Linear Mixed Effects Models

The hierarchical model idea may be applied to regression settings.

We have several groups of observations. Each group has its own linear regression model.

We assume that the regression coefficients from the different models are drawn from some prior distribution.

In this way, *information across groups is shared in estimating the whole set of regression coefficients.*

Suppose there are m groups and assume that

$$\boldsymbol{Y}_j = \boldsymbol{X}_j \boldsymbol{\beta}_j + \boldsymbol{\epsilon}_j, \quad j = 1, \ldots, m.$$

- $\boldsymbol{Y}_j$: an $n_j \times 1$ vector of observations

- $\boldsymbol{X}_j$: an $n_j \times p$ matrix of covariates

- $\boldsymbol{\beta}_j$: a $p \times 1$ vector of unknown regression coefficients

- $\boldsymbol{\epsilon}_j$: an $n_j \times 1$ vector of unobserved error terms whose components are i.i.d. $N(0, \sigma^2)$.

Therefore, $\boldsymbol{Y}_j \sim N(\boldsymbol{X}_j\boldsymbol{\beta}_j, \sigma^2\boldsymbol{I})$. We also assume that $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_m$ are independent, conditional on $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m$ and σ^2.

As a prior for the regression coefficients, we assume that, given $\boldsymbol{\theta}$, $\boldsymbol{\Sigma}$ and σ^2, *$\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m$ are i.i.d. $N(\boldsymbol{\theta}, \boldsymbol{\Sigma})$.*

Why does the name *mixed effects model* apply here?

Conditional on $\boldsymbol{\theta}$, $\boldsymbol{\Sigma}$ and σ^2, we may write

$$\boldsymbol{\beta}_j = \boldsymbol{\theta} + \boldsymbol{\gamma}_j,$$

where $\boldsymbol{\gamma}_j \sim N(\mathbf{0}, \boldsymbol{\Sigma})$.

Substituting the last expression into the linear model, we have

$$\boldsymbol{Y}_j = \boldsymbol{X}_j\boldsymbol{\theta} + \boldsymbol{X}_j\boldsymbol{\gamma}_j + \boldsymbol{\epsilon}_j.$$

The parameter $\boldsymbol{\theta}$ is called a *fixed effect*, since it is the same across groups, whereas $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_m$ are *random effects*, since they vary randomly from one group to the next.

Since the model contains both types of effects, it is called a *mixed effects* model.

A generalization of the mixed effects model on the previous page is one in which *the covariates associated with the fixed effect may be different than those associated with the random effects.*

In this case we would have

$$\boldsymbol{Y}_j = \boldsymbol{X}_j\boldsymbol{\theta} + \boldsymbol{Z}_j\gamma_j + \boldsymbol{\epsilon}_j.$$

The matrices $\boldsymbol{X}_j$ and $\boldsymbol{Z}_j$ could have columns in common, meaning that the model on the previous page is a special case of the one immediately above.

In a proper Bayesian analysis, we need a prior for $\boldsymbol{\theta}$, $\boldsymbol{\Sigma}$ and σ^2.

We assume that these parameters are a priori independent with

$$\boldsymbol{\theta} \sim N(\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0),$$

$$\boldsymbol{\Sigma} \sim \text{inverse-Wishart}(\eta_0, \boldsymbol{S}_0^{-1}) \text{ and}$$

$$\sigma^2 \sim \text{inverse-gamma}(\nu_0/2, \nu_0\sigma_0^2/2).$$

With these choices for priors, all the full conditionals have familiar forms, as given on pp. 198-199 of Hoff.

So we may use Gibbs sampling to explore the posterior.

Example 21 Analysis of data from a randomized block design

Consider the data used in Exercise 11.2 of Hoff. These are in `pdensity.dat`, in the usual repository.

- There are ten plots of land, which are the blocks.
- Of interest is the effect of *planting density*, x, on the the yield, y, of a certain type of perennial grass.
- Each plot is divided into 8 subplots.
- Planting densities 2, 4, 6 and 8 (plants per square meter) are randomly assigned to the eight subplots, with each planting density appearing twice within a block.

The regression model

$$y = \beta_1 + \beta_2 x + \beta_3 x^2 + \epsilon$$

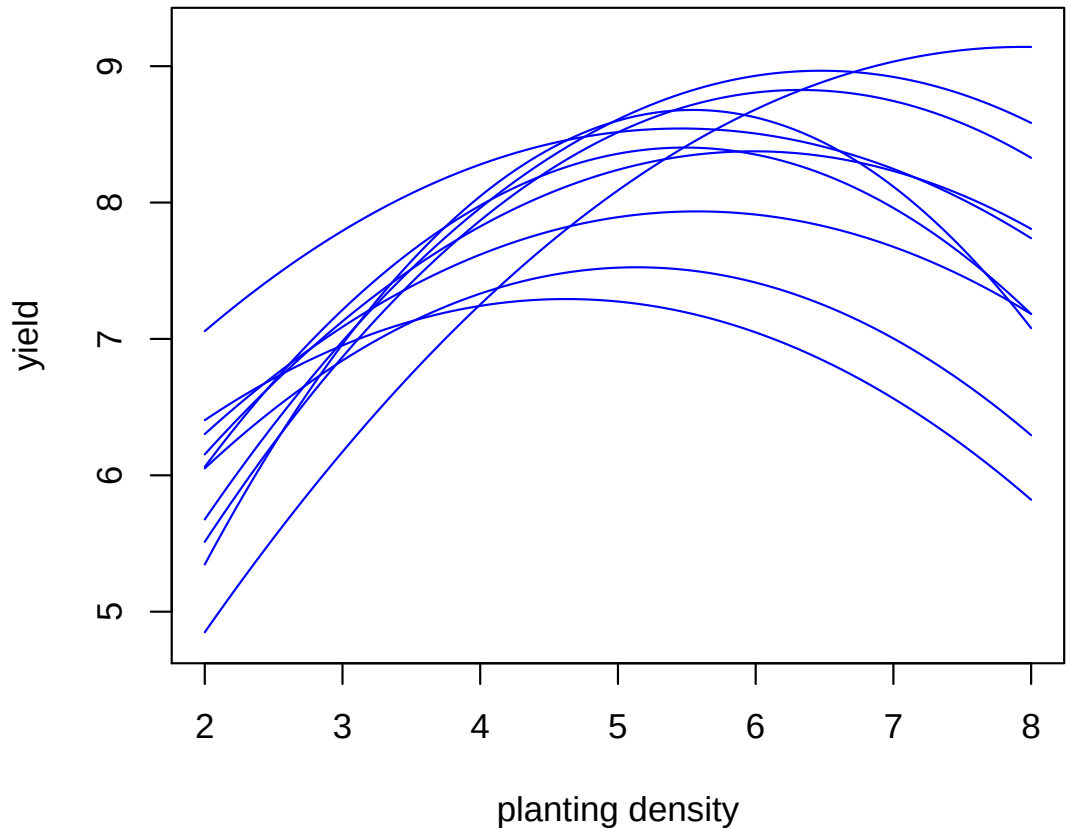
is to be fitted.

The researchers feel that there may be a substantial block effect, *so they will fit separate models for the ten blocks.*

In our previous notation, $m = 10$ and $n_j = 8$, $j = 1, \ldots, 10$.

To get an impression of the data, least squares estimates of the ten quadratic models were obtained.

Least squares quadratic fits for the ten different blocks



There appears to be heterogeneity among the fitted curves. Note, however, that the planting density that maximizes yield appears to be between 4 and 7.

Per the suggestion of Exercise 11.2, we will obtain parameters for our "priors" by using information from the fitted least squares curves.

We have $\boldsymbol{\beta}_j = (\beta_{1j}, \beta_{2j}, \beta_{3j})^T$, $j = 1, \ldots, 10$, and the least squares estimates $\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_{10}$ estimate these.

We can estimate $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$ by the sample mean and covariance matrix, $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\Sigma}}$, of $\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_{10}$.

$$\widehat{\boldsymbol{\theta}}^T = (2.869, 1.855, -0.159)$$

and

$$\widehat{\boldsymbol{\Sigma}} = \begin{bmatrix} 2.001 & -0.693 & 0.044 \\ -0.693 & 0.276 & -0.021 \\ 0.044 & -0.021 & 0.002 \end{bmatrix}.$$

We may estimate σ^2 by pooling residuals from all ten least squares fits. This yields $\hat{\sigma}^2 = 0.4924$.

We use the following priors:

$$\boldsymbol{\theta} \sim N(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\Sigma}}),$$

$$\boldsymbol{\Sigma}^{-1} \sim \text{Wishart}(4, \hat{\boldsymbol{\Sigma}}^{-1}) \text{ and}$$

$$\sigma^2 \sim \text{inverse-gamma}(1, \hat{\sigma}^2).$$

We'll now use Gibbs sampling to estimate the parameters.

Ten thousand observations from the posterior were generated.

Mixing was pretty good.

- All acfs died out after about 20 lags.
- The largest first lag correlation was 0.8, and most were no bigger than 0.4.
- If HPD intervals are desired, I would still use thinning.

The point estimates that follow were obtained by averaging over all the output.

$$\widetilde{\boldsymbol{\theta}}^T = (2.841, 1.869, -0.161)$$

$$\widetilde{\boldsymbol{\Sigma}} = \begin{bmatrix} 1.211 & -0.411 & 0.023 \\ -0.411 & 0.157 & -0.011 \\ 0.023 & -0.011 & 0.001 \end{bmatrix}$$

Remarks:

- The estimate of $\boldsymbol{\theta} = E(\boldsymbol{\beta}_j)$ is very close to the estimate $\widehat{\boldsymbol{\theta}}$ gotten by averaging least squares estimates.

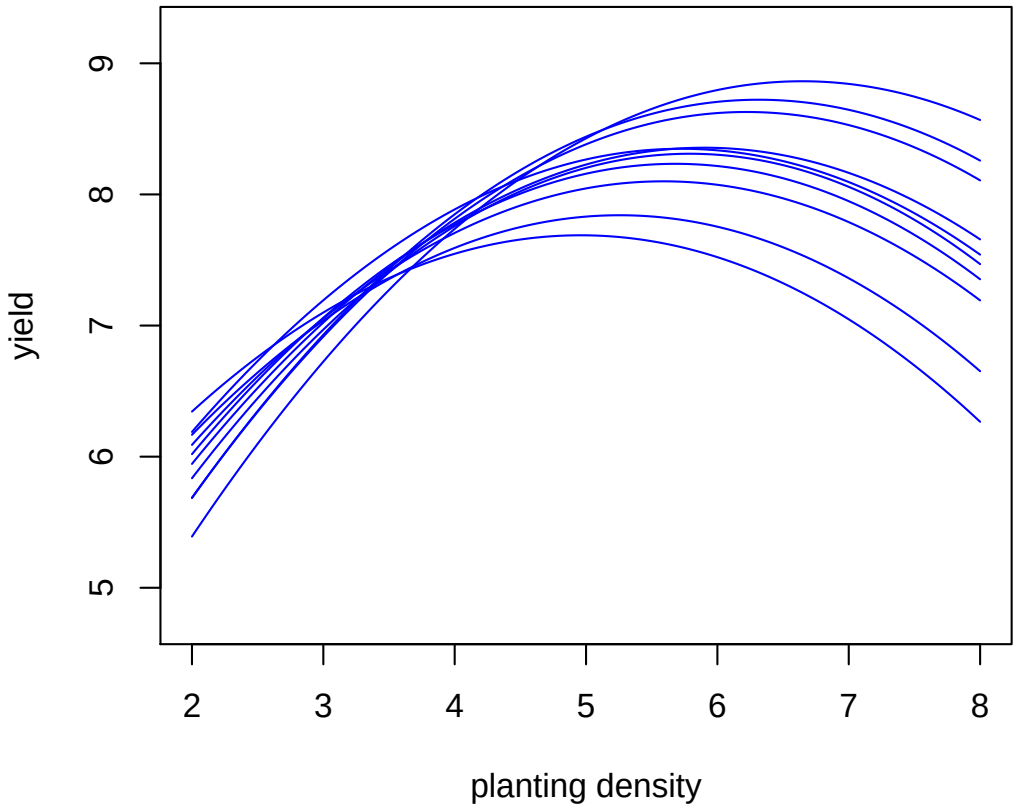
- The variances on the main diagonal of $\widetilde{\boldsymbol{\Sigma}}$ are smaller than those in $\widehat{\boldsymbol{\Sigma}}$. This makes sense because

$$\text{Var}(\widehat{\boldsymbol{\beta}}_j) = \text{Var}(\boldsymbol{\beta}_j) + \sigma^2(\boldsymbol{X}_j^T \boldsymbol{X}_j)^{-1}.$$

Estimates of regression coefficients

Block	β_1	β_2	β_3
1	3.245	1.783	-0.156
2	3.911	1.524	-0.154
3	2.215	2.061	-0.163
4	2.958	1.858	-0.164
5	3.478	1.660	-0.158
6	1.757	2.139	-0.161
7	3.225	1.745	-0.156
8	2.534	1.995	-0.172
9	2.838	1.871	-0.159
10	2.240	2.054	-0.165

Plotting the 10 estimated curves is revealing.



The shapes of the curves are more uniform than are the least squares estimates on p. 306N.

This analysis has "concluded" that *borrowing information in such a way that the estimates are relatively uniform is better supported by the data than 10 independent estimates of regression curves.*

The average of all 10,000 generated values of σ^2 was $\tilde{\sigma}^2 = 0.715$.

This is larger than the estimate $\hat{\sigma}^2 = 0.4924$ obtained by pooling residuals from all ten least squares fits.

This makes sense because of the greater uniformity among curve estimates in the Bayesian analysis.

The Bayesian analysis takes out some of the variability between curves and puts it into the error variance.