Modeling Ordinal Categorical Variables

Bayesian Modeling for Socio-Environmental Data

Chris Che-Castaldo, Mary B. Collins, N. Thompson Hobbs

Summer 2018

How confident are you in your ability use Bayesian models?

1 2 3 4 5

We use *ordinal regression* to deal with data where the dependent variable is measured in ordered categories. Examples of such variables include:

- Psyschometric Likert scales
- Tumor grading
- General quantities (i.e. insurance level: none, adequate, full; index of environmental concern: none, low, moderate, high)
- Cover classes (i.e., Daubenmire classes)

Ordered categorical data can be

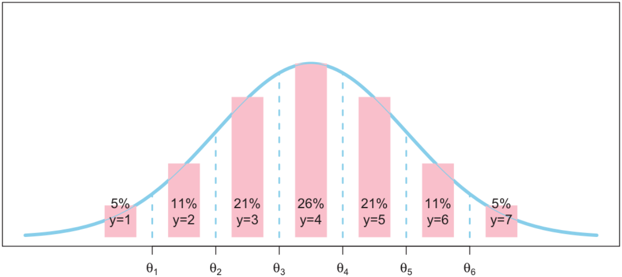
- unscaled (e.g. attitudes/opinions, etc.)
- scaled (e.g. cover/size classes, etc.)

Useful reference

Kruschke, J. (2014). Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan. Academic Press.

“How do people generate a descrete ordered response?”

- Imagine that your true Bayesian abilities vary on a continuous scale, but you also have some sense of which categorical threshold you would report
- **Central idea**: there is a latent continuous metric that underlies the observed ordinal response
- Categories or *thresholds* partition regions of this continuous metric



Crutial bit: *the probabiliy of a particular ordinal outcome is the area under the normal curve between the thresholds of that outcome*.

Therefore, the probability of outcome 2 is the area under the normal curve between thresholds θ_1 and θ_2. How?

A general, Bayesian model for ordinal data

$$[\boldsymbol{\theta}, \beta, \sigma^2|\mathbf{y}] \propto \prod_{i=1}^{n} \left[y_i \mid \underbrace{\int_{\theta_{k-1}}^{\theta_k} \overbrace{[z_i|g(\beta, \mathbf{x}_i), \sigma^2]}^{\text{pdf of latent response}} \, \mathrm{d}z_i}_{Pr(\theta_{k-1} < z_i < \theta_k)} \right] [\beta][\boldsymbol{\theta}][\sigma^2]$$

- y_i is *ith* observation in categories $= k = 1, ...K$
- $\boldsymbol{\theta}$ is an *ordered* vector of cutpoints
- $\theta_0 = -\infty$
- $\theta_K = +\infty$

Why is $\boldsymbol{z}$ missing from the posterior?

What is $Pr(\theta_{k_{i-1}} < z_i < \ \theta_{k_i})$?

What is the quantity between the large brackets?

An general algorithm for implementation

Let $F(\theta_k, \mu, \sigma^2)$ be a properly moment matched, cummulative distribution function for the distribution of the latent quantity z_i. The function F() returns the proability that $z_i < \theta_k$. For notational convenience, we let $\mu_i = g(\boldsymbol{\beta}, \mathbf{x}_i)$. Compute:

$$p[1,i] = F(\theta_1, \mu_i, \sigma^2) \tag{1}$$
$$p[2,i] = F(\theta_2, \mu_i, \sigma^2) - F(\theta_1, \mu, \sigma^2) \tag{2}$$
$$. \tag{3}$$
$$. \tag{4}$$
$$p[K-1] = F(\theta_{K-1}, \mu, \sigma^2) - F(\theta_{K-2}, \mu, \sigma^2) \tag{5}$$
$$p[K] = 1 - F(\theta_K, \mu, \sigma^2) \tag{6}$$

The likelihood of the data conditional on the parameters is then:

$$y_i \sim \text{categorical}(\mathbf{p}_i)$$

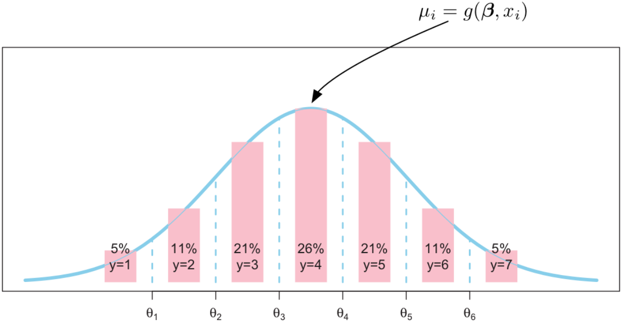
The categorical distribution

$$y_i \sim \text{categorical}(\mathbf{p}_i)$$

Let y_i be an observation that can take on values $k = 1, .., K$. $\mathbf{p}$ is a vector of length K with elements $p_i = \Pr(y_i = k_i)$, which is the same as $\Pr(y_i = i)$.

You can use *any continuous distribution* appropirate to the support of the random variable, y_i.

Issues of identifiability and what to do about it

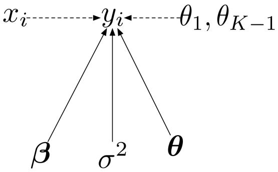


- The likelihood will not result in a unique solution.
- Both β and θ are "location" parameters that calibrate the mapping from what is observed, y_i to the latent z_i.
- In other workds, there is no unique combination of θ and β that produce equally informative posterior distributions.
- Put differently, for any given β there exists a θ that produces a likelihood equal to that obtained from at least one other β and θ.

Potential Identification Contraints to Apply

Options	β	σ	θ
1	unconstrained	fixed	fix one of θ_j
2	drop intercept, β_0	fixed	unconstrained
3	unconstrained	unconstrained	fix two of θ_j

xample: Predicting A *Unscaled* Ordinal Quantity



$$\mu = \beta_1 + \beta_2 x_i$$

$$[\boldsymbol{\theta}, \boldsymbol{\beta}, \sigma^2 | \mathbf{y}] \propto \prod_{i=1}^{n} \left[y_i \mid \int_{\theta_{k-1}}^{\theta_k} [z_i \mid g(\boldsymbol{\beta}, x_i), \sigma^2] dz \right.$$

$$\times [\beta_1][\beta_2] \prod_{j=2}^{K-2} [\theta_j][\sigma]$$

$$y_i \sim \left[y_i \mid \int_{\theta_{k-1}}^{\theta_k} [z_i \mid g(\boldsymbol{\beta}, x_i), \sigma^2] dz_i \right]$$

$$\boldsymbol{\beta} \sim \text{normal}(0, 0.001)$$

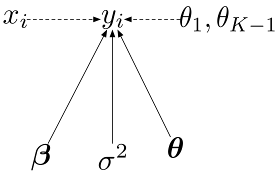
$$\sigma \sim \text{uniform}(0, 100)$$

$$\theta_j \sim \text{uniform}(0, 10)$$

```
for (i in 1:length(y)) {
  mu[i] = beta[1] + beta[2]*x[i]
  y[i] ~ dcat( pr[i,1:nYlevels])
  y.sim[i] ~ dcat( pr[i,1:nYlevels])
  pr[i,1] <- pnorm( thresh[1], mu[i] , tau)

  for ( k in 2:(nYlevels-1) ) {
    pr[i,k] <- max(.00001, pnorm( thresh[ k ] , mu[i] , tau) - pnorm( thresh[k-1] , mu[i] , tau ))
    }
  pr[i,nYlevels] <- 1 - pnorm( thresh[nYlevels-1] , mu[i] , tau )
  }
```

Example: Predicting A *Scaled* Ordinal Quantity



$$\mu = \frac{e^{\beta_1 + \beta_2 x_i}}{1 + e^{\beta_1 + \beta_2 x_i}} = g(\boldsymbol{\beta}, x_i)$$

$$[\boldsymbol{\theta}, \boldsymbol{\beta}, \sigma^2 | \mathbf{y}] \propto \prod_{i=1}^{n} \left[y_i \mid \int_{\theta_{k-1}}^{\theta_k} [z_i \mid m(g(\boldsymbol{\beta}, x_i), \sigma^2)] dz_i \right]$$
$$\times [\beta_1][\beta_2] \prod_{j=2}^{K-2} [\theta_j][\sigma]$$

```
for (i in 1:length(y)) {
      mu[i] = ilogit(beta[1] + beta[2]*x[i])
      a[i] <-max(.00001,(mu[i]^2-mu[i]^3-mu[i]*sigma^2)/sigma^2)
      b[i] <- max(.00001,(mu[i]-2*mu[i]^2+mu[i]^3-sigma^2+mu[i]*sigma^2)/sigma^2 )
      y[i] ~ dcat( pr[i,1:nYlevels])
      pr[i,1] <- pbeta( theta[1], a[i] , b[i])
      for ( k in 2:(nYlevels-1) ) {
        pr[i,k] <- max(.00001, pbeta( theta[ k ] , a[i] , b[i]) - pbeta( theta[k-1] , a[i] , b[i] ))
      }
      pr[i,nYlevels] <- 1 - pbeta( theta[nYlevels-1] , a[i] , b[i] )
      }
```

$$y_i \sim \left[y_i \mid \int_{\theta_{k-1}}^{\theta_k} [z_i \mid m(g(\boldsymbol{\beta}, x_i), \sigma^2)] dz_i \right]$$

$$\boldsymbol{\beta} \sim \text{normal}(0, 0.0001)$$
$$\sigma \sim \text{uniform}(0.01, .5)$$
$$\theta_j \sim \text{uniform}(0, 1)$$

Other notables

- Referred to as *ordinal regression* or *ordered probit regression*.
- Cut points are often specified using τ.
- The latent quantity that we are calling z_i is also specified as y_i^*
- Often in the unscaled case, the standard normal is used ($\beta_0 = 0$ and $\sigma = 1$) with the probabily of outcome θ_k being:

$$p(\tau = k \mid \mu, \sigma, \theta_j) = \Phi((\theta_k - \mu)/\sigma) - \Phi((\theta_{k-1} - \mu)/\sigma)$$

Table 15.2: For the generalized linear model: typical noise distributions and inverse-link functions for describing various scale types of the predicted variable y. The value μ is a central tendency of the predicted data (not necessarily the mean). The predictor variable is x, and $\text{lin}(x)$ is a linear function of x, such as those shown in Table 15.1. Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

Scale Type of Predicted y	Typical Noise Distribution $y \sim \text{pdf}(\mu, [\text{parameters}])$	Typical Inverse-Link Function $\mu = f(\text{lin}(x), [\text{parameters}])$
Metric	$y \sim \text{normal}(\mu, \sigma)$	$\mu = \text{lin}(x)$
Dichotomous	$y \sim \text{bernoulli}(\mu)$	$\mu = \text{logistic}(\text{lin}(x))$
Nominal	$y \sim \text{categorical}(\ldots, \mu_k, \ldots)$	$\mu_k = \dfrac{\exp(\text{lin}_k(x))}{\sum_c \exp(\text{lin}_c(x))}$
Ordinal	$y \sim \text{categorical}(\ldots, \mu_k, \ldots)$	$\mu_k = \Phi((\theta_k - \text{lin}(x))/\sigma) - \Phi((\theta_{k-1} - \text{lin}(x))/\sigma)$
Count	$y \sim \text{poisson}(\mu)$	$\mu = \exp(\text{lin}(x))$