# Bayesian SAE using Complex Survey Data
# Lecture 3A: Hierarchical Bayes Modeling

**Jon Wakefield**

Departments of Statistics and Biostatistics
University of Washington

# Course Content

Motivation

Non-Spatial Hierarchical Model for Normal Data

Non-Spatial Hierarchical Model for Binomial Data

Discussion

Technical Appendix: Computation for Beta Binomial Model

Technical Appendix: Prior Choice for Binomial GLMM

Technical Appendix: Model Comparison

# Motivation

In a SAE context, we are often faced with situations in which the data are sparse in space, which leads to great uncertainty in calculated estimates.

Hierarchical models[1] are designed to alleviate this problem, by modeling the totality of data from all areas, in order to leverage similarities in the data.

The key element is coupling the different areas, by assuming this parameters in these areas are linked through a common probability distribution.

In this lecture we describe hierarchical models for normal and binomial data.

[1] also known as random or mixed effects models

# Simulated Data with Constant Risk Across Areas

Suppose we are collecting samples in King County HRAs to see if the prevalence, $p$, of some condition is dangerously high in some areas.

We simulate data (via simple random sampling) with $p = 0.2$ in every area (so there is no between-area variability in the true prevalence).

We take sample sizes of $n_i = 10, 25, 50, 200$ in each area and simulate data from the model

$$Y_i | p = 0.2 \sim \text{Binomial}(n_i, p),$$

for $i = 1, \dots, 48$ areas.

# MLEs of Prevalences $p_i$, $i = 1, \ldots, 48$

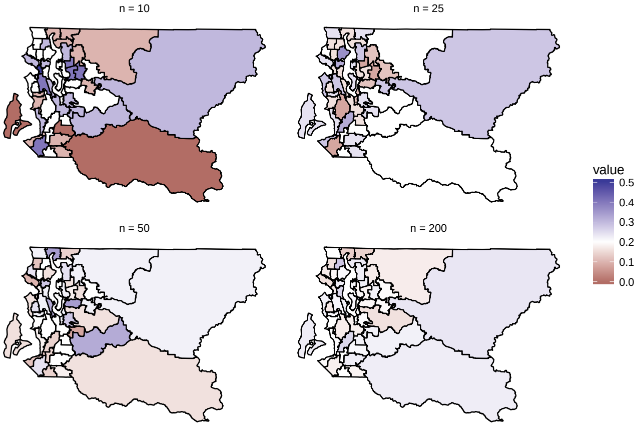

Figure 1: Realized proportions (MLEs) in the 48 areas, with different sample sizes.

For the low sample sizes we see lots of areas that suggest intervention is required, but this is just sampling variability.

# Motivating Examples: Normal and Binomial Data

We now return to the simulated data we saw in the first lecture.

We have two outcomes, one continuous and one binary, again using the King County geography.

These data were simulated with non-constant risk across HRAs (areas).

# Motivating Examples: Normal and Binomial Data

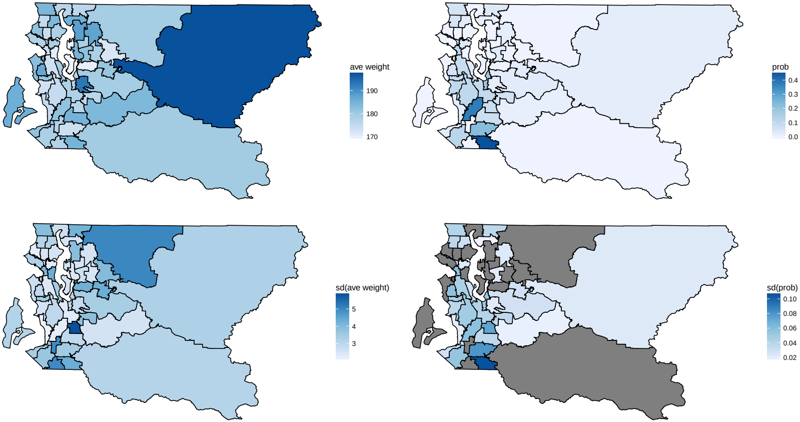

Figure 2: Sample mean weights (top left) and fractions with diabetes (top right) Standard errors of: mean weights (bottom left) and fractions with diabetes (bottom right). Gray areas in the right map correspond to areas with zero counts, and hence an estimated standard error of zero.

# Non-Spatial Hierarchical Model for Normal Data

# Smoothing Models

Instability of estimates has lead to methods being developed to impose smoothness on the underlying parameters using hierarchical/random effects models that use the data from the totality of areas to provide more reliable estimates in each of the constituent areas.

Overview of Models:

- Basic Normal Model: No smoothing.
- Random Effects Models:
  - Normal likelihood. Random effects with no spatial structure (known as IID[2]).
  - Normal Likeilhood, Two sets of random effects, one set with no spatial structure, one set with spatial structure.
- Covariates may be added to each of these in order to smooth over covariate space.
- Estimation in these models is a separate issue.

[2] Independent and Identically Distributed

# Basic Hierarchical Model

Let $Y_{ik}$ be the weight of the $k$-th sampled individual in area $i$.

Three possible models:

1. No between-area variability:

$$Y_{ik} = \underbrace{\beta_0}_{\text{Common Mean}} + \epsilon_{ik},$$

   with $\epsilon_{ik} \sim \mathrm{N}(0, \sigma_\epsilon^2)$.

2. Distinct between-area variability:

$$Y_{ik} = \underbrace{\beta_i}_{\text{Mean of Area } i} + \epsilon_{ik},$$

   with $\epsilon_{ik} \sim \mathrm{N}(0, \sigma_\epsilon^2)$. Known as a fixed effects model. Note: no way to link different areas.

3. Linked between-area variability:

$$Y_{ik} = \underbrace{\beta_0 + \delta_i}_{\text{Mean of Area } i} + \epsilon_{ik},$$

   with $\delta_i \sim \mathrm{N}(0, \sigma_\delta^2)$, $\epsilon_{ik} \sim \mathrm{N}(0, \sigma_\epsilon^2)$. Known as a random effects model.

# Basic Normal Hierarchical Model

We will concentrate on the linked between-area random effects model:

$$Y_{ik} = \underbrace{\beta_0 + \delta_i}_{\text{Mean of Area } i} + \epsilon_{ik},$$

with $\delta_i \sim \mathrm{N}(0, \sigma^2_\delta)$ – these are the area-specific deviations (the random effects) from the overall level $\beta_0$ – and $\epsilon_{ik} \sim \mathrm{N}(0, \sigma^2_\epsilon)$, is the measurement error.

This model is also known as a Linear Mixed Effects Model (LMEM).

In this model, the totality of data are used to inform on the overall level $\beta_0$ and between-area variability $\sigma^2_\delta$.

The unknown parameters are:

| | |
|---|---|
| Overall mean | $\beta_0$ |
| Between-Area Variance | $\sigma^2_\delta$ |
| Measurement Error Variance | $\sigma^2_\epsilon$ |
| Random Effects | $\delta_1, \ldots, \delta_n$ |

# The Linear Hierarchical Model

We can write the above model hierarchically as

$$
\begin{aligned}
Y_{ik}|\beta_0, \delta_i, \sigma^2_\epsilon &\sim_{iid} \mathrm{N}(\beta + \delta_i, \sigma^2_\epsilon) \\
\delta_i|\sigma^2_\delta &\sim_{iid} \mathrm{N}(0, \sigma^2_\delta) \\
\beta_0, \sigma^2_\epsilon, \sigma^2_\delta &\sim \text{Priors}
\end{aligned}
$$

The posterior, given data $\boldsymbol{y}$, is obtained as

$$
\begin{aligned}
p(\beta_0, \delta_1, \ldots, \delta_n, \sigma^2_\epsilon, \sigma^2_\delta|\boldsymbol{y}) &= p(\boldsymbol{y}|\beta_0, \delta_1, \ldots, \delta_n, \sigma^2_\epsilon, \sigma^2_\delta) \\
&\times p(\beta_0, \delta_1, \ldots, \delta_n, \sigma^2_\epsilon, \sigma^2_\delta)/p(\boldsymbol{y}) \\
&= \prod_{i=1}^{n} p(\boldsymbol{y}_i|\beta_0, \delta_i, \sigma^2_\epsilon) \times p(\delta_i|\sigma^2_\delta) \\
&\times p(\beta_0, \sigma^2_\epsilon, \sigma^2_\delta)/p(\boldsymbol{y})
\end{aligned}
$$

Marginal distributions, such as $p(\beta_0|\boldsymbol{y})$, are obtained by integration.

# Estimation in the Normal Hierarchical Model

In general, there are no closed-form (i.e., explicit) forms for the estimates of the parameters.

Suppose, for simplicity, the variances $\sigma_\delta^2$ and $\sigma_\epsilon^2$ are known; this allows some insight into inference.

The posterior mean for $\beta_0$ is a weighted least squares estimator, with the weights depending on the sample sizes in the areas; denote this by $\widehat{\beta}_0$.

# Estimation in the Normal Hierarchical Model

The posterior mean of the random effect (area-specific adjustment) is:

$$\begin{aligned}\widehat{\delta}_i &= \mathrm{E}[\delta_i|y_i] \\ &= \frac{n_i\sigma_\delta^2}{\sigma_\epsilon^2 + n_i\sigma_\delta^2}(\overline{y}_i - \widehat{\beta}_0) \\ &= w_i(\overline{y}_i - \widehat{\beta}_0)\end{aligned}$$

where

$$w_i = \frac{n_i\sigma_\delta^2}{\sigma_\epsilon^2 + n_i\sigma_\delta^2} \leq 1$$

and is small (so more shrinkage) if:

- $n_i$ is small (not much data in the area), or
- $\sigma_\delta^2$ is small (between-area variability is small), or
- $\sigma_\epsilon^2$ is large (within-area variability is large).

# Predicting the Population Total and Mean

Let $s_i$ and $r_i$ denote, respectively, the set of indices of the sampled and unsampled individuals in area $i$.

Let $T_i = \sum_{k=1}^{N_i} y_{ik}$ be the total for the population in area $i$, where $N_i$ is the population size.

The average for the population in area $i$ is

$$\begin{aligned}
\overline{Y}_i &= \frac{T_i}{N_i} \\
&= \frac{\sum_{k=1}^{N_i} y_{ik}}{N_i} \\
&= \frac{\sum_{k\in s_i} y_{ik} + \sum_{k\in r_i} y_{ik}}{N_i} \\
&= \underbrace{\frac{\sum_{k\in s_i} y_{ik}}{n_i}}_{\text{Mean of Sampled}} \times \frac{n_i}{N_i} + \underbrace{\frac{\sum_{k\in r_i} y_{ik}}{N_i - n_i}}_{\text{Mean of Unsampled}} \times \frac{N_i - n_i}{N_i}
\end{aligned}$$

# Predicting the Population Total and Mean

Suppose now we have fitted the linear mixed effects model and obtained posterior medians $\widehat{\beta}_0$ and $\widehat{\delta}_i$.

The obvious estimate is:

$$\widehat{\overline{Y}}_i = \underbrace{\frac{\sum_{k \in s_i} y_{ik}}{n_i}}_{\text{Mean of Sampled}} \times \frac{n_i}{N_i} + \underbrace{(\widehat{\beta}_0 + \widehat{\delta}_i)}_{\text{Estimated Mean}} \times \frac{N_i - n_i}{N_i}$$

If $N_i \gg n_i$, then the sampled data provide a small fraction of the total population in the area and we can estimate the area mean by

$$\widehat{\overline{Y}}_i = \widehat{\beta}_0 + \widehat{\delta}_i. \tag{1}$$

If an area contains no data, then its mean is assumed to be $\beta_0 + \delta^\star$, where $\delta^\star \sim \text{N}(0, \sigma_\delta^2)$.

# Motivating Example: Continuous Outcome

In the simulated data example we have very large populations, so we can neglect finite sampling correction factors (see later).

We can also neglect the observed sample mean in the estimate of the area mean, and use (1).

Totals can be similarly estimated:

$$\begin{aligned} \widehat{T}_i &= \sum_{k \in s_i} y_{ik} + \sum_{k \in r_i} y_{ik} \\ &= \underbrace{\sum_{k \in s_i} y_{ik}}_{\text{Total of Sampled}} + \underbrace{(N_i - n_i) \times (\widehat{\beta}_0 + \widehat{\delta}_i)}_{\text{Estimated Total for Unsampled}} . \end{aligned}$$

If $N_i \gg n_i$,

$$\widehat{T}_i \approx N_i \times (\widehat{\beta}_0 + \widehat{\delta}_i).$$

# MLEs of Prevalences $p_i$, $i = 1, \ldots, 48$

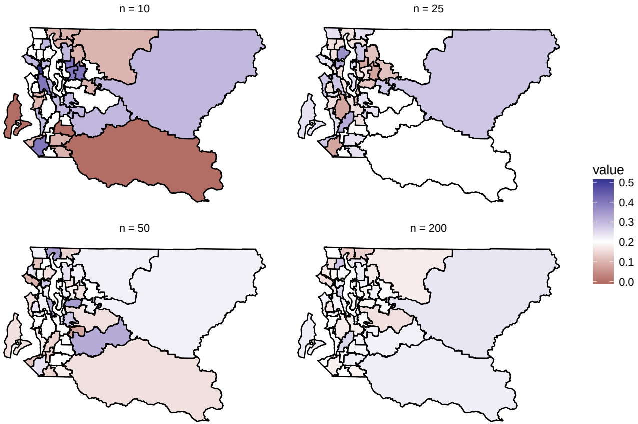

Figure 3: Realized proportions (MLEs) in the 48 areas, with different sample sizes.

For the low sample sizes we see lots of areas that suggest intervention is required, but this is just sampling variability.

# Posterior medians of Prevalences $p_i$, $i = 1, \ldots, 48$

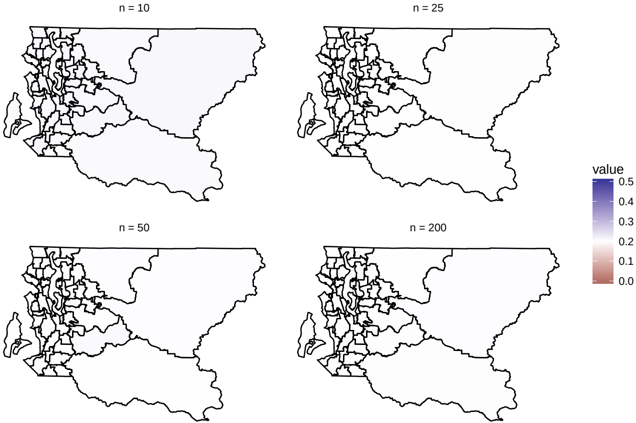

Figure 4: Posterior medians in the 48 areas, with different sample sizes.

We clearly see the effect of the shrinkage!

# Posterior Medians of Prevalences $p_i$, $i = 1, \ldots, 48$

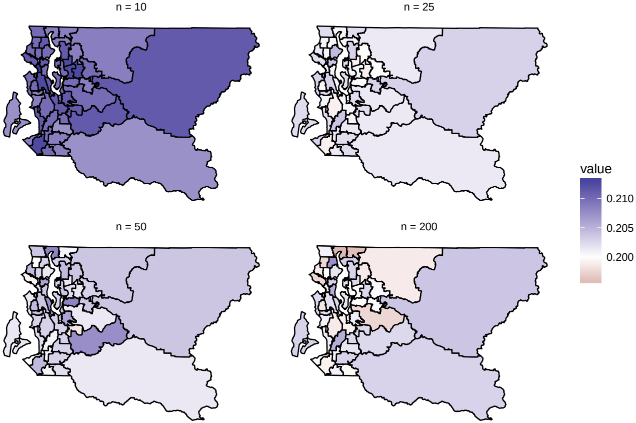

Figure 5: Posterior medians in the 48 areas, with different sample sizes, now on a different scale.

# Motivating Example: Continuous Outcome

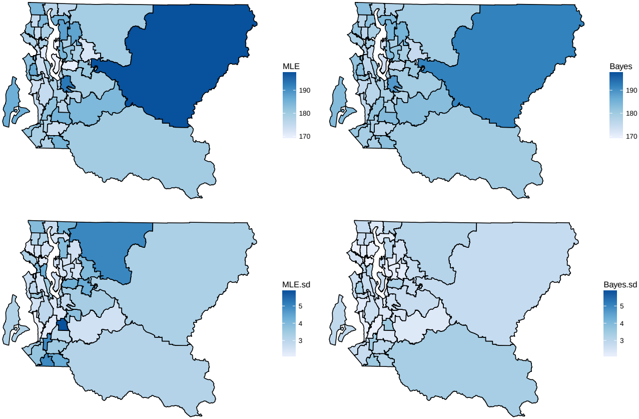

Figure 6: Top row: Estimates of area averages of weight via MLE's (left) and posterior medians (right). Bottom row: Uncertainty of estimates with standard errors (left) and posterior standard deviations (right).

# Motivating Example: Continuous Outcome

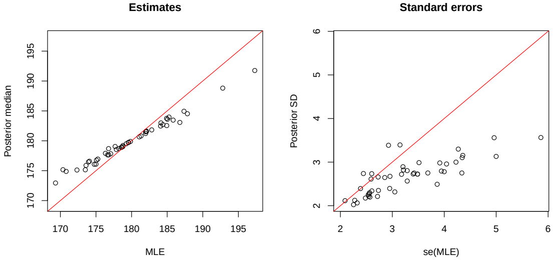

Figure 7: Comparison of area averages: Posterior medians versus MLEs (left). Posterior standard deviations versus standard errors associated with the MLEs (right).

The posterior medians are shrunk from the MLEs towards the overall mean, with the extreme values undergoing the most shrinkage.

In general, the Bayes measures of uncertainty (the posterior standard deviations) are smaller than the standard errors of the MLEs, with the greatest difference occurring for those areas with the large standard errors (which have the smallest sample sizes).

# Basic Normal Hierarchical Model

See (Rao and Molina, 2015, Section 4.3) for a description of the model

$$Y_{ik} = \underbrace{\beta_0 + \delta_i}_{\text{Mean of Area } i} + \epsilon_{ik}, \tag{2}$$

under the heading "Basic Unit Level Model" (also referred to as a nested error model).

The driving assumption is that this model is appropriate for the individuals (units) in area $i$, and in particular for those that were sampled $y_{ik}, k \in s_i$; specifically, there is no selection bias.

Let $\boldsymbol{x}_i = (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{iN_i})$ denote a set of covariates measured on all members of the population in area $i$.

Also let $\boldsymbol{r}_i = (r_{i1}, \ldots, r_{iN_i})$ denote response indicators, i.e., $r_{ik} = 1$ if $i \in s_i$ and $= 0$, otherwise.

# Basic Normal Hierarchical Model

Model (2) is assumed for the population and let $\boldsymbol{\theta} = (\beta_0, \boldsymbol{\delta}, \sigma^2_\delta, \sigma^2_\epsilon)$, with $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)$.

The sampling model for the observed data is,

$$\begin{aligned} p(\boldsymbol{y}_i^{\text{OBS}}, \boldsymbol{r}_i | \boldsymbol{x}_i, \boldsymbol{\theta}, \phi) &= \int p(\boldsymbol{y}_i^{\text{OBS}}, \boldsymbol{y}_i^{\text{MIS}}, \boldsymbol{r}_i | \boldsymbol{x}_i, \boldsymbol{\theta}, \phi)\, d\boldsymbol{y}_i^{\text{MIS}} \\ &= \int p(\boldsymbol{y}_i^{\text{OBS}}, \boldsymbol{y}_i^{\text{MIS}} | \boldsymbol{x}_i, \boldsymbol{\theta}) \times \underbrace{p(\boldsymbol{r}_i | \boldsymbol{y}_i^{\text{OBS}}, \boldsymbol{y}_i^{\text{MIS}}, \boldsymbol{x}_i, \phi)}_{\text{Selection Model}}\, d\boldsymbol{y}_i^{\text{MIS}} \end{aligned}$$

If we assume that selection does not depend on the data,

$$p(\boldsymbol{r}_i | \boldsymbol{y}_i^{\text{OBS}}, \boldsymbol{y}_i^{\text{MIS}}, \boldsymbol{x}_i, \phi) = p(\boldsymbol{r}_i | \boldsymbol{x}_i, \phi),$$

then there is no selection bias, and

$$\begin{aligned} p(\boldsymbol{y}_i^{\text{OBS}}, \boldsymbol{r}_i | \boldsymbol{x}_i, \boldsymbol{\theta}) &= \int \underbrace{p(\boldsymbol{y}_i^{\text{OBS}}, \boldsymbol{y}_i^{\text{MIS}} | \boldsymbol{x}_i, \boldsymbol{\theta})}_{\text{Population Model (2)}}\, d\boldsymbol{y}_i^{\text{MIS}} \times p(\boldsymbol{r}_i | \boldsymbol{x}_i, \phi) \\ &= p(\boldsymbol{y}_i^{\text{OBS}} | \boldsymbol{x}_i, \boldsymbol{\theta}) \times p(\boldsymbol{r}_i | \boldsymbol{x}_i, \phi), \end{aligned}$$

so conditionally independent and can concentrate on $p(\boldsymbol{y}_i^{\text{OBS}} | \boldsymbol{x}_i, \boldsymbol{\theta})$.

# Basic Normal Hierarchical Model

Under simple random sampling the selection does not depend on the response and we can use (2), i.e.,

$$Y_{ik} = \beta_0 + \delta_i + \epsilon_{ik}.$$

Under stratified simple random sampling within area *i* (for example, based on urban/rural) then we would not want to fit model (2).

Suppose we oversample individuals in urban regions (say) in area *i*, and the response is associated with urbanicity, then if we ignore this aspect when modeling the responses, we will introduce bias.

This is because the assumption that selection is independent of outcome is violated.

If the mean response in urban regions is lower than in rural regions and we oversample urban regions, we will underestimate the mean.

# Basic Normal Hierarchical Model

In the stratified SRS case, with sampling based on urban/rural, we could use the covariate version of this model,

$$Y_{ik} = \underbrace{\beta_0 + \boldsymbol{x}_{ik}^{\mathsf{T}}\boldsymbol{\beta}_1}_{\text{Regression Model for Individual } k} + \underbrace{\delta_i}_{\text{Adjustment for Area } i} + \epsilon_{ik},$$

where in this case $\boldsymbol{x}_{ik}$ is univariate and consists of an urban/rural indicator.

Now we are fine, because

$$p(\boldsymbol{r}_i | \boldsymbol{y}_i^{\text{OBS}}, \boldsymbol{y}_i^{\text{MIS}} \boldsymbol{x}_i, \phi) = p(\boldsymbol{r}_i | \boldsymbol{x}_i, \phi).$$

# Basic Normal Hierarchical Model

Suppose there are $N_{i0}$ and $N_{i1}$ individuals in rural and urban regions in area $i$.

The population mean in area $i$ can be written as

$$\overline{Y}_i = \underbrace{\overline{Y}_{i0}}_{\text{Rural Mean}} \times \frac{N_{i0}}{N_i} + \underbrace{\overline{Y}_{i1}}_{\text{Urban Mean}} \times \frac{N_{i1}}{N_i},$$

where $\overline{Y}_{i0}$ and $\overline{Y}_{i1}$ are the population means over the rural and urban regions.

# Basic Normal Hierarchical Model

If $N_{i0} \gg n_{i0}$ and $N_{i1} \gg n_{i1}$, then we can estimate this mean by

$$\begin{aligned}\widehat{\overline{Y}}_i &= \widehat{\overline{Y}}_{i0}\frac{N_{i0}}{N_i} + \widehat{\overline{Y}}_{i1}\frac{N_{i1}}{N_i} \\ &= (\widehat{\beta}_0 + \widehat{\delta}_i)\frac{N_{i0}}{N_i} + (\widehat{\beta}_0 + \widehat{\beta}_1 + \widehat{\delta}_i)\frac{N_{i1}}{N_i}\end{aligned}$$

Note: a key assumption here is that the random effect $\delta_i$ is common to both rural and urban individuals.

[Aside: we can contrast this with the weighted estimator

$$\frac{\sum_{k \in s_i} w_{ik} y_{ik}}{\sum_{k \in s_i} w_{ik}} = \overline{y}_{i0}\frac{N_{i0}}{N_i} + \overline{y}_{i1}\frac{N_{i1}}{N_i},$$

where $w_{ik} = N_{i0}/n_{i0}$ for sampled urban individuals and $= N_{i1}/n_{i1}$ for sampled rural individuals, which we will see later.]

# Non-Spatial Hierarchical Model for Binomial Data

# Smoothing Models

The above considerations of instability led to methods being developed to *smooth* the risks using hierarchical/random effects models that use the data from the totality of areas to provide more reliable estimates in each of the constituent areas.

Overview of Models:

- Basic Binomial Model: No smoothing.
- Random Effects Models:
    - Binomial Beta: Non-spatial smoothing.
    - Binomial IID GLMM[3]: Non-spatial smoothing.
    - Binomial Spatial GLMM: Spatial and non-spatial smoothing.
- Covariates may be added to each of these in order to smooth over covariate space.
- Estimation in these models is a separate issue.

[3] Generalized Linear Mixed Model

# Overview of Models

The individual responses are the binary random variables $Y_{ik}$, for the sampled individuals $k \in s_i$.

If we take $Y_i = \sum_{k \in s_i} Y_{ik}$, the obvious sampling model (likelihood) is:

$$Y_i | \theta_i \sim \text{Binomial}\left(n_i, \theta_i\right).$$

Having unconstrained $\theta_i$ and taking the MLE's leads to $\widehat{\theta_i} = Y_i / n_i$.

We briefly describe a Beta Binomial model in which the probabilities are assumed to arise from a common beta distribution:

$$\theta_i \sim \text{Beta}(a, b).$$

# Overview of Models

The simplest GLMM assumes the odds in area *i* are of the form

$$\frac{\theta_i}{1-\theta_i} = \exp(\beta_0) \times \exp(\delta_i),$$

with $\exp(\delta_i)$ are area-specific adjustments that multiply the overall odds $\exp(\beta_0)$.

This model is equivalent to a linear model on the logistic scale:

$$\log\left(\frac{\theta_i}{1-\theta_i}\right) = \beta_0 + \delta_i.$$

The random effects $\delta_i$ are assumed to follow a normal distribution, that is, $\delta_i \sim_{iid} \mathrm{N}(0, \sigma_\delta^2)$.

# Beta-Binomial Model

We begin by describing a simple Beta Binomial two-stage model that offers analytic tractability and ease of estimation.

- A very simple model is

  $$Y_i|\theta \sim_{ind} \text{Binomial}\,(n_i, \theta)\,,$$

  so that the risk is constant across all areas, and equal to $\theta$ (the overall risk).

- We would like a model between the above form with one parameter and the $n$ distinct, unrelated risks model, which leads to the empirical risks, $\widehat{\theta}_i = Y_i/n_i$; a random effects model provides this compromise.

# Beta-Binomial Model

We assume there are no covariates and assume the first stage likelihood is given by

$$Y_i|\theta_i \sim_{ind} \text{Binomial}\left(n_i, \theta_i\right). \tag{3}$$

At the second stage the random effects $\theta_i$ are assigned a distribution.

We initially assume that $\theta_i$ are modeled by

$$\theta_i|a, b \sim_{iid} \text{Beta}(a, b), \tag{4}$$

a beta distribution with mean

$$\text{E}[\theta_i] = \frac{a}{a+b},$$

and variance

$$\text{var}(\theta_i) = \frac{\text{E}[\theta_i](1 - \text{E}[\theta_i])}{a+b+1}.$$

If $a + b$ is small we have a narrow beta distribution, and we would expect large shrinkage of an area's risk estimate to the overall level, but if $a + b$ is large we have a more spread out distribution and low shrinkage is anticipated.

# Beta-Binomial Model

The rationale here is that we expect some similarity of risks $\theta_i$ across the map.

How do we decide upon values for *a* and *b*, which determines the location and spread of the $\theta_i$?

- We might hope that the totality of data might aid in estimating the $\theta_i$ in each area.
- One possibility would be to simply fix *a*, *b*, based on the context/historical data.
- However, estimating *a* and *b* from the data will often lead to an appropriate measure of the spread of the distribution.
- Estimation may be carried out using empirical Bayes or full Bayes methods.

Before we discuss estimation of *a* and *b* we see how we would proceed, if they were known.

# Beta-Binomial Model

The model is

$$\begin{aligned} Y_i|\theta_i &\sim_{ind} \text{Binomial}\,(n_i,\theta_i) \\ \theta_i|a,b &\sim_{iid} \text{Beta}(a,b) \end{aligned}$$

This leads to a beta posterior for $\theta_i$:

$$\theta_i|y_i,a,b \sim \text{Beta}(a+y_i, b+n_i-y_i).$$

Hence, the posterior mean risk estimate is

$$\begin{aligned} \widehat{\theta}_i &= \frac{a+y_i}{a+b+n_i} \\ &= \underbrace{\frac{a}{a+b}}_{\text{Prior Mean}} \frac{a+b}{a+b+n_i} + \underbrace{\frac{y_i}{n_i}}_{\text{Observed Risk}} \frac{n_i}{a+b+n_i} \end{aligned}$$

Notice behavior at $y_i = 0$ or $y_i = n_i$.

# Beta-Binomial Model

The estimated variance of the sample average in area *i* is estimated as

$$\frac{\widehat{\theta}_i(1-\widehat{\theta}_i)}{n_i},$$

so the variance can grow without bound as $n_i$ decreases.

Also, problems when $\widehat{\theta}_i = 0/1$.

For the smoothed estimate the variance is obtained from the biasposterior – recall the variance of a Beta$(a, b)$ is

$$\frac{\mathsf{E}[\theta_i] \times (1 - \mathsf{E}[\theta_i])}{a+b+1}.$$

# Beta-Binomial Model

The posterior variance is

$$\frac{\mathrm{E}[\theta_i|y_i](1-\mathrm{E}[\theta_i|y_i])}{a+b+n_i+1}$$

showing that the posterior variances are bounded above.

Also,

$$\mathrm{E}[\theta_i|y_i] = \frac{a+y_i}{a+b+n_i},$$

can't equal 0 or 1 when $a, b > 0$.

The question of how we estimate *a* and *b* is considered in a Technical Appendix.

# Binomial GLMM Model

The beta prior model is computationally convenient but cannot easily be extended to allow for residual spatial dependence.

A Binomial GLMM non-spatial random effect model is given by:

$$\begin{aligned} Y_i|\theta_i &\sim_{ind} \text{Binomial}(n_i, \theta_i) \\ \log\left(\frac{\theta_i}{1-\theta_i}\right) &= \beta_0 + \delta_i \\ \delta_i|\sigma_\delta^2 &\sim_{iid} \text{N}(0, \sigma_\delta^2) \end{aligned}$$

where $\delta_i$ are area-specific random effects that capture the residual or unexplained (logit) risk in area $i$, $i = 1, \ldots, n$.

It is straightforward to add area-level covariates to this model via

$$\log\left(\frac{\theta_i}{1-\theta_i}\right) = \beta_0 + \boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}_1 + \delta_i.$$

Same arguments for selection bias given for normal-normal hold for Beta-Binomial and Binomial GLMM models.

# SAE Inference

We may wish to estimate the total number of cases or the average (prevalence) in area *i*:

$$\begin{aligned} T_i &= \sum_{k \in s_i} y_{ik} + \sum_{k \in r_i} y_{ik} \\ \frac{T_i}{N_i} &= \frac{\sum_{k \in s_i} y_{ik} + \sum_{k \in r_i} y_{ik}}{N_i}. \end{aligned}$$

Estimates:

$$\begin{aligned} \widehat{T}_i &= \sum_{k \in s_i} y_{ik} + (N_i - n_i) \times \widehat{\theta}_i \\ \frac{\widehat{T}_i}{N_i} &= \frac{\sum_{k \in s_i} y_{ik}}{n_i} \times \frac{n_i}{N_i} + \widehat{\theta}_i \times \frac{N_i - n_i}{N_i}. \end{aligned}$$

# SAE Inference

If $N_i \gg n_i$:

$$\begin{aligned}\widehat{T}_i &= N_i\widehat{\theta}_i \\ \frac{\widehat{T}_i}{N_i} &= \widehat{\theta}_i.\end{aligned}$$

In the Binomial GLMM:

$$\widehat{\theta}_i = \frac{\exp(\widehat{\beta}_0 + \widehat{\delta}_i)}{1 + \exp(\widehat{\beta}_0 + \widehat{\delta}_i)}.$$

# Motivating Example: Binary Outcome

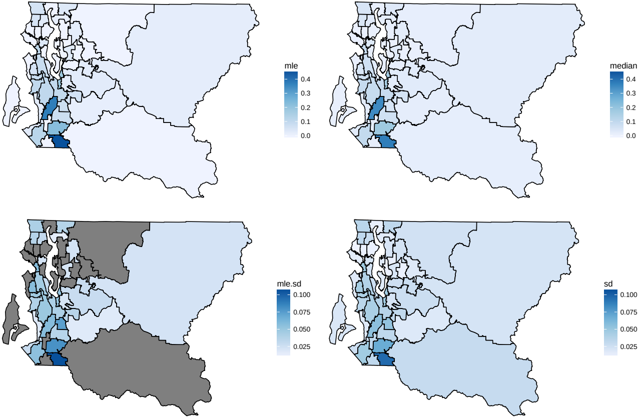

Figure 8: Top row: Estimates of area proportions with diabetes via MLE's (left) and posterior medians (right). Bottom row: Uncertainty of estimates with standard errors (left) and posterior standard deviations (right).

# Motivating Example: Binary Outcome

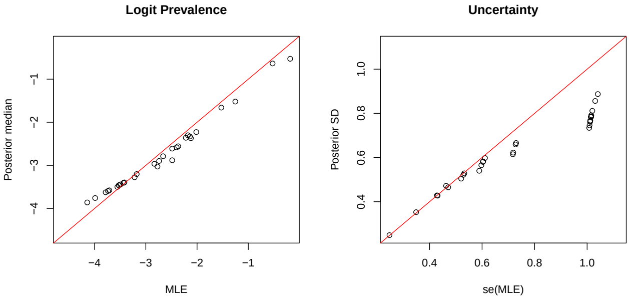

Figure 9: Comparison of area averages: Posterior medians versus MLEs on the logistic scale (left). Posterior standard deviations versus standard errors associated with the MLEs on the logistic scale (right).

As in the normal model case, we see that the Bayesian estimates are shrunk relative to the MLEs, and the uncertainty of the Bayesian estimates is in general smaller.

# Discussion

# Discussion

Random effects models:

- use all the data to shrink area-level estimates,
- this introduces bias,
- but the use of all the data, usually gives a reduction in variance, and this can be substantial.

The Beta-Binomial model is useful to introduce the smoothing concept and for non-spatial random effects, but cannot be extended easily to the spatial case; hence, in practice I would use the Binomial-GLMM model.

We haven't talked about priors, the analyses reported here were obtained using the INLA R package, and the default priors in this implementation are usually reliable (but see the technical appendix for a discussion of prior choice in the Binomial GLMM model.

# Discussion

We have also not talked about model comparison or model checking.

Model comparison may be compared out with `INLA` using a variety of measures:

- Bayes factors.
- Deviance Information Criteria (DIC).
- Widely Applicable Information Criteria (WAIC).
- Conditional Predictive Ordinate (CPO).

These are described in a Technical Appendix; Bayes factors, DIC and CPO were used in the context of SAE for U5MR by Mercer *et al.* (2015).

Model checking for hierarchical models is described in Chapters 8 and 9 of Wakefield (2013).

# References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In P. B.N. and C. F., editors, *Second International Symposium on Information Theory*, pages 267–281. Akademia Kiadó, Budapest.

Ando, T. (2007). Bayesian predictive information criterion for the evaluation of hierarchical Bayesian and empirical Bayes models. *Biometrika*, **94**, 443–458.

Besag, J., York, J., and Mollié, A. (1991). Bayesian image restoration with two applications in spatial statistics. *Annals of the Institute of Statistics and Mathematics*, **43**, 1–59.

Held, L., Schrödle, B., and Rue, H. (2010). Posterior and cross-validatory predictive checks: A comparison of MCMC and INLA. In T. Kneib and G. Tutz, editors, *Statistical Modeling and Regression Structures – Festschrift in Honour of Ludwig Fahrmeir*, pages 91–110. Physica-Verlag.

Kass, E. and Wasserman, L. (1995). A reference Bayesian test for nested hypothesis and its relationship to the Schwarz criterion. *Journal of the American Statistical Association*, **90**(431), 928–934.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models, Second Edition*. Chapman and Hall, London.

Mercer, L., Wakefield, J., Pantazis, A., Lutambi, A., Mosanja, H., and Clark, S. (2015). Small area estimation of childhood of childhood mortality in the absence of vital registration. *Annals of Applied Statistics*, **9**, 1889–1905.

Plummer, M. (2008). Penalized loss functions for Bayesian model comparison. *Biostatistics*, **9**, 523–539.

Rao, J. and Molina, I. (2015). *Small Area Estimation, Second Edition*. John Wiley, New York.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society, Series B*, **71**, 319–392.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464.

Spiegelhalter, D., Best, N., Carlin, B., and Linde, A. V. D. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B*, **64**, 583–639.

Spiegelhalter, D., Best, N., Carlin, B., and Linde, A. V. D. (2014). The deviance information criterion: 12 years on (with discussion). *Journal of the Royal Statistical Society: Series B*, **64**, 485–493.

Wakefield, J. (2013). *Bayesian and Frequentist Regression Methods*. Springer, New York.

Watanabe, S. (2013). A widely applicable Bayesian information criterion. *Journal of Machine Learning Research*, **14**, 867–897.

# Technical Appendix: Computation for Beta Binomial Model

# Empirical Bayes Estimation in the Beta Binomial Model

In an empirical Bayes approach the random effects $\delta_i$ are eliminated from the model to give a negative binomial likelihood that depends on $a$ and $b$ only:

$$\begin{aligned}\Pr(Y_i|a,b) &= \int \Pr(Y_i|\theta_i) \times p(\theta_i|a,b)d\theta_i \\ &= \binom{n_i}{y_i}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\frac{\Gamma(a+y_i)\Gamma(b+n_i-y_i)}{\Gamma(a+b+n_i)}.\end{aligned}$$

The likelihood is

$$L(a,b) = \prod_{i=1}^{n} \Pr(Y_i|a,b),$$

which is maximized as a function of $a$ and $b$.

We then proceed as if $a, b$ are known, i.e., the posterior mean estimates are

$$\mathrm{E}[\theta_i|y_i,\widehat{a},\widehat{b}].$$

# Full Bayes Estimation in the Beta Binomial Model

The full Bayes approach assigns a (hyper) prior to the (hyper) parameters $a, b$ to give the three stage hierarchical model:

Stage 1: $Y_i|\theta_i \sim_{ind}$ Binomial$(n_i, \theta_i)$, $i = 1, \ldots, n$.

Stage 2: $\theta_i|a, b \sim_{iid}$ Beta$(a, b)$, $i = 1, \ldots, n$.

Stage 3: Priors for $a, b$.

The posterior is

$$p(\theta_1, \ldots, \theta_n, a, b|\boldsymbol{y}) \propto \left[\prod_{i=1}^{n} p(y_i|\theta_i)p(\theta_i|a, b)\right] p(a, b).$$

This model is not analytically tractable and we do not discuss further, since the Binomial GLMM model we describe shortly is more flexible.

What do we gain by full Bayes? Uncertainty in $a, b$ can be acknowledged.

# Full Bayes Estimation in the Beta Binomial Model

In general, the posterior distribution is analytically intractable but can be implemented using:

- Markov chain Monte Carlo (MCMC). `WinBUGS` and more specifically the `GeoBUGS` module is a convenient way to do this. Other (generic) MCMC environments include `JAGS` (very similar to `WinBUGS`) and `Stan`.
- This is the method that has been used since the early 1990s (Besag *et al.*, 1991).
- More recently (Rue *et al.*, 2009) the integrated nested Laplace approximation (INLA) has been developed — can't be used for this model, but for the all the GLMM models we will see later.

# Technical Appendix: Prior Choice for Binomial GLMM

# Prior Choice for Binomial GLMM

We need to specify priors for:

- The intercept $\beta_0$ and regression coefficient $\beta_1$.
- The variance of the normal random effects $\sigma_\epsilon^2$.

An improper prior[4]

$$p(\beta_0, \beta_1) \propto 1$$

may often be used, but in some circumstances such a choice may lead to an improper posterior.

If there are a large numbers of covariates, or high dependence amongst multiple covariates then more informative priors will be beneficial.

If an informative prior is required, then a multivariate normal distribution is the natural choice.

This is equivalent to a multivariate lognormal distribution for the relative risks.

[4] This means that it doesn't integrate to 1

# Prior Choice for Binomial GLMM

It is convenient to specify lognormal priors for a positive parameter $\exp(\beta)$ (i.e., the odds $\exp(\beta_0)$ or odds ratio $\exp(\beta_1)$), since one may specify two quantiles of the distribution, and directly solve for the two parameters of the lognormal.

Denote by LogNormal$(\mu, \sigma)$ the lognormal distribution for a generic parameter $\theta$ with

$$\mathsf{E}[\log(\theta)] = \mu, \qquad \mathsf{var}(\log(\theta)) = \sigma^2,$$

and let $\theta_1$ and $\theta_2$ be the $q_1$ and $q_2$ quantiles of this prior.

In our example, $\theta = \exp(\beta)$.

Then it is straightforward to show that

$$\begin{aligned} \mu &= \log(\theta_1)\left(\frac{z_{q_2}}{z_{q_2} - z_{q_1}}\right) - \log(\theta_2)\left(\frac{z_{q_1}}{z_{q_2} - z_{q_1}}\right), \\ \sigma &= \frac{\log(\theta_1) - \log(\theta_2)}{z_{q_1} - z_{q_2}}. \end{aligned}$$

# $\exp(\beta_0)$

As an example, suppose that for the ecological relative risk

$$\theta = \exp(\beta)$$

we believe there is a 50% chance that the odds ratio is less than 1 and a 95% chance that it is less than 5.

This gives

$$q_1 = 0.5, \qquad \theta_1 = 1.0, \qquad q_2 = 0.95, \qquad \theta_2 = 5.0,$$

we obtain lognormal parameters

$$\mu = 0, \qquad \sigma = \frac{\log 5}{1.645} = 0.98.$$

The density is shown in Figure 10.

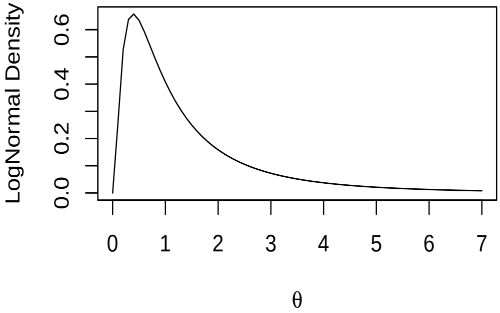

Figure 10: Lognormal density with 50% point 1 and 95% point 5.

# Prior Choice for Binomial GLMM

The priors $\tau_\epsilon = \sigma_\epsilon^{-2} \sim \text{Ga}(1, 0.0260)$ or $\tau_\epsilon = \sigma_\epsilon^{-2} \sim \text{Ga}(0.5, 0.0005)$ will often be suitable in a mapping context.

$\tau_\epsilon$ is the precision, i.e., the reciprocal variance.

For the Ga(1,0.026) prior the 2.5%, 50% (median) and 97.5% quantiles for $\sigma_\epsilon$ are:

$$(0.014, \quad 0.047, \quad 1.01).$$

For the Ga(0.5,0.0005) prior the 2.5%, 50% (median) and 97.5% quantiles for $\sigma_\epsilon$ are:

$$(0.084, \quad 0.194, \quad 1.01).$$

So the Ga(1,0.026) prior favors smaller values, i.e., more shrinkage is anticipated.

# Prior Choice for Binomial GLMM

Interpretation is helped by approximation of the residual odds ratio

$$\exp(\epsilon) \approx 1 + \epsilon$$

for small $\epsilon$ and so

$$\text{s.d}(e^{\epsilon}) = \sigma_\epsilon$$

is approximately the standard deviation of the residual relative risks.

Sensitivity of the results to the specification should be carried out, particularly if the number of areas is not large.

# Empirical Bayes Estimation in the Poisson-Gamma Model Without Covariates

In an empirical Bayes approach the random effects $\delta_i$ are eliminated from the model to give a negative binomial likelihood that depends on $\beta_0$ and $\alpha$ only:

$$
\begin{aligned}
\Pr(Y_i|\beta_0,\alpha) &= \int \Pr(Y_i|\beta_0,\delta_i) \times p(\delta_i|\alpha) d\delta_i \\
&= \frac{\Gamma(y_i+\alpha)}{\Gamma(\alpha)} \left( \frac{E_i e^{\beta_0}}{E_i e^{\beta_0}+\alpha} \right)^{y_i} \left( \frac{\alpha}{E_i e^{\beta_0}+\alpha} \right)^{\alpha}.
\end{aligned}
$$

The likelihood is

$$L(\beta_0,\alpha) = \prod_{i=1}^{n} \Pr(Y_i|\beta_0,\alpha),$$

which is maximized as a function of $\beta_0$ and $\alpha$ – `R` can do this for us using the `glm.nb()` function in the `MASS` library.

We then proceed as if $\alpha$ and $\beta_0$ are known, i.e. the estimates are $\text{E}[\delta_i|y_i,\widehat{\alpha},\widehat{\beta_0}]$.

# Full Bayes Estimation in the Poisson-Gamma Model Without Covariates

The full Bayes approach assigns a (hyper) prior to the (hyper) parameters $\alpha, \beta_0$ to give the three stage hierarchical model:

Stage 1: $Y_i | \delta_i, \beta_0 \sim_{ind} \text{Poisson}(e^{\beta_0} E_i \delta_i)$, $i = 1, \dots, n$.

Stage 2: $\delta_i | \alpha \sim_{iid} \text{Ga}(\alpha, \alpha)$, $i = 1, \dots, n$.

Stage 3: Priors for $\alpha, \beta_0$.

The posterior is

$$p(\delta_1, ..., \delta_n, \alpha, \beta_0 | \boldsymbol{y}) \propto \left[ \prod_{i=1}^{n} p(y_i | \delta_i, \beta_0) p(\delta_i | \alpha) \right] p(\alpha, \beta_0).$$

This model is not analytically tractable and we do not discuss further (including the issue of prior choice), since the Poisson-Lognormal model we describe shortly is more flexible.

# Full Bayes Estimation in the Poisson-Gamma Model Without Covariates

What do we gain by full Bayes? Uncertainty in $\alpha, \beta_0$ can be acknowledged.

The posterior distribution is analytically intractable but can be implemented using

- Markov chain Monte Carlo (MCMC). `WinBUGS` and more specifically the `GeoBUGS` module is a convenient way to do this. Other (generic) MCMC environments include `JAGS` (very similar to `WinBUGS`) and `Stan`.
- This is the method that has been used since the early 1990s (Besag *et al.*, 1991).
- More recently (Rue *et al.*, 2009) the integrated nested Laplace approximation (INLA) has been developed — can't be used for this model, but for the lognormal models we will see later.

Note: the Poisson-Gamma model is useful to introduce the smoothing concept and for non-spatially dependent random effects, but cannot be extended easily.

# Technical Appendix: Model Comparison

# Model Comparison

Markov chain Monte Carlo in particular has allowed the fitting of more and more complex models, often hierarchical in nature with layers of random effects.

The search for a method to find the "best" of a set of candidate models has also grown.

Let $p(\boldsymbol{y}|\boldsymbol{\theta})$ represent a generic likelihood for $\boldsymbol{y} = [y_1, \dots, y_n]$ and let

$$D(\boldsymbol{\theta}) = -2\log[p(\boldsymbol{y}|\boldsymbol{\theta})]$$

represent the deviance.

For example, in an iid $\text{N}(\mu_i(\boldsymbol{\theta}), \sigma^2)$ normal the deviance is

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}[y_i - \mu_i(\boldsymbol{\theta})]^2.$$

Frequentist model comparison for nested models is often carried out using likelihood ratio statistics, which corresponds to the comparison of deviances in generalized linear models (GLMs), see for example McCullagh and Nelder (1989).

# Model Comparison: AIC

One approach to model comparison is based on a model's ability to make good predictions.

Such an objective, and predicting the actual observed data, leads to Akaike's an information criterion (AIC), derived in Akaike (1973).

In AIC one tries to estimate the (Kullback-Leibler) distance between the true distribution of the data, and the modeled distribution of the data.

# Model Comparison: AIC

AIC is given by

$$\text{AIC} = -2\log[p(y|\widehat{\boldsymbol{\theta}})] + 2k$$

where $\widehat{\theta}$ is the MLE and $k$ is the number of parameters in the model, i.e. the size of $\boldsymbol{\theta}$.

Small values of the AIC are favored, since they suggest low prediction error.

The penalty term $2k$ penalizes the double use of the data.

In general for prediction: overly complex models are penalized since redundant parameters "use up" information in the data.

# Model Comparison: BIC

Another approach is based on trying to identify the “true” model.

Schwarz (1978) developed the Bayesian Information Criterion (BIC) which is given by

$$\text{BIC} = -2 \log[p(y|\widehat{\theta})] + k \log n.$$

BIC approximates $-2 \log p(\boldsymbol{y}|\boldsymbol{\theta})$ under a certain unit information prior (Kass and Wasserman, 1995).

BIC is consistent[5] for finding the true model, if that model lies in the set being compared.

AIC is not consistent for finding the true model, but recall is intended for prediction.

[5] meaning the BIC hones in on the true model as the sample size increases

# Model Comparison: DIC

Spiegelhalter *et al.* (2002) introduced what has proved to be a very popular model comparison statistic, the deviance information criterion (DIC).

To define the DIC, define an “effective number of parameters” as

$$\begin{aligned} p_i &= E_{\theta|y}\{-2\log[p(\boldsymbol{y}|\boldsymbol{\theta})]\} + 2\log[p(\boldsymbol{y}|\overline{\boldsymbol{\theta}})] \\ &= \overline{D} + D(\overline{\boldsymbol{\theta}}) \end{aligned}$$

where $\overline{\boldsymbol{\theta}} = E[\theta|\boldsymbol{y}]$ is the posterior mean, $D(\overline{\boldsymbol{\theta}})$ is the deviance evaluated at the posterior mean and $\overline{D} = E[D|\boldsymbol{y}]$.

Hence, $p_i$ is the

posterior mean deviance – deviance of posterior means.

# Model Comparison: DIC

The DIC is given by

$$\begin{aligned} \text{DIC} &= D(\overline{\boldsymbol{\theta}}) + 2p_i \\ &= \overline{D} + p_i, \end{aligned}$$

so that we have a measure of goodness of fit $+$ complexity.

DIC is straightforward to evaluate using MCMC or INLA.

# Model Comparison: DIC

DIC has been heavily criticized (Spiegelhalter *et al.*, 2014):

- $p_i$ is not invariant to parameterization.
- DIC is not consistent for choosing the correct model.
- DIC has a weak theoretical justification and is not universally applicable.
- DIC has been shown to under penalize complex models (Plummer, 2008; Ando, 2007).
- See Spiegelhalter *et al.* (2014) for an interesting discussion of the history of DIC, including a summary of attempts to improve DIC.
- According to Google Scholar, as of June 20th, 2014, Spiegelhalter *et al.* (2002) has 5251 citations. . .

WAIC (Watanabe, 2013) is growing in popularity.

# Model Comparison: CPO

Another approach based on prediction uses the conditional predictive ordinate (CPO).

Let

$$\boldsymbol{y}_{-i} = [y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n]$$

represent the vector of data with the *i*-th observation removed.

The idea is to predict the density ordinate of the left-out observation, based on those that remain.

Specifically, the CPO for observation *i* is defined as:

$$\begin{aligned} \text{CPO}_i &= p(y_i|\boldsymbol{y}_{-i}) \\ &= \int p(y_i|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{y}_{-i})\, d\boldsymbol{\theta} \\ &= E_{\theta|y_{-i}}\left[p(y_i|\boldsymbol{\theta})\right] \end{aligned}$$

# Model Comparison: CPO

The CPOs can be used to look at local fit, or one can define an overall score for each model:

$$\log(\text{CPO}) = \sum_{i=1}^{n} \log \text{CPO}_i.$$

Good models will have relatively high values of $\log(\text{CPO})$.

See Held *et al.* (2010) for a discussion of shortcuts for estimation (i.e. avoidance of fitting the model *n* times) using MCMC and INLA.