is the binary indicator of whether either parent of i obtained a college degree.

Example (continued)

Suppose we fit a regular linear regression model:

$$DEG_i = \beta_0 + \beta_1 CHILD_i + \beta_2 PDEG_i + \beta_3 CHILD_i \times PDEG_i + \varepsilon_i$$

where $\varepsilon_i \sim iid\ N(0, \sigma^2)$

What's wrong with that?

- Empirical distribution of DEG is over the digits 1, … , 5. Since the variable DEG takes on only a small set of discrete values, the normality assumption of the residuals will certainly be violated.
- The regression model imposes a numerical scale to the data that is not present: A bachelor's degree is not "twice as much" as a high school degree, and an associate's degree is not "two less" than a graduate degree. There is an order to the categories in the sense that a graduate degree is "higher" than a bachelor's degree, but otherwise the scale of DEG is not meaningful.

Ordinal Regression

Ordinal logistic regression is a statistical analysis method that can be used to model the relationship between an ordinal response variable and one or more explanatory variables.

The model is:

$$\log \frac{P(Y \leq j)}{1 - P(Y \leq j)} = \beta_{j0} - \beta_1 x_1 - \beta_2 x_2 - \cdots - \beta_p x_p,$$
$$j = 1, \ldots, J - 1$$

The model incorporates a negative sign so that there is a direct correspondence between the slope and the ranking. Thus, a positive coefficient indicates that as the value of the explanatory variable increases, the likelihood of a higher-ranking increases.

Example 2:
Applying to graduate school

This dataset has a three-level variable called apply, with levels "unlikely", "somewhat likely", and "very likely", coded 1, 2, and 3, respectively, that we will use as our outcome variable. We also have three variables that we will use as predictors: pared, which is a 0/1 variable indicating whether at least one parent has a graduate degree; public, which is a 0/1 variable where 1 indicates that the undergraduate institution is public and 0 private, and gpa, which is the student's grade point average.

We will fit the model in R with polr function (proportional odds logistic regression)

Estimated model:

$$logit\big(P(Y \leq 1)\big) = 2.20 - 1.05 * PARED - (-0.06) * PUBLIC - 0.616 * GPA$$
$$logit(P(Y \leq 2)) = 4.30 - 1.05 * PARED - (-0.06) * PUBLIC - 0.616 * GPA$$

Probit regression

It is natural to think of many ordinal, non-numeric variables as arising from some underlying numeric process. For example, the severity of a disease might be described "low", "moderate" or "high", although we imagine a patient's actual condition lies within a continuum. Similarly, the amount of effort a person puts into formal education may lie within a continuum, but a survey may only record a rough, categorized version of this variable, such as DEG. This idea motivates a modeling technique known as ordered probit regression, in which we relate a variable *Y* to a vector of predictors *x* via a regression in terms of a latent variable *Z*:

$$Y_i = g(Z_i)$$
$$Z_i = \beta' x_i + \varepsilon_i$$
$$\varepsilon_1, \ldots, \varepsilon_n \sim iid\ N(0,1)$$

Probit regression (continued)

The function g in $Y_i = g(Z_i)$ is taken to be non-decreasing, so that we can interpret small and large values of Z as corresponding to small and large values of Y. The scale of the distribution of Y can already be represented by g, so we don't need variance parameter.

If the sample space for Y takes on K values, say {1,...,K}, then the function g can be described with only K − 1 ordered parameters

$g_1 < g_2 < \cdots < g_{k-1}$

$$y = g(z) = \begin{cases} 1, if \ -\infty = g_0 < z < g_1 \\ 2, \qquad if \ z_1 < z < g_2 \\ \qquad \vdots \\ K, if \ g_{K-1} < z < g_k = \infty \end{cases}$$

The values $\{g_1,...,g_{K-1}\}$ are to be estimated, and can be thought of as "thresholds," so that moving z past a threshold moves y into the next highest category.

Gibbs Sampler

If we use normal prior distributions for the unknown parameters, the joint posterior distribution of $\{\boldsymbol{\beta}, g_1, \ldots, g_{K-1}, Z_1, \ldots, Z_n\}$ given $\boldsymbol{Y} = \boldsymbol{y} = (y_1, \ldots, y_n)$ can be approximated using a Gibbs sampler.

- If $\boldsymbol{\beta} \sim N(0, n(\boldsymbol{X}'\boldsymbol{X})^{-1})$, then

$$\boldsymbol{\beta}|\boldsymbol{z} \sim N\left(\frac{n}{n+1}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{z}, \frac{n}{n+1}(\boldsymbol{X}'\boldsymbol{X})^{-1}\right)$$

- Given g, observing $Y_i = y_i$ means Z_i must lie inside the interval $(a, b) = \left(g_{y_i-1}, g_{y_i}\right)$:

$$Z_i|\boldsymbol{\beta}, \boldsymbol{y}, g \sim N(z_i, \boldsymbol{\beta}'\boldsymbol{x_i}, 1) \times \delta_{(a,b)}(z_i)$$

There are methods to sample from such constrained (truncated) Gaussian distributions. For example, the inverse cdf can be used.

- The full conditional density of g_k is a $N(\mu_k, \sigma_k^2)$ density constrained to some intervals as well (see R code).

Back to the example

The lines suggest that for people whose parents did not go to college, the number of children they have is indeed weakly negatively associated with their educational outcome. However, the opposite seems to be true among people whose parents went to college. The posterior distribution of β_3 is given in the second panel of the fig ure, along with the prior distribution for comparison.