E C O N O M I C S B U L L E T I N

On calculating estimates of stratified error-components models

Robert Phillips
George Washington University

Abstract

This note provides an AECM (alternating expectation conditional maximization) algorithm for calculating maximum-likelihood estimates of stratified error-components models. An advantage it has over other algorithms is that it can be easily modified to incorporate useful restrictions on the variance components. The new algorithm is applied in an example that illustrates the variance restrictions.

Citation: Phillips, Robert, (2008) "On calculating estimates of stratified error-components models." *Economics Bulletin*, Vol. 3, No. 75 pp. 1-10
Submitted: October 17, 2008. **Accepted:** November 28, 2008.
URL: http://economicsbulletin.vanderbilt.edu/2008/volume3/EB-08C20066A.pdf

1. Introduction

In a stratified error-components model the conditional regression error variance changes across some but not all cross sections. Specifically, consider the error-components model

$$y_{it} = \alpha + x_{it}'\beta + u_{it}, \qquad u_{it} = \mu_i + \upsilon_{it} \qquad (t = 1, \ldots, T,\ \ i = 1, \ldots, N),$$

where x_{it} is a $K \times 1$ vector of regressors, $b = (\alpha, \beta')'$ is a $(K + 1) \times 1$ vector of regression parameters, and μ_i and υ_{it} are error components. The distinctive feature of a stratified error-components model is that μ_i and υ_{it} are assumed to have conditional variances $\sigma^2_{\mu j}$ and $\sigma^2_{\upsilon j}$ given the ith cross section is drawn from the jth subpopulation or stratum ($j = 1, \ldots, q$), implying that cross-sectional errors from different cross sections can exhibit different dispersion. This model captures dispersion heterogeneity across cross sections while avoiding the incidental parameters problem (see Neyman and Scott, 1948) that would arise if the conditional variances of μ_i and υ_{it} were allowed to vary without restriction across cross sections (see Phillips, 2003).

There are at least two types of applications where allowing for such dispersion heterogeneity can be important. One is when forecast intervals for future values of y_{it} are sought and the amount of dispersion in u_{it} differs across cross sections. Another is when one wishes to classify cross sections in terms of the dispersion in u_{it}. Such an exercise might be useful, for example, in applications in which dispersion can be interpreted in terms of risk and the researcher wants to classify cross sections into risk categories. In general it will not be known *a priori* which cross sections belong to which strata, but after the model is estimated one can use posterior probabilities to assign cross sections to strata.

In the model studied in Phillips (2003), both the conditional variance of μ_i and the conditional variance of υ_{it} are allowed to change across strata. However, differing conditional variances of the μ_is across strata has a different interpretation than when the conditional variances of the υ_{it}s differ. Specifically, if the conditional variance of υ_{it} changes across two strata that says the dispersion of the remainder term υ_{it} differs across some cross sections, whereas if the conditional variance of μ_i changes across strata, then those cross sections belonging to the stratum with the largest conditional variance for μ_i have cross-sectional specific effects, μ_is, that are outliers.

In a given application there may be outlying μ_is, or cross sections with more dispersion in the remainder terms, or both. The model considered in Phillips (2003) allows for both. But in some applications there may be only outlying μ_is. In other applications there may be no outlying μ_is, but the remainder terms may exhibit more dispersion for some cross sections than for others. In other words, in a given application, it may be of interest to check whether or not the restrictions $\sigma^2_{\upsilon 1} = \sigma^2_{\upsilon 2} = \cdots = \sigma^2_{\upsilon q}$ are satisfied or whether the restrictions $\sigma^2_{\mu 1} = \sigma^2_{\mu 2} = \cdots = \sigma^2_{\mu q}$ hold.

Furthermore, other equality restrictions may be of interest. For example, consider a model that allows for four strata that are characterized as follows: a small $\sigma^2_{\upsilon j}$ and small $\sigma^2_{\mu j}$ stratum, a small $\sigma^2_{\upsilon j}$ and large $\sigma^2_{\mu j}$ stratum, a large $\sigma^2_{\upsilon j}$ and small $\sigma^2_{\mu j}$ stratum, and, finally, a large $\sigma^2_{\upsilon j}$ and large $\sigma^2_{\mu j}$ stratum. These four possibilities can be captured by setting $q = 4$ and using the restrictions $\sigma^2_{\upsilon 1} = \sigma^2_{\upsilon 2}$ and $\sigma^2_{\upsilon 3} = \sigma^2_{\upsilon 4}$ (where $\sigma^2_{\upsilon 1} < \sigma^2_{\upsilon 3}$) and $\sigma^2_{\mu 1} = \sigma^2_{\mu 3}$ and $\sigma^2_{\mu 2} = \sigma^2_{\mu 4}$ (with $\sigma^2_{\mu 1} < \sigma^2_{\mu 2}$), from which we see that four possible outcomes can be modeled with only four distinct variance components.

Unfortunately, however, the EM (expectation-maximization) algorithm proposed by Phillips (2003) for computing maximum-likelihood estimates of the parameters of a stratified error-

components model is not easily modified to incorporate equality restrictions on the variance components. This note rectifies this shortcoming. In the next section an AECM (alternating expecation conditional maximization) algorithm is provided. This algorithm, like the previous algorithm suggested in Phillips (2003), calculates maximum-likelihood estimates with fitted variance components that are guaranteed to be non-negative, but, unlike that algorithm, it can also be easily modifed to incorporate equality restrictions on the variance components. In Section 3 the algorithm is applied in an example.

2. An AECM algorithm

The presence of latent strata implies that if we draw randomly across strata, then $y_i = (y_{i1}, \ldots, y_{iT})'$ comes from a mixture of distributions. In particular, when μ_i and the components of $v_i = (v_{i1}, \ldots, v_{iT})'$ are independent, mean zero normal random variables conditional on both $x_i' = [x_{i1} \cdots x_{iT}]$ and on their being drawn from the jth stratum, the joint density of y_i conditional on only x_i is a finite mixture of multivariate normal densities:

$$p(y_i|x_i;\, \psi) = \sum_{j=1}^{q} \lambda_j f(y_i|x_i;\, b, \sigma_{vj}^2, \sigma_{\mu j}^2)$$

(see Phillips 2003). Here

$$f(y_i|x_i;\, b, \sigma_{vj}^2, \sigma_{\mu j}^2) = (2\pi)^{-T/2}|\,\Sigma_j|^{-1/2} \exp\left[-(y_i - X_i b)'\Sigma_j^{-1}(y_i - X_i b)/2\right],$$

$\psi = (b', \theta')' = (b', \sigma_{v1}^2, \ldots, \sigma_{vq}^2, \sigma_{\mu 1}^2, \ldots, \sigma_{\mu q}^2, \lambda_1, \ldots, \lambda_q)'$, $X_i = [\iota_T \;\; x_i]$, $\Sigma_j = \sigma_{vj}^2 I_T + \sigma_{\mu j}^2 \iota_T \iota_T'$, I_T is a T-dimensional identity matrix, ι_T is a $T \times 1$ vector of ones, and λ_j is the fraction of cross sections in the population belonging to the jth stratum. Phillips (2003) showed that the likelihood $\ell(\psi) = \prod_{i=1}^{N} p(y_i|x_i;\, \psi)$ is bounded provided the variance components are constrained to be non-negative. Moreover, that paper provided a constrained EM (expectation maximization) algorithm for maximizing $\ell(\cdot)$ subject to these constraints.

Although the objective of this note is to provide an algorithm for maximizing the likelihood $\ell(\psi)$ subject to equality restrictions on the variance comonents, the computational approach described here yields relatively simple and stable algorithms regardless of whether or not restrictions are applied. This section therefore first describes how to calculate estimates without imposing restrictions on the variance components and then shows how these calculations are modified in order to calculate estimates subject to equality restrictions on the variance components.

The computational strategy suggested in this note relies on the AECM algorithm (see Meng and van Dyk, 1997), an extension of the EM algorithm. Like the EM algorithm, the AECM algorithm simplifies computations via data augmentation. The data are augmented during the E (expectation) step, a step that builds an imputed log-likelihood by taking the expectation of the log-likelihood based on the augmented or complete data while conditioning on the observed or incomplete data and while using the current fit of the parameters as the parameters of the conditional distribution. A standard EM algorithm then applies the M (maximization) step, which maximizes this imputed log-likelihood, and this, in turn, produces an increase in the actual log-likelihood (see, e.g., Meng and van Dyk, 1997). An AECM algorithm, on the other hand, replaces the M step with a sequence

of conditional maximization (CM) steps. Moreover, the CM steps may rely on different amounts of data augmentation.

Two CM steps suffice to calculate maximum-likelihood estimates of the stratified error-components model. In the first CM step the observed data $y = (y_1', \ldots, y_N')'$ are augmented with the unobserved "data" $\mu = (\mu_1, \ldots, \mu_N)'$ and $d = (d_1', \ldots, d_N')'$, where the $q \times 1$ vector $d_i = (d_{i1}, \ldots, d_{iq})'$ equals ω_j—a vector of zeros except for a one in the jth position—if the ith cross section is drawn from the jth stratum. Like μ, the vector d is unobserved, for we do not know *a priori* which cross sections are drawn from which strata.

Using the complete-data—y, μ, and d—execution of the E step consists of taking the expectation of the complete-data log-likelihood—that is, the log-likelihood for y, μ, and d—while conditioning on y (and on $x = [x_1' \cdots x_N']'$) and while using the current fit of the parameters as the parameters of the conditional distribution. This E step produces the imputed log-likelihood

$$\begin{aligned} Q_1(\psi; \psi^c) \;=\; & const + \sum_{j=1}^{q} \ln(\lambda_j) \sum_{i=1}^{N} P_{ij}(\psi^c) - \frac{1}{2} \sum_{j=1}^{q} \ln(\sigma_{\mu j}^2) \sum_{i=1}^{N} P_{ij}(\psi^c) \\ & - \frac{1}{2} \sum_{j=1}^{q} \sum_{i=1}^{N} E_{\psi^c}(d_{ij} \mu_i^2 | y_i, x_i) / \sigma_{\mu j}^2 - \frac{T}{2} \sum_{j=1}^{q} \ln(\sigma_{\upsilon j}^2) \sum_{i=1}^{N} P_{ij}(\psi^c) \\ & - \frac{1}{2} \sum_{j=1}^{q} \sum_{i=1}^{N} E_{\psi^c}(d_{ij} \upsilon_i' \upsilon_i | y_i, x_i) / \sigma_{\upsilon j}^2. \end{aligned}$$

Here ψ^c denotes the current fit of the parameter vector ψ, $P_{ij}(\psi^c)$ is the posterior probability $P_{ij}(\psi) = \lambda_j f(y_i | x_i; b, \sigma_{\upsilon j}^2, \sigma_{\mu j}^2) / p(y_i | x_i; \psi)$ evaluated at ψ^c, and $E_\psi(\cdot | y_i, x_i)$ denotes a conditional expectation using ψ as the parameter vector of the conditional distribution.[1]

The first CM step consists of maximizing $Q_1(\cdot; \psi^c)$ conditional on $b = b^c$ while also imposing the restriction $\sum_{j=1}^{q} \lambda_j = 1$. The details of this step are provided in CM Step 1. (See the appendix for the derivations leading to the formulas appearing in CM Step 1.)

CM Step 1: Compute the residuals $u_i^c = y_i - X_i b^c$ $(i = 1, \ldots, N)$, the posterior probabilities $P_{ij}(\psi^c)$ $(j = 1, \ldots, q,\ \ i = 1, \ldots, N)$, and $(\sigma_j^2)^c = (\sigma_{\mu j}^2)^c + (\sigma_{\upsilon j}^2)^c / T$ and $a_j = (\sigma_{\mu j}^2)^c (\sigma_{\upsilon j}^2)^c / [T(\sigma_j^2)^c]$ $(j = 1, \ldots, q)$. Then, for $j = 1, \ldots, q$, compute

$$\lambda_j^+ = \frac{1}{N} \sum_{i=1}^{N} P_{ij}(\psi^c), \tag{1}$$

$$(\sigma_{\mu j}^2)^+ = \frac{1}{N \lambda_j^+} \sum_{i=1}^{N} P_{ij}(\psi^c) \left[(\sigma_{\mu j}^2)^c \iota_T' u_i^c / (T(\sigma_j^2)^c) \right]^2 + a_j, \tag{2}$$

[1] Derivation of the imputed log-likelihood $Q_1(\psi; \psi^c)$ relies on the observation that the conditional density of y_i given μ_i, d_i, and x_i is $\prod_{j=1}^{q} \{(2\pi)^{-T/2} (\sigma_{\upsilon j}^2)^{-T/2} \exp[-\upsilon_i' \upsilon_i / (2\sigma_{\upsilon j}^2)]\}^{d_{ij}}$, the conditional density of μ_i given d_i and x_i is $\prod_{j=1}^{q} \{(2\pi)^{-1/2} (\sigma_{\mu j}^2)^{-1/2} \exp[-\mu_i^2 / (2\sigma_{\mu j}^2)]\}^{d_{ij}}$, and the probability mass function of d_i given x_i is $\prod_{j=1}^{q} \lambda_j^{d_{ij}}$.

and

$$(\sigma^2_{\upsilon j})^+ = \frac{1}{NT\lambda_j^+}\sum_{i=1}^{N} P_{ij}(\psi^c)\left\{u_i^{c\prime}Qu_i^c + \left[(\sigma^2_{\upsilon j})^c \iota_T' u_i^c/(T(\sigma_j^2)^c)\right]^2/T\right\} + a_j, \tag{3}$$

where $Q = I_T - \iota_T\iota_T'/T$.

The second CM step relies on less data augmentation. For this step, the observed data y are augmented with only d. This is the amount of data augmentation used to derive the EM algorithm described in Phillips (2003), and thus, for this step, the imputed log-likelihood is similar to that used in Phillips (2003). Specifically, upon setting $\sigma_j^2 = \sigma^2_{\mu j} + \sigma^2_{\upsilon j}/T$ and $W_i(\theta) = \sum_{j=1}^{q} P_{ij}(\psi^{+/2})[Q/\sigma^2_{\upsilon j} + \iota_T\iota_T'/(\sigma_j^2 T^2)]$, the imputed log-likelihood is

$$\begin{aligned} Q_2(\psi;\psi^{+/2}) &= const + \sum_{i=1}^{N}\sum_{j=1}^{q} P_{ij}(\psi^{+/2})\ln(\lambda_j) - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{q} P_{ij}(\psi^{+/2})\ln(\sigma_j^2) \\ &\quad -\frac{T-1}{2}\sum_{i=1}^{N}\sum_{j=1}^{q} P_{ij}(\psi^{+/2})\ln(\sigma^2_{\upsilon j}) - \frac{1}{2}\sum_{i=1}^{N}(y_i - X_i b)'W_i(\theta)(y_i - X_i b) \end{aligned}$$

(see Phillips 2003). Note that the current fit is now taken to be $\psi^{+/2} = (b^{c\prime}, \theta^{+\prime})' = (b^{c\prime}, (\sigma^2_{\upsilon 1})^+, \ldots, (\sigma^2_{\upsilon q})^+, (\sigma^2_{\mu 1})^+, \ldots, (\sigma^2_{\mu q})^+, \lambda_1^+, \ldots, \lambda_q^+)'$.

Maximizing $Q_2(\cdot;\psi^{+/2})$ while conditioning on $\theta = \theta^+$ produces

$$b^+ = \left(\sum_{i=1}^{N} X_i'W_i(\theta^+)X_i\right)^{-1}\sum_{i=1}^{N} X_i'W_i(\theta^+)y_i \tag{4}$$

This is CM Step 2.

After the new fit $\psi^+ = (b^{+\prime}, \theta^{+\prime})'$ is calculated, it is made the current fit, i.e., we set $\psi^c = \psi^+$, and the two CM steps are repeated, and so on, until convergence.

This algorithm has important advantages over available alternatives. Consider, for example, the Newton-Raphson algorithm, an obvious candidate for calculating the extremum of a nonlinear function. The Newton-Raphson algorithm does not always exhibit stable convergence (see, e.g., Greene 2003, p. 938), and, when fitting an error-components model, it can produce negative variance estimates (see Meng and van Dyk, 1998). This is a potentially serious drawback when fitting a stratified error-components model, for there may be several component variances and the likelihood becomes unbounded should the algorithm stray into a region of the parameter space where one or more variance components are negative (see Phillips 2003). On the other hand, the constrained EM algorithm described in Phillips (2003) guarantees the actual log-likelihood does not decrease from one iteration to the next and the fitted variance components must be non-negative. The AECM algorithm has these properties as well. But the AECM algorithm has an important advantage over the constrained EM algorithm: it can be easily modified to incorporate equality restrictions on the variance components.

In order to see how equality restrictions can be incorporated, let $\{j_1, j_2, \ldots, j_r\} \subset \{1, 2, \ldots, q\}$ be a collection of indexes such that $\sigma^2_{\mu j_1} = \sigma^2_{\mu j_2} = \cdots = \sigma^2_{\mu j_r}$ and let $\{k_1, k_2, \ldots, k_s\} \subset \{1, 2, \ldots, q\}$ be a set of indexes such that $\sigma^2_{\upsilon k_1} = \sigma^2_{\upsilon k_2} = \cdots = \sigma^2_{\upsilon k_s}$. Moreover, let ω_n

$(n = 2, \ldots, r)$ and z_n $(n = 2, \ldots, s)$ denote $r - 1$ and $s - 1$ Lagrangean multipliers and set

$$Q_1^*(\psi; \psi^c) = Q_1(\psi; \psi^c) + \sum_{n=2}^{r} w_n(\sigma^2_{\mu j_{n-1}} - \sigma^2_{\mu j_n}) + \sum_{n=2}^{s} z_n(\sigma^2_{\upsilon k_{n-1}} - \sigma^2_{\upsilon k_n}).$$

Then upon applying the method of Lagrangean multipliers and exploiting the fact that $\sigma^2_{\mu j_n} = \sigma^2_{\mu j_1}$ $(n = 2, \ldots, r)$, one obtains

$$\begin{aligned}(\sigma^2_{\mu j_1})^+ &= \frac{1}{N\sum_{n=1}^{r}\lambda^+_{j_n}} \sum_{i=1}^{N}\sum_{n=1}^{r} P_{ij_n}(\psi^c)\left\{(\sigma^2_{\mu j_n})^c \iota_T' u_i^c / [T(\sigma^2_{j_n})^c]\right\}^2 \\ &\quad + \frac{1}{\sum_{n=1}^{r}\lambda^+_{j_n}} \sum_{n=1}^{r} \lambda^+_{j_n} a_{j_n} \end{aligned} \tag{5}$$

and $(\sigma^2_{\mu j_n})^+ = (\sigma^2_{\mu j_1})^+$ $(n = 2, \ldots, r)$, where the formulas for λ^+_j, $P_{ij}(\psi^c)$, u_i^c, $(\sigma^2_j)^c$, and a_j are the same as before. Also, upon using the fact that $\sigma^2_{\upsilon k_n} = \sigma^2_{\upsilon k_1}$ $(n = 2, \ldots, s)$, we get

$$\begin{aligned}(\sigma^2_{\upsilon k_1})^+ &= \frac{1}{NT\sum_{n=1}^{s}\lambda^+_{k_n}} \sum_{i=1}^{N}\sum_{n=1}^{s} P_{ik_n}(\psi^c)\left\{u_i^{c\prime} Q u_i^c + [(\sigma^2_{\upsilon k_n})^c \iota_T' u_i^c / (T(\sigma^2_{k_n})^c)]^2 / T\right\} \\ &\quad + \frac{1}{\sum_{n=1}^{s}\lambda^+_{k_n}} \sum_{n=1}^{s} \lambda^+_{k_n} a_{k_n} \end{aligned} \tag{6}$$

and $(\sigma^2_{\upsilon k_n})^+ = (\sigma^2_{\upsilon k_1})^+$ $(n = 2, \ldots, s)$. (The derivations leading to equations (5) and (6) are provided in the appendix.) Moreover, if a particular $\sigma^2_{\mu j}$ is not restricted to be equal to any other cross-sectional specific effect variance, then $(\sigma^2_{\mu j})^+$ is calculated according to the formula in (2); similarly, if $\sigma^2_{\upsilon k}$ is unrestricted, then $(\sigma^2_{\upsilon k})^+$ is calculated according to (3).

The new fit for b^+ is still given by (4).

3. Application

The AECM algorithm, with and without equality restrictions imposed on the variance components, was applied to calculate estimates of a model previously considered by Baltagi and Griffin (1983, 1988) and Phillips (2003). The model relates the logarithm of gasoline consumption per car (Gas/Car) to the logarithms of real per capita income (Y/N), lagged real gasoline prices (P_{MG}/P_{GDP}), and cars per capita (Car/N):

$$\ln\left(\frac{Gas}{Car}\right)_{it} = \alpha + \beta_1 \ln\left(\frac{Y}{N}\right)_{it} + \beta_2 \sum_{j=1}^{n} \omega_j \ln\left(\frac{P_{MG}}{P_{GDP}}\right)_{i,t-j} + \beta_3 \ln\left(\frac{Car}{N}\right)_{it} + u_{it}.$$

Using annual data for 18 OECD countries covering the period 1969 to 1978 this model was estimated using several different stratified error-components models.[2]

[2]For a description of data sources and the construction of the variables see Phillips (2003).

Table 1: Strata Membership and Maximum Posterior Probabilities

Country	Maximum Posterior Probability	Country	Maximum Posterior Probability
	First Stratum ($\widehat{\sigma}_{\upsilon 1} = 0.032$, $\widehat{\sigma}_{\mu 1} = 0.060$, $\widehat{\lambda}_1 = 0.293$)		
Belgium	0.834	Norway	0.836
France	0.815	Switzerland	0.751
Germany	0.857	U.K.	0.721
	Second Stratum ($\widehat{\sigma}_{\upsilon 2} = 0.032$, $\widehat{\sigma}_{\mu 2} = 0.493$, $\widehat{\lambda}_2 = 0.278$)		
Canada	1.000	Spain	0.996
Ireland	1.000	U.S.A.	1.000
	Third Stratum ($\widehat{\sigma}_{\upsilon 3} = 0.067$, $\widehat{\sigma}_{\mu 3} = 0.060$, $\widehat{\lambda}_3 = 0.429$)		
Austria	0.997	Japan	1.000
Denmark	1.000	The Netherlands	0.995
Greece	1.000	Sweden	0.597
Italy	0.861	Turkey	1.000

In Phillips (2003) maximum-likelihood estimates were calculated for this model with $q = 2$ and with the $\sigma^2_{\mu j}$s and $\sigma^2_{\upsilon j}$s left unconstrainted using the EM algorithm described in that paper. When the AECM algorithm was applied to the same model, I obtained estimates that were the same as those reported in Phillips (2003). The estimates indicated that the large $\sigma^2_{\mu j}$ is associated with the small $\sigma^2_{\upsilon j}$; in other words, those cross sections with more dispersion in μ_i have less dispersion in υ_{it}.

A model that allows for more possibilities while increasing the number of free parameters by only two (specifically, it introduces λ_3 and λ_4) is obtained by setting $q = 4$ and imposing the restrictions $\sigma^2_{\upsilon 1} = \sigma^2_{\upsilon 2}$, $\sigma^2_{\upsilon 3} = \sigma^2_{\upsilon 4}$, $\sigma^2_{\mu 1} = \sigma^2_{\mu 3}$, and $\sigma^2_{\mu 2} = \sigma^2_{\mu 4}$ (with $\sigma^2_{\upsilon 1} < \sigma^2_{\upsilon 3}$ and $\sigma^2_{\mu 1} < \sigma^2_{\mu 2}$). This model allows for a small $\sigma^2_{\upsilon j}$ and small $\sigma^2_{\mu j}$ stratum, a small $\sigma^2_{\upsilon j}$ and large $\sigma^2_{\mu j}$ stratum, a large $\sigma^2_{\upsilon j}$ and small $\sigma^2_{\mu j}$ stratum, and a large $\sigma^2_{\upsilon j}$ and large $\sigma^2_{\mu j}$ stratum. However, when this model was estimated, there was no evidence supporting the presence of a large $\sigma^2_{\upsilon j}$ and large $\sigma^2_{\mu j}$ stratum. In particular, the estimate of λ_4 was virtually zero.

Therefore, a more parsimonious model with $q = 3$ and the restrictions $\sigma^2_{\upsilon 1} = \sigma^2_{\upsilon 2}$ and $\sigma^2_{\mu 1} = \sigma^2_{\mu 3}$ was estimated. An unrestricted model with $q = 3$ was also estimated, but the log-likelihood value for the unrestricted model was only marginally larger than the log-likelihood of the model with the restrictions $\sigma^2_{\upsilon 1} = \sigma^2_{\upsilon 2}$ and $\sigma^2_{\mu 1} = \sigma^2_{\mu 3}$ imposed. When these restrictions were imposed, the estimates of long-run demand elasticity with respect to per capita income, price, and cars per capita were 0.472, -0.486, and -0.627. As for the estimates of $\sigma_{\upsilon j}$, $\sigma_{\mu j}$, and λ_j ($j = 1, 2, 3$), they are provided in Table 1 along with estimated maximum posterior probabilities.

Estimates of the posterior probablities $P_{ij}(\psi)$ ($i = 1, \ldots, N,\ \ j = 1, \ldots, q$) can be used to

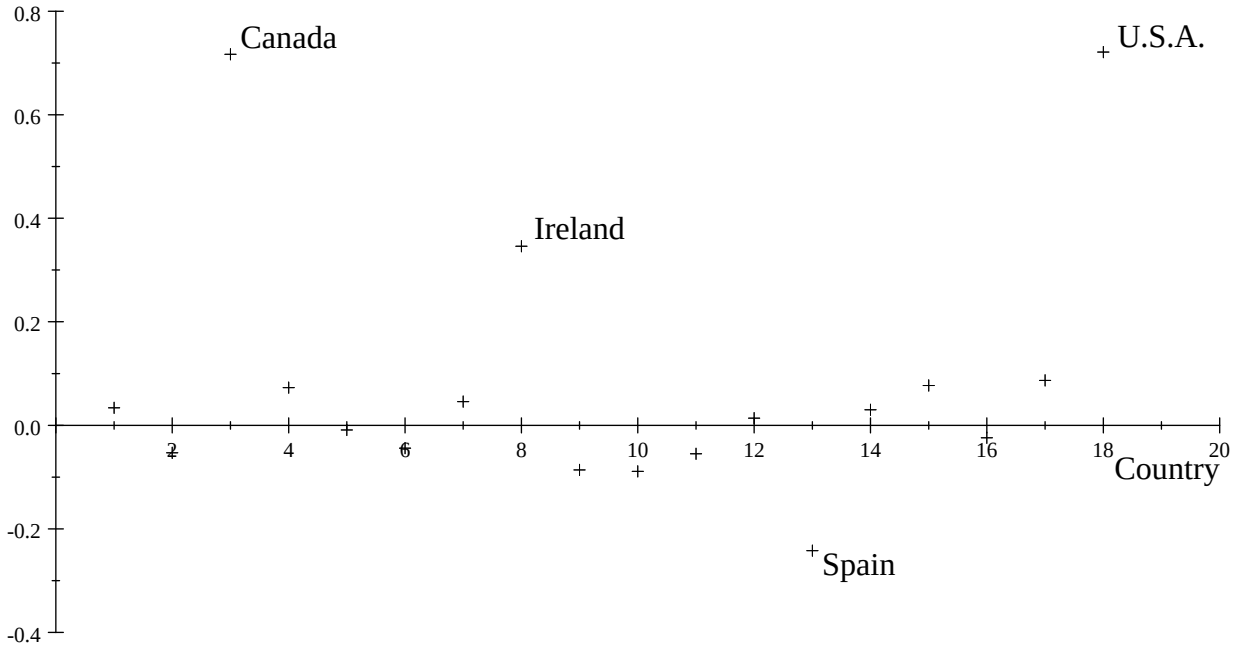

Figure 1: Conditional Means of Cross-Sectional Specific Effects

classify the sample countries into strata. Assigning each country to that stratum for which the posterior probability $P_{ij}(\psi)$ is largest minimizes the assignment error rate (see McLachlan and Basford, 1988, p. 11). According to the estimates provided in Table 1, Canada, Ireland, Spain, and the U.S.A. are outliers in terms of the country specific effects (μ_is), for they are assigned to the stratum corresponding to $\widehat{\sigma}_{\mu 2}$, which is over eight times the size of $\widehat{\sigma}_{\mu 1}$ $(= \widehat{\sigma}_{\mu 3})$.

We can estimate the country specific effects for Canada, Ireland, Spain, and the U.S.A., as well as for the other sample countries, with estimates of the posterior means $E_\psi(\mu_i|y_i, x_i)$ $(i = 1, \ldots, q)$. Straightforward calculations give that

$$E_\psi(\mu_i|y_i, x_i) = \sum_{j=1}^{q} P_{ij}(\psi)\sigma^2_{\mu j}\iota'_T(y_i - X_i b)/(T\sigma^2_j). \tag{7}$$

An estimate of $E_\psi(\mu_i|y_i, x_i)$ can therefore be obtained by replacing the unknown elements of ψ on the right-hand side of (7) with maximum-likelihood estimates.

Figure 1 plots the estimated posterior means obtained when the model was estimated with $q = 3$ while imposing the restrictions $\sigma^2_{v1} = \sigma^2_{v2}$ and $\sigma^2_{\mu 1} = \sigma^2_{\mu 3}$. According to these estimates, during the period 1969 to 1978 Canadians and Americans consumed about 72 percent more gasoline per car, on average, than the average amount consumed in the other OECD countries even after controlling for per capita income, gasoline prices, and cars per capita. Gasoline consumption per car was also higher in Ireland by 35 percent, but lower in Spain by 24 percent on average.

Appendix

This appendix provides the derivations for the formulas of the AECM algorithm.

In order to maximize $Q_1(\psi;\psi^c)$ with respect to the λ_js while imposing the restriction $\sum_{j=1}^q \lambda_j = 1$ consider the Lagrangean function $\sum_{j=1}^q \ln(\lambda_j)\sum_{i=1}^N P_{ij}(\psi^c) + z(1-\sum_{j=1}^q \lambda_j)$, where z is the Lagrangean multiplier. The first-order conditions for maximizing this function yields the solutions $\lambda_j^+ = \sum_{i=1}^N P_{ij}(\psi^c)/z^+$ $(j=1,\ldots,q)$. The restriction that $\sum_{j=1}^q \lambda_j^+ = 1$ implies $z^+ = N$.

Maximizing $Q_1(\psi;\psi^c)$ with respect to $\sigma^2_{\mu j}$ gives $(\sigma^2_{\mu j})^+ = \sum_{i=1}^N E_{\psi^c}(d_{ij}\mu_i^2|y_i,x_i)/(N\lambda_j^+)$. And, on exploiting the law of iterated expectations we find that $E_\psi(d_{ij}\mu_i^2|y_i,x_i) = E_\psi[d_{ij}E_\psi(\mu_i^2|y_i,x_i,d_i)|y_i,x_i] = E_\psi(\mu_i^2|y_i,x_i,d_i=\omega_j)\Pr_\psi(d_i=\omega_j|y_i,x_i)$, and $\Pr_\psi(d_i=\omega_j|y_i,x_i) = P_{ij}(\psi)$. Also, $E_\psi(\mu_i^2|y_i,x_i,d_i=\omega_j) = [E_\psi(\mu_i|y_i,x_i,d_i=\omega_j)]^2 + \upsilon ar_\psi(\mu_i|y_i,x_i,d_i=\omega_j)$, and since μ_i and y_i are jointly normal conditional on x_i and $d_i=\omega_j$, it follows from multivariate normal theory (see, e.g., Greene, 2003, p. 872) that $E_\psi(\mu_i|y_i,x_i,d_i=\omega_j) = \sigma^2_{\mu j}\iota_T'\Sigma_j^{-1}u_i$ (with $u_i = y_i - X_i b$) and $\upsilon ar_\psi(\mu_i|y_i,x_i,d_i=\omega_j) = \sigma^2_{\mu j}\left(1-\sigma^2_{\mu j}\iota_T'\Sigma_j^{-1}\iota_T\right)$. Upon using $\Sigma_j^{-1} = Q/\sigma^2_{\upsilon j} + \iota_T\iota_T'/(\sigma_j^2T^2)$ (see, e.g., Hsiao, 1990, p. 35, Eq. (3.3.8)), $\iota_T'Q = \mathbf{0}$, and some manipulations, we obtain $E_\psi(\mu_i|y_i,x_i,d_i=\omega_j) = \sigma^2_{\mu j}\iota_T'u_i/(T\sigma_j^2)$ and $\upsilon ar_\psi(\mu_i|y_i,x_i,d_i=\omega_j) = \sigma^2_{\mu j}\sigma^2_{\upsilon j}/(T\sigma_j^2)$. These observations imply Eq. (2).

Maximizing $Q_1(\psi;\psi^c)$ with respect to $\sigma^2_{\upsilon j}$ gives $(\sigma^2_{\upsilon j})^+ = \sum_{i=1}^N E_{\psi^c}(d_{ij}\upsilon_i'\upsilon_i|y_i,x_i)/(NT\lambda_j^+)$. Applying the law of iterated expectations we obtain $E_\psi(d_{ij}\upsilon_i'\upsilon_i|y_i,x_i) = E_\psi(\upsilon_i'\upsilon_i|y_i,x_i,d_i=\omega_j)P_{ij}(\psi)$. Furthermore, $E_\psi(\upsilon_i'\upsilon_i|y_i,x_i,d_i=\omega_j) = E_\psi(\upsilon_i'|y_i,x_i,d_i=\omega_j)E_\psi(\upsilon_i|y_i,x_i,d_i=\omega_j) + tr[Var_\psi(\upsilon_i|y_i,x_i,d_i=\omega_j)]$, where $Var_\psi(\cdot|y_i,x_i,d_i=\omega_j)$ denotes a conditional variance-covariance matrix using ψ as the parameter vector of the conditional distribution. It follows from multivariate normal theory that $E_\psi(\upsilon_i|y_i,x_i,d_i=\omega_j)] = \sigma^2_{\upsilon j}\Sigma_j^{-1}u_i$ and $Var_\psi(\upsilon_i|y_i,x_i,d_i=\omega_j) = \sigma^2_{\upsilon j}(I_T - \sigma^2_{\upsilon j}\Sigma_j^{-1})$. And, some manipulations give $\sigma^4_{\upsilon j}u_i'\Sigma_j^{-1}\Sigma_j^{-1}u_i = u_i'Qu_i + [\sigma^2_{\upsilon j}\iota_T'u_i/(T\sigma_j^2)]^2/T$, while $tr[\sigma^2_{\upsilon j}(I_T - \sigma^2_{\upsilon j}\Sigma_j^{-1})] = \sigma^2_{\upsilon j}[1-\sigma^2_{\upsilon j}/(T\sigma_j^2)] = \sigma^2_{\upsilon j}\sigma^2_{\mu j}/\sigma_j^2$. These results imply Eq. (3).

To obtain (5) first observe that $\sum_{n=1}^r \partial Q_1^*(\psi;\psi^c)/\partial\sigma^2_{\mu j_n} = \sum_{n=1}^r \partial Q_1(\psi;\psi^c)/\partial\sigma^2_{\mu j_n}$. Next, let $\psi^{+/2}$ satisfy the first-order condtions that $\partial Q_1^*(\psi^{+/2};\psi^c)/\partial\sigma^2_{\mu j_n} = 0$ $(n=1,\ldots,r)$. Then it follows from the foregoing that

$$\sum_{n=1}^r \partial Q_1(\psi^{+/2};\psi^c)/\partial\sigma^2_{\mu j_n} = 0. \tag{8}$$

Upon setting $(\sigma^2_{\mu j_n})^+ = (\sigma^2_{\mu j_1})^+$ $(n=2,\ldots,r)$, we can solve equation (8) for $(\sigma^2_{\mu j_1})^+$, which

gives

$$
\begin{aligned}
(\sigma^2_{\mu j_1})^+ &= \frac{1}{N\sum_{n=1}^r \lambda^+_{j_n}} \sum_{i=1}^N \sum_{n=1}^r E_{\psi^c}(d_{ij_n}\mu_i^2|y_i, x_i) \\
&= \frac{1}{N\sum_{n=1}^r \lambda^+_{j_n}} \sum_{i=1}^N \sum_{n=1}^r P_{ij_n}(\psi^c) E_{\psi^c}(\mu_i^2|y_i, x_i, d_i = \omega_{j_n}) \\
&= \frac{1}{N\sum_{n=1}^r \lambda^+_{j_n}} \sum_{i=1}^N \sum_{n=1}^r P_{ij_n}(\psi^c) \left\{(\sigma^2_{vj_n})^c \iota_T' u_i^c / [T((\sigma^2_{j_n})^c]\right\}^2 \\
&\quad + \frac{1}{\sum_{n=1}^r \lambda^+_{j_n}} \sum_{n=1}^r \lambda^+_{j_n} a_{j_n}.
\end{aligned}
$$

Verification of (6) is similar.

References

Baltagi, B. H., and J. M. Griffin (1983) "Gasoline demand in the OECD: An application of pooling and testing procedures" *European Economic Review* 22, 117-137.

Baltagi, B. H., and J. M. Griffin (1988) "A generalized error component model with heteroscedastic disturbances" *International Economic Review* 29, 745-753.

Greene, W. H. (2003) *Econometric Analysis*, 5th ed., Prentice Hall: Upper Saddle River.

Hsiao, C. (1990) *Analysis of Panel Data*, Cambridge University Press: New York.

McLachlan, G. J., and K. E. Basford (1988) *Mixture Models: Inference and Applications to Clustering*, Marcel Dekker: New York.

Meng, X. L, and D. van Dyk (1997) "The EM algorithm—an old folk-song sung to a fast new tune" *Journal of the Royal Statistical Society* B59, 511-567.

Meng, X. L., and D. van Dyk (1998) "Fast EM-type implementations for mixed effects models" *Journal of the Royal Statistical Society* B60, 559-578.

Neyman, J., and E. L. Scott (1948) "Consistent estimates based on partially consistent observations" *Econometrica* 16, 1-32.

Phillips, R. F. (2003) "Estimation of a stratified error-components model" *International Economic Review* 44, 501-521.