# ADVANCED BAYESIAN MODELING

# Loglinear Regression Modeling

# Data Model

$$X_i = (x_{i1}, \ldots, x_{ik}), \quad y_i \qquad\qquad i = 1, \ldots, n$$

$X$ has $k$ explanatory variables (usually including an intercept)

$$y_i \mid \beta, X_i \quad \sim \quad \text{indep. Poisson}(\lambda_i)$$

$$\log \lambda_i \;=\; X_i\beta \qquad\qquad \lambda_i \;=\; e^{X_i\beta}$$

This is a **(Poisson) loglinear regression model**.

Poisson models are appropriate for responses that are counts with no obvious maximum value, especially counts of rare, unrelated events, such as:

- Cases of a rare genetic disease
- Shark attacks on swimmers
- Numbers of lottery winners

If all counts are large enough (hundreds or more), a transformed normal-theory model might be a good approximation.

If some counts are small (especially, if some are zero), normal-theory models are not appropriate.

# Priors

Conjugate (gamma-type) priors exist for special cases, but are not very useful in general.

Flat priors on coefficients (in $\beta$) may be safe, provided data are not too "sparse" (too many zeros). Otherwise, may need to make priors informative.

We consider normal priors, which may be diffuse (large variance).

As in ordinary regression, standardizing quantitative explanatory variables is recommended.

# Overdispersion

Mean and variance of Poisson are equal:

$$\mathrm{E}(y_i \mid \beta, X_i) = \lambda_i \qquad \mathrm{var}(y_i \mid \beta, X_i) = \lambda_i$$

Common problem: Data exhibit variances that do not equal the means – typically larger than they should be: **overdispersion**.

Possible causes:

- Explanatory variables are missing
- Events being counted are dependent
- Events of different rates are combined into a single count

The **chi-square discrepancy** measures overdispersion:

$$T(y, \theta, X) \;=\; \sum_{i=1}^{n} \frac{(y_i - \lambda_i)^2}{\lambda_i}$$

(When the model is correct and the $\lambda_i$s are large enough, it has an approximate $\chi^2_n$ distribution.)

When there is overdispersion, it tends to be larger.

Can be used to compute a posterior predictive $p$-value for overdispersion:

$$\Pr\big(T(y^{\mathsf{rep}}, \theta, X) \;\geq\; T(y, \theta, X) \mid y\big)$$

One possible way to account for overdispersion:

Add a separate random effect to the linear combination of each observation.

$$\log \lambda_i \;=\; X_i\beta \,+\, \epsilon_i \qquad\qquad \epsilon_i \;\sim\; \text{iid } \mathrm{N}(0, \sigma_\epsilon^2)$$

Then add a reasonably noninformative prior on $\sigma_\epsilon^2$ (e.g. flat on $\sigma_\epsilon$).

(See BDA3, Sec. 16.4, for example.)

Alternative: Use a *negative binomial* response distribution, which naturally allows for overdispersion.

(BDA3, Sec. 17.2)