1. A Gauss-Markov Theorem for Dependent Data

Suppose \( E[Y] = x\beta \) and \( \text{var}(Y) = V \), where \( Y = (Y_1, \ldots, Y_m) \) with \( Y_i = (Y_{i1}, \ldots, Y_{ini})^\top \), \( x = (x_1, \ldots, x_m)^\top \) is \( N \times p \) with \( x_i = (x_{i1}, \ldots, x_{ini}) \), \( x_{ij} = (1, x_{ij1} \ldots x_{ijk}) \), \( N = \sum i n_i \) and \( \beta \) is \((k + 1) \times 1\).

Consider linear estimators of the form

\[
\tilde{\beta}_W = (x^\top W^{-1} x)^{-1} x^\top W^{-1} Y,
\]

where \( W \) is symmetric and positive definite.

(a) Show that \( E[\tilde{\beta}_W] = \beta \).

(b) Show that \( \text{var}(\tilde{\beta}_V) \leq \text{var}(\tilde{\beta}_W) \).

[Hint: You can show that \( \text{var}(\tilde{\beta}_V) - \text{var}(\tilde{\beta}_W) \) is positive semi-definite.]

2. Consider the data in Table 1 (from Davies, 1967, Statistical Methods in Research and Production, Third Edition, Olive and Boyd, London) which contain the yield in grams from six randomly chosen batches of raw material, with five replicates each. The aim of this experiment was to find out to what extent batch-to-batch variation was responsible for variation in the final product yield.

<table>
<thead>
<tr>
<th>Replicate</th>
<th>Observation</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1545</td>
<td>1540</td>
<td>1595</td>
<td>1445</td>
<td>1595</td>
<td>1520</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1440</td>
<td>1555</td>
<td>1550</td>
<td>1440</td>
<td>1630</td>
<td>1455</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1440</td>
<td>1490</td>
<td>1605</td>
<td>1595</td>
<td>1515</td>
<td>1450</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1520</td>
<td>1560</td>
<td>1510</td>
<td>1465</td>
<td>1635</td>
<td>1480</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1580</td>
<td>1495</td>
<td>1560</td>
<td>1545</td>
<td>1625</td>
<td>1445</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Dyestuff data: yield of dyestuff in grams of standard color.

One possibility for a model for these data is the one-way analysis of variance with:

\[
y_{ij} = \mu + b_i + \epsilon_{ij},
\]

with \( j = 1, \ldots, n \), replicates on \( i = 1, \ldots, m \), batches, \( b_i \sim_{i.i.d} N(0, \sigma_b^2) \), \( \epsilon_{ij} \sim_{i.i.d} N(0, \sigma_\epsilon^2) \).

In this question you will find the following useful: let \( I_n \) denote the \( n \times n \) identity matrix and \( J_n \) the \( n \times n \) matrix of 1’s. Then

\[
(a I_n + b J_n)^{-1} = \frac{1}{a} \left( I_n - \frac{b}{a + nb} J_n \right), \quad a \neq 0, \quad a \neq -nb,
\]
and

$$|aI_n + bJ_n| = a^{n-1}(a + nb).$$

(a) Derive the log-likelihood for $\mu, \sigma_0^2, \sigma_\epsilon^2$.

(b) Differentiate the log-likelihood and hence show that the MLEs are given by

$$\hat{\mu} = \bar{y},$$

$$\hat{\sigma}_\epsilon^2 = \text{MSE},$$

$$\hat{\sigma}_0^2 = \frac{(1 - 1/m)\text{MSA} - \text{MSE}}{n},$$

where $\text{MSA} = n \sum_{i=1}^m (\bar{y}_i - \bar{y}.)^2 / (m - 1)$, and $\text{MSE} = \sum_{i=1}^m \sum_{j=1}^n (y_{ij} - \bar{y}_i)^2 / [m(n - 1)]$.

[Hint: you might find it useful to write $\lambda = \sigma_\epsilon^2 + n\sigma_0^2$].

(c) Obtain the form of $\text{var}(\hat{\mu})$ and hence an estimate of this quantity.

(d) Attempt to find the REML estimators of $\sigma_0^2$ and $\sigma_\epsilon^2$.

(e) In the one-way random effects model with balanced data, it can be shown that

$$\frac{\text{MSA}/(n\sigma_0^2 + \sigma_\epsilon^2)}{\text{MSE}/\sigma_\epsilon^2} \sim F_{m-1,m(n-1)},$$

the $F$ distribution on $m - 1$ and $m(n - 1)$ degrees of freedom. Hence explain why $F = \text{MSA}/\text{MSE}$ may be compared with an $F_{m-1,m(n-1)}$ to provide a test of $H_0 : \sigma_0^2 = 0$.

(f) Using the last part, show that the probability that the REML estimator $\hat{\sigma}_0^2$ is negative is the probability that an $F_{m(n-1),(m-1)}$ random variable is bigger than $1 + n\sigma_0^2/\sigma_\epsilon^2$.

(g) Within R write code to obtain an estimate and associated standard error for $\mu$, and ML and REML estimates of $\sigma_0^2$ and $\sigma_\epsilon^2$.

(h) Confirm these estimates using the R function `lme()`.

3. Consider the so-called Neymann-Scott problem in which we have

$$Y_{ij} \sim_{\text{ind}} N(\mu_i, \sigma^2),$$

for $i = 1, \ldots, n$, $j = 1, 2$.

(a) Obtain the MLE for $\sigma^2$ and show that it is inconsistent. Why are there problems here?

(b) Consider a REML approach. Assign an improper uniform prior to $\mu_1, \ldots, \mu_n$, and integrate out these parameters. Hence obtain the REML of $\sigma^2$ and show that it is an unbiased estimator.