Likelihood, Prior and Posterior Distributions for Bayesian Exchangeable Models with Respect to Finite Populations

Ed Stanek and Julio Singer

Introduction

We review Ericson’s (1969, 1988) development of Bayesian methods. We begin by discussing what is meant by a finite population and a sample, and notation for each. Subsequently, we discuss the likelihood, prior, and posterior distributions relative to a target parameter for a population in these settings.

The Population

We define a finite population, \( f = \{1, 2, ..., N\} \), as a set of \( N \) subjects labeled \( j = 1, ..., N \). We represent response for subject \( j \) by \( y^{(j)} \), and an auxiliary variable for subject \( j \) by \( a^{(j)} \). The response and auxiliary variables could possible be vector valued; and the observed value of these variables could include measurement error. We assume that \( y^{(j)} \) and \( a^{(j)} \) are measured without error. We summarize data for the population by the set of label, response, auxiliary variable triples, \( \{j \ y^{(j)} \ a^{(j)}\} \) for \( j = 1, ..., N \).

The subjects in population \( f \) may be considered to be a realization of a sample of \( N \) subjects from a superpopulation of \( \mathbb{N} > N \) subjects. Let the labels for subjects in the superpopulation be represented by \( v = 1, ..., \mathbb{N} \). With this understanding, the labels in population \( f \) replace the superpopulation labels in order \( v_1 < v_2 < ... < v_N \) by the labels \( j = 1, ..., N \).

Let \( j = (1 \ 2 \ \cdots \ N)' \) represent a vector of labels for population \( f \),
\[
y^{(j)} = \begin{pmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(N)} \end{pmatrix}
\]
represent a corresponding response vector, and
\[
a^{(j)} = \begin{pmatrix} a^{(1)} \\ a^{(2)} \\ \vdots \\ a^{(N)} \end{pmatrix}
\]
a corresponding auxiliary variable vector. The label, response, auxiliary variable information for the population is contained in \( \{j \ y^{(j)} \ a^{(j)}\} \). The order of the subjects in \( \{j \ y^{(j)} \ a^{(j)}\} \) is not important and conveys no additional information. This means that same label, response, auxiliary variable information is contained in a permutation of rows of \( \{j \ y^{(j)} \ a^{(j)}\} \) given by \( \mathbf{r}_p(j \ y^{(j)} \ a^{(j)}) \), where \( p = 1, ..., P = N! \) represents an \( N \times N \) permutation matrix.

The Sample

Comment: This seems good. The population is a set of subjects and the data for the population is a set of triplets. We know the population without knowing the data (except for the labels and possibly, the values of the auxiliary variables); otherwise there would be nothing to do.

Comment: I believe this is done to simplify notation for the population.

Deleted: C10ed05v3.doc
Deleted: 4/28/2010 5:47 PM
Deleted: 4/28/2010 3:12 PM
Inserted: 4/28/2010 5:47 PM
Ericson defines a sample by the pair \((s, y_s)\) where \(s = \{i_1, i_2, \ldots, i_n\} \subseteq \eta\) is the set of distinct subject labels comprising the sample, and \(y_s = (y_{i_1}, y_{i_2}, \ldots, y_{i_n})^t\) is a sequence of responses. For definiteness, Ericson assumes that \(i_1 < i_2 < \ldots < i_n\). With this assumption, the relative position of a subject in \(y_s\) is determined. Let the labels for the subjects in sample \(s\) (where \(i_1 < i_2 < \ldots < i_n \)) be given by \(1, 2, \ldots, n\), and response for subject \(i\) in sample \(s\) of population \(f\) be represented by \(y^{(s)}_i\) where \(y^{(s)} = (y^{(s)}_1, y^{(s)}_2, \ldots, y^{(s)}_n)^t\), with the auxiliary variable represented by \(a^{(s)}_i\), where \(a^{(s)} = (a^{(s)}_1, a^{(s)}_2, \ldots, a^{(s)}_n)^t\).

The label, response, auxiliary variable information for sample \(s\) is contained in \((i, y^{(s)}, a^{(s)})\). The order of the subjects in \((i, y^{(s)}, a^{(s)})\) is not important and conveys no additional information. This means that same label, response, auxiliary variable information is contained in a permutation of rows of \((i, y^{(s)}, a^{(s)})\) given by \(b^t(i, y^{(s)}, a^{(s)})\), where \(b_1, b_2, \ldots, b_n\) represent an \(n \times n\) permutation matrix.

We define the relationship between the information in sample \(s\) and the population \(f\) by \(\delta_s\) such that

\[
(i, y^{(s)}, a^{(s)}) = \delta_s(j, y^{(f)}, a^{(f)}).
\]

The matrix \(\delta_s\) is an \(N \times n\) matrix where all elements are either zero or one, and all columns sum to one. The columns of \(\delta_s\) are determined such that the order of the subject labels in the sample is the same as the order of the subject labels in the population (i.e., \(j_1 < j_2 < \ldots < j_n\)). Notice that previously the subscript \(h\) was used for \(\delta\) instead of \(s\).

**An Example**

We discuss an example to make these ideas clear. Suppose that the label, response, auxiliary information in population \(f\) is given by

\[
\begin{bmatrix}
Roseland \\
Diety \\
Lords
\end{bmatrix}
\begin{bmatrix}
y_{Roseland} \\
y_{Diety} \\
y_{Lords}
\end{bmatrix}
\begin{bmatrix}
a_{Roseland} \\
a_{Diety} \\
a_{Lords}
\end{bmatrix}
= \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}
\begin{bmatrix} y^{(f)}_1 \\ y^{(f)}_2 \\ y^{(f)}_3 \end{bmatrix}
= \begin{bmatrix} j_1 & j_2 & j_3 \end{bmatrix}
\begin{bmatrix} y^{(f)} \\ a^{(f)} \end{bmatrix},
\]

and the label, response, auxiliary information in sample \(s\) is given by

\[
\begin{bmatrix}
Roseland \\
Lords
\end{bmatrix}
\begin{bmatrix}
y_{Roseland} \\
y_{Lords}
\end{bmatrix}
\begin{bmatrix}
a_{Roseland} \\
a_{Lords}
\end{bmatrix}
= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}
\begin{bmatrix} y^{(s)}_1 \\ y^{(s)}_2 \end{bmatrix}
= \begin{bmatrix} i_1 & i_2 \end{bmatrix}
\begin{bmatrix} y^{(s)} \\ a^{(s)} \end{bmatrix}.
\]

We define the relationship between the sample and the population by \(\delta_s = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\) such that

\[
\begin{bmatrix} y^{(s)} \\ a^{(s)} \end{bmatrix} = \delta_s
\begin{bmatrix} y^{(f)} \\ a^{(f)} \end{bmatrix}.
\]
The population is defined by $f = \{ \text{Diety, Roseland, Lords} \}$, while the sample is defined by $s = \{ \text{Lords, Roseland} \}$. Since both the population and sample are sets, when the information is represented by a sequence (as in the matrix $\begin{pmatrix} i & y^{(i)} & a^{(i)} \end{pmatrix}$ or $\begin{pmatrix} i & y^{(i)} & a^{(i)} \end{pmatrix}$), the ordering of the rows in the matrix does not matter. For example, we could represent the information in the sample $s$ with response $\begin{pmatrix} y_{\text{Roseland}} & a_{\text{Roseland}} \end{pmatrix}$. If we wish to represent the sample label, response, auxiliary variable information in the order given by the elements in this set, we define $b_s = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}$ such that

$$b_s^t \begin{pmatrix} i & y^{(i)} & a^{(i)} \end{pmatrix} = \begin{pmatrix} \text{Lords} & y_{\text{Lords}} & a_{\text{Lords}} \\ \text{Roseland} & y_{\text{Roseland}} & a_{\text{Roseland}} \end{pmatrix}.$$ 

Similarly, suppose that the response for population $f$ is given by $\{ \begin{pmatrix} y_{\text{Diety}} & a_{\text{Diety}} \\ y_{\text{Roseland}} & a_{\text{Roseland}} \end{pmatrix} \}$. If we wish to represent the label, response, auxiliary variable information in a matrix using the order given by the elements in the set $f$, we define $r_p = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ such that

$$r_p^t \begin{pmatrix} j & y^{(j)} & a^{(j)} \end{pmatrix} = \begin{pmatrix} \text{Diety} & y_{\text{Diety}} & a_{\text{Diety}} \\ \text{Roseland} & y_{\text{Roseland}} & a_{\text{Roseland}} \\ \text{Lords} & y_{\text{Lords}} & a_{\text{Lords}} \end{pmatrix}.$$ 

The Probability of a Sample and the Likelihood

The term $p_X(s)$ used by Ericson is the probability of sample set $s$ associated with a sample design. Let all possible sample sets be given by $\eta$. Then $p_X(s) \geq 0$ and $\sum_{\eta} p_X(s) = 1$.

These probabilities may be a function of $X$, the auxiliary variables (which we represent by $\begin{pmatrix} a^{(i)} \end{pmatrix}$). As a result, we represent $p_X(s)$ by $p_{X\mid s}(s)$. Ericson’s likelihood function is given by

$$l(\{Y \mid s, y_s\}) = p_X(s) k \text{ for } Y \mid S(Y) = y_s = 0 \text{ otherwise}$$

Using the notation previously defined, the likelihood function corresponds to
This likelihood function is non-zero for a given population, $f$, and sample, $s$. Ericson states (p215, 1988) that the likelihood function $\pi$ is constant for all $Y \ldots$, which we interpret as meaning that regardless of which population is identified, (what the value of $f$ is), the likelihood is constant. The term $y_I$ is not defined in the likelihood. We discuss the definition of this term next.

**Definition of $y_I$.**

We define $y_I$ in reference to a set of $n$ triples,

$$\{(v_1, z_{v_1}, a_{v_1}), (v_2, z_{v_2}, a_{v_2}), \ldots, (v_n, z_{v_n}, a_{v_n})\},$$

where the label $v_i$ for $i = 1, \ldots, n$ identifies the subjects in the set. These labels may correspond to labels of subjects in a superpopulation (where $v_i$ are a subset of $\nu = 1, \ldots, N$), or they may correspond to labels of subjects in a population (where $v_i = j_i$ are a subset of $j = 1, \ldots, N$). Since $y_I$ is a vector, it is defined as a sequence.

We assume for definiteness that the labels are placed in ascending order, such that $v_1 < v_2 < \ldots < v_n$, and define $y_I$ as \( (z_{v_1}, z_{v_2}, \ldots, z_{v_n}) \). If the labels $v_1 < v_2 < \ldots < v_n$ (or $j_1 < j_2 < \ldots < j_n$) are given explicitly, we could use the labels to identify the sample set $s$.

Replacing these labels by $i = 1, \ldots, n$, we define $y_I$ as

$$y_I = (y_1, y_2, \ldots, y_n) = (z_{v_1}, z_{v_2}, \ldots, z_{v_n}).$$

The definition may imply some knowledge about the population (or superpopulation), which we illustrate next.

If the labels relate to the superpopulation, then the way in which population labels are assigned relative to the superpopulation labels may be important. Suppose, for example, that the population is a simple random sample without replacement of $N$ subjects from $N > N$ subjects in the superpopulation, and the labels $j = 1, \ldots, N$ are assigned to subjects in the population after placing population subjects in the set in increasing superpopulation label order. In such a setting, $y_I$ could represent the vector of sample responses from the population. On the other hand, suppose the labels of the subjects in the population, $j = 1, \ldots, N$, are assigned after placing the subjects in descending superpopulation label order. The vector of sample responses, $y_I$ could not have arisen from a population defined in this manner, since the response in $y_I$ is for subjects placed in increasing superpopulation label order. This implies that defining response for a sample set as a sequence of responses relative to labels in a larger set (either a population or superpopulation) places restrictions on the relationship between the labels for the population and superpopulation.
Note that both the population and the superpopulation may be completely artificial (hypothetical), and we may not know the subjects or labels of the subjects in either. In such a setting, we will not know the order of the subjects in the population or superpopulation. In fact, we do not have to imagine either a population or a superpopulation exists in order to define $y_j$. Consider a set $\{(x_w, a_w), (x_1, a_1), \ldots, (x_n, a_n)\}$ where the pairs $(x_w, a_w)$ indicate the response, auxiliary variable pair for a subject. The label distinguishes the subjects in the sample set, but does not specify an order. Let us place the subjects in some (arbitrary) order, and label this order via $i = 1, \ldots, n$. Using this idea, we represent the label, response, auxiliary variable triples for the sample set by $\{(n, x_n, a_n), (1, x_1, a_1), \ldots, (n, x_n, a_n)\}$. Arrange the subject labels in increasing order by $i = 1, \ldots, n$. The vector of response for the sample, $y_j$, can be defined by

$$y_j = (y_1, y_2, \ldots, y_n) = (x_1, x_2, \ldots, x_n)^T.$$ 

No population or superpopulation need be defined.

Since the sample is defined by the set of subjects, $s = \{1, 2, \ldots, n\}$, and data for the sample is defined by the set of triples, $\{(1, y_1, a_1), (2, y_2, a_2), \ldots, (n, y_n, a_n)\}$, different sequences of subject’s response may describe response for the sample. These different sequences do not define response for a different sample set, even though the sequence may not equal $y_j$. By specifying the appropriate re-ordering of a sequence of responses via a permutation matrix $b$, it is possible to re-order a response sequence to obtain the same response vector as in $y_j$.

**Prior Distribution**

Ericson represents the probability associated with a vector of parameters for a finite population, $f = \{1, 2, \ldots, N\}$, as $p(Y | X)$. This is the prior probability, and is used along with the accompanying realizations of $Y$ to specify the prior distribution. Ericson (1969, p197) considers the simplest useful class of prior distributions to be one where the random variables in $Y$ are exchangeable. We first discuss this idea, and what is mean by conditioning on $X$ using the notation that we have introduced.

Recall that we have represented the label, response, auxiliary variable triples for population $f$ in a matrix $\begin{pmatrix} j & y^{(f)} & a^{(f)} \end{pmatrix}$ and noted that the same label, response, auxiliary variable information is contained in a permutation of $\begin{pmatrix} j & y^{(f)} & a^{(f)} \end{pmatrix}$ given by

$$r^{(p)} \begin{pmatrix} j & y^{(f)} & a^{(f)} \end{pmatrix} = \begin{pmatrix} j^{(p)} & y^{(p)} & a^{(p)} \end{pmatrix},$$

where $j^{(p)} = r^{(p)} j$ such that $j^{(p)} = \begin{pmatrix} 1^{(p)} & 2^{(p)} & \cdots & N^{(p)} \end{pmatrix}$, $y^{(p)} = r^{(p)} y^{(f)}$ such that

$$y^{(p)} = \begin{pmatrix} y_1^{(p)} & y_2^{(p)} & \cdots & y_N^{(p)} \end{pmatrix},$$

and $a^{(p)} = r^{(p)} a^{(f)}$ such that

$$a^{(p)} = \begin{pmatrix} a_1^{(p)} & a_2^{(p)} & \cdots & a_N^{(p)} \end{pmatrix}.$$ 

The sequence of responses for subjects in order $j$ of population $f = \{1, 2, \ldots, N\}$ is $y^{(f)}$. When subjects are placed in a different order, the sequence of responses may be represented by $y^{(p)}$. Let us define an indicator random variable, $I^{(f)}_p$, Comment: This is viewed as a random variable and it is used in the expression of the likelihood. This is why I think we should have something similar in our expression for the likelihood.
which has a value of one if permutation $p$ is selected, and zero otherwise. We define $\xi$ as expectation over $I_p^{(f)}$, and $E_{I_p^{(f)}}(I_p^{(f)}) = \pi_p^{(f)}$.

We define $Y^{(f)} = \sum_p I_p^{(f)} y^{(p)}$ and $X = a^{(f)}$. For a realized value of $I_p^{(f)} = 1$, we define

$$ (Y^{(f)} | I_p^{(f)} = 1) = \sum_p I_p^{(f)} y^{(p)} y_p^{(f)} = 1 $$

as the realized population response. Associated with the realization of $I_p^{(f)} = 1$ is a realized sequence of labels, $j^{(p)}$ and a realized sequence of auxiliary values, $a^{(p)}$. We refer to the matrix $(j^{(p)} y^{(p)} a^{(p)})$ as the realized population (corresponding to the realization $I_p^{(f)} = 1$).

The prior distribution is the set of probabilities associated with the realized populations. Using this definition, we represent Ericson’s $p(Y | X)$ by $p(Y^{(f)} | a^{(f)})$. It is not necessary to have auxiliary variables to define such a prior distribution. Also notice that by defining the prior distribution in this way, the elements in the realization of $Y^{(f)}$ (given by $y^{(p)}$) are not necessarily in the same order as the elements in $a^{(f)}$. The order of the elements in $y^{(p)}$ is given by $j^{(p)}$, while the order of the elements in $a^{(f)}$ is $j$. The order differs since we have conditioned on $a^{(f)}$, not $y^{(p)}$. There is an advantage to this since $a^{(f)}$ does not depend in any way on the prior distribution.

**Posterior Distribution**

For any given prior distribution, Ericson defines the posterior distribution by (see page 214, 1988)

$$ p(Y | (s, y_s)) = \frac{p(Y | X)}{p_{a(Y)}(y_s | X)} \text{ for } Y | S(Y) = y_s $$

$$ = 0 \text{ otherwise} $$

where $p_{a(Y)}(y_s | X)$ is the marginal prior on $S(Y) = y_s$. We discuss the posterior distribution defined by Ericson relative to our notation.

First, note that $p(Y | X)$ correspond to $p(Y^{(f)} | a^{(f)})$ where $Y^{(f)} = \sum_p I_p^{(f)} y^{(p)}$. Now $p(Y | X)$ is given by $p(Y^{(f)} = y^{(p)} | a^{(f)}) = \pi_p^{(f)}$. In words, this is the probability of a particular permutation of the response vector, $y^{(f)}$, identified by $I_p^{(f)} = 1$ and given by $y^{(p)}$.

Let us introduce the notation

$$ (i, y^{(p)}, a^{(p)}) = \delta^{(p)}(j^{(p)} y^{(p)} a^{(p)}) $$

Comment: If our superpopulation is the set of all permutations of a given population, the superscript (f) could be eliminated.
for sample \( h \) from the realized population \( p \), where \( h \) is a set of \( n \) subjects, 
\[ h = \{ j^{(p)}, j^{(p)}, ..., j^{(p)} \} \]

such that \( j^{(p)} < j^{(p)} < ... < j^{(p)} \). The matrix \( \delta^{(p)}_u \) is an \( N \times n \) matrix where all elements are either zero or one, and all columns sum to one. Explicitly, \( \delta^{(p)}_u = \delta^{(p)}(r) \) while \( j^{(p)} = r^j \). The columns of \( \delta^{(p)}_u \) are determined such that the order of the subject labels in the sample is the same as the order of the subject labels in the realized population (i.e., \( j^{(p)} = \{ ^1(p), 2^{(p)}, ... , N^{(p)} \} \)). For a realized population with response vector \( y^{(p)} \), the response vector for the subjects identified by \( \delta^{(p)}_u j \) is \( y^{(p)} \). When considering a realization of \( Y^{(f)} = \sum_{j=1}^{n} j^{(f)} \), the vector of sample responses for a realized population comparable to Ericson’s \((s, \ y_j)\) is \( (h \ y^{(p)}) \).

Let \( S^{(p)}_k \) represent an indicator random variable associated with sample \( h \), such that \( P\{ S^{(p)}_k = 1 \} = \pi^{(p)}_k \). We represent a randomly selected sample response vector as
\[ y^{(p)} = \sum_{j=1}^{n} S^{(p)}_k \delta^{(p)}_u (y^{(p)}) \] for the realized population \( p \). Unconditionally, we represent a randomly selected sample response vector as 
\[ y^{(f)} = \sum_{p=1}^{P} \sum_{h=1}^{H} \sum_{j=1}^{n} j^{(f)} \pi^{(p)}_k \delta^{(p)}_u (y^{(p)}) \].

We discuss the additional statement, “when \( Y \mid S(Y) = y_j \)” next. First, notice that in the context of a prior distribution, we represent Ericson’s \( y_j \) by \( y^{(p)} \) (not \( y_j \)). We represent the function \( S(Y) \) by \( Y^{(f)} \). As a result, we express “when \( Y \mid S(Y) = y_j \)” as “when \( Y^{(f)} = y^{(p)} \) where we require that the subjects in \( h \) and \( s \) are the same as in the sample set \( s = \{ i_1, i_2, ..., i_n \} \). When the set of subjects in \( h \) and \( s \) are the same, there will be a permutation matrix \( b \) such that \( y^{(p)} = b^j \). We include this condition to indicate that the same subjects are in the two sample sets. For different realized populations, i.e., different \( p \neq p \), the sample set \( s \) will be identified by different \( \delta^{(p)}_u \).

Using these expressions, \( P(Y \mid X) \) when \( Y \mid S(Y) = y_j \) is given by \( P\{ Y^{(f)} \mid Y^{(p)} = 1, S^{(p)}_k = 1 \} = \pi^{(p)}_k \sum_{h^{(p)} \text{ to be the same subjects as in the sample set } s} b^j \). As a result, the numerator in the posterior distribution corresponding to \( P(Y \mid X) \) when \( Y \mid S(Y) = y_j \) is given by
\[ P\{ Y^{(f)} \mid Y^{(p)} = 1, S^{(p)}_k = 1 \} = \pi^{(p)}_k \sum_{h^{(p)} \text{ to be the same subjects as in the sample set } s} b^j \].
It remains to express \( p_{\text{Y|X}}(y_j | X) \).

Ericson’s expression \( p_{\text{Y|X}}(y_j | X) \) represents the marginal prior on \( S(Y) = y_j \). This represents the sum over the prior distribution of
\[
\sum_{p=1}^{p} p(Y_j^f | Y_j^f = y^{\text{obs}}) = b_j^f \text{Y}_j^f \), so that
\[
p_{\text{Y|X}}(y_j | X) \text{ is given by } \sum_{p=1}^{p} p(Y_j^f | Y_j^f = y^{\text{obs}}) = \sum_{p=1}^{p} \pi_p \sum_{i_{\text{obs}}} p_i^r \).
\]

Combining these expressions, the posterior distribution is given by
\[
p(Y_j^f | h, y^{\text{obs}}) = \frac{\pi^r_i \sum_{i_{\text{obs}}} p_i^r}{\sum_{i_{\text{obs}}} p_i^r} \text{ for } y^{\text{obs}} = b_j^f \text{Y}_j^f ,
\]
\[
= 0 \quad \text{otherwise}
\]

We discuss a simple example to try to make these ideas more clear. Consider a population which we index by \( f = 3 \), where the label, response, auxiliary information in population \( f \) is given by
\[
\begin{array}{ccc}
\text{Roseland} & 1 & 5 & 6 \\
\text{Victory} & Y_{\text{Roseland}} & a_{\text{Roseland}} \\
\text{Lords} & Y_{\text{Victory}} & a_{\text{Victory}} \\
\end{array}
\]
\[
\begin{array}{ccc}
\text{Roseland} & 1 & 5 & 6 \\
\text{Victory} & Y_{\text{Roseland}} & a_{\text{Roseland}} \\
\text{Lords} & Y_{\text{Victory}} & a_{\text{Victory}} \\
\end{array}
\]

Let the different orderings of the elements in the population correspond to the possible realized populations (Table 1). The particular population, \( f = 3 \), could have been selected from a superpopulation. In any case, the population corresponds to a particular set of subjects.

Table 1. Realized Populations defined as Permutations of a Set of \( N = 3 \) subjects with simple label, response pairs when \( f = 3 \).

<table>
<thead>
<tr>
<th>( pf = 13 )</th>
<th>( pf = 23 )</th>
<th>( pf = 33 )</th>
<th>( pf = 43 )</th>
<th>( pf = 53 )</th>
<th>( pf = 63 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( f^1 ) ( (v \ z_1) )</td>
<td>( f^2 ) ( (v \ z_1) )</td>
<td>( f^3 ) ( (v \ z_1) )</td>
<td>( f^4 ) ( (v \ z_1) )</td>
<td>( f^5 ) ( (v \ z_1) )</td>
<td>( f^6 ) ( (v \ z_1) )</td>
</tr>
<tr>
<td>\begin{pmatrix} 1 \ 5 \ 2 \ 6 \ 3 \ 2 \end{pmatrix}</td>
<td>\begin{pmatrix} 1 \ 5 \ 2 \ 6 \ 3 \ 2 \end{pmatrix}</td>
<td>\begin{pmatrix} 1 \ 5 \ 2 \ 6 \ 3 \ 2 \end{pmatrix}</td>
<td>\begin{pmatrix} 1 \ 5 \ 2 \ 6 \ 3 \ 2 \end{pmatrix}</td>
<td>\begin{pmatrix} 1 \ 5 \ 2 \ 6 \ 3 \ 2 \end{pmatrix}</td>
<td>\begin{pmatrix} 1 \ 5 \ 2 \ 6 \ 3 \ 2 \end{pmatrix}</td>
</tr>
<tr>
<td>( f^j ) ( y_j^{(r)} )</td>
<td>( f^j ) ( y_j^{(r)} )</td>
<td>( f^j ) ( y_j^{(r)} )</td>
<td>( f^j ) ( y_j^{(r)} )</td>
<td>( f^j ) ( y_j^{(r)} )</td>
<td>( f^j ) ( y_j^{(r)} )</td>
</tr>
<tr>
<td>\begin{pmatrix} 1 \ 5 \ 2 \ 6 \ 3 \ 2 \end{pmatrix}</td>
<td>\begin{pmatrix} 1 \ 5 \ 2 \ 6 \ 3 \ 2 \end{pmatrix}</td>
<td>\begin{pmatrix} 1 \ 5 \ 2 \ 6 \ 3 \ 2 \end{pmatrix}</td>
<td>\begin{pmatrix} 1 \ 5 \ 2 \ 6 \ 3 \ 2 \end{pmatrix}</td>
<td>\begin{pmatrix} 1 \ 5 \ 2 \ 6 \ 3 \ 2 \end{pmatrix}</td>
<td>\begin{pmatrix} 1 \ 5 \ 2 \ 6 \ 3 \ 2 \end{pmatrix}</td>
</tr>
</tbody>
</table>
Notice that associated with the set \( f = 3 \) are the auxiliary variables

\[
\begin{pmatrix}
\gamma \\
a_i
\end{pmatrix},
\begin{pmatrix}
R \\
6
\end{pmatrix},
\begin{pmatrix}
L \\
6
\end{pmatrix},
\begin{pmatrix}
V \\
10
\end{pmatrix}
\]

such that \( A^f = (6, 6, 10) \). This vector of auxiliary variables is the same for all realized populations listed in Table 1.

Let the sample set of subjects be given by \( s = \{\text{Roseland, Lords}\} \). Note that \( y_j = (y_{\text{Roseland}}, y_{\text{Lords}}) \). Let us define \( p_{hs}^{f} = \delta_{h}^{f} (y_{s}) \) for \( f = 3 \) where \( Y_{s}^{f} = \delta_{h}^{f} (y_{s}) \), or \( Y_{s}^{f} = \delta_{h}^{f} (y_{s}) \), and for each of these realized populations, the sample response vector is given as \( \binom{5}{6} \). For the sample vectors \( Y_{s}^{f} = \delta_{h}^{f} (y_{s}) \), \( Y_{s}^{f} = \delta_{h}^{f} (y_{s}) \), and \( Y_{s}^{f} = \delta_{h}^{f} (y_{s}) \), the sample response vector is given as \( \binom{6}{5} \). We list these samples for the realized populations in Table 2.

**Table 2.** Samples from Realized Populations defined as Permutations of a Set of \( N = 3 \) subjects with simple label, response pairs when \( f = 3 \) where \( Y_{s}^{f} = b_i \) and \( y_j = (y_{\text{Roseland}}, y_{\text{Lords}}) \).

<table>
<thead>
<tr>
<th>Prior</th>
<th>Sample</th>
<th>( y_{(12)} )</th>
<th>( y_{(21)} )</th>
<th>( y_{(31)} )</th>
<th>( y_{(32)} )</th>
<th>( y_{(43)} )</th>
<th>( y_{(53)} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_{1}^{(1)} )</td>
<td>( p_{1}^{(1)} )</td>
<td>( 5 )</td>
<td>( 6 )</td>
<td>( 5 )</td>
<td>( 6 )</td>
<td>( 6 )</td>
<td>( 6 )</td>
</tr>
<tr>
<td>( x_{2}^{(2)} )</td>
<td>( p_{2}^{(2)} )</td>
<td>( 5 )</td>
<td>( 6 )</td>
<td>( 5 )</td>
<td>( 6 )</td>
<td>( 6 )</td>
<td>( 6 )</td>
</tr>
<tr>
<td>( x_{3}^{(3)} )</td>
<td>( p_{3}^{(3)} )</td>
<td>( 5 )</td>
<td>( 6 )</td>
<td>( 5 )</td>
<td>( 6 )</td>
<td>( 6 )</td>
<td>( 6 )</td>
</tr>
<tr>
<td>( \pi_{h}^{(3)} )</td>
<td>( p_{h}^{(3)} )</td>
<td>( 5 )</td>
<td>( 6 )</td>
<td>( 5 )</td>
<td>( 6 )</td>
<td>( 6 )</td>
<td>( 6 )</td>
</tr>
</tbody>
</table>

The posterior probability is given by
\[ p(\mathbf{Y}_i^{(l)} \mid (h, \mathbf{y}^{(m)})) = \frac{\pi_p^{(l)} \sum_{h^{(m)}} p_h^{(r)}}{\sum_{h^{(m)}} \pi_p^{(l)} \sum_{h^{(m)}} p_h^{(r)}} \quad \text{for } \mathbf{y}^{(m)} = \mathbf{b}'_j \mathbf{y}_j \]

\[ = 0 \quad \text{otherwise} \]

For \( \mathbf{y}^{(12)} \), the posterior probability is given by

\[ p(\mathbf{Y}_i^{(l)} \mid (2, \mathbf{y}^{(12)})) = \frac{\pi_p^{(l)} p_i^{(l)}}{\pi_p^{(l)} p_2^{(l)} + \pi_p^{(l)} p_i^{(l)}} \quad \text{for } \mathbf{y}^{(12)} = \mathbf{b}'_j \mathbf{y}_j \]

\[ = 0 \quad \text{otherwise} \]

Similar expressions can be given for other posterior probabilities.

When the random variables in the prior distribution are exchangeable, the sample probability will usually be the same for a sample set, regardless of the realized population. In such a setting, \( p_i^{(l)} = p_i^{(m)} = p_i^{(s)} = p_i^{(3)} = p_i^{(5)} = p_i^{(6)} = c \). In this case, the posterior distribution for \( \mathbf{y}^{(12)} \) will be given by

\[ p(\mathbf{Y}_i^{(l)} \mid (2, \mathbf{y}^{(12)})) = \frac{\pi_p^{(l)} c}{\pi_p^{(l)} c + \pi_2^{(l)} c} = \pi_p^{(l)} \quad \text{for } \mathbf{y}^{(12)} = \mathbf{b}'_j \mathbf{y}_j \]

\[ = 0 \quad \text{otherwise} \]

The posterior distribution will not depend on the sample probabilities, and will be equal to the prior distribution.

TO HERE 4/28/2010

**Basic Results on Linear Bayes Estimators**

We discuss results summarized by Ericson (1988, p216-217) on linear Bayes estimators in this context. The results rely on low order moments of the prior distribution. We discuss an exchangeable prior distribution. Recall that the prior distribution is the distribution of \( \mathbf{y}^{(l)} = \sum_{j=1}^{\mathbf{Y}} f_j^{(l)} y_j^{(i)} \) where \( E_i \left( f_j^{(l)} \right) = \pi_p^{(l)} \) and \( y_j^{(m)} = r_j^{(m)} \). We represent \( \mathbf{Y}^{(l)} = \sum_{j=1}^{\mathbf{Y}} f_j^{(l)} r_j^{(l)} y_j^{(l)} \) or \( \mathbf{Y}^{(l)} = \sum_{j=1}^{\mathbf{Y}} f_j^{(l)} r_j^{(l)} y_j^{(l)} \). Let \( \mathbf{U}^{(l)} = \sum_{j=1}^{\mathbf{Y}} f_j^{(l)} r_j^{(l)} \) so that \( \mathbf{Y}^{(l)} = \mathbf{U}^{(l)} y_j^{(l)} \). Subjects in \( \mathbf{y}^{(l)} \) are labeled by \( j \), with elements given by \( y_j^{(l)} \). We have represented the labels of subjects in permutation \( p \) of \( y_j^{(l)} \) by \( j^{(p)} = (1^{(p)}, 2^{(p)}, \ldots, \mathbf{N}^{(p)}) \). We have not yet defined notation for elements of \( \mathbf{y}^{(l)} \). Let us define this as \( \mathbf{Y}^{(l)} = \begin{pmatrix} y_1^{(l)} & y_2^{(l)} & \ldots & y_N^{(l)} \end{pmatrix} \) for \( y_j^{(l)} \). Using this
notation, let $E_{\xi} \left( Y_{i}^{(f)} \right) = \mu_{f}$, while the variance is $\text{var}_{\xi} \left( Y_{i}^{(f)} \right) = \frac{N-1}{N} \sigma_{f}^{2}$, where $\mu_{f} = \frac{1}{N} \sum_{i=1}^{N} Y_{i}^{(f)}$, $\sigma_{f}^{2} = \frac{1}{N-1} \sum_{i=1}^{N} \left( Y_{i}^{(f)} - \mu_{f} \right)^{2}$, $E_{\xi} \left( Y^{(f)} \right) = \mu_{f} \mathbf{1}_{\delta}$, and $\text{var}_{\xi} \left( Y^{(f)} \right) = \sigma_{f}^{2} \left( \mathbf{1}_{\delta} - \frac{1}{N} \mathbf{J}_{\delta} \right)$. The random variables $Y_{i}^{(f)}$, $i = 1, \ldots, N$ may define the exchangeable set of random variables $Y_{1}^{(f)}$, $Y_{2}^{(f)}$, ..., $Y_{N}^{(f)}$.

**Exchangeability**

Ericson (1988, p218, section 4.1) describes the connection between sampling and exchangeability. We discuss this using our notation. The target is $\mu_{f} = \frac{1}{N} \sum_{i=1}^{N} Y_{i}^{(f)}$ or $\mu_{f} = \frac{1}{N} \mathbf{1}_{\delta}^{\top} Y^{(f)}$, the finite population mean. The random variables $Y^{(f)} = \left( Y_{1}^{(f)}, Y_{2}^{(f)}, \ldots, Y_{N}^{(f)} \right)$ are assumed to be exchangeable. Furthermore, as discussed by Ericson (1969, p198), we can view $Y^{(f)}$ as conditional on $\mu_{f}$ and $\sigma_{f}^{2} = \frac{N-1}{N} \sigma_{f}^{2}$, where these parameters correspond to a realization of $Y^{(f)}$ from a superpopulation. We have defined an indicator random variable $I_{\delta}$ corresponding to the set of subjects in $\delta$, and let us define $\xi_{\delta}$ as the expectation over $I_{\delta}$ (see c10ed03.doc). As in Ericson (1969), suppose the population is a simple random sample from a large $\mathbb{N} > N$ (which may be infinite) superpopulation. With this understanding, the random variables whose realization is set $f$ are exchangeable, where

$$E_{\xi_{\delta}} \left( \sum_{j=1}^{N} I_{j} Y_{i}^{(f)} \right) = m, \quad \text{var}_{\xi_{\delta}} \left( \sum_{j=1}^{N} I_{j} Y_{i}^{(f)} \right) = v, \quad \text{for all } i = 1, \ldots, N \cos_{\xi_{\delta}} \left( \sum_{j=1}^{N} I_{j} Y_{i}^{(f)} \sum_{j=1}^{N} I_{j} Y_{k}^{(f)} \right) = c$$

for $i \neq k$. We do not know the values of $m$, $v$, and $c$. They are subjective values that may be selected arbitrarily by the analyst. Also, let $\hat{\sigma}_{f}^{2} = \frac{1}{N} \sum_{i=1}^{N} \left( Y_{i}^{(f)} - \mu_{f} \right)^{2}$ or $\hat{\sigma}_{f}^{2} = \frac{1}{N} Y^{(f)} \left( \mathbf{1}_{\delta} - \frac{1}{N} \mathbf{J}_{\delta} \right) Y^{(f)}$, and the mean for a sample set $s$ be given by $\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_{i}^{(f)}$ or $\bar{Y} = \frac{1}{n} Y^{(f)} \delta \bar{Y}^{(f)}$. Notice that we have expressed $\bar{Y}$ as a random variable even though we specify a particular set of subjects via $\delta$. This is due to the fact that we have not specified the order, $r_{s}$, of the subjects in $Y^{(f)}$. With these assumptions and notation,

$$E_{\xi_{\delta}} \left( \sum_{j=1}^{N} I_{j} \mu_{f} \right) = \frac{1}{N} \sum_{j=1}^{N} E_{\xi_{\delta}} \left( \sum_{j=1}^{N} I_{j} Y_{i}^{(f)} \right) = m,$$
\[ E_{\tilde{\xi}t} \left( \sum_{j} I_{j} \tilde{c}_{j} \right) = \frac{1}{n} \sum_{t} E_{\tilde{\xi}t} \left( \sum_{j} I_{j} ^{\delta(j)} \right) = m, \]

so that

\[ \text{var}_{\tilde{\xi}t} \left( \sum_{j} I_{j} \mu_{j} \right) = \frac{1}{N} \text{var}_{\tilde{\xi}t} \left( \sum_{j} I_{j} Y^{(j)} \right) \frac{1}{N} I_{N} \]

\[ = \frac{1}{N^2} I_{N} \left( (v-c) I_{N} + c J_{N} \right) I_{N} \]

\[ = \frac{1}{N^2} \left[ N (v-c) + c N^2 \right] \]

\[ = \frac{1}{N} \left[ v + (N-1) c \right] \]

Since \( \tilde{\sigma}_{j} = \frac{1}{N} \gamma^{\delta(j)} \left( I_{N} - \frac{1}{N} J_{N} \right) \gamma^{(j)} \) or \( \tilde{\sigma}_{j} = \frac{1}{N} tr \left[ \gamma^{(j)} \gamma^{\delta(j)} \left( I_{N} - \frac{1}{N} J_{N} \right) \right] \) and

\[ E_{\tilde{\xi}t} \left( \sum_{j} I_{j} Y^{(j)} \gamma^{\delta(j)} \right) = \text{var}_{\tilde{\xi}t} \left( \sum_{j} I_{j} Y^{(j)} \right) + E_{\tilde{\xi}t} \left( \sum_{j} I_{j} Y^{(j)} \right) E_{\tilde{\xi}t} \left( \sum_{j} I_{j} ^{\delta(j)} \right), \]

then

\[ E_{\tilde{\xi}t} \left( \sum_{j} I_{j} \tilde{\sigma}_{j} \right) = \frac{1}{N} tr \left[ E_{\tilde{\xi}t} \left( \sum_{j} I_{j} Y^{(j)} \gamma^{\delta(j)} \right) \left( I_{N} - \frac{1}{N} J_{N} \right) \right] \]

\[ = \frac{1}{N} tr \left[ (v-c) I_{N} + (c + m^2) J_{N} \left( I_{N} - \frac{1}{N} J_{N} \right) \right] \]

\[ = \frac{(v-c)}{N} \left( I_{N} - \frac{1}{N} J_{N} \right) \]

\[ = \frac{(N-1) (v-c)}{N} \]

Also, \( \bar{E}_{j} = \frac{1}{n} \sum_{t} Y_{i,t} ^{\delta(j)} \) or \( \bar{E}_{j} = \frac{1}{n} Y_{i,\delta^{j}} Y^{\delta(j)} \), so that
\[ \text{var}_{i \in \mathcal{I}} \left( \sum_j I_j \hat{F}_j \right) = \frac{1}{n} \left( \sum_j I_j \delta_j \right) \left( \sum_j I_j Y^{(f)} \right) \frac{1}{n} \delta_j 1_s \]

\[ = \frac{1}{n} \left( \sum_j I_j \delta_j \right) \left( (v-c) I_s + c J_s \right) \delta_j 1_s \]

\[ = \frac{1}{n} \left[ (v-c) \left( \sum_j I_j \delta_j \right) \delta_j 1_s + c \left( \sum_j I_j \delta_j \right) J_s 1_s \right] \]

\[ = \frac{1}{n} \left[ n(v-c) + n^2 c \right] \]

\[ = \frac{1}{n} \left[ v + (n-1)c \right] \]

Finally, we consider

\[ E_{i \in \mathcal{I}} \left( \sum_j I_j \nabla_j Y^{(f)} \right) = E_{i \in \mathcal{I}} \left( \frac{1}{n} \sum_j I_j \nabla_j Y^{(f)} \right) = \mu_j, \sum_j \nabla_j = \nabla_f \]

I have expressed \( \nabla_f = \frac{1}{n} \nabla_j Y^{(f)} \). It is valuable to see how this could arise from different realizations of \( Y^{(f)} \). We represent a realization of \( Y^{(f)} \) by \( Y^{(m)} = r_f Y^{(f)} \). Of course, depending upon the realization, the values that will be included in the vector \( \frac{1}{n} \nabla_j Y^{(m)} \) will vary.

When defining a sample set, the set is identical for any ordering of the subjects in the set. If this is the case, we can not represent the same sample set as \( \frac{1}{n} \nabla_j Y^{(m)} \) for different \( p = 1, \ldots, P = N! \). This is because when \( \delta_j \) is non-stochastic, the different orderings, \( Y^{(m)} \), will result in different sets of subjects in the sample. The set of subject labels, identified by \( \delta_j \), is the same, since the same element of \( \delta_j \) are identified by \( \delta_j \). However, the subject that corresponds to a label for \( p \neq p^* \) may be different. For example, a sample set may be labeled by \( \{4^{(p)}, 8^{(p)}, \ldots, 13^{(p)}\} \), but the actual subject identified by \( 4^{(p)} \) may be different from the subject identified by \( 4^{(p^*)} \) for \( p \neq p^* \). This subtle difference is of interest, since it provides insight as to how you could have a sample, and yet consider the response to be a random variable.

We now consider

\[ E_{i \in \mathcal{I}} \left( \sum_j I_j \nabla_j Y^{(f)} \right) = E_{i \in \mathcal{I}} \left( \frac{1}{n} \sum_j I_j \nabla_j Y^{(f)} \right) = \mu_j, \sum_j \nabla_j = \nabla_f \]

Notice that for all realizations of \( Y^{(f)} \), \( \frac{1}{N} \nabla_j Y^{(f)} = \mu_j \) and \( \sum_j \nabla_j = \nabla_f \). As a result, \( \frac{1}{N} \sum_j I_j \nabla_j Y^{(f)} = \sum_j I_j \mu_j \). If this quantity is equal to \( \mu_j \), and \( \mu_j \neq \mu_f \) when \( f \neq f^* \),
then \( I_f = 1 \). We assume this is the case, so that \( \frac{1}{N} \sum_{f} I_f Y^{(f)} = \mu_f \) is the same as
\[
\frac{1}{N} Y^{(f)} = \mu_f. 
\]
Similarly, we assume that \( \sum_{f} I_f \hat{\sigma}_f^2 = \hat{\sigma}_f^2 \) implies that \( I_f = 1 \). With these assumptions, \( E_{\hat{\theta}_i} \left( \sum_{f} I_f Y^{(f)} \mid \mu_f, \hat{\sigma}_f^2 \right) = E_{\hat{\theta}_i} \left( Y^{(f)} \mid \mu_f, \hat{\sigma}_f^2 \right) \), and
\[
\text{var}_{\hat{\theta}_i} \left( \sum_{f} I_f Y^{(f)} \mid \mu_f, \hat{\sigma}_f^2 \right) = \text{var}_{\hat{\theta}_i} \left( Y^{(f)} \mid \mu_f, \hat{\sigma}_f^2 \right). 
\]
Now \( Y_f = \frac{1}{n} Y^{(f)} \) and from the earlier section,
\[
E_{\hat{\theta}_i} \left( Y^{(f)} \right) = \mu_f 1_n, \quad \text{and} \quad \text{var}_{\hat{\theta}_i} \left( Y^{(f)} \right) = \sigma_f^2 \left( 1_n - \frac{1}{N} 1_n \right). \]
As a result,
\[
E_{\hat{\theta}_i} \left( Y_f \mid \mu_f, \sigma_f^2 \right) = E_{\hat{\theta}_i} \left( \frac{1}{n} Y^{(f)} \mid \mu_f, \sigma_f^2 \right) = \frac{1}{n} Y^{(f)} \left( \mu_f, \sigma_f^2 \right). 
\]
Also,
\[
\text{var}_{\hat{\theta}_i} \left( Y_f \mid \mu_f, \sigma_f^2 \right) = \text{var}_{\hat{\theta}_i} \left( \frac{1}{n} Y^{(f)} \mid \mu_f, \sigma_f^2 \right) = \frac{1}{n^2} Y^{(f)} \left( \mu_f, \sigma_f^2 \right) = \frac{1}{n^2} \left( 1_n - \frac{1}{N} 1_n \right) \delta 1_n. 
\]
Since \( \hat{\sigma}_f^2 = \frac{N-1}{N} \sigma_f^2 \), \( \text{var}_{\hat{\theta}_i} \left( Y_f \mid \mu_f, \hat{\sigma}_f^2 \right) = \frac{\hat{\sigma}_f^2}{n} \left( \frac{N-n}{N-1} \right) \). These results match the results given by Ericson (1988, section 4.1, p218-219). Ericson notes that we have not used the probability of obtaining a sample set \( s \) from population, \( p(s) \), given by \( p^{(f)} \). This comment is noteworthy in light of Gelman et al. (2004, p7) comment that Bayes’ rule stems from the product of the prior distribution and the ‘sampling distribution (or data distribution)’. There is no role of for the ‘sampling distribution’ in Ericson’s development.

**Development of a BLUP**
Ericson describes linear Bayes estimators using Result 3.1 (p216), attributing these results to several authors, the earliest of which is Ericson (1969) and Hartigan (1969). We first describe the basic idea of BLUP in the finite population mixed models, and subsequently identify how the development is a special case of Ericson’s Result 3.1.

Review of Development of Finite Mixed Model BLUP

Suppose that
\[
\begin{bmatrix}
Y_1 \\
Y_2
\end{bmatrix} = \alpha
\]
where \( Y_i \ (n \times 1) \) represents a vector of response for a sample, and \( Y_2 \ ((N-n) \times 1) \) represents potential response for the remainder of a population, and assume that the target of interest is \( P = g' \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} \) where we partition \( g \) as \( \begin{bmatrix} g'_1 \\ g'_2 \end{bmatrix} \) so that \( P = g'_1 Y_1 + g'_2 Y_2 \). We wish to predict the target conditional on realizing \( Y_i \) (which corresponds to selecting the sample and observing the results). The basic inference problem is prediction of \( g'_2 Y_2 \). We require the predictor to be:
- Linear in the sample: \( \hat{P} = (g'_1 + a') Y_1 \)
- Unbiased: \( E(\hat{P} - P) = 0 \)
- Minimum MSE: \( \text{var} \{\hat{P} - P\} \) is minimized.

The unbiased constraint implies that \( a'X_1 - g'_2 X_2 = 0 \). We minimizing the variance,
\[
\text{var} \{\hat{P} - P\} = a'\Sigma_1 a - 2g'_2 \Sigma_1 a + g'_2 \Sigma_2 g_2,
\]
subject to this unbiased constraint using Lagrangian multipliers. The quantity to be minimized is
\[
f(a, \lambda) = a'\Sigma_1 a - 2g'_2 \Sigma_1 a + g'_2 \Sigma_2 g_2 + 2(a' X_1 - g'_2 X_2) \lambda.
\]
The solution is given by the best linear unbiased predictor,
\[
\hat{P} = g'_1 Y_1 + g'_2 \left[ X_1 \Sigma_{11}^{-1} X_1' \right]^{-1} X_1' \Sigma_{11}^{-1} \left( X_1 \Sigma_{11}^{-1} X_1' \right)^{-1} \left( X_1 \Sigma_{11}^{-1} X_1' \right)^{-1} X_1' \Sigma_{11}^{-1} X_1 \hat{\alpha}.
\]

Development of Finite Mixed Model BLUP with Reversed Notation adapted to Ericson’s setting

In order to see how these results relate to Ericson’s Result 3.1, we first re-arrange the order of the vectors to express them similar to the order given by Ericson. Ericson lists the sample as the second vector, and the remainder as the first part of the vector. Using this re-arrangement, suppose that \( E \begin{bmatrix} Y_2 \\ Y_1 \end{bmatrix} = \begin{bmatrix} X_2 \\ X_1 \end{bmatrix} \alpha \) where \( Y_i \ (n \times 1) \) represents a vector of response for a sample, and \( Y_2 \ ((N-n) \times 1) \) represents potential response for the remainder of a population,
\[
\text{var}\left( \begin{bmatrix} Y_1 \\ Y_j \end{bmatrix} \right) = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix},
\]
and assume that the target of interest is \( P = g \left( \begin{bmatrix} Y_1 \\ Y_j \end{bmatrix} \right) \) where we partition \( g \) as \( g = (g'_1, g'_j) \) so that \( P = g(Y_1 + g'_j Y_j). \) We assume that \( g = \begin{bmatrix} 1_{n-a} & 0 \end{bmatrix}^{\prime} \) so that \( P = Y_j. \)

We wish to predict the target conditional on realizing \( Y_1 \) (which corresponds to selecting the sample and observing the results). The basic inference problem is prediction of \( g'_j Y_j = Y_j. \) We require the predictor to be:

Linear in the sample:
\[
\hat{P} = (g'_j + a') Y_j
\]

Possibly Unbiased:
\[
E_{Y,S_j} \left( \hat{P} - P \right) = \gamma \left( = 0_{n-a} \right)
\]
or
\[
E_{Y,Y_j} (a' Y_j - Y_j) = \gamma \left( = 0_{n-a} \right)
\]

so that \( E_{Y,S_j} (a' Y_j - Y_j) = (a' X_j - X_j) a = \gamma \left( = 0_{n-a} \right) \) where we assume that we minimize:

\[
E_Y \left[ \left( E(Y_j | Y_j) - (g'_j + a') Y_j \right)^2 \right] = E_Y \left[ \left( E(Y_j | Y_j) - a' Y_j \right)^2 \right] = E_Y \left[ \left( E(Y_j | Y_j) - a' Y_j \right) \left( E(Y_j | Y_j) - a' Y_j \right)^\prime \right]
\]

which is the Euclidean norm given by \( \|x\| = \sqrt{\sum_{i=1}^{n} x_i^2}. \) Let us express \( E(Y_j | Y_j) = E_{Y,Y_j} (Y_j). \)

Aside: If we have a clustered setting where some of the units in a cluster are in \( Y_1 \), and others are in \( Y_j \), then there will be a difference between \( E_{Y,S_j} (Y_j) \) and \( E_{Y,Y_j} (Y_j). \) This might be an important consideration in understanding the estimator.

We want to find values of \( a' \) that minimize

\[
E_X \left[ \left( E_{Y,Y_j} (Y_j) - a' Y_j \right) \left( E_{Y,Y_j} (Y_j) - a' Y_j \right)^\prime \right] =
\]

subject to the unbiased constraint. Now \( E_Y \left[ Y_j E_{Y,S_j} (Y_j) \right] = E_{Y,Y_j} (Y_j). \) Also,
\[
\Sigma_{11} = E_{Y,S_j} (Y_j Y_j') - E_{Y,S_j} (Y_j) E_{Y,S_j} (Y_j') \quad \Sigma_{12} = E_{Y,S_j} (Y_j Y_j') - E_{Y,S_j} (Y_j) \left[ E_{Y,S_j} (Y_j') \right] E_{Y,S_j} \left[ E_{Y,S_j} (Y_j') \right]
\]

\[
\Sigma_{21} = E_{Y,S_j} (Y_j Y_j') - E_{Y,S_j} (Y_j) \left[ E_{Y,S_j} (Y_j) \right] E_{Y,S_j} \left[ E_{Y,S_j} (Y_j) \right]
\]

\[
\Sigma_{22} = E_{Y,S_j} (Y_j Y_j') - E_{Y,S_j} (Y_j) \left[ E_{Y,S_j} (Y_j') \right] E_{Y,S_j} \left[ E_{Y,S_j} (Y_j) \right]
\]

Deleted: C10ed05v3.doc
Deleted: 4/28/2010 5:47 PM
Deleted: 4/28/2010 3:12 PM
Inserted: 4/28/2010 5:47 PM
and $\Sigma_{ij} = E_{X,Y_i}(Y_j) - E_{X,Y_i}(Y_j) E_{X,Y_i}(Y_j')$. Also, $E_{X,Y_i}(Y_i) = E_x \left[ E_{Y,X_i}(Y_i) \right]$ and $E_{X,Y_i}(Y_j) = E_x \left[ E_{Y,X_i}(Y_j) \right]$. As a result,

$$E_x \left[ E_{Y,X_i}(Y_j) - a'Y_i \right] \left( E_{Y,X_i}(Y_j) - a'Y_i \right) =$$

$$= tr \left[ a'E_x \left( Y_iY_j \right) a - 2E_x \left[ Y_iE_{Y,X_i}(Y_j) \right] a' \right] + tr \left[ E_x \left( E_{Y,X_i}(Y_j) E_{Y,X_i}(Y_j') \right) \right]$$

$$= tr \left[ a' \left[ \Sigma_{ii} + E_x \left( Y_i \right) E_x \left( Y_j \right) \right] a \right] - 2 \left[ \Sigma_{ij} + E_x \left( Y_i \right) E_x \left( Y_j' \right) \right] a' + tr \left[ E_x \left( E_{Y,X_i}(Y_j) E_{Y,X_i}(Y_j') \right) \right]$$

We express this as

$$E_x \left[ E_{Y,X_i}(Y_j) - a'Y_i \right] \left( E_{Y,X_i}(Y_j) - a'Y_i \right) = tr \left[ a' \left[ E_x \left( Y_i \right) E_x \left( Y_j \right) \right] a \right]$$

$$- 2 \left[ E_x \left[ E_{Y,X_i}(Y_j) \right] E_x \left[ E_{Y,X_i}(Y_j') \right] \right] a'$$

$$+ tr \left[ E_x \left( E_{Y,X_i}(Y_j) E_{Y,X_i}(Y_j') \right) \right]$$

Now $E_x \left[ E_{Y,X_i}(Y_j) \right] = E_x \left( Y_i \right)$. We have assumed that $E_{X,Y_i}(a'Y_i - Y_i) = \gamma = 0$ which implies that $E_{X,Y_i}(Y_j) = E_{X,Y_i}(a'Y_i) - \gamma$ so that $E_x \left[ E_{Y,X_i}(Y_j) \right] = E_{X,Y_i}(a'Y_i) - \gamma$. Using these expressions,

$$E_x \left[ E_{Y,X_i}(Y_j) - a'Y_i \right] \left( E_{Y,X_i}(Y_j) - a'Y_i \right) = tr \left[ a' \left[ E_x \left( Y_i \right) E_x \left( Y_j \right) \right] a \right]$$

$$- 2 \left[ E_x \left[ E_{Y,X_i}(Y_j) \right] E_x \left[ E_{Y,X_i}(Y_j') \right] \right] a'$$

$$+ tr \left[ E_x \left( E_{Y,X_i}(Y_j) E_{Y,X_i}(Y_j') \right) \right]$$

or

$$E_x \left[ E_{Y,X_i}(Y_j) - a'Y_i \right] \left( E_{Y,X_i}(Y_j) - a'Y_i \right) = tr \left[ a' \left[ E_x \left( Y_i \right) \right] a \right]$$

$$- 2 \left[ E_x \left[ E_{Y,X_i}(Y_j) \right] a - E_x \left( Y_j \right) \gamma \right] a'$$

$$+ tr \left[ E_x \left( E_{Y,X_i}(Y_j) E_{Y,X_i}(Y_j') \right) \right]$$

or

$$E_x \left[ E_{Y,X_i}(Y_j) - a'Y_i \right] \left( E_{Y,X_i}(Y_j) - a'Y_i \right) = tr \left[ a' \left[ E_x \left( Y_i \right) \right] a \right]$$

$$+ 2 \left[ E_x \left( Y_i \right) \gamma \right]$$

$$+ tr \left[ E_x \left( E_{Y,X_i}(Y_j) E_{Y,X_i}(Y_j') \right) \right]$$
I would like to say that $E_\mathbf{Y} (\mathbf{a}^t \mathbf{Y}_1) = E_\mathbf{Y} (\mathbf{a}^t \mathbf{Y}_1)$. Is this true? If it is, then

$$E_\mathbf{Y} (\mathbf{a}^t \mathbf{Y}_1) E_\mathbf{Y} (\mathbf{Y}_1) - E_\mathbf{Y} (\mathbf{a}^t \mathbf{Y}_1) E_\mathbf{Y} (\mathbf{Y}_1) = E_\mathbf{Y} (\mathbf{a}^t \mathbf{Y}_1) E_\mathbf{Y} (\mathbf{Y}_1) - E_\mathbf{Y} (\mathbf{a}^t \mathbf{Y}_1) E_\mathbf{Y} (\mathbf{Y}_1).$$

Also, I would like to say that $E_{\mathbf{Y}, \mathbf{Y}} (\mathbf{Y}_1^2) = E_{\mathbf{Y}, \mathbf{Y}} [E_{\mathbf{Y}} (\mathbf{Y}_1^2)]$. If this is true, then

$$E_{\mathbf{Y}} (\mathbf{a}^t \mathbf{Y}_1) E_{\mathbf{Y}} (\mathbf{Y}_1) - E_{\mathbf{Y}} (\mathbf{a}^t \mathbf{Y}_1) E_{\mathbf{Y}} (\mathbf{Y}_1) = E_{\mathbf{Y}} (\mathbf{a}^t \mathbf{Y}_1) (E_{\mathbf{Y}} (\mathbf{Y}_1) - E_{\mathbf{Y}} [E_{\mathbf{Y}} (\mathbf{Y}_1^2)]).$$

Also, using

$$+E_{\mathbf{Y}} (E_{\mathbf{Y}} (\mathbf{Y}_1) E_{\mathbf{Y}} (\mathbf{Y}_1^2)) - E_{\mathbf{Y}} [E_{\mathbf{Y}} (\mathbf{a}^t \mathbf{Y}_1)] E_{\mathbf{Y}} (\mathbf{Y}_1^2).$$

I would like to simply this also.

4/25/2010
References


