

ST 762, HOMEWORK 6 EXTRA PROBLEMS, FALL 2009

These problems are from previous years and are for you to work on or not as you choose; they are not to be turned in. There are no “regular” problems to be turned in this time. Solutions to these problems will be posted on the last day of class.

1. Consider a two-stage subject-specific model of the form

$$E(\mathbf{Y}_i | \mathbf{z}_i, \boldsymbol{\beta}_i) = \mathbf{f}_i(\mathbf{z}_i, \boldsymbol{\beta}_i), \quad \text{var}(\mathbf{Y}_i | \mathbf{z}_i, \boldsymbol{\beta}_i) = \mathbf{R}_i(\boldsymbol{\beta}_i, \boldsymbol{\gamma}, \mathbf{z}_i),$$

$$\boldsymbol{\beta}_i = \mathbf{A}_i \boldsymbol{\beta} + \mathbf{b}_i,$$

where \mathbf{b}_i are independent of $\mathbf{x}_i = (\mathbf{z}_i^T, \mathbf{a}_i^T)^T$ and satisfy $E(\mathbf{b}_i) = \mathbf{0}$ and $\text{var}(\mathbf{b}_i) = \mathbf{D}$. Let $\hat{\boldsymbol{\beta}}_i$ be an individual estimator for $\boldsymbol{\beta}_i$ that is assumed to satisfy the approximate model

$$E(\hat{\boldsymbol{\beta}}_i | \mathbf{x}_i) = \mathbf{A}_i \boldsymbol{\beta}, \quad \text{var}(\hat{\boldsymbol{\beta}}_i | \mathbf{x}_i) = \mathbf{D} + \mathbf{C}_i. \quad (1)$$

Let the “Standard Two Stage” estimators for $\boldsymbol{\beta}$ and \mathbf{D} be as on page 438, namely,

$$\hat{\boldsymbol{\beta}}_{STS} = \left(\sum_{i=1}^m \mathbf{A}_i^T \mathbf{A}_i \right)^{-1} \sum_{i=1}^m \mathbf{A}_i^T \hat{\boldsymbol{\beta}}_i,$$

$$\hat{\mathbf{D}}_{STS} = (m-1)^{-1} \sum_{i=1}^m (\hat{\boldsymbol{\beta}}_i - \mathbf{A}_i \hat{\boldsymbol{\beta}}_{STS})(\hat{\boldsymbol{\beta}}_i - \mathbf{A}_i \hat{\boldsymbol{\beta}}_{STS})^T.$$

(a) Find the expectation of $\hat{\mathbf{D}}_{STS}$ assuming that (1) holds exactly, and argue that $\hat{\mathbf{D}}_{STS}$ is not an unbiased estimator for \mathbf{D} in general

(b) In the special case where $\mathbf{A}_i = \mathbf{I}_b$ for all i , find the value of the bias; i.e., find the matrix \mathbf{B} , where $E(\hat{\mathbf{D}}_{STS}) = \mathbf{D} + \mathbf{B}$. Will $\hat{\mathbf{D}}_{STS}$ over- or underestimate \mathbf{D} in general under these conditions?

2. Again consider a two-stage subject-specific model of the form

$$E(\mathbf{Y}_i | \mathbf{z}_i, \boldsymbol{\beta}_i) = \mathbf{f}_i(\mathbf{z}_i, \boldsymbol{\beta}_i), \quad \text{var}(\mathbf{Y}_i | \mathbf{z}_i, \boldsymbol{\beta}_i) = \mathbf{R}_i(\boldsymbol{\beta}_i, \boldsymbol{\gamma}, \mathbf{z}_i),$$

$$\boldsymbol{\beta}_i = \mathbf{A}_i \boldsymbol{\beta} + \mathbf{b}_i,$$

where \mathbf{b}_i are independent of $\mathbf{x}_i = (\mathbf{z}_i^T, \mathbf{a}_i^T)^T$ and satisfy $\mathbf{b}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{D})$. Let $\hat{\boldsymbol{\beta}}_i$ be an individual estimator for $\boldsymbol{\beta}_i$ that is assumed to satisfy the approximate model given in (15.20) and (15.21) on pages 437–438 of the notes for some \mathbf{C}_i ; i.e.,

$$\hat{\boldsymbol{\beta}}_i \approx \mathbf{A}_i \boldsymbol{\beta} + \mathbf{b}_i + \mathbf{e}_i, \quad \mathbf{e}_i | \mathbf{b}_i, \mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_i). \quad (2)$$

Of course, (2) is a special case of a linear mixed effects model, as discussed on pages 443–444. Here, we will treat the estimated covariance matrices \mathbf{C}_i as known.

As mentioned on page 444, one way to fit a linear mixed models is via the so-called EM algorithm. In this problem, you will derive the steps (ii) and (iii) of the EM algorithm given on pages 444–445.

The usual EM algorithm may be described as follows. Suppose \mathbf{U} and \mathbf{b} have joint density $p(\mathbf{U}, \mathbf{b}; \boldsymbol{\theta})$, depending on a parameter $\boldsymbol{\theta}$, which may be factorized as $p(\mathbf{U}|\mathbf{b}; \boldsymbol{\theta})p(\mathbf{b}; \boldsymbol{\theta})$. Suppose further that we only observe \mathbf{U} . Then a likelihood for $\boldsymbol{\theta}$ may be based on the marginal density

$$p(\mathbf{U}; \boldsymbol{\theta}) = \int p(\mathbf{U}|\mathbf{b}; \boldsymbol{\theta})p(\mathbf{b}; \boldsymbol{\theta}) d\mathbf{b}. \quad (3)$$

We wish to maximize (3) in $\boldsymbol{\theta}$ to obtain an estimator $\hat{\boldsymbol{\theta}}$. The EM algorithm is a numerical technique to carry out this task. Given a starting value $\hat{\boldsymbol{\theta}}_{(0)}$, one iterates the following steps. At iteration $k + 1$, with current estimate $\hat{\boldsymbol{\theta}}_{(k)}$

- (i) *E-step.* Find $Q(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}_{(k)}) = E\{\log p(\mathbf{U}, \mathbf{b}; \boldsymbol{\theta})|\mathbf{U}; \hat{\boldsymbol{\theta}}_{(k)}\}$. Here, the expectation is taken with respect to the conditional distribution of \mathbf{b} given \mathbf{U} , which also depends on $\boldsymbol{\theta}$, which is replaced by $\hat{\boldsymbol{\theta}}_{(k)}$ when evaluating the integral. Thus,

$$Q(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}_{(k)}) = \int \log p(\mathbf{U}, \mathbf{b}; \boldsymbol{\theta})p(\mathbf{b}|\mathbf{U}; \hat{\boldsymbol{\theta}}_{(k)}) d\mathbf{b}. \quad (4)$$

It is very important to note that the integrand is a function of $\boldsymbol{\theta}$, while the conditional density $p(\mathbf{b}|\mathbf{U}; \hat{\boldsymbol{\theta}}_{(k)})$ is evaluated at $\hat{\boldsymbol{\theta}}_{(k)}$. Thus, $Q(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}_{(k)})$ is a function of $\boldsymbol{\theta}$ only through the dependence of the integrand on $\boldsymbol{\theta}$.

- (ii) *M-step.* Maximize $Q(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}_{(k)})$ in $\boldsymbol{\theta}$. Note that $\hat{\boldsymbol{\theta}}_{(k)}$ in (4) is held fixed; the maximization is only with respect to $\boldsymbol{\theta}$ in the integrand.
 (iii) Set $k = k + 1$ and return to (i).

The algorithm is iterated until convergence.

It is possible to show that iterating (i)–(iii) to convergence indeed results in a final iterate $\hat{\boldsymbol{\theta}}$ that maximizes (3). In fact, it may be shown that the value of (3) evaluated at each successive iterate k increases, so that each step results in getting closer to the maximum value of (3).

The EM algorithm on pages 444–445 is a slight variation of this scheme, where $\boldsymbol{\theta}$ consists of \mathbf{D} and $\boldsymbol{\beta}$, but $\boldsymbol{\beta}$ is estimated separately at each iteration, as you will now show. Identify $\mathbf{U} = (\hat{\boldsymbol{\beta}}_1^T, \dots, \hat{\boldsymbol{\beta}}_m^T)^T$ and $\mathbf{b} = (\mathbf{b}_1^T, \dots, \mathbf{b}_m^T)^T$. The \mathbf{b}_i are assumed independent in the two-stage model; in the following, take the $\hat{\boldsymbol{\beta}}_i$ to be independent as well (although they may depend on, for example, a pooled variance parameter estimate).

- (a) Let $\hat{\boldsymbol{\theta}}_{(k)} = (\hat{\boldsymbol{\beta}}_{(k)}, \hat{\mathbf{D}}_{(k)})$ from the k th iteration. Assuming that the model (2) is exact and the above conventions on independence, find $Q(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}_{(k)})$. Hint: Denote the expression for the conditional mean of \mathbf{b}_i given $\hat{\boldsymbol{\beta}}_i$ evaluated at $\hat{\boldsymbol{\theta}}_{(k)}$ as $\hat{\mathbf{b}}_i$ and note that, for random vector \mathbf{Y} with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, $E(\mathbf{Y}^T \mathbf{A} \mathbf{Y}) = \text{tr}(\mathbf{A} \boldsymbol{\Sigma}) + \boldsymbol{\mu}^T \mathbf{A} \boldsymbol{\mu}$ for matrix \mathbf{A} .

- (b) Show that $\tilde{\boldsymbol{\beta}}_{i,(k+1)} = \mathbf{A}_i \hat{\boldsymbol{\beta}}_{(k)} + \hat{\mathbf{b}}_i$, where $\hat{\mathbf{b}}_i$ is as defined in the hint in (a).

- (c) From page 442, the usual estimator for $\boldsymbol{\beta}$ if we solved the estimating equations (15.30) and (15.31) directly would have the “GLS” form given in the second-to-last equation on this page. Show that, with \mathbf{D} fixed at $\hat{\mathbf{D}}_{(k)}$, the expression for $\hat{\boldsymbol{\beta}}_{(k+1)}$ given at the top of page 435 may be written alternatively in the “GLS” form if we take $\hat{\boldsymbol{\beta}}_{(k+1)} = \hat{\boldsymbol{\beta}}_{(k)}$ (so if we were at convergence of the algorithm).

- (d) Show that $\hat{\mathbf{D}}_{(k+1)}$ given at the top of page 445 has the same form as the value of \mathbf{D} maximizing $Q(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}_{(k)})$ you found in (a), except that $\hat{\boldsymbol{\beta}}_{(k+1)}$ is replaced by $\hat{\boldsymbol{\beta}}_{(k)}$ in the second term on page 445. Hint: It may be shown that, for two square matrices \mathbf{A} and \mathbf{B} , $\log |\mathbf{A}| + \text{tr}(\mathbf{A}^{-1} \mathbf{B})$ is minimized as a function of \mathbf{A} at $\mathbf{A} = \mathbf{B}$.

3. Consider the derivation of the approximate marginal likelihood contribution (15.53) for i via Laplace's approximation for the general subject-specific model with first stage

$$E(\mathbf{Y}_i | \mathbf{x}_i, \mathbf{b}_i) = \mathbf{f}_i(\mathbf{x}_i, \boldsymbol{\beta}, \mathbf{b}_i), \quad \text{var}(\mathbf{Y}_i | \mathbf{x}_i, \mathbf{b}_i) = \mathbf{R}_i(\boldsymbol{\gamma}, \mathbf{x}_i),$$

so that the within-individual covariance matrix does not depend on $\boldsymbol{\beta}$ or \mathbf{b}_i , as on page 445. Assume as on page 445 that $\mathbf{b}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{D})$ independently of all covariates and the \mathbf{Y}_i are conditionally normal as in the fourth paragraph of page 445.

Via tedious matrix algebra as described on page 446, show that (15.51) may be rewritten as (15.53) on page 446. Thus, you will have shown that the contribution from individual i to the marginal likelihood may be approximated by a normal density with mean and covariance matrix as given in the first bullet on page 446.

Note: This is a “bread and butter” calculation that anyone who is interested in mixed effects models will have to do at some point!

Hints: You will probably have to look up some relationships having to do with determinants. Also, it will behoove you to define

$$\mathbf{u}_i = \mathbf{Y}_i + \mathbf{Z}_i(\mathbf{x}_i, \boldsymbol{\beta}, \hat{\mathbf{b}}_i) \hat{\mathbf{b}}_i.$$

You can use shorthand notation such as $\mathbf{Z}_i = \mathbf{Z}_i(\mathbf{x}_i, \boldsymbol{\beta}, \hat{\mathbf{b}}_i)$, $\mathbf{f}_i = \mathbf{f}_i(\mathbf{x}_i, \boldsymbol{\beta}, \hat{\mathbf{b}}_i)$, etc.