STA 360/602L: Module 7.2

Metropolis in action

Dr. Olanrewaju Michael Akande

Count data

- We will use the Metropolis sampler on count data with predictors, so let's first do some general review.
- Suppose you have count data as your response variable.
- For example, we may want to explain the number of c-sections carried out in hospitals using potential predictors such as hospital type, (that is, private vs public), location, size of the hospital, etc.
- The models we have covered so far are not (completely) adequate for count data with predictors.
- Of course there are instances where linear regression, with some transformations (especially taking logs) on the response variable, might still work reasonably well for count data.
- That's not the focus here, so we won't cover that.

Poisson regression

- As we have seen so far, a good distribution for modeling count data with no limit on the total number of counts is the Poisson distribution.

- As a reminder, the Poisson pmf is given by

$$\Pr[Y = y|\lambda] = \frac{\lambda^y e^{-\lambda}}{y!}; \quad y = 0, 1, 2, \ldots; \quad \lambda > 0.$$

- Remember that

$$\mathbb{E}[Y = y] = \mathbb{V}[Y = y] = \lambda.$$

- When our data fails this assumption, we may have what is known as over-dispersion and may want to consider the Negative Binomial distribution instead (actually easy to fit within the Bayesian framework!).

- With predictors, index λ with i, so that each λ_i is a function of $\boldsymbol{X}$. Therefore, the random component of the glm is

$$p(y_i|\lambda_i) = \text{Poisson}(\lambda_i); \quad i = 1, \ldots, n.$$

Poisson regression

- We must ensure that $\lambda_i > 0$ at any value of $\boldsymbol{X}$, therefore, we need a link function that enforces this. A natural choice is

$$\log(\lambda_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}.$$

- Combining these pieces give us our full mathematical representation for the Poisson regression.

- Clearly, λ_i has a natural interpretation as the "expected count", and

$$\lambda_i = e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}}$$

  so the e^{β_j}'s are multiplicative effects on the expected counts.

- For the frequentist version, in R, use the `glm` command but set the option `family = “poisson”`.

Analysis of horseshoe crabs

- We have data from a study of nesting horseshoe crabs (J. Brockmann, Ethology, 102: 1–21, 1996). The data has been discussed in Agresti (2002).
- Each female horseshoe crab in the study had a male crab attached to her in her nest.
- The study investigated factors that affect whether the female crab had any other males, called satellites, residing nearby her.
- The response outcome for each female crab is her number of satellites.
- We have several factors (including the female crab's color, spine condition, weight, and carapace width) which may influence the presence/absence of satellite males.
- The data is called `hcrabs` in the R package `rsq`.

Analysis of horseshoe crabs

- Let's fit the Poisson regression model to the data. In vector form, we have

$$y_i \sim \text{Poisson}(\lambda_i); \quad i = 1, \ldots, n;$$

$$\log[\lambda_i] = \boldsymbol{\beta}^T \boldsymbol{x}_i$$

where y_i is the number of satellites for female crab i, and $\boldsymbol{x}_i$ contains the intercept and female crab i's

  - color;
  - spine condition;
  - weight; and
  - carapace width.

- Suppose we specify a normal prior for $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \ldots, \beta_{p-1})$, $\pi(\boldsymbol{\beta}) = \mathcal{N}_p(\boldsymbol{\beta}_0, \Sigma_0)$.

- Can you write down the posterior for β? Can you sample directly from it?

Analysis of horseshoe crabs

- We can use Metropolis to generate samples from the posterior.
- First, we need a "symmetric" proposal density $\boldsymbol{\beta}^{\star} \sim g[\boldsymbol{\beta}^{\star}|\boldsymbol{\beta}^{(s)}]$; a reasonable choice is usually a multivariate normal centered on $\boldsymbol{\beta}^{(s)}$.
- What about the variance of the proposal density? We can use the variance of the ols estimate, that is, $\hat{\sigma}^2 \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1}$, which we can scale using δ, to tune the acceptance ratio.
- Here, $\hat{\sigma}^2$ is calculated as the sample variance of $\log[y_i + c]$, for some small constant c, to avoid problems when $y_i = 0$.
- So we have $g[\boldsymbol{\beta}^{\star}|\boldsymbol{\beta}^{(s)}] = \mathcal{N}_p \left(\boldsymbol{\beta}^{(s)}, \delta\hat{\sigma}^2 \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1}\right)$.
- Finally, since we do not have any information apriori about $\boldsymbol{\beta}$, let's set the prior for it to be $\pi(\boldsymbol{\beta}) = \mathcal{N}_p(\boldsymbol{\beta}_0 = \mathbf{0}, \Sigma_0 = \boldsymbol{I})$.

Analysis of horseshoe crabs

- The Metropolis algorithm for this model is:

  1. Given a current $\boldsymbol{\beta}^{(s)}$, generate a *candidate* value $\boldsymbol{\beta}^\star \sim g[\boldsymbol{\beta}^\star | \boldsymbol{\beta}^{(s)}] = \mathcal{N}_p\left(\boldsymbol{\beta}^{(s)}, \delta\hat{\sigma}^2 \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\right)$.

  2. Compute the acceptance ratio

$$
\begin{aligned}
r &= \frac{\pi(\boldsymbol{\beta}^\star | Y)}{\pi(\boldsymbol{\beta}^{(s)} | Y)} = \frac{\pi(\boldsymbol{\beta}^\star) \cdot p(Y | \boldsymbol{\beta}^\star)}{\pi(\boldsymbol{\beta}^{(s)}) \cdot p(Y | \boldsymbol{\beta}^{(s)})} \\
&= \frac{\mathcal{N}_p(\boldsymbol{\beta}^\star | \boldsymbol{\beta}_0 = \mathbf{0}, \Sigma_0 = \boldsymbol{I}) \cdot \prod_{i=1}^{n} \text{Poisson}\left(Y_i | \lambda_i = \exp\left\{(\boldsymbol{\beta}^\star)^T \boldsymbol{x}_i\right\}\right)}{\mathcal{N}_p(\boldsymbol{\beta}^{(s)} | \boldsymbol{\beta}_0 = \mathbf{0}, \Sigma_0 = \boldsymbol{I}) \cdot \prod_{i=1}^{n} \text{Poisson}\left(Y_i | \lambda_i = \exp\left\{\left(\boldsymbol{\beta}^{(s)}\right)^T \boldsymbol{x}_i\right\}\right)}.
\end{aligned}
$$

  3. Sample $u \sim U(0,1)$ and set

$$
\boldsymbol{\beta}^{(s+1)} = \begin{cases} \boldsymbol{\beta}^\star & \text{if} \quad u < r \\ \boldsymbol{\beta}^{(s)} & \text{if} \quad \text{otherwise} \end{cases}.
$$

Move to the R script here.

What's next?

Move on to the readings for the next module!