Ex: Sparrows

Y: # fledglings (offspring)

X: ages

idea: The # of Fledglings a sparrow has in a mating season depends on the age of the sparrow.

We could model this count data Y with:

$\boxed{Y|x \sim \text{Poisson}(\theta_x)}$ since support Poisson r.v. Y is $\{0,1,2,\dots\}$

Here, $\mathbb{E}Y|X = \theta_x$ is age-specific.

Possible ways to model θ_x:

(1) discrete: $\theta_1, \theta_2, \dots, \theta_6$
a specific θ_i for each age i of sparrow.

* Note: if we don't collect much data on certain age sparrows then our estimates for some θ_i will be poor.

(2) $\theta_x = f(x) = \underbrace{\beta_1 + \beta_2 x + \beta_3 x^2}_{\text{equation of a hyperplane}}$

* Note θ_x should not be negative yet it can be under model (2).

(3) soln: log-transform.

$$\boxed{\log \mathbb{E}Y|X = \log \theta_{x_i} = \beta_1 + \beta_2 x_i + \beta_3 x_i^2}$$

i.e. $\mathbb{E}Y|X = \exp\{\beta_1 + \beta_2 x + \beta_3 x^2\} > 0$

Terminology

$Y|x \sim \text{Poisson}(\exp\{\beta^T \vec{x}\})$ "Poisson regression"

$\beta^T x$ "linear predictor"

$\mathbb{E}\,Y|X$ is <u>linked</u> to the linear predictor w/ the <u>log</u> function

So we say this model has a <u>log-link.</u>

In general $f(\mathbb{E}\,Y|X) = \beta^T x$ is a <u>generalized linear model</u> (GLM) w/ link function f.

A note on $\beta^T \vec{x}$

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} \quad \text{Dimension } 3 \times 1$$

$$\vec{x}_i = \begin{bmatrix} 1 \\ x_i \\ x_i^2 \end{bmatrix} \quad \text{Dimension } 3 \times 1 \quad \text{i.e. corresponds to 1 sparrow.}$$

So for any individual sparrow,

$$\mathbb{E}\,y_i|x_i = \exp\{\underset{1\times3}{\beta^T}\underset{3\times1}{x_i}\} = \exp\{\beta_1 \cdot 1 + \beta_2 x_i + \beta_3 x_i^2\}$$

Recap:

we have a data generative model:

$$Y|x \sim \text{Poisson}(\exp\{\beta^T x\})$$

and if we know a bird's age and if we knew β, we could predict y.

The trouble is, we don't know β.

$\beta_1, \beta_2, \beta_3$ are unknown.

We want to estimate them from some data.

We want $p(\beta|y, x)$.

We need to specify priors on β.

Common prior:

$\beta \sim MVN(0, \Sigma)$

Ex What does $p(\beta | y, X)$ look like?

Sol

$$p(\beta | y, x) \propto p(\beta) \prod_{i=1}^{n} p(y_i | \beta, x_i)$$

$$\propto \underbrace{\exp\{-\tfrac{1}{2}\beta^T \Sigma \beta\}}_{\text{normal}} \underbrace{\prod_{i=1}^{n} \exp\{(\beta^T x_i) \cdot y_i\} \cdot \exp\{-\exp\{\beta^T x_i\}\}}_{\underbrace{\exp\{\sum_{i=1}^{n}[\beta^T x_i y_i - \exp\{\beta^T x_i\}]\}}_{?}}$$

Does not look like a known kernel.

NOT conjugate

NOT semi-conjugate

NOT easy to sample from using known methods.

But still, we want to sample from it because

$$p(\beta | y, X) = \frac{p(\beta)\, p(y | X, \beta)}{\underbrace{\iiint p(\beta)\, p(y | X, \beta)\, d\beta_1\, d\beta_2\, d\beta_3}_{\text{"}p(y)\text{"}}}$$

is a nasty integral to numerically approx in higher dimensions.

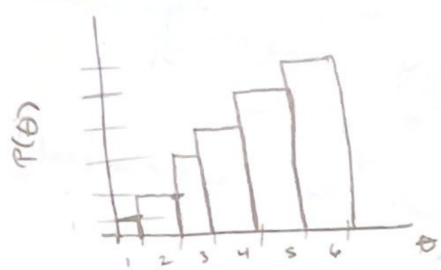
Goal: construct a Markov chain that approximates the posterior. We need to avoid computing $p(y)$.

Metropolis algo.

Intuition: weighted die ex.

Let $p(\theta = i) = i/w$ for $i \in \{1, \dots, 6\}$

so the pmf of θ looks like:



We want to sample from $p(\theta)$.

Start @ $\theta = 3$

propose $\theta^* = 2$

We need more $\theta = 3$ than $\theta = 2$ in our sample, if our sample is to approximate $p(\theta)$ well.

How many more?

$$\frac{p(\theta^*)}{p(\theta)} = \frac{2/w}{3/w} = 2/3.$$

I need 2/3 as many samples of $\theta = 2$ in my Markov chain.

So accept new state $\theta^* = 2$ w/ probability 2/3.

Say, I propose $\theta^* = 4$. Current state is $\theta = 3$

$$\frac{p(\theta^*)}{p(\theta)} = \frac{4/w}{3/w} > 1.$$

So the state $\theta = 4$ is more probable. We should accept it as the next state in our Markov chain.

Eventually we will have:

$$\begin{array}{c} \theta \\ \begin{bmatrix} 3 \\ 2 \\ 3 \\ 4 \\ 4 \\ 5 \\ 4 \\ 5 \\ 6 \\ 5 \\ \vdots \end{bmatrix} \end{array}$$

that will approximate the target distr. well.

$\boxed{\text{* Note we do not need to know } w \text{ to implement this.}}$

Metropolis Algorithm

1. Sample $\theta^* | \theta^{(s)} \sim J(\theta | \theta^{(s)})$
2. Compute acceptance ratio

$$r = \frac{p(\theta^* | y)}{p(\theta | y)}$$

3. Let $\theta^{(s+1)} = \begin{cases} \theta^* & \text{w/ prob } \min(r,1) \\ \theta^{(s)} & \text{w/ prob } 1-\min(r,1) \end{cases}$

$J(\theta | \theta^{(s)})$ is the proposal distribution. It proposes a new value θ given our current $\theta^{(s)}$.

For this to be the "Metropolis algo.", J is symmetric.
i.e. $J(\theta_a | \theta_b) = J(\theta_b | \theta_a)$.

Practice

(1) implement die example, w/ $J(\theta = j | \theta^{(s)} = i) = 1/6$ for all j.
i.e. propose a new state j uniformly.