ANNALES D'ÉCONOMIE ET DE STATISTIQUE. – N° 69 – 2003

Random Coefficients in Unbalanced Panels: An Application on Data From Chemical Plants

Erik BIØRN, Kjersti-Gro LINDQUIST, Terje SKJERPEN *

ABSTRACT. – A framework for analyzing substitution and scale properties from plant-level panel data is presented. Focus is on comparing the constant and random coefficient specification of the substitution and scale parameters and investigating the variation of the parameters across plants. Characteristics of the model framework are *(i)* an equation system consisting of a three-factor translog cost function and the corresponding cost-share equations, *(ii)* random plant specific heterogeneity in coefficients, and *(iii)* a Maximum Likelihood procedure allowing for unbalanced panel data. The empirical results, based on data from Norwegian chemical plants, indicate pronounced plant specific heterogeneity in substitution and scale properties. Substantial parts of the variances of the cost and the cost shares can be ascribed to variation in the coefficient vector. The estimated mean scale properties are considerably influenced by the choice of model specification, while conclusions regarding price effects are more robust.

Coefficients aléatoires et données de panel non-cylindrées : une application sur des données provenant d'entreprises chimiques

RÉSUMÉ. – Dans cet article, nous présentons une méthode économétrique permettant d'analyser les effets de substitution et d'échelle de production en utilisant des données de panel d'entreprises. La méthode consiste à comparer des spécifications à coefficients fixés avec des spécifications à coefficients aléatoires en modélisant les paramètres de substitution et d'échelle de production, et à examiner la variabilité des paramètres entre les entreprises. Les modèles sont caractérisés par *(i)* un système d'équations composé d'une fonction du coût à trois facteurs de forme translog et des fonctions de demande de facteurs correspondantes, *(ii)* la spécification de l'hétérogénéité des coefficients entre les entreprises et *(iii)* une procédure d'estimation par maximum vraisemblance permettant l'utilisation de données de panel non cylindrées. Les résultats empiriques, obtenues sur des données d'entreprises chimiques norvégiennes, indiquent une hétérogénéité prononcée dans les effets de substitution et d'échelle de production. Une part substantielle des variances des coûts observées et des parts de chacun des facteurs dans le coût total s'explique par l'hétérogénéité des coefficients. Les effets d'échelle de production estimés sont très sensibles à la spécification du modèle, en revanche les conclusions concernant les effets prix sont plus robustes.

* E. BIØRN: University of Oslo, Department of Economics; and Statistics Norway, Research Department; K.-G. LINDQUIST: Central Bank of Norway, Research Department. Work done while the author was at Statistics Norway; T. SKJERPEN: Statistics Norway, Research Department.
Previous versions of the paper were presented at the *Eighth International Conference on Panel Data*, Göteborg, 1998, and at the *Econometric Society European Meeting*, Berlin, 1998. The detailed and very helpful comments of two referees are gratefully acknowledged. We also thank Petter FRENGER, Stéphane GREGOIR, Jan LARSSON, and conference participants for comments.

1 Introduction

A common challenge in microeconometric analyses of economic relationships is how to treat heterogeneity with respect to the functional form across the micro units. Such heterogeneity may be modelled and analyzed when panel data are available, unlike the situation when only cross-section data are at hand. However, even in a panel data context, most researchers have tended to assume a common coefficient structure, possibly allowing for unit specific (or time specific) differences in the intercept terms of the equations ("fixed" or "random" effects) only. If the heterogeneity has a more complex form, this modelling approach may lead to inefficient estimation of the slope coefficients and invalid inference.

A more general approach is to allow for heterogeneity also in slope coefficients. The challenge then becomes how to obtain a model which is sufficiently flexible while avoiding overparametrization. The fixed effects slope coefficients approach, in which each unit has its distinct coefficient vector, with no *a priori* assumptions made about its variation across units, is very flexible, but may easily suffer from the degrees of freedom problem. The *random coefficients* approach, in which specific assumptions are made about the distribution from which the unit specific coefficients are drawn, is far more parsimonious in this respect. The expectation vector in this distribution represents, in a precise way, the coefficients of an average unit, while its second-order moments matrix gives easily interpretable measures of the degree of heterogeneity.

While there is a growing body of methodological articles in the econometric literature dealing with this random coefficient problem for balanced panel data, far less has been done with *unbalanced panel data*. This is somewhat surprising, since in practice the latter is rather the rule than the exception. Because working with complete panels is mathematically more convenient, a common procedure for researchers is to leave out the units for which the time series are incomplete and use the balanced sub-sample of the original, less tidy data-set. This may, however, involve a loss of efficiency, see MÁTYÁS and LOVRICS [1991] and BALTAGI and CHANG [1994].

In this paper, we consider a general and rather flexible framework for analyzing *heterogeneity in the production process*, including factor substitution and scale properties, from unbalanced plant-level panel data. The technology of the average plant is described by a parametric cost function with variable returns to scale and technological trends. The paper focuses on alternative specifications of random plant heterogeneity of the coefficients and compares these with specifications under full homogeneity. This modelling framework is accommodated to an application on plant-level panel data for Norwegian chemical industries for the years 1972-1993.

The framework of the analysis has the following basic characteristics: *(i)* A translog cost-function with three variable inputs and derived cost-share equations. Capital is treated as a "quasi-fixed" input, reflecting the long lead-time needed to build new capacity. Modelling the optimal capital accumulation mechanism is beyond the scope of the investigation. *(ii)* Plant specific heterogeneity in coefficients of the cost function is allowed for and treated as random. *(iii)* The model is designed for an unbalanced plant-level panel data set. While *(i)* is rather standard, the combination of *(ii)* and *(iii)* is not, at least in applied econometrics. Mixed regression models with unbalanced design have to some extent been discussed in the statistical literature, see, *eg,* AMEMIYA [1994] and SHIN [1995]. Random coefficients in regression equations in econometrics are treated in the pioneering studies of SWAMY [1970, 1971, 1974]; an extension to simultaneous equation systems is discussed in BALESTRA and NEGASSI [1992]. See also HSIAO [1975, 1996] and LONGFORD [1995a,b]. MAIRESSE and GRILICHES [1990] analyze constant returns to scale two-factor Cobb-Douglas production functions in a random coefficients setting, using balanced panel data for 13 years for manufacturing firms in France, Japan, and the United States. To our knowledge, no previous study has considered the more flexible translog parametrization of the cost function in a similar framework.

We find that substantial improvement in model fit is obtained when allowing for random coefficient heterogeneity – even within such a flexible parametrization of the average technology as the non-constant returns to scale translog model. In addition, conclusions regarding scale properties and the effect of the capital stock on variable costs are considerably influenced by the choice of model specification, while conclusions regarding price effects are more robust. The results also indicate that none of the variable inputs are price-elastic, and that labour, energy, and materials are all substitutes.

The rest of the paper is organized as follows. In Section 2, we present the model and its econometric specification with random cost function coefficients. In Section 3, we present the estimation procedure and the data, and in Section 4, we report and discuss the empirical results. Concluding remarks are given in Section 5.

2 Model and Notation

We assume that three inputs are subject to cost minimization: labour (L), energy (E), and materials (M), with prices and quantities denoted as P_g and Q_g $(g = L, E, M)$, and their total factor cost as $C = \sum P_g Q_g$. The quantities of output and capital are denoted as X and K respectively, and both, like the input prices, are treated as exogenous variables. A deterministic trend, τ, representing, *inter alia*, technological change, is included.

Some comments should be made about the exogeneity assumptions. While the capital stock is treated as a predetermined variable due to the long lead-time needed to build new capacity, wages are assumed to be exogenous due to the centralized wage formation system in Norway and a very high degree of unionization in this industry. Wages vary across plants, however, due to variations in skill structure and seniority. We use industry rather than plant specific prices of materials (see Appendix A), which should reduce the potential endogeneity problem with respect to this variable. The energy price is closely linked to the price of electricity, and our sample consists of a mixture of plants trading in the spot-market for electricity and plants with long-term electricity contracts (frequently up to several years), specifying both price and quantity in advance. In general, the latter group covers chemical plants that use electricity heavily and which therefore may influence the market price. Hence, one can argue that chemical plants are either price takers in the electricity market or face predetermined electricity prices. Ideally, for plants operating with long-term electricity contracts, information on the time at which electricity prices were negotiated and the length of the contract periods would have been useful, but our data base does not include this kind of information. Output is probably the most problematic variable with respect to the exogeneity assumption, even if there are strong arguments also for this assumption. Over time, it has become increasingly common among Norwegian chemical plants to make agreements about future deliveries, which implies that they are committed to produce an *a priori* given output level. The motivation for long-term contracting is that chemical plants produce not only staple goods, but also specialized products that are developed to satisfy specific customers; *cf.* PIL [1996].

The production technology, represented by its dual cost function, is parametrized as a *translog* function in $(X, K, \tau, P_L, P_E, P_M)$ (see CHRISTENSEN, JORGENSON and LAU [1971, 1973] and JORGENSON [1986]). This is a flexible functional form in the sense that it could provide a second-order differential approximation to an arbitrary twice continuously differentiable cost function that satisfies linear homogeneity in prices at any point in an admissible domain (see DIEWERT and WALES [1987, p. 45]).[1] From an examination of the data it is clear that a flexible functional form should be considered, since relative input values vary significantly across plants and also over time for each plant. *A priori*, it was not obvious that this variation could be explained by the variation in input prices and other explanatory variables within a simple functional form.[2]

Moreover, in our context, the translog parametrization has some attractive econometric properties. The cost-shares have constant elasticities both with respect to prices (see JORGENSON [1986, p. 1856]) and with respect to output and capital. Also, the inverse scale elasticity of the variable inputs (which is the elasticity of the cost function with respect to output) is allowed to depend linearly on the logs of output, capital, and the input prices. Constant scale elasticity emerges as a special case. These properties are the main reasons why we prefer this specification to *eg,* the Generalized Leontief (see DIEWERT [1971]) or the Generalized McFadden cost function (see DIEWERT and WALES [1987]), which possess a similar degree of flexibility. See also LAU [1986].

Using lower-case letters to symbolize logarithms, *ie,* $c = \ln C$, $x = \ln X$, $p_M = \ln P_M$ etc., the translog cost function, after imposing linear homogeneity in prices, can be written as:

1. Such Taylor series approximations to flexible functional forms may, however, be criticized; see GALLANT [1981].
2. An inspection of the Annual Reports of large Norwegian chemical companies also made it clear to us that assuming a cost minimizing behaviour contingent on the capital stock seems sensible.

$$(1)\quad (c - p_M) = \beta_0 + \beta_X x + \beta_K k + \beta_\tau \tau \\ + \gamma_L (p_L - p_M) + \gamma_E (p_E - p_M) \\ + \beta_{XL} x (p_L - p_M) + \beta_{XE} x (p_E - p_M) \\ + \beta_{KL} k (p_L - p_M) + \beta_{KE} k (p_E - p_M) \\ + \beta_{\tau L} \tau (p_L - p_M) + \beta_{\tau E} \tau (p_E - p_M) \\ + \frac{1}{2}\gamma_{LL}(p_L - p_M)^2 + \gamma_{LE}(p_L - p_M)(p_E - p_M) \\ + \frac{1}{2}\gamma_{EE}(p_E - p_M)^2 \\ + \frac{1}{2}\beta_{XX} x^2 + \frac{1}{2}\beta_{KK} k^2 + \frac{1}{2}\beta_{\tau\tau}\tau^2 \\ + \beta_{XK} xk + \beta_{\tau X}\tau x + \beta_{\tau K}\tau k + u_C,$$

where we have added the disturbance u_C. Application of Shephard's lemma then gives the following cost-share equations of labour and energy,[3]

$$(2)\quad s_L = \gamma_L + \beta_{XL} x + \beta_{KL} k + \beta_{\tau L}\tau \\ + \gamma_{LL}(p_L - p_M) + \gamma_{LE}(p_E - p_M) + u_L,$$

$$(3)\quad s_E = \gamma_E + \beta_{XE} x + \beta_{KE} k + \beta_{\tau E}\tau \\ + \gamma_{LE}(p_L - p_M) + \gamma_{EE}(p_E - p_M) + u_E,$$

where u_L and u_E are disturbances. We define the following coefficients for the material input, exploiting the adding-up, homogeneity, and symmetry conditions on the cost function,

$$\gamma_M = 1 - \gamma_L - \gamma_E,$$
$$\beta_{XM} = -\beta_{XL} - \beta_{XE}, \quad \beta_{KM} = -\beta_{KL} - \beta_{KE}, \quad \beta_{\tau M} = -\beta_{\tau L} - \beta_{\tau E},$$
$$\gamma_{LM} = -\gamma_{LL} - \gamma_{LE}, \quad \gamma_{EM} = -\gamma_{EE} - \gamma_{LE}, \quad \gamma_{MM} = \gamma_{LL} + 2\gamma_{LE} + \gamma_{EE}.$$

Cross-price and own-price elasticities of substitution in the demand for factor g with respect to the price of factor h, defined as the Slutsky analogues (output constrained price elasticities of input quantities), are:

$$(4)\quad \varepsilon_{gh} = \begin{cases} \dfrac{\gamma_{gh}}{s_g} + s_h, & g \neq h, \\ \dfrac{\gamma_{gg}}{s_g} + s_g - 1, & g = h, \end{cases}$$

3. To these equations may be added, for symmetry reasons, the corresponding equation for materials, but to avoid singularity of the disturbance covariance matrix, it is omitted from the model.

which satisfy $\sum_h \varepsilon_{gh} = 0$. The corresponding Allen Uzawa cross-price elasticities are

$$\eta_{gh} = \eta_{hg} = \frac{\varepsilon_{gh}}{s_h} = \frac{\gamma_{gh}}{s_g s_h} + 1, \qquad g \neq h. \tag{5}$$

The cost elasticity of output and capital, and the rate of increase of cost with time are, respectively,

$$(\partial c)/(\partial x) = \varepsilon_X = \beta_X + \beta_{XX} x + \beta_{XK} k + \beta_{\tau X} \tau + \textstyle\sum_g \beta_{Xg}(p_g - p_M), \tag{6}$$

$$(\partial c)/(\partial k) = \varepsilon_K = \beta_K + \beta_{XK} x + \beta_{KK} k + \beta_{\tau K} \tau + \textstyle\sum_g \beta_{Kg}(p_g - p_M), \tag{7}$$

$$(\partial c)/(\partial \tau) = \varepsilon_\tau = \beta_\tau + \beta_{\tau X} x + \beta_{\tau K} k + \beta_{\tau\tau} \tau + \textstyle\sum_g \beta_{\tau g}(p_g - p_M). \tag{8}$$

Two *model classes* will be considered:

Model A: The two cost-share equations, (2) and (3).

Model B: The cost function, (1), and the two cost-share equations, (2) and (3).

Identification of the scale properties of the technology and the trend effects is possible within Model B only. The substitution properties can be identified from both models. Model A contains 11 and Model B contains 21 (fixed or random) coefficients. Since Model B incorporates more prior information, it leads to more efficient estimation within a full Maximum Likelihood procedure, provided that the restrictions are valid. Estimation of cost-share equation systems without including the cost function in the econometric model, is a common approach, and it is therefore of interest to compare the estimated price responses from Models A and B.

The data are from an unbalanced panel, in which the plants are observed in at least 1 and at most P years. We assume that the selection rules for the unbalanced panels are ignorable, *ie,* the way in which the plants enter or exit is not related to the endogenous variables in the model; see VERBEEK and NIJMAN [1996, section 18.2]. This assumption is discussed in Sections 3 and 4. The plants are arranged in groups according to the number of years they are observed. Let N_p be the number of plants which are observed in p years (not necessarily the same and not necessarily consecutive), let (ip) index the i'th plant among those observed in p years ($i = 1, \ldots, N_p$; $p = 1, \ldots, P$), and let t index the observation number ($t = 1, \ldots, p$). The total number of plants in the panel is $N = \sum_{p=1}^{P} N_p$ and the total number of observations is $n = \sum_{p=1}^{P} N_p p$.

The two models can be written compactly as:

$$\boldsymbol{y}_{(ip)t} = \boldsymbol{X}_{(ip)t} \boldsymbol{\beta}_{(ip)} + \boldsymbol{u}_{(ip)t}, \quad p = 1, \ldots, P;\ i = 1, \ldots, N_p;\ t = 1, \ldots, p, \tag{9}$$

where $\boldsymbol{\beta}_{(ip)}$ is the coefficient vector of plant (ip). *Model B* is characterized by:

$$\boldsymbol{X}'=\begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & x \\ 0 & 0 & k \\ 0 & 0 & \tau \\ 1 & 0 & (p_L-p_M) \\ 0 & 1 & (p_E-p_M) \\ x & 0 & x(p_L-p_M) \\ 0 & x & x(p_E-p_M) \\ k & 0 & k(p_L-p_M) \\ 0 & k & k(p_E-p_M) \\ (p_L-p_M) & 0 & \frac{1}{2}(p_L-p_M)^2 \\ (p_E-p_M) & (p_L-p_M) & (p_L-p_M)(p_E-p_M) \\ 0 & (p_E-p_M) & \frac{1}{2}(p_E-p_M)^2 \\ \tau & 0 & \tau(p_L-p_M) \\ 0 & \tau & \tau(p_E-p_M) \\ 0 & 0 & \frac{1}{2}x^2 \\ 0 & 0 & \frac{1}{2}k^2 \\ 0 & 0 & \frac{1}{2}\tau^2 \\ 0 & 0 & xk \\ 0 & 0 & \tau x \\ 0 & 0 & \tau k \end{bmatrix}, \boldsymbol{\beta}=\begin{bmatrix} \beta_0 \\ \beta_X \\ \beta_K \\ \beta_\tau \\ \gamma_L \\ \gamma_E \\ \beta_{XL} \\ \beta_{XE} \\ \beta_{KL} \\ \beta_{KE} \\ \gamma_{LL} \\ \gamma_{LE} \\ \gamma_{EE} \\ \beta_{\tau L} \\ \beta_{\tau E} \\ \beta_{XX} \\ \beta_{KK} \\ \beta_{\tau\tau} \\ \beta_{XK} \\ \beta_{\tau X} \\ \beta_{\tau K} \end{bmatrix},$$

$$\boldsymbol{y}' = [s_L \;\; s_E \;\; c - p_M], \qquad \boldsymbol{u}' = [u_L \;\; u_E \;\; u_C].$$

In *Model A*, the matrices $\boldsymbol{X}'$, $\boldsymbol{\beta}$, $\boldsymbol{y}'$, and $\boldsymbol{u}'$ are appropriate submatrices of these matrices.

Initially, we specified all coefficients of the translog cost function, and hence all coefficients of the cost-share equations, as random with constant second-order moments. It turned out, however, that estimating such general models raised problems with the numerical computation of the Maximum Likelihood estimates (*cf.* Section 3), probably due to instability of the higher order moments of the sample distribution of the squared and interaction terms in the regressor matrix. Considering the translog cost function as a second order approximation to the underlying cost function, we therefore decided to treat, at most, the intercepts and first order coefficients as random and restricting the second order coefficients, representing the second order derivatives of the log of the cost function, to be constants. Specifically, we consider the following *model versions* in the empirical applications:

Model A1: All 11 coefficients are fixed.

Model A2: γ_L, γ_E are random, the other 9 coefficients are fixed.

Model B1: All 21 coefficients are fixed.

Model B2: $\beta_0, \gamma_L, \gamma_E$ are random, the other 18 coefficients are fixed.

Model B3: $\beta_0, \beta_X, \beta_K, \beta_\tau, \gamma_L, \gamma_E$ are random, the other 15 coefficients are fixed.

In addition, a "constant returns to scale" version of Model B2, denoted as B2R, is considered. Model B3, for example, implies that in the two cost-share equations, only the intercepts are random, whereas the cost equation have five random slope coefficients, and a random intercept. Hence, the derived elasticity functions (4) – (8) will contain random coefficients.

The models are formally systems of G regression equations with random coefficients and with a total of H (fixed or random) coefficients. In Model A, $G = 2, H = 11$, in Model B, $G = 3, H = 21$. The $(G \times 1)$ vector of observations of the regressands in the G equations from plant (ip), observation t is $\boldsymbol{y}_{(ip)t}$, and the corresponding $(G \times H)$ regressor matrix is $\boldsymbol{X}_{(ip)t}$. The $(H \times 1)$ coefficient vector of plant (ip) is:

$$\boldsymbol{\beta}_{(ip)} = \boldsymbol{\beta} + \boldsymbol{\delta}_{(ip)}, \tag{10}$$

where $\boldsymbol{\beta}$ is the common expectation vector of $\boldsymbol{\beta}_{(ip)}$ for all plants, and $\boldsymbol{\delta}_{(ip)}$ is a zero mean vector specific to plant (ip). By inserting (10) in (9), the G equations for plant (ip), observation t, can thus be written as:

$$\boldsymbol{y}_{(ip)t} = \boldsymbol{X}_{(ip)t}\boldsymbol{\beta} + \boldsymbol{\eta}_{(ip)t}, \quad \boldsymbol{\eta}_{(ip)t} = \boldsymbol{X}_{(ip)t}\boldsymbol{\delta}_{(ip)} + \boldsymbol{u}_{(ip)t}. \tag{11}$$

We further assume as our main hypothesis that:

$$\boldsymbol{X}_{(ip)t},\ \boldsymbol{u}_{(ip)t},\ \boldsymbol{\delta}_{(ip)} \ \text{are all independent,} \tag{12}$$

$$\boldsymbol{u}_{(ip)t} \sim \mathsf{IIN}(\mathbf{0}_{G1}, \boldsymbol{\Sigma}^u), \qquad \boldsymbol{\delta}_{(ip)} \sim \mathsf{IIN}(\mathbf{0}_{H1}, \boldsymbol{\Sigma}^\delta), \tag{13}$$

where IIN signifies independently, identically, normally distributed, $\mathbf{0}_{mn}$ is a $(m \times n)$ zero matrix and:

$$\boldsymbol{\Sigma}^u = \begin{bmatrix} \sigma^u_{11} & \cdots & \sigma^u_{1G} \\ \vdots & & \vdots \\ \sigma^u_{G1} & \cdots & \sigma^u_{GG} \end{bmatrix}, \quad \boldsymbol{\Sigma}^\delta = \begin{bmatrix} \boldsymbol{\sigma}^\delta_{11} & \cdots & \boldsymbol{\sigma}^\delta_{1H} \\ \vdots & & \vdots \\ \boldsymbol{\sigma}^\delta_{H1} & \cdots & \boldsymbol{\sigma}^\delta_{HH} \end{bmatrix}.$$

The latter matrix may be singular, with zero columns/rows, reflecting that some coefficients are fixed. In all model versions we consider, $\boldsymbol{\Sigma}^u$ is a full positive definite matrix, while $\boldsymbol{\Sigma}^\delta$ has reduced rank.

This specification assumes that all coefficient variation across plants is purely random. Arguments may be given for allowing systematic coefficient variation as well, by for instance letting some of the coefficients vary linearly with (some of) the covariates. To some extent, such systematic coefficient variation may be accounted for by the translog cost function we have specified, since linear variation of its first-order coefficients with the covariates may be absorbed by the second-order terms of the function.

We stack the p realizations from plant (ip) in:

$$\boldsymbol{y}_{(ip)} = \begin{bmatrix} \boldsymbol{y}_{(ip)1} \\ \vdots \\ \boldsymbol{y}_{(ip)p} \end{bmatrix}, \boldsymbol{X}_{(ip)} = \begin{bmatrix} \boldsymbol{X}_{(ip)1} \\ \vdots \\ \boldsymbol{X}_{(ip)p} \end{bmatrix}, \boldsymbol{u}_{(ip)} = \begin{bmatrix} \boldsymbol{u}_{(ip)1} \\ \vdots \\ \boldsymbol{u}_{(ip)p} \end{bmatrix}, \boldsymbol{\eta}_{(ip)} = \begin{bmatrix} \boldsymbol{\eta}_{(ip)1} \\ \vdots \\ \boldsymbol{\eta}_{(ip)p} \end{bmatrix},$$

which have dimensions $(Gp \times 1)$, $(Gp \times H)$, $(Gp \times 1)$, and $(Gp \times 1)$, respectively. Then we can write (11) as:

$$\text{(14)} \qquad \boldsymbol{y}_{(ip)} = \boldsymbol{X}_{(ip)}\boldsymbol{\beta} + \boldsymbol{\eta}_{(ip)}, \quad \boldsymbol{\eta}_{(ip)} = \boldsymbol{X}_{(ip)}\boldsymbol{\delta}_{(ip)} + \boldsymbol{u}_{(ip)}.$$

It follows from (12) – (14) that:

(15) All $\boldsymbol{\eta}_{(ip)}|\boldsymbol{X}_{(ip)}$ are independent and $\boldsymbol{\eta}_{(ip)}|\boldsymbol{X}_{(ip)} \sim \mathsf{N}(\boldsymbol{0}_{Gp,1}, \boldsymbol{\Omega}_{(ip)})$,

where the $(Gp \times Gp)$ covariance matrix of $\boldsymbol{\eta}_{(ip)}$ is:

$$\text{(16)} \qquad \boldsymbol{\Omega}_{(ip)} = \boldsymbol{X}_{(ip)}\boldsymbol{\Sigma}^{\delta}\boldsymbol{X}'_{(ip)} + \boldsymbol{I}_p \otimes \boldsymbol{\Sigma}^{u}.$$

The gross disturbance $\boldsymbol{\eta}_{(ip)}$ exhibits a particular kind of heteroskedasticity, since the random effects enter multiplicatively to the covariates of the equation. The random coefficients approach may therefore be considered a way of representing disturbance heteroskedasticity.

3 Estimation Procedure and Data

From (14) and (15) it follows that the joint log-density function of plant (ip), *ie,* of $\boldsymbol{y}_{(ip)}$ conditional on $\boldsymbol{X}_{(ip)}$, is

$$L_{(ip)} = -\frac{Gp}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{\Omega}_{(ip)}| - \frac{1}{2}[\boldsymbol{y}_{(ip)} - \boldsymbol{X}_{(ip)}\boldsymbol{\beta}]'\,\boldsymbol{\Omega}^{-1}_{(ip)}\,[\boldsymbol{y}_{(ip)} - \boldsymbol{X}_{(ip)}\boldsymbol{\beta}],$$

so that by utilizing the ordering of the observations in the P groups, we can write the log-likelihood function of all observations on $\boldsymbol{y}$ conditional on all observations on $\boldsymbol{X}$ as:

$$\text{(17)} \qquad L = \sum_{p=1}^{P}\sum_{i=1}^{N_p} L_{(ip)} = -\frac{Gn}{2}\ln(2\pi) - \frac{1}{2}\sum_{p=1}^{P}\sum_{i=1}^{N_p}\ln|\boldsymbol{\Omega}_{(ip)}| - \frac{1}{2}\sum_{p=1}^{P}\sum_{i=1}^{N_p}[\boldsymbol{y}_{(ip)} - \boldsymbol{X}_{(ip)}\boldsymbol{\beta}]'\,\boldsymbol{\Omega}^{-1}_{(ip)}\,[\boldsymbol{y}_{(ip)} - \boldsymbol{X}_{(ip)}\boldsymbol{\beta}].$$

The *Maximum Likelihood (ML)* estimators of $(\beta, \mathbf{\Sigma}^u, \mathbf{\Sigma}^\delta)$ are obtained by maximizing L with respect to (the unknown elements of) these parameter matrices, as given in Section 2.[4] The structure of this problem is more complicated than the ML problem for systems of regression equations in more standard situations with balanced panel data sets and fixed slope coefficients for two (related) reasons. First, the various $\boldsymbol{y}$, $\boldsymbol{X}$, and $\mathbf{\Omega}$ matrices have different number of rows, reflecting the different number of observations of the plants in the panel. Although the dimensions of $\mathbf{\Sigma}^u$ and $\mathbf{\Sigma}^\delta$ in (16) are the same for all plants, the dimensions of $\boldsymbol{X}_{(ip)}$, and hence of $\mathbf{\Omega}_{(ip)}$, differ. Second, different plants have different (gross) disturbance covariance matrices, since $\mathbf{\Omega}_{(ip)}$ depends on $\boldsymbol{X}_{(ip)}$ when $\mathbf{\Sigma}^\delta$ is non-zero.

Primarily, we use data from the *Manufacturing Statistics* database of Statistics Norway, supplemented, to a minor extent, by data from the Norwegian *National Accounts*. All industries classified under SIC-code 351 *Manufacture of Industrial Chemicals* are included. The data set is unbalanced and covers $T = 22$ years (1972-1993), with a total number of plants $N = \sum_p N_p = 90$ and a total number of observations $n = \sum_p N_p p = 1265$. Some plants have gaps in their time series, and the corresponding numbers when only plants with contiguous series are included are $N = \sum_p N_p = 77$ and $n = \sum_p N_p p = 1084$; see Table 1. On average, the plants are observed in 12-13 years. Of these plants, $N_{22} = 29$ are observed in all the 22 years, representing 638 observations (about 60 per cent), and $N_1 = 7$ plants are observed in one year only. Most of the estimation and test results reported below are based on the complete data set, but supplementary results based, *inter alia*, on only the contiguous time series are reported in order to illustrate the sensitivity of the results to, *eg,* entry and exit.

TABLE 1

Number of Plants Classified by Number of Replications

p = no. of observations per plant, N_p = no. of plants observed p times, c = contiguous time series, nc = non-contiguous time series

p	22	21	20	19	18	17	16	15	14	13	12
N_p, c	29	0	3	0	2	3	3	6	0	2	1
N_p, nc	0	0	0	0	0	1	6	0	1	1	0
$N_p p$	638	0	60	0	36	68	144	90	14	39	12

p	11	10	09	08	07	06	05	04	03	02	01	Sum
N_p, c	1	1	1	2	2	3	3	2	3	3	7	77
N_p, nc	1	3	0	0	0	0	0	0	0	0	0	13
$N_p p$	22	40	9	16	14	18	15	8	9	6	7	1265

4. The solution conditions may be simplified by concentrating L over β and maximizing the resulting function with respect to the unknown elements of the $\mathbf{\Omega}$ matrices.

Some remarks on the unbalancedness of the data set and reasons for the gaps in the times series seem appropriate. All large plants are obliged by law to report information on a large number of variables to Statistics Norway. Missing observations due to non-response can therefore be expected to be a minor problem. Three reasons for gaps occurring in the series may be given: *(i)* Only large plants, defined as plants with in general at least 5 employees, are obliged to report. If a plant switches between being "large" and "small" according to this criterion, there may be gaps in its time series. This may cause a potential endogenous selection problem, and ideally, our data set should have included these "missing" observations. An inspection of the data revealed, however, that this was not an important cause for gaps. *(ii)* The plants in our sample are in general multi-output plants and are defined (by Statistics Norway) as belonging to a specific industry depending on their most important products. Although not very common, a plant can switch between two industries due to major shifts in their output structure, and hence go into and out of our sample. With respect to identifying the production technology of true chemical plants, these plants represent a potential problem. However, by reproducing the estimation with the non-contiguous time series removed from the data set, we get an indication of whether such plants tend to "pollute" our estimation results. *(iii)* Gaps may be due to dramatic events such as insolvency. If the same type of production continues at the same location after an inactive period, the plant may re-enter the data base with the same plant-number (but probably with a new company number). Again, by excluding the plants with gaps in the time series, we get an indication of whether this is a major problem with respect to identifying the production technology of true chemical plants.[5]

The output measure is tonnes output. This implies that systematic, gradual quality changes in output over time will be captured by the trend variable in the model and hence, estimated trend effects should not be interpreted as representing technological change solely. The Manufacturing Statistics include very detailed information about plants' output. In theory, it is therefore possible to trace changes in plants' output structure both between and within commodity groups. The latter might pick up changing degree of processing over time, and ideally, we would like to control for such changes. However, the detailed output data suffer from major shifts in the nomenclature in addition to noise due to inconsistency in how plants classify output at the detailed level, and attempts to trace "true" changes in the output structure at the plant level had to be given up. The capital input data are constructed from information on fire insurance values, in combination with information on gross investment flows from the Manufacturing Statistics and the National Accounts. Details on the data and data construction are given in Appendix A.[6]

---

5. DIONNE, GAGNÉ and VANASSE [1998] use a similar translog cost function/cost share model and an unbalanced panel data set of Canadian trucking firms. They extend this framework by endogenizing the entry and exit decisions of the firms. On the other hand, their model includes randomness in the intercept term only. In principle, our random coefficients approach could have been extended along similar lines, but probably at the cost of heavier numerical problems. Their motivation for incorporating endogenous selection, significant regulatory changes, is not relevant to our analysis. Hence, we decided not to follow up this possible extension.

6. When analysing panel data at the plant level one may face a problem with retooling, *ie,* that a plant changes its capital structure and output mix substantially over time. A way of controlling for this phenomenon might be to combine detailed output data with data on investments in attempting to identify when an eventual retooling took place. However, as pointed out above, the detailed output data contain substantial noise. To identify retooling from data on investments alone, we would need information on investment in "new" versus "old" capital, which is not available.

4 Empirical Results

Maximum Likelihood estimates of the six model versions described in Section 2 have been obtained by using the PROC MIXED-procedure in the SAS/STAT software (SAS [1992]).[7] The results are given in Tables 2-12. Except for parts of Table 8, all the results below refer to the full data set from the 90 plants with both contiguous and non-contiguous time series.

Tables 2 (A-Models) and 4 (B-Models) contain the estimated elements of the β vector, *cf.* eq. (11). The estimates of the remaining coefficients are obtained by using the adding-up conditions described in section 2. A comparison of the models that include coefficient heterogeneity shows that several coefficient estimates are relatively stable across models. A clear majority of the coefficient estimates are significant[8] according to the asymptotic standard-error estimates.

TABLE 2
Coefficient Estimates and Standard Errors in A-Models

	Model A1[a]		Model A2[b]	
Coefficient	Estimate	Standard error	Estimate	Standard error
γ_L	0.6192	0.0754	0.4470	0.0790
γ_E	0.3455	0.0431	– 0.1068	0.0481
γ_M	0.9647	0.0993	0.3402	0.0990
β_{XL}	– 0.0100	0.0028	– 0.0052	0.0024
β_{XE}	0.0060	0.0021	0.0010	0.0014
β_{XM}	0.0040	0.0039	0.0042	0.0027
β_{KL}	– 0.0202	0.0036	– 0.0252	0.0054
β_{KE}	0.0071	0.0026	0.0187	0.0037
β_{KM}	0.0131	0.0049	0.0066	0.0068
γ_{LL}	0.0375	0.0166	0.0552	0.0122
γ_{LE}	– 0.0639	0.0077	– 0.0169	0.0052
γ_{LM}	0.0264	0.0193	– 0.0383	0.0135
γ_{EE}	– 0.0573	0.0071	0.0214	0.0042
γ_{EM}	0.1212	0.0114	– 0.0045	0.0068
γ_{MM}	– 0.1476	0.0262	0.0428	0.0171
$\beta_{\tau L}$	– 0.0007	0.0010	– 0.0018	0.0006
$\beta_{\tau E}$	0.0058	0.0007	0.0014	0.0003
$\beta_{\tau M}$	– 0.0050	0.0014	0.0004	0.0008

Note : [a] *Model A1: None of the coefficients are assumed to be random.*
[b] *Model A2:* γ_L, γ_E *and* γ_M *are the expectations of* $\gamma_{L(i,p)}$, $\gamma_{E(i,p)}$ *and* $\gamma_{M(i,p)}$ *respectively.*

7. Various applications of this procedure are discussed in VERBEKE and MOLENBERGHS [1997].
8. A 5 per cent significance level is used throughout the paper.

Measures of the overall model fit are given in Tables 3 and 5. Imposing homogeneity leads to a loss of efficiency, and both the Akaike and the Schwarz Bayesian information criteria support the models with most heterogeneity allowed, *ie,* Models A2 and B3. Furthermore, it is evident that the fit is substantially worsened when the output-elasticity is restricted to one *a priori, ie,* in Model B2R (Table 5, column 4). The Log-Likelihood values of the different models are also given.[9]

In Table 6, own-price and cross-price elasticities of (variable) input demand [*cf.* (4)] and cross-price Allen-Uzawa elasticities [*cf.* (5)] at the global sample mean[10] of the exogenous variables are reported. The latter, unlike the former, are symmetric. The translog functional form implies that the elasticities are functions not only of the random coefficients, but also of the exogenous variables. Typically, the elasticities will therefore vary across plants and over time. When calculating the elasticity estimates in Table 6, we use the estimated expectations of the random coefficients. Because the predicted cost-shares depend on unknown coefficients, all the estimated elasticities are non-linear functions of the coefficients. To calculate standard-error estimates of the elasticities, we utilized a first-order Taylor-expansion, starting the expansion at the global sample mean of the exogenous variables (*cf.* KMENTA [1986, p. 486]).[11]

From Table 6 we conclude that none of the inputs are price-elastic. The only exception is energy in Models A1 and B1. All the cross-price Allen-Uzawa elasticities are positive (at the global mean of the exogenous variables) for the models with heterogeneity (A2, B2, B2R, and B3), which means that all the three (variable) inputs, on average, are substitutes. For the two homogeneous models (A1 and B1), labour and energy come out as complements. Except for

TABLE 3
Overall Measures of Fit[a] in A-Models

	Model A1	Model A2
Number of estimated parameters	14	17
Log-likelihood value	1798.348	3402.395
Akaike's information criterion	1784.348	3385.395
Schwarz's Bayesian criterion	1743.496	3335.789

Note : [a] The Akaike and the Schwarz Bayesian criterion are defined, for a model m, by, respectively, $AIC_m = l_m - v_m$ *and* $SBC = l_m - 0.5 v_m ln(r_m N_m)$*, where* l_m *is the log-likelihood value of model m,* v_m *is its number of parameters,* r_m *its number of equations, and* N_m *its number of observations.*

9. Likelihood Ratio test statistics can be easily calculated from the tables. These statistics are, however, not asymptotically χ^2-distributed under the null hypothesis of full coefficient homogeneity, because the parameters in $\mathbf{\Sigma}^{\delta}$ then are on the border of the admissible parameter space, see SHIN [1995, p. 321]. Thus, for making formal inference of coefficient heterogeneity versus homogeneity, other test procedures may be needed, see the recent papers by KHURI, MATHEW and SINHA [1998] and ANDREWS [1999]. We have not followed up these ideas in the present paper, however.

10. Throughout, global sample means are defined as the logs of arithmetic means.

11. We recomputed these standard-error estimates by replacing the estimated expectations with the predicted coefficient values (*cf.* Appendix B) for a plant with covariate vector equal to the global mean of the exogenous variables. This gave only minor changes in the standard error estimates and t-values, which suggests that the procedure chosen is sufficiently robust.

the latter models, the price elasticities are fairly robust to the choice of model. Even the price elasticities from the rather restrictive Model B2R are similiar to those obtained for the other models. Hence, our results suggest *(i)* that little gain is obtained by adding the cost function to the cost-share equations if the sole interest is in factor-price elasticities and *(ii)* that the representation of heterogeneity has an effect.

TABLE 4
Coefficient Estimates and Standard Errors in B-Models

	Model B1[a]		Model B2[b]		Model B3[c]		Model B2R[b]	
Coef.	Estimate	Std. err.	Estimate	Std. err.	Estimate	Std. err.	Estimate	Std. err.
β_0	– 1.5704	0.5151	5.5267	1.0346	6.5823	1.5694	– 6.5565	2.2035
β_X	0.4879	0.0794	– 0.0621	0.0710	0.2224	0.1518	1[d]	
β_K	0.4634	0.1185	0.0595	0.1742	– 0.4945	0.2800	1.2442	0.3940
β_τ	– 0.0162	0.0183	0.0159	0.0122	– 0.0377	0.0236	0.0969	0.0288
γ_L	0.5732	0.0645	0.2756	0.0747	0.3827	0.0733	0.4583	0.0789
γ_E	0.2355	0.0371	– 0.1666	0.0473	– 0.0939	0.0467	– 0.1080	0.0478
γ_M	0.8087	0.0825	0.1090	0.0940	0.2889	0.0924	0.3502	0.0983
β_{XL}	– 0.0090	0.0028	– 0.0049	0.0024	– 0.0050	0.0024	0[d]	
β_{XE}	0.0062	0.0021	0.0011	0.0014	0.0010	0.0014	0[d]	
β_{XM}	0.0028	0.0039	0.0038	0.0027	0.0040	0.0027	0[d]	
β_{KL}	– 0.0201	0.0035	– 0.0154	0.0053	– 0.0206	0.0053	– 0.0283	0.0052
β_{KE}	0.0071	0.0026	0.0240	0.0036	0.0185	0.0037	0.0194	0.0036
β_{KM}	0.0129	0.0048	– 0.0086	0.0067	0.0021	0.0067	0.0089	0.0066
γ_{LL}	0.0367	0.0143	0.0690	0.0109	0.0584	0.0104	0.0506	0.0120
γ_{LE}	– 0.0444	0.0066	– 0.0160	0.0049	– 0.0178	0.0047	– 0.0167	0.0051
γ_{LM}	0.0078	0.0158	– 0.0529	0.0120	– 0.0406	0.0113	– 0.0339	0.0133
γ_{EE}	– 0.0445	0.0067	0.0192	0.0042	0.0180	0.0041	0.0213	0.0042
γ_{EM}	0.0889	0.0095	– 0.0032	0.0066	– 0.0002	0.0062	– 0.0047	0.0067
γ_{MM}	– 0.0967	0.0208	0.0561	0.0154	0.0407	0.0144	0.0386	0.0168
$\beta_{\tau L}$	– 0.0016	0.0009	– 0.0025	0.0006	– 0.0020	0.0006	– 0.0019	0.0006
$\beta_{\tau E}$	0.0045	0.0006	0.0014	0.0003	0.0015	0.0003	0.0015	0.0003
$\beta_{\tau M}$	– 0.0029	0.0012	0.0010	0.0007	0.0005	0.0007	0.0005	0.0008
β_{XX}	– 0.0025	0.0068	– 0.0061	0.0067	– 0.0205	0.0122	0[d]	
β_{XK}	– 0.0178	0.0097	0.0365	0.0092	0.0219	0.0159	0[d]	
β_{KK}	0.0313	0.0159	– 0.0048	0.0179	0.0660	0.0301	– 0.1100	0.0351
$\beta_{\tau\tau}$	– 0.0025	0.0008	– 0.0017	0.0005	– 0.0033	0.0004	– 0.0034	0.0013
$\beta_{\tau X}$	– 0.0098	0.0016	– 0.0080	0.0013	– 0.0039	0.0016	0[d]	
$\beta_{\tau K}$	0.0121	0.0021	0.0076	0.0016	0.0103	0.0024	– 0.0073	0.0023

Note : [a] *Model B1: None of the coefficients are assumed to be random.*

[b] *Model B2 and B2R:* β_0, γ_L, γ_E *and* γ_M *are the expectations of* $\beta_{0(i,p)}$, $\gamma_{L(i,p)}$, $\gamma_{E(i,p)}$ *and* $\gamma_{M(i,p)}$ *respectively.*

[c] *Model B3:* β_0, β_X, β_K, β_τ, γ_L, γ_E *and* γ_M *are the expectations of* $\beta_{0(i,p)}$, $\beta_{X(i,p)}$, $\beta_{K(i,p)}$, $\beta_{\tau(i,p)}$, $\gamma_{L(i,p)}$, $\gamma_{E(i,p)}$ *and* $\gamma_{M(i,p)}$ *respectively.*

[d] *A priori restriction.*

In the models discussed above, price homogeneity and symmetry have been imposed *a priori*. Within the A models it is pertinent to test the validity of these restrictions. When homogeneity and symmetry are not imposed there are no cross-equation restrictions (beyond those following from the adding-up condition) and p_M is included as an additional regressor in the cost-share equations. To test for homogeneity, *ie,* for the significance of p_M in the regressions, we employ a Likelihood Ratio test. In both A-models a clear rejection of the homogeneity hypothesis is obtained. The significance probabilities, based on the $\chi^2(2)$-distribution, are less than 10^{-4}.[12] Testing for symmetry given that homogeneity is imposed involves one cross-equation restriction. We cannot reject this hypothesis in either of the A models. The significance probabilities, according to the Likelihood Ratio statistics which are $\chi^2(1)$-distributed, are 0.43 in Model A2 and 0.24 in Model A1.

The random coefficient approach allows us to predict plant specific coefficients (*cf.* Appendix B), which enables us to calculate plant specific elasticities evaluated at the plant specific predicted coefficients and plant specific means of the explanatory variables. Table 7 displays quartiles of the (univariate) distribution of the different price elasticities obtained for Model B2. With respect to the direct price elasticities, all the quartiles are negative and below unity in absolute value. The inter-quartile range is larger for the direct price elasticity of energy than for labour and materials. With respect to the cross-price elasticities and the Hicks-Allen substitution elasticities, the quartiles are generally positive, the only exception is the cross-price elasticities between labour and energy. Thus, for a substantial part of the plants, energy and labour are complements in the production process.

Table 8 contains the sample mean estimates of the output elasticity in the B-models (columns 1-3), the elasticity of variable costs with respect to capital (columns 4-6), and the derivative of the log-cost with respect to the trend variable (columns 7-9) [*cf.* eqs. (6)-(8)], and illustrates the sensitivity of the results to changes in the main assumptions. Rows 1-5 refer to the *base*

TABLE 5
Overall Measures of Fit[a] in B-Models

	Model B1	Model B2	Model B3	Model B2R
Number of estimated parameters	27	33	48	27
Log-likelihood value	916.084	3023.298	3477.367	1820.094
Akaike's information criterion	889.084	2990.298	3429.367	1793.094
Schwarz's Bayesian criterion	804.824	2887.314	3279.572	1708.835

Note : [a] The Akaike and the Schwarz Bayesian criterion are defined, for a model m, by, respectively, $AIC_m = l_m - v_m$ *and* $SBC = l_m - 0.5 v_m ln(r_m N_m)$, *where* l_m *is the log-likelihood value of model m,* v_m *is its number of parameters,* r_m *its number of equations, and* N_m *its number of observations.*

12. The rejection of price homogeneity may be due to a relatively large price variability in our data. Furthermore, the rejection may reflect the way in which we measure the electricity price. The electricity prices are plant-specific annual average prices while "annual marginal prices" might have been preferable. Since the data contain no information on the latter, we have decided that the theory restriction of price homogeneity should be imposed, even if it is not supported by the test applied.

specification of the model estimated, with different data sets. This serves to illustrate possible biases following from entry or exit of plants, plants with temporary interruption in their production activity, etc. Rows 6-10 refer to specific changes in the model assumptions when the full data set is used.

The output elasticity (ε_X) estimate obtained from the alternative data sets in rows 1-3 is between 0.15 and 0.34 in the random coefficient Models B2 and B3. In Model B1, with full coefficient homogeneity, the estimated output

TABLE 6

Own- and Cross Price-Elasticities of Input Demand and Allen-Uzawa Partial Elasticities of Substitution.[a] Standard errors in parentheses

Elasticity	A1	A2	B1	B2	B3	B2R
Own- and cross-price						
ε_{LL}	– 0.606	– 0.529	– 0.607	– 0.475	– 0.516	– 0.547
	(0.071)	(0.050)	(0.060)	(0.041)	(0.041)	(0.047)
ε_{EE}	– 1.343	– 0.707	– 1.223	– 0.720	– 0.731	– 0.708
	(0.071)	(0.029)	(0.061)	(0.030)	(0.029)	(0.029)
ε_{MM}	– 0.586	– 0.321	– 0.519	– 0.330	– 0.334	– 0.335
	(0.039)	(0.035)	(0.032)	(0.036)	(0.032)	(0.035)
ε_{EL}	– 0.286	0.130	– 0.110	0.169	0.135	0.140
	(0.071)	(0.043)	(0.056)	(0.042)	(0.041)	(0.043)
ε_{LE}	– 0.151	0.077	– 0.059	0.098	0.076	0.079
	(0.037)	(0.027)	(0.030)	(0.027)	(0.034)	(0.026)
ε_{ML}	0.274	0.183	0.251	0.178	0.190	0.199
	(0.029)	(0.027)	(0.025)	(0.027)	(0.024)	(0.027)
ε_{LM}	0.757	0.453	0.666	0.376	0.440	0.468
	(0.085)	(0.058)	(0.068)	(0.049)	(0.048)	(0.056)
ε_{ME}	0.311	0.138	0.267	0.152	0.144	0.136
	(0.017)	(0.017)	(0.014)	(0.020)	(0.017)	(0.017)
ε_{EM}	1.629	0.578	1.334	0.551	0.597	0.568
	(0.117)	(0.053)	(0.089)	(0.053)	(0.050)	(0.053)
Allen-Uzawa						
η_{EL}	– 1.224	0.527	– 0.462	0.623	0.523	0.547
	(0.319)	(0.160)	(0.238)	(0.135)	(0.143)	(0.153)
η_{ML}	1.176	0.744	1.051	0.658	0.736	0.779
	(0.129)	(0.091)	(0.105)	(0.080)	(0.074)	(0.087)
η_{ME}	2.531	0.950	2.105	0.964	0.998	0.946
	(0.163)	(0.077)	(0.128)	(0.077)	(0.071)	(0.078)
s_L	0.233	0.246	0.239	0.271	0.257	0.256
s_E	0.123	0.146	0.127	0.158	0.145	0.144
s_M	0.644	0.609	0.634	0.572	0.598	0.600

Note : [a] All elasticities and cost shares (s_L, s_E and s_M) are computed at the global mean of the exogenous variables.

elasticity is as low as 0.07 – 0.09. The inverse of this elasticity can be interpreted as "a variable-input scale elasticity" of the underlying (three factor) production function, and these estimated "economies of scale effects" are rather high. Excluding the plants with non-contiguous series only leads to very small changes in the output elasticity estimates. The change is somewhat larger, but still modest, when only the balanced, full-length sub-panel is used. From this we conclude that entry and exit of plants in our unbalanced panel data set do not affect our output elasticity estimates substantially. The same is true for the price elasticities.

To be well-behaved, the cost function should be non-increasing in capital, *cf.* BROWN and CHRISTENSEN [1981, p. 217-218]. The intuition is that as capital stock increases, one needs less variable inputs to maintain the given output. This will accordingly reduce variable costs. Furthermore, the expected sign of the trend derivative (ε_τ) should be negative if it represents technological progress. However, for Models B1, B2, and B3 and the data sets in rows 1-5 the capital elasticity (ε_K) has the "wrong" sign. The trend derivative (ε_τ) is uniformly negative in Model B2 only. Correct signs are generally obtained when constant returns to scale is imposed (row 6). On the other hand, as remarked in Section 3, our trend may pick up increasing input requirement following from quality changes in output (not represented by our output measure), which may well affect the variable cost positively, so *a priori*, the net effect may be of either sign.

Our interpretation of the "theory-inconsistent" results with respect to capital and trend in Table 8 is that we, probably, are unable to properly identify simultaneously the effect of changes in production, capital stock, and technological progress on variable costs from our data. The problem of identifying the different properties of the production process is not particular to our study; *cf.*, *eg,* MORRISON [1988, p. 278] and LINDQUIST [2001], and it is common to assume a

TABLE 7
Distribution of Plant Specific Predicted Elasticities of Model B2. Quartiles[a]

	First quartile	Median	Third quartile
ε_{LL}	– 0.469	– 0.454	– 0.386
ε_{EE}	– 0.707	– 0.605	– 0.370
ε_{MM}	– 0.420	– 0.284	– 0.206
ε_{EL}	– 0.252	0.060	0.247
ε_{LE}	– 0.039	0.014	0.107
ε_{ML}	0.135	0.209	0.272
ε_{LM}	0.292	0.400	0.480
ε_{ME}	0.027	0.055	0.136
ε_{EM}	0.390	0.518	0.618
η_{EL}	– 1.307	0.227	0.693
η_{ML}	0.611	0.678	0.733
η_{ME}	0.850	0.911	0.954

Note : [a] The elasticities (and cost shares s_L, s_E and s_M) are computed at the global mean of the exogenous variables.

TABLE 8

Output Elasticity (ε_X)***, Capital Elasticity*** (ε_K) ***and Trend Effect*** (ε_τ) ***at Sample Mean. Different Models and Data Sets. Standard errors in parentheses***

	ε_X			ε_K			ε_τ		
	B1	B2	B3	B1	B2	B3	B1	B2	B3
1) Base specification (BS), all plants	0.087	0.245	0.206	0.726	0.505	0.713	0.005	– 0.005	0.011
	(0.019)	(0.017)	(0.048)	(0.022)	(0.049)	(0.066)	(0.004)	(0.003)	(0.006)
2) BS, plants with contiguous series	0.066	0.250	0.199	0.716	0.439	0.620	0.006	– 0.007	0.013
	(0.021)	(0.017)	(0.052)	(0.024)	(0.051)	(0.078)	(0.004)	(0.003)	(0.007)
3) BS, balanced subpanel, 22 years	0.088	0.337	0.147	0.641	0.582	0.300	– 0.002	– 0.007	0.027
	(0.026)	(0.023)	(0.084)	(0.028)	(0.079)	(0.166)	(0.005)	(0.003)	(0.008)
4) BS, plants with at least 11 obs.	0.052	0.236	0.211	0.726	0.466	0.781	0.007	– 0.002	0.013
	(0.019)	(0.017)	(0.052)	(0.022)	(0.054)	(0.082)	(0.004)	(0.002)	(0.006)
5) BS, plants with at least 11 obs.,	0.052	0.240	0.066	0.726	0.262	0.908	0.007	– 0.004	0.056
fixed coefficients	(0.019)	(0.017)	(0.082)	(0.022)	(0.065)	(0.363)	(0.004)	(0.002)	(0.015)
6) Constant returns to scale	1^a	1^{ab}	1^a	– 0.293	– 0.392[b]	0.160	– 0.027	– 0.043[b]	– 0.006
				(0.036)	(0.094)	(0.062)	(0.008)	(0.005)	(0.005)
7) No capital effect	0.660	0.248	0.192	0^a	0^a	0^a	– 0.029	– 0.006	0.007
	(0.022)	(0.016)	(0.052)				(0.007)	(0.002)	(0.005)
8) No trend effect	0.067	0.213	0.342	0.746	0.510	0.620	0^a	0^a	0^a
	(0.018)	(0.015)	(0.053)	(0.022)	(0.047)	(0.089)			
9) Break in trend[c]	0.090	0.251	0.189	0.724	0.489	0.778	– 0.027	– 0.015	– 0.000
	(0.024)	(0.019)	(0.048)	(0.027)	(0.049)	(0.070)	(0.030)	(0.018)	(0.012)
10) Autocorrelation in the genuine	0.090	0.252	0.207	0.731	0.620	0.758	0.004	– 0.007	0.014
disturbances	(0.008)	(0.013)	(0.048)	(0.009)	(0.040)	(0.070)	(0.002)	(0.002)	(0.006)

Note : [a] *A priori restriction.*

[b] *Corresponds to B2R in Tables 4, 5, 6 and 12.*

[c] *Trend coefficients are estimated separately for three sub-periods: 1972-1978, 1979-1985, 1986-1993. The estimates are weighted averages of the three periods.*

constant returns to scale technology *a priori* and estimate the remaining elasticities contingent on this restriction. Rather than focusing on one particular restriction, we have tested a set of alternative restrictions on the output, capital, and trend effects. The consequences of these restrictions are illustrated in rows 6-8 of Table 8, respectively. Our conclusion is that the impact of the restrictions on the capital or trend effects on the estimates of the other coefficients is relatively small, while the opposite in general is true when the constant returns to scale restriction is imposed. With this latter restriction, we get "theory consistent" signs in Models B1 and B2.

The random coefficient specification may suffer from correlation between the random coefficients/intercepts and the regressors, *eg,* output, which implies that assumption (12) is violated and the estimated elasticities will be biased. The effect of replacing the random coefficient heterogeneity with fixed coefficient heterogeneity, which is robust to this potential problem, is illustrated in rows 4 and 5 of Table 8. We performed this analysis for the plants with at least 11 observations only, due to the loss of degrees of freedom. No conclusions are qualitatively altered, but the estimates for Model B3, which has a more flexible random coefficient specification than Model B2, are most strongly affected. A comparison of the elasticities in the random and fixed coefficient specification of these two models does not show any clear patterns, however, and while for example ε_K is smaller in the fixed coefficient case in Model B2, it is higher in Model B3. Hence, we do not find clear indications of estimation biases in the random coefficient model.

The sensitivity of the results to the dynamic specification of the models is investigated in rows 9 and 10 of Table 8. In row 9, we consider models in which coefficients of the trend are allowed to change over time. Specifically, we allow $\beta_\tau, \beta_{\tau\tau}, \beta_{\tau X}, \beta_{\tau K}, \beta_{\tau L}, \beta_{\tau E}$ to take on different values for the three sub-periods 1972-1978, 1979-1985, and 1986-1993 (the break years coinciding roughly with the two OPEC induced oil price shocks and the subsequent sharp decline in the oil prices), whereas the base specification assumes time-invariance.[13] A purpose of this is to check the robustness of the coefficient estimates to changes in world market growth, since the Chemical industry is heavily export oriented. The estimates of ε_X and ε_K depart only modestly from those for the base specification (row 1). For the average trend coefficient ε_τ, the estimates have the "theory consistent" negative sign, but are not significantly different from zero.

In row 10 of Table 8 we finally report the estimation results for model versions in which the no autocorrelation assumption for the genuine disturbances [*cf.* (13)] is replaced by the first order vector autoregression $\boldsymbol{u}_{(ip)t} = \text{diag}(\rho, \rho, \rho_c)\boldsymbol{u}_{(ip)t-1} + \boldsymbol{\xi}_{(ip)t}$, where $\text{diag}(\rho, \rho, \rho_c)$ denotes the (3×3) diagonal matrix with the (non-random) coefficients ρ, ρ, ρ_c along the diagonal. For the vector $\boldsymbol{\xi}_{(ip)t}$ we adopt the same set of assumptions as for $\boldsymbol{u}_{(ip)t}$ in (13).[14] First, the genuine residuals for plant (ip) in year t are calcu-

13. Recall that in the B3-type model $\beta_\tau = E(\beta_{\tau(ip)})$. To calculate output and capital elasticities and trend effects for the models in row 9 of Table 7 we used an average of coefficient estimates in the three sub-periods, weighted by their number of years.

14. The lagged residuals from the two share equations are restricted to enter their respective equations with the same autoregressive coefficient in order to retain adding-up.

lated from $\widehat{\boldsymbol{u}}_{(ip)t} = \boldsymbol{y}_{(ip)t} - \boldsymbol{X}_{(ip)t}\widehat{\boldsymbol{\beta}}_{(ip)}$, where $\widehat{\boldsymbol{\beta}}_{(ip)}$ may contain either estimated plant invariant or predicted plant specific coefficients. Second, exemplifying the procedure by model B3, we estimate the following "augmented" model:

$$\boldsymbol{y}_{(ip)t} = \boldsymbol{X}_{(ip)t}\boldsymbol{\beta}_{(ip)} + \text{diag}(\rho,\rho,\rho_c)\widehat{\boldsymbol{u}}_{(ip)t-1} + \boldsymbol{\xi}_{(ip)t}.$$

Because of the lag in the latter equation and because of the occurrence of non-contiguous time series for some plants, the estimation of the augmented model is based on 105 fewer observations than the reference B3 model. The estimated autoregression coefficients in the B1-, B2-, and B3-type models are reported in Table 9. They are as large as 0.9 in the homogeneous Model B1,

TABLE 9
***Estimated Autoregression Coefficients of the Disturbances in the Cost-Share Equations* (ρ) *and the Cost Function* (ρ_c)**

	Model B1	Model B2	Model B3
ρ	0.907	0.569	0.543
ρ_c	0.923	0.724	0.440

TABLE 10
The Estimated Covariance Matrix of the Random Coefficients[a]

Model A2	$\gamma_{L(i,p)}$	$\gamma_{E(i,p)}$				
$\gamma_{L(i,p)}$	1.72					
$\gamma_{E(i,p)}$	0.36	1.14				
Model B2	$\beta_{0(i,p)}$	$\gamma_{L(i,p)}$	$\gamma_{E(i,p)}$			
$\beta_{0(i,p)}$	212.25					
$\gamma_{L(i,p)}$	– 16.68	1.79				
$\gamma_{E(i,p)}$	– 8.90	0.43	1.16			
Model B3	$\beta_{0(i,p)}$	$\beta_{X(i,p)}$	$\beta_{K(i,p)}$	$\beta_{\tau(i,p)}$	$\gamma_{L(i,p)}$	$\gamma_{E(i,p)}$
$\beta_{0(i,p)}$	829.79					
$\beta_{X(i,p)}$	– 9.20	6.46				
$\beta_{K(i,p)}$	– 45.13	– 4.03	6.46			
$\beta_{\tau(i,p)}$	1.31	– 0.30	0.09	0.12		
$\gamma_{L(i,p)}$	– 22.39	– 0.10	0.80	– 0.11	1.72	
$\gamma_{E(i,p)}$	– 5.70	0.79	– 0.72	– 0.08	0.37	1.13
Model B2R	$\beta_{0(i,p)}$	$\gamma_{L(i,p)}$	$\gamma_{E(i,p)}$			
$\beta_{0(i,p)}$	270.37					
$\gamma_{L(i,p)}$	– 12.21	1.74				
$\gamma_{E(i,p)}$	– 9.84	0.34	1.15			

Note : [a] All entries are multiplied by 100.

and are lower in Model B3 than in Model B2. The estimates of ε_X, ε_K, and ε_τ, however, are not much affected (compare rows 1 and 10 in Table 8). Hence, although there may be signs that our reference model is dynamically mis-specified [*cf.* rows 9 and 10 of Table 8], this does not affect the estimated mean cost function coefficients significantly.

Table 10 contains the estimated covariance matrix of the random coefficient vector for Models A2, B2, B3, and B2R. The covariance matrices of the genuine disturbances for all models are given in Table 11. Both for the A-class and B-class of models we find that assuming coefficient homogeneity leads to higher estimated variances of the genuine disturbances than the more flexible models. This is as expected when coefficient heterogeneity is present, because Models A1 and B1 are then mis-specified and the estimated disturbance variances captures the coefficient heterogeneity. Constraining the output elasticity to unity has a notable impact on both the covariance matrix of the random coefficients and the covariance matrix of the genuine error

TABLE 11
The Estimated Covariance Matrix of the Genuine Disturbances[a]

Model A1	u_L	u_E	
u_L	2.03		
u_E	0.38	1.06	
Model A2	u_L	u_E	
u_L	0.50		
u_E	– 0.02	0.18	
Model B1	u_L	u_E	u_C
u_L	2.03		
u_E	0.38	1.06	
u_C	– 5.57	– 2.72	41.58
Model B2	u_L	u_E	u_C
u_L	0.51		
u_E	– 0.01	0.18	
u_C	– 0.85	– 0.06	9.48
Model B3	u_L	u_E	u_C
u_L	0.50		
u_E	– 0.02	0.18	
u_C	– 0.82	– 0.18	4.29
Model B2R	u_L	u_E	u_C
u_L	0.51		
u_E	– 0.01	0.18	
u_C	– 0.07	– 0.14	57.27

Note : [a] *All entries are multiplied by 100.*

terms (compare the results for Models B2 and B2R in Tables 10 and 11). However, the sub-matrix containing only the second-order moments of the genuine errors in the two cost-share equations is very similiar to that in Model B2 (Table 11).

Finally, by exploiting the estimation results, we can decompose the variation in the endogenous variables, into parts which represent *(i)* variation in the exogenous variables, *(ii)* coefficient heterogeneity, *(iii)* eventual interaction between *(i)* and *(ii)*, and *(iv)* genuine disturbance variation. The results from such a decomposition for the model versions with random coefficients are given in Table 12. The interaction term represents the combined effect of the randomness of the slope coefficients and the explanatory variables in the cost function. Such interaction terms are absent from the cost-share equations, which include random intercepts only. For computational details, see Appendix B.

As much as 70-80 per cent of the variation in the cost shares can be ascribed to variation in the coefficient vector, while only 6-15 per cent is due to variation in the log of the inputs. As is common in panel data, the between variation in our data dominates the within variation, and our interpretation of

TABLE 12
Decomposition of Variation in Endogenous Variables. Shares[a]

Endogenous variable	Variation in input vector	Variation in coefficient vector[b]	Interaction term: Input-Coefficient vectors[c]	Genuine disturbance variation
Model A2				
Labour cost share	0.125	0.677	0	0.198
Energy cost share	0.103	0.774	0	0.123
Model B2				
Labour cost share	0.060	0.732	0	0.208
Energy cost share	0.152	0.734	0	0.114
Cost function	0.675	0.284	0.002	0.039
Model B3				
Labour cost share	0.091	0.703	0	0.206
Energy cost share	0.102	0.774	0	0.124
Cost function	0.734	0.150	0.103	0.013
Model B2R				
Labour cost share	0.109	0.691	0	0.200
Energy cost share	0.101	0.776	0	0.122
Cost function	0.673	0.240	0.001	0.086

Note : [a] *The decomposition is described in Appendix B.*
[b] *Including intercept heterogeneity.*
[c] *The contribution from the interaction term is, by restriction, zero in the share equations.*

the results is that the random intercept terms in the cost-share equations capture much of the between variation in the cost shares. The variation in the genuine disturbance accounts for 11-20 per cent of the variation in the cost shares. With respect to the cost function, the picture is somewhat different. Around 70 per cent of the variation in the log of the deflated total variable costs is due to variation in the exogenous variables, 15-28 per cent to coefficient heterogeneity, and only 1-9 per cent to genuine disturbance variation. For the general Model B3, the interaction between the variation in the exogenous variables and the coefficient heterogeneity is not negligible.

5 Concluding Remarks

In this paper, a general framework for analyzing substitution and scale properties from unbalanced plant-level panel data is put forward. The production technology is represented by its dual cost function. The most general econometric model class consists of a three-factor translog cost function and the corresponding cost-share equations. A simpler model class containing only the two cost-share equations is also considered. Labour, energy, and materials are treated as variable inputs, while capital is assumed to be quasi-fixed. Focus has been on the importance of choosing alternative random coefficient specifications of plant heterogeneity in comparison with the corresponding specifications with no heterogeneity in coefficients. A Maximum Likelihood estimation procedure for the situation with unbalanced panel data is applied.

The empirical results, based on data from Norwegian chemical plants, show a clear improvement in model fit when random coefficient heterogeneity is allowed for. About 75 per cent of the variation in the cost shares can be ascribed to variation in the coefficient vector (including intercept heterogeneity), while only about 10 per cent is due to variation in the exogenous variables and about 15 per cent to the genuine disturbance. Around 70 per cent of the variation in the log of the deflated total variable costs is due to variation in the exogenous variables, around 25 per cent to coefficient heterogeneity, and only around 5 per cent to genuine disturbance variation. In addition, conclusions regarding scale properties and the effect of the capital stock on variable costs are considerably influenced by the choice of model specification, while conclusions regarding price effects are more robust. The results also indicate that, although there is substantial variation in the predicted elasticities between plants, none of the variable inputs are price-elastic, and that labour, energy, and materials are all substitutes. Our results further suggest that little gain is obtained by adding the cost function to the cost-share equations if the sole interest is in factor-price elasticities, but that the representation of heterogeneity has an effect.

We find it difficult, however, from our micro panel data, to identify separately the impact of technology, output, and the capital stock in the translog cost function when no *a priori* restrictions on this function are imposed, and some "theory inconsistent" results are found. Other researchers have made similar experiences. When restricting the elasticity of costs with respect to output to unity, this problem in general vanishes, but at the cost of a considerably worsened goodness of fit. This restriction affects the estimated price elasticities only marginally. ▼

• References

AMEMIYA Y. (1994). – On Multivariate Mixed Model Analysis, *Multivariate Analysis and its Applications*, IMS Lecture Notes – Monograph Series, vol. 24, p. 83-95.

ANDREWS D.W.K. (1999). – « Estimation when a Parameter is on a Boundary », *Econometrica*, 67, p. 1341-1383.

BALESTRA P., NEGASSI S. (1992). – « A Random Coefficient Simultaneous Equation System with an Application to Direct Foreign Investment by French Firms », *Empirical Economics*, 17, p. 205-220.

BALTAGI B.H., CHANG Y. (1994). – « Incomplete Panels: A Comparative Study of Alternative Estimators for the Unbalanced One-Way Error Component Regression Model », *Journal of Econometrics*, 62, p. 67-89.

BIØRN E., LINDQUIST K.-G., SKJERPEN T. (2000). – « Micro Data on Capital Inputs: Attempts to Reconcile Stock and Flow Information », *Statistics Norway, Discussion Papers* No. 268.

BROWN R.S., CHRISTENSEN L.R. (1981). – Estimating Elasticities of Substitution in a Model of Partial Static Equilibrium: An Application to U.S. Agriculture, 1947 to 1974, in Berndt E.R. and Field B.C. (eds.), *Modelling and Measuring Natural Resource Substitution*, Cambridge, Mass.: MIT Press, p. 209-229.

CHRISTENSEN L.R., JORGENSON D.W., LAU L. (1971). – « Conjugate Duality and the Transcendental Logarithmic Production Function », *Econometrica*, 39, p. 255-256.

CHRISTENSEN L.R., JORGENSON D.W., LAU L. (1973). – « Transcendental Logarithmic Production Frontiers », *Review of Economics and Statistics*, 55, p. 28-45.

DIEWERT W.E. (1971). – « An Application of the Shephard Duality Theorem: A Generalized Leontief Production Function », *Journal of Political Economy*, 79, p. 481-507.

DIEWERT W.E., WALES T.J. (1987). – « Flexible Functional Forms and Global Curvature Conditions », *Econometrica*, 55, p. 43-68.

DIONNE G., GAGNÉ R., VANASSE C. (1998). – « Inferring Technological Parameters from Incomplete Panel Data », *Journal of Econometrics*, 87, p. 303-327.

GALLANT A.R. (1981). – « On the Bias in Flexible Functional Forms and an Essentially Unbiased Form. The Fourier Flexible Form », *Journal of Econometrics*, 15, p. 211-245.

HSIAO C. (1975). – « Some Estimation Methods for a Random Coefficient Model », *Econometrica*, 43, p. 305-325.

HSIAO C. (1986). – *Analysis of Panel Data*, Cambridge: Cambridge University Press.

HSIAO C. (1996). – Random Coefficients Models. Chapter 5 in: Mátyás L. and Sevestre P. (eds.): *The Econometrics of Panel Data. Handbook of the Theory with Applications*, Dordrecht: Kluwer.

JORGENSON D.W. (1986). – Econometric Methods for Modeling Producer Behavior, Chapter 31 in: Griliches Z. and Intriligator M.D. (eds.): *Handbook of Econometrics, Vol. III*, Amsterdam: North-Holland.

JUDGE G.G., HILL R.C., GRIFFITHS W.E., LÜTKEPOHL H., LEE T.-C. (1985). – *The Theory and Practice of Econometrics*, Second edition, New York: Wiley.

KHURI A.I., MATHEW T., SINHA B.K. (1998). – *Statistical Tests for Mixed Linear Models*, New York: Wiley.

KMENTA J. (1986). – *Elements of Econometrics*, Second edition, New York: Macmillan.

LAU L.J. (1986). – Functional Forms in Econometric Model Building, Chapter 26 in: Griliches Z. and Intriligator M.D. (eds.): *Handbook of Econometrics*, *Vol. III*, Amsterdam: North-Holland.

LEE L.F., GRIFFITHS W.E. (1979). – « The Prior Likelihood and Best Unbiased Prediction in Stochastic Coefficient Linear Models », *Working Papers in Econometrics and Applied Statistics* No. 1, Department of Economic Statistics, University of New England.

LINDQUIST K.-G. (2001). – « The Response by the Norwegian Aluminium Industry to Changing Market Structure », *International Journal of Industrial Organization*, 19, p. 79-98.

LONGFORD N.T. (1995a). – *Random Coefficient Models*, Oxford: Clarendon Press.

LONGFORD N.T. (1995b). – *Random Coefficient Models*, Chapter 10 in: Arminger G., Clogg C.C., and Sobel M.E. (eds.): *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, New York: Plenum Press.

MAIRESSE J., GRILICHES Z. (1990). – Heterogeneity in Panel Data: Are There Stable Production Functions? Chapter 33 in: Champsaur, P. *et al.* (eds.): *Essays in Honor of Edmond Malinvaud*, Cambridge, Mass.: MIT Press.

MÁTYÁS L., LOVRICS L. (1991). – « Missing Observations and Panel Data: A Monte-Carlo Analysis », *Economics Letters*, 37, p. 39-44.

MORRISON C. (1988). – « Quasi-Fixed Inputs in US and Japanese Manufacturing: A Generalized Leontief Restricted Cost Function Approach », *Review of Economics and Statistics*, 70, p. 275-287.

PIL (1996). – *The Norwegian Chemical Industry*, Oslo: PIL, The Federation of Norwegian Process Industries.

SAS (1992). – SAS Technical Report P-229, SAS/STAT Software: *Changes and Enhancements*, Release 6.07, Cary, NC: SAS Institute, Inc.

SHIN C. (1995). – « On the Multivariate Random Coefficient Model », *American Statistical Association: 1995 Proceedings of the Business and Economic Statistics Section*, p. 320-325.

STATISTICS NORWAY (1995). – *Energy Statistics 1995*, NOS C 347, Oslo: Statistics Norway.

SWAMY P.A.V.B. (1970). – « Efficient Estimation in a Random Coefficient Regression Model », *Econometrica*, 38, p. 311-323.

SWAMY P.A.V.B. (1971). – *Statistical Inference in Random Coefficient Regression Models*, New York: Springer.

SWAMY P.A.V.B. (1974). – Linear Models with Random Coefficients, Chapter 5 in: Zarembka P. (ed.): *Frontiers in Econometrics*, New York: Academic Press.

VERBEEK M., NIJMAN T.E. (1996). – Incomplete Panels and Selection Bias, Chapter 18 in: Mátyás L. and Sevestre P. (eds.): *The Econometrics of Panel Data. A Handbook of the Theory with Applications*, Dordrecht: Kluwer.

VERBEKE G., MOLENBERGHS G. (eds.) (1997). – *Linear Mixed Models in Practice. A SAS Oriented Approach*, New York: Springer.

APPENDIX A

A1. *Definition of Variables*

Most variables are observed directly or easily calculated from the available data. The exception is capital stock. Further details on all variables, including the capital stock calculations, are given in Section A2. MS indicates that the data are from the Manufacturing Statistics database of Statistics Norway, and the data are plant specific. NA indicates that the data are from the Norwegian National Accounts. In this case, the data are identical for all plants classified in the same National Account industry, of which our data set specifies three. Data in value terms are measured in Norwegian kroner (NOK).

CL: Total labour cost, 1000 NOK (MS)

CM: Total material cost (incl. motor gasoline), 1000 NOK (MS)

CE: Total energy cost, 1000 NOK (MS)

$C = CL + CM + CE$: Total factor cost, excluding capital, 1000 NOK

QL: Labour input, man-hours (MS)

$PL = 1000 * CL/QL$: Labour cost, NOK per man-hour

PM: Price of materials (incl. motor gasoline), 1991=1 (NA)

$QM = CM/PM$: Input of materials (incl. motor gasoline), 1000 1991-kroner

QE: Energy input, electricity plus fuels (excl. motor gasoline), 1000 kWh (MS)

$PE = 100 * CE/QE$: Price of energy, øre per kWh (100 øre = 1 NOK) X: Output, tonnes (MS)

$K = KB + KM$: Total capital stock (buildings plus machinery/transport equipment), 1000 1991-kroner (MS/NA).

The calculations of capital stock data are based on the perpetual inventory method assuming constant and plant invariant retirement rates. We combine plant data on gross investment with fire insurance values for each of the categories KB=buildings and structures and KM=machinery and transport equipment (MS). The data on investment and fire insurance are deflated by means of price indices (1991=1) for investment in the two categories from the Norwegian National Accounts (NA).

A2. *Further Description*

The Manufacturing Statistics follow the Standard Industrial Classification (SIC) and gives annual data for large plants at the 5-digit code. Until 1992, plants with at least five employees were defined as large, while from 1992 on, the limit is 10 employees. In 1993, the activity classification was revised according to EU's NACE Rev. 1 and UN's SIC Rev. 3, while previously based on UN's SIC Rev. 2. For this analysis, we use data from the period 1972-1993. While the revision of the activity classification does not cause inconsistency problems, the change in the definition of large plants causes a reduction in the number of plants with 5-9 employees in 1992. Our data include all industries classified under SIC-code 351 Manufacture of industrial chemicals. The large plants, regardless of definition, represent about 99 per cent (each year) of the total activity in the sector, whether measured by total gross output value or by employment.

Some minor data cleaning has been done. We exclude observations where no output, labour input or energy input is reported. This reduces the initial panel data set by 56 observations, and leaves us with 1,265 observations. The number of plants per year, which ranges from 47 to 66, shows a negative trend over the sample period.

Output: The plants are in general multi-output plants and report output of a number of products. Two output-measures are available: *(i)* total output in value terms (NOK) and *(ii)* output of each product in both value terms and physical units. We use measure *(ii)* to calculate an average price per tonne output for each plant, and deflate measure *(i)* by these prices to get the output-measure in tonnes. It was necessary to utilize both output measures, because some plants do not report their production of all products in both value and volume terms.

Factor input costs and prices: The Manufacturing Statistics gives the number of man-hours used, total labour costs in NOK, total electricity consumption in kWh and in NOK, the consumption of different fuels in various denominations and NOK, and total material costs in NOK for each plant. We use this to calculate labour costs per man-hour and total energy costs (excl. motor gasoline) in NOK for each plant. The different fuels, such as coal, coke, fuelwood, petroleum oils and gases, and aerated waters, are transformed to the common denominator kWh by using estimated average energy content of each fuel (STATISTICS NORWAY [1995, p. 124]). This enables us to calculate an energy price per kWh for each plant. For most plants, the energy aggregate is dominated by electricity. The price of material inputs (incl. motor gasoline) is from the Norwegian National Accounts (1991=1), and the price is identical for all plants classified in the same National Account industry.

Capital stock: We calculate capital stock data for KB = buildings and KM=machinery and transport equipment. Plant data from the Manufacturing Statistics on gross investment flows and on fire insurance values are input in the calculations. The data on investment and fire insurance are deflated by means of industry specific price indices of investment goods from the Norwegian National Accounts (1991=1). Fire insurance values in a chosen reference year – which varies across plants – are deflated and used to determine benchmark levels of capital stocks. These insurance values are assumed to reflect the replacement values of the existing capital stock. We found it necessary to choose plant specific reference years for each capital category for two reasons: First, the fact that the panel data set is unbalanced made it impossible to choose one common reference year. Second, we have clear indications that the quality of the fire insurance values varies over time for each plant and capital category. This data quality problem involves partly extreme values, partly missing observations, and partly an increasing trend in the insurance coverage for many plants. To minimize the effect of this data quality problem when choosing reference years, we sorted the plant observations by increasing fire insurance values and chose as a plant specific reference year the observation that is closest to the 75 per cent fractile. From the chosen reference year, the level of capital stock for each category is extrapolated backwards and forwards, using the perpetual inventory method. The depreciation rates are 0.020 for buildings and 0.068 for machinery and transport equipment. The procedures for capital stock calculation are discussed in more detail in BIØRN, LINDQUIST and SKJERPEN [2000].

The table below reports the overall means and standard deviations of variables used in this analysis.

Overall Means and Standard Deviations of Basic Variables

Variable	Mean	Log of mean	Std. dev.
$\ln(C)$	10.278		1.919
C		11.695	
$\ln(X)$	9.355		2.443
X		11.955	
$\ln(K)$	11.821		1.941
K		13.381	
τ	10.960		6.194
s_L	0.296		0.154
s_E	0.107		0.112
s_M	0.597		0.202
$\ln(P_L/P_M)$	4.702		0.353
P_L/P_M		4.758	
$\ln(P_E/P_M)$	2.482		0.489
P_E/P_M		2.598	

APPENDIX B

B1. *Prediction of Coefficients*

The coefficient vector of plant (ip), $\boldsymbol{\beta}_{(ip)}$, can be predicted as follows:

$$\text{(B.1)}\quad \boldsymbol{\beta}^*_{(ip)} = \widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\Sigma}}^{\delta}\boldsymbol{X}'_{(ip)}(\boldsymbol{X}_{(ip)}\widehat{\boldsymbol{\Sigma}}^{\delta}\boldsymbol{X}'_{(ip)} + \boldsymbol{I}_p \otimes \widehat{\boldsymbol{\Sigma}}^{u})^{-1}(\boldsymbol{y}_{(ip)} - \boldsymbol{X}_{(ip)}\widehat{\boldsymbol{\beta}}),$$

where $\widehat{\boldsymbol{\beta}}$ is the ML (strictly, the Feasible GLS) estimator of the expected coefficient vector $\boldsymbol{\beta}$ (*cf.* LEE and GRIFFITHS [1979, section 4], JUDGE *et al.* [1985, p. 541], and HSIAO [1986, p. 134]), and $\widehat{\boldsymbol{\Sigma}}^{\delta}$ and $\widehat{\boldsymbol{\Sigma}}^{u}$ are the corresponding estimates of $\boldsymbol{\Sigma}^{\delta}$ and $\boldsymbol{\Sigma}^{u}$. Apart from the fact that $\boldsymbol{\Sigma}^{\delta}$ and $\boldsymbol{\Sigma}^{u}$ have been estimated, this is the best linear unbiased predictor (BLUP) of $\boldsymbol{\beta}_{(ip)}$.

Note that (B.1) may also be used to predict coefficients for (synthetic) plants not represented in the data set, for example a plant with $\boldsymbol{X}$ and $\boldsymbol{y}$ values equal to their sample means.

B2. *Decomposition of the Shares of Variability*

Let us write the part of (11) which relates to eq. g $(g = 1,\ldots,G)$ as:

$$\text{(B.2)}\quad y_{g(ip)t} = \boldsymbol{x}_{g(ip)t}\boldsymbol{\beta} + \eta_{g(ip)t}, \quad \eta_{g(ip)t} = \boldsymbol{x}_{g(ip)t}\boldsymbol{\delta}_{(ip)} + u_{g(ip)t},$$

where $\boldsymbol{x}_{g(ip)t}$ is a $(1 \times H)$ vector and $y_{g(ip)t}$, $\eta_{g(ip)t}$, and $u_{g(ip)t}$ are scalars. We want to decompose $\text{var}(y_{g(ip)t})$ for each of the G equations into parts which represent (a) variation in the exogenous variables, (b) variation in the coefficients, and (c) genuine disturbance variation. From the law of iterated expectations we have:

$$\text{(B.3)}\quad \text{var}(y_{g(ip)t}) = \mathsf{E}[\text{var}(y_{g(ip)t}|\boldsymbol{x}_{g(ip)t})] + \text{var}[\mathsf{E}(y_{g(ip)t}|\boldsymbol{x}_{g(ip)t})].$$

Using (B.2), (12), and (13), the first term on the right hand side of (B.3) can be written as:

$$\begin{aligned}\text{(B.4)}\quad \mathsf{E}[\text{var}(y_{g(ip)t}|\boldsymbol{x}_{g(ip)t})] &= \mathsf{E}[\text{var}(\eta_{g(ip)t}|\boldsymbol{x}_{g(ip)t})]\\ &= \mathsf{E}[\boldsymbol{x}_{g(ip)t}\boldsymbol{\Sigma}^{\delta}\boldsymbol{x}_{g(ip)t} + u^2_{g(ip)t}]\\ &= \mathsf{E}[\text{tr}(\boldsymbol{x}'_{g(ip)t}\boldsymbol{x}_{g(ip)t}\boldsymbol{\Sigma}^{\delta})] + \sigma^u_{gg}\\ &= \text{tr}[\mathsf{V}(\boldsymbol{x}_{g(ip)t})\boldsymbol{\Sigma}^{\delta}]\\ &\quad + \text{tr}[\mathsf{E}(\boldsymbol{x}'_{g(ip)t})\mathsf{E}(\boldsymbol{x}_{g(ip)t})\boldsymbol{\Sigma}^{\delta}] + \sigma^u_{gg},\end{aligned}$$

where $\mathsf{V}(\boldsymbol{x}_{g(ip)t})$ is the population covariance matrix of $\boldsymbol{x}_{g(ip)t}$. The second term on the right hand side of (B.3) can be written as:

$$\begin{aligned}\text{(B.5)}\quad \text{var}[\mathsf{E}(y_{g(ip)t}|\boldsymbol{x}_{g(ip)t})] &= \text{var}(\boldsymbol{x}_{g(ip)t}\boldsymbol{\beta}) \;=\; \boldsymbol{\beta}'\mathsf{V}(\boldsymbol{x}_{g(ip)t})\boldsymbol{\beta}\\ &= \text{tr}[\boldsymbol{\beta}'\mathsf{V}(\boldsymbol{x}_{g(ip)t})\boldsymbol{\beta}] \;=\; \text{tr}[\mathsf{V}(\boldsymbol{x}_{g(ip)t})\boldsymbol{\beta}\boldsymbol{\beta}'].\end{aligned}$$

Inserting (B.4) and (B.5) in (B.3) we obtain the decomposition formula:

$$\begin{aligned}\text{(B.6)}\quad \text{var}(y_{g(ip)t}) &= \text{tr}[\mathsf{V}(\boldsymbol{x}_{g(ip)t})\boldsymbol{\beta}\boldsymbol{\beta}'] + \text{tr}[\mathsf{E}(\boldsymbol{x}'_{g(ip)t})\mathsf{E}(\boldsymbol{x}_{g(ip)t})\boldsymbol{\Sigma}^{\delta}]\\ &\quad + \text{tr}[\mathsf{V}(\boldsymbol{x}_{g(ip)t})\boldsymbol{\Sigma}^{\delta}] + \sigma^u_{gg}.\end{aligned}$$

Here, the first component represents the variation in the exogenous variables, the second component represents the variation in the coefficient vector (including the intercept term), the third component is an interaction term between the variation in the exogenous variables in the coefficients, and the fourth component represents the genuine disturbance variation. This formula underlies the decomposition given in Table 12.