STAT 530: Ordinal Data Models

April 7, 2010

Text Ex.

Want to do regression, with a response variable such as:

$$Y = \begin{cases} 1 & \text{no high school completion} \\ 2 & \text{high school completion} \\ 3 & \text{`associate' degree} \\ 4 & \text{bachelor's degree} \\ 5 & \text{graduate degree} \end{cases}$$

Can't (or at least shouldn't) model as $Y = \beta^T X + \text{noise}$

Instead, think of 'latent' continuous response Z

$Z \sim N(\beta^T X, 1^2)$

For $k = 1, \ldots, K$:

$Y = k \leftrightarrow g_{k-1} < Z < g_k$.

Likelihood methods require $Pr(Y = k|X)$ as a function of $(\beta, g) = (\beta_1, \ldots, \beta_p, g_1, \ldots, g_{K-1})$.

Bayesian analysis implemented via Gibbs sampling avoids this explicitly

Regarding $x_{1:n}$ as fixed, joint density of everything looks like...

$\prod_{i=1}^{n} \{p(y_i|z_i, g)p(z_i|\beta)\}p(\beta)p(g)$

Nice full conditionals

What sort of inferences to report?

On β directly, so are infering the relationship between X and Z

E.g. text ex., with Y=DEG, X=(CHILD, PDEG, CHILD×PDEG).
Interesting that $\hat{\beta}_1 < 0$, $\hat{\beta}_1 + \hat{\beta}_3 > 0$.

Or could try to relate more directly to the (X, Y) relationship...

Alternatively, can bypass g

A priori

$$p(z, \beta) = \left\{ \prod_{i=1}^{n} p(z_i|\beta) \right\} p(\beta)$$

Equate observing $y = y_{1:n}$ with partial knowledge of $z = z_{1:n}$, expressed as $z \in R(y)$, i.e., (z, y) relationship must be monotone.

So *a posteriori*

$$p(z, \beta | z \in R(y)) = \left\{ \prod_{i=1}^{n} p(z_i|\beta) \right\} p(\beta) I\{z \in R(y)\}$$

Ammenable full conditionals again

Pros and cons

- Rank likelihood method frees one from choosing a prior on g
- Rank likelihood method simplifies Gibbs sampling (no update to g)
- Rank likelihood method only leads to inference on (X, Z) relationship, can't extend to (X, Y) relationship.

Text ex: virtually the same inferences on $\beta = (\beta_1, \beta_2, \beta_3)$ from either approach

These ideas extend from regression to multivariate analysis

$Y_{i,j}$ is the i-th observation on the j-th variable.

$Z_{i,j}$ is the corresponding latent variable with standard normal distribution.

$Z_{i,1:p} \overset{iid}{\sim} N_p(0, \Psi)$

$Y_{i,j} = g_j(Z_{i,j})$

Again the rank likelihood method allows inference about Ψ without explicit modelling of $g_j()$, $j = 1, \ldots, p$.

So can infer dependence amongst the p variables whilst ignorning the scale on which each variable lives...