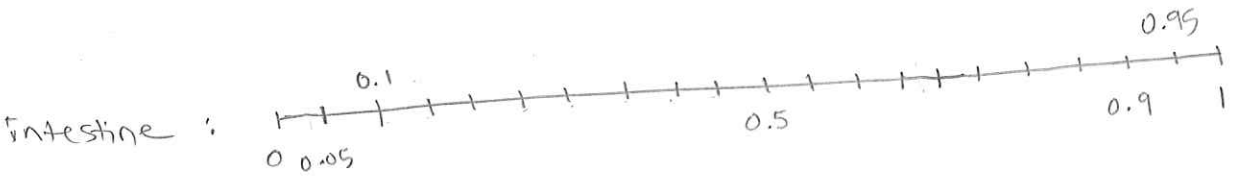
Bayesian hierarchical regression modeling

Data example: Haigis et al. (2004) (& see Hoff ch. 11)

- Load data & take a look from script on website.
- Data look like this:

	location 1	location 2	...	location 20
mouse 1	1	0	...	3
mouse 2	0	1	...	1
⋮	⋮			
mouse 21	1	...		

- entries of the matrix, $Y[j,k]$, are the tumor counts for mouse j at location k.
- There are 21 mice & 20 locations.
- Scientists examine tumor counts at 20 equally spaced regions of the intestine of 21 mice:



- Question of interest: how does position on intestine relate to tumor count?

In other words, can we understand spatial variability in tumor count?

Rephrased again (in a slightly more specific & statistically suggestive way):

- what's the expected # of tumors at location k?

Task: Let's write down a **model** to describe how the data could plausibly have been generated.

Notation: let Y_{jk} be the number of tumors mouse j has at position k.

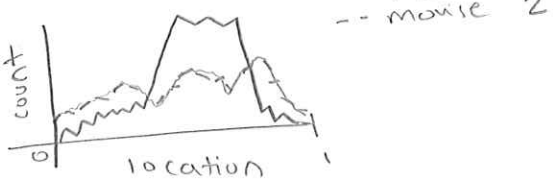
We are interested in modeling $\mathbb{E}[Y_{jk} \mid k]$.

Since Y is a count, a **Poisson regression model** is sensible:

$$Y_{jk} \mid \lambda_j(k) \sim \text{Poisson}(\lambda_j(k)) \; ; \; \mathbb{E}[Y_{jk} \mid k] = \lambda_j(k)$$

Q: **Why** have a specific mean for each mouse? (suggested by the j index I have attached to "λ" above).

A: **EDA**: examine the plot of tumor count vs location for several mice.

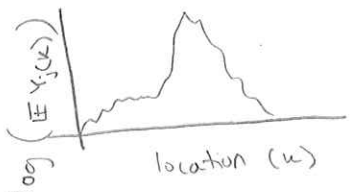


* Some curves almost entirely above / below mean curve.

Q: how to specify $\lambda_j(k)$?

A: Remember, the rate (mean) of a Poisson model must be positive, $\lambda_j(k) > 0$.

Idea: write $\log(\lambda_j(k))$ as a **polynomial** in location k.



Q: how to choose polynomial degree?

A: Compare fit of different degrees, e.g. R code for polynomial regression degree 4 looks good (empirically).

Putting it all together:

$$Y_{j,k} \mid \lambda_j(k) \sim \text{Poisson}(\lambda_j(k))$$

$$\log(\lambda_j(k)) = \beta_{0j} + \beta_{1j}k + \beta_{2j}k^2 + \beta_{3j}k^3 + \beta_{4j}k^4 \quad \left.\right\} \text{Model}$$

Let $\underline{\beta}_j = [\beta_{0j}, \ldots, \beta_{4j}]^T \in \mathbb{R}^5 \Rightarrow$ 5 unknown parameters per mouse!

$$\underline{\beta}_j \mid \underline{\theta}, \underline{\underline{\Sigma}} \overset{iid}{\sim} \text{MVN}(\underline{\theta}, \underline{\underline{\Sigma}})$$

$$\underline{\theta} \sim \text{MVN}(\underline{\mu}, \underline{\underline{\Lambda}}_0)$$

$$\underline{\underline{\Sigma}} \sim \text{inv-Wishart}(\eta_0, \underline{\underline{S}}_0)$$

} between "group" sampling model.

} priors

Q: how to choose $\underline{\mu}, \underline{\underline{\Lambda}}_0, \eta_0, \underline{\underline{S}}_0$?

A: one option: unit information

Quiz: Describe in words / pseudo-code an MCMC procedure to sample all unknowns.

Soln:

Initialize $\underline{\beta}^{(0)}, \underline{\theta}^{(0)}$

for (s in 1:S) {

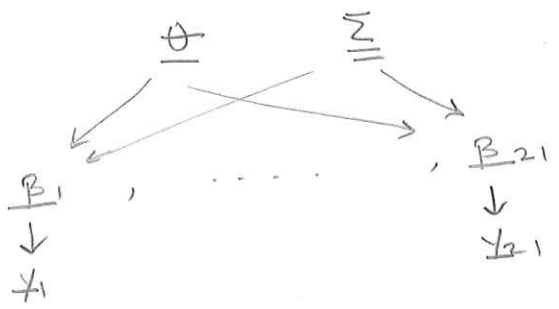
- sample $\underline{\underline{\Sigma}}^{(s+1)} \sim$ inv Wishart(,) ← parameters of full condl post: $p(\underline{\underline{\Sigma}} \mid \cdot)$ — Gibbs update
- sample $\underline{\theta}^{(s+1)} \sim$ MVN(,) ← parameters of full condl: $p(\underline{\theta} \mid \cdot)$ — Gibbs update
- propose $\underline{\beta}^* \sim \text{MVN}(\underline{\beta}^{(s)}, S_0)$, compute $r = \frac{p(\underline{\beta}^* \mid \cdot)}{p(\underline{\beta}^{(s)} \mid \cdot)}$ — Metropolis update

accept / reject w/ prob $\min(1, r)$.

}

Q: Why can't we Gibbs update $\underline{\beta}$?

A: Can't compute full cond'l in closed form. Look:



Let $\underline{\beta} = \begin{bmatrix} | & | & & | \\ \underline{\beta}_1 & \underline{\beta}_2 & \cdots & \underline{\beta}_{21} \\ | & | & & | \end{bmatrix}$

$$p(\underline{\beta}, \underline{\theta}, \underline{\Sigma} \mid \underline{y}) \propto p(\underline{y} \mid \underline{\beta}) \cdot p(\underline{\beta} \mid \underline{\theta}, \underline{\Sigma})\, p(\underline{\theta})\, p(\underline{\Sigma})$$

$$\Rightarrow p(\underline{\beta}_j \mid \cdot) \propto p(\underline{y}_j \mid \underline{\beta}_j)\, p(\underline{\beta}_j \mid \underline{\theta}, \underline{\Sigma})$$

Poisson: $\prod_{k=1}^{20} \frac{\left(e^{\underline{x}^T \underline{\beta}_j}\right)^{y_{jk}} e^{-e^{\underline{x}^T \underline{\beta}_j}}}{y_{jk}!}$

normal: $e^{-\frac{1}{2}(\underline{\beta}_j - \underline{\theta})^T \underline{\Sigma} (\underline{\beta}_j - \underline{\theta})}$

here, $\underline{x} = \begin{bmatrix} 1 \\ k \\ k^2 \\ k^3 \\ k^4 \end{bmatrix}$ identical $\forall\, j$.

no way to combine, complete the square, etc.

$\Rightarrow$ can't identify full cond'l distr.

$\Rightarrow$ can't Gibbs sample.

$\Rightarrow$ Metropolis algo. or some other M-H proposal.

Final thoughts:

* We can change the efficiency of our sampler (the speed we move through parameter space) by transforming our parameters.
  - Common method: "non-central parameterization"

  Notice: $\underline{\beta}_j = \underline{\theta} + \underline{\gamma}_j$, $\underline{\gamma}_j \sim MVN(\underline{0}, \underline{\Sigma})$
  - can decorrelate posterior to sample from $\underline{\theta}, \underline{\Sigma}, \underline{\gamma}$ space instead.

Final thoughts continued.

* hierarchical models are useful when
  - small samples in some groups.
    → fixes issues w/ partial pooling (shrinkage)
  - want to understand btwn group variability (e.g. when we want to make predictions about new unobserved groups, i.e. mice in this example)

The major cost is that it can be much more computationally expensive to make inferences under such a model.

* We can map back to classic GLM w/ random effects framework (i.e. view a random-effect model as a hierarchical model):

If we parameterize $\underline{\beta}_j = \underline{\theta} + \underline{\gamma}_j$ & $\underline{\gamma}_j \sim MVN(0, \Sigma)$

then

$$\underline{X}^T \underline{\beta}_j = \underline{X}^T (\underline{\theta} + \underline{\gamma}_j) = \theta_0 + \gamma_{0j} + \theta_1 x_1 + \gamma_{1j} x_1 + \cdots + \theta_4 x_4 + \gamma_{4j} x_4$$

The $\underline{\theta}$ are called "fixed effects"

$\underline{\gamma}_j$ are called "random effects"

The general syntax you might see is:

$$Y_{ij} = \underline{\theta}^T \underline{X}_{ij} + \underline{\gamma}_j^T \underline{Z}_{ij} + \varepsilon_{ij}$$

- Y_{ij}: where "ij" here is ith obs from jth group
- $\underline{\theta}$: fixed effect
- $\underline{\gamma}_j$: random effect
- $\underline{X}$: subset of predictors corresponding to fixed effects
- $\underline{Z}$: subset of predictors corresponding to random effects

($\underline{X}$ & $\underline{Z}$ can be the same)