ADVANCED BAYESIAN MODELING

Flint Data Example: Model

Flint Water Crisis

Citizen data set (from `http://flintwaterstudy.org`):

- 271 observations (tap water sampling kits)
- Three lead readings (ppb) for each observation:
  - First draw
  - After flushing 45 seconds
  - After flushing 2 minutes

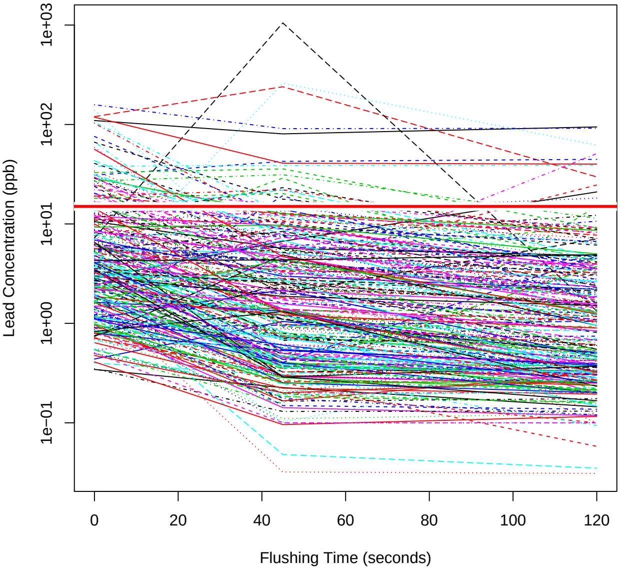
Regard each observation as a separate "household."

Federal Lead and Copper Rule: Action required if lead level of 15 ppb is exceeded in more than 10% of homes.

For each household, profile of measured lead level (ppb) versus flushing time.

Note log scale on vertical axis.

Solid horizontal red line marks 15 ppb level.



Note:

- Mean lead levels seem to decrease as flushing time increases.
- Substantial fraction of households exceed 15 ppb at first draw.
- Households seem to vary in overall lead level.

Sampling Model

$$y_{ij} \quad = \quad \text{logarithm of lead level (ppb) in household } i \text{ at draw } j$$

Regard households as randomly sampled – their overall levels will be random effects.

Thus, the "group" index is i (not j).

Allow each draw to have its own mean.

$$
\begin{aligned}
y_{ij} \mid \beta^d, \beta^h, \sigma_y^2 \quad &\sim \quad \text{indep. N}\big(\beta_j^d + \beta_i^h,\ \sigma_y^2\big) \\
\beta_i^h \mid \sigma_h^2 \quad &\sim \quad \text{iid N}(0,\ \sigma_h^2)
\end{aligned}
$$

β_i^h = household i effect (random) β_j^d = draw j effect (fixed)

σ_h^2 = variance between households

σ_y^2 = measurement variance (within households)

This is a mixed model: Matrix X_f would have indicators for the three draws, and matrix X_r would have indicators for households.

Note: Because draw effects β_j^d are fixed and unrestricted, there is an implicit intercept – no need for an explicit one.

Based on earlier figure, we expect mean lead levels to fall as flushing time increases:

$$\beta_1^d \; > \; \beta_2^d \; > \; \beta_3^d$$

We also expect substantial variation among households:

$$\text{within-house correlation} \quad \frac{\sigma_h^2}{\sigma_h^2 + \sigma_y^2} \quad \text{will be high}$$

Priors

We would like to be noninformative:

$$\begin{aligned}
p(\sigma_y^2) &\propto (\sigma_y^2)^{-1} \qquad \sigma_y^2 > 0 \\
p(\sigma_h^2) &\propto (\sigma_h^2)^{-1/2} \qquad \sigma_h^2 > 0 \\
p(\beta_1^d, \beta_2^d, \beta_3^d) &\propto 1
\end{aligned}$$

By formal transformation of variables, the σ_h^2 prior is equivalent to

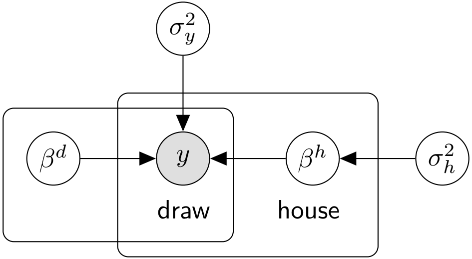
$$\sigma_h \sim \text{flat on } (0, \infty)$$

To use JAGS, we must replace these with proper (but diffuse) priors:

$$\begin{aligned}
\sigma_y^2 &\sim \text{Inv-gamma}(0.0001, 0.0001) \\
\sigma_h &\sim \text{U}(0, 1000) \\
\beta_1^d, \beta_2^d, \beta_3^d &\sim \text{iid N}(0, 10000)
\end{aligned}$$

(Diffuse enough? Probably, based on figure, but could change if needed.)

DAG Model



Note: Plates can overlap when there is multiple indexing.

Generalizations

Could extend model to allow for:

- Measurement variances σ_y^2 that vary by household

  (These would need a distribution with hyperparameters.)

- Non-constant correlations within households

  (For example, the first draw measurement from a household might be more correlated with the second than with the third.)

For simplicity, we skip these generalizations – may not be needed anyway.