Likelihood Inference for Dynamic Panel Models

by
Marc Nerlove

WP 98-16

Department of Agricultural and Resource Economics
The University of Maryland, College Park

files/papers97/Paris/LikelihoodPanel.doc **FINAL REVISION 03.31.98** © Marc Nerlove 1998
To be published in *L'Annales d'Économie et de Statistique* de l'INSEE in late 1998.

Comments welcome.

Likelihood Inference for Dynamic Panel Models

Marc Nerlove
Department of Agricultural and Resource Economics
University of Maryland
Tel: (301) 405-1388 Fax: (301) 314-9032
e-mail: mnerlove@arec.umd.edu
homepage: http://www.arec.umd.edu/mnerlove/mnerlove.htm

ABSTRACT

The *likelihood principle* is applied to the problem of inference in dynamic panel models. The principle states that the likelihood function contains "...*all* the information which the data provide concerning the relative merits of..." alternative parametric hypotheses. The usual asymptotic theory of maximum likelihood is based on a quadratic approximation to the likelihood function in the nearby neighborhood of a local maximum of the function. One needs to look at the entire function more broadly in order to ascertain the true significance of the data for the hypotheses under consideration, not only because of the possibilities of multiple local maxima and boundary solutions, but also because the data are typically differentially informative with respect to different regions of the parameter space. In order to handle cases in which the likelihood function depends on more than two parameters, The devices of "concentrating" and of "slicing" or sectioning the function in the direction of a hyperplane or surface reflecting the variation of all but two of the parameters are introduced. The likelihood functions for two basic dynamic panel models: (1) a model involving individual-specific effects which reflect the influence of latent time-persistent variables; (2) a model involving individual-specific time trends which reflect the nonstationarity introduced by trending latent variables, are derived. The methods are applied to the analysis of cross-country economic growth. The findings demonstrate the power and feasibility of general methods of likelihood inference, especially to reveal problems of inference and areas of ignorance.

---

*This paper was prepared for the Seventh Conference on Panel Data Econometrics, 19-20 June 1997, Paris. The research on which it is based was supported by the Maryland Agricultural Experiment Station, Project A-53. The paper will appear as MAES Paper No. ****.

I am indebted to G. S. Maddala for his comments, to two anonymous referees for theirs, and to Anke Meyer for numerous helpful suggestions and for her wise expository counsel. Jinkyo Suh provided able computational assistance.

What has now appeared is that the mathematical concept of probability is ... inadequate to express our mental confidence or diffidence in making ... inferences, and that the mathematical quantity which usually appears to be appropriate for measuring our order of preference among different possible populations does not in fact obey the laws of probability. To distinguish it from probability, I have used the term "**Likelihood**" to designate this quantity; since both the words "likelihood" and "probability" are loosely used in common speech to cover both kinds of relationship.

R. A. Fisher, *Statistical Methods for Research Workers*, 1925.

Within the framework of a statistical model, *all* the information which the data provide concerning the relative merits of two hypotheses is contained in the likelihood ratio of those hypotheses on the data. ...For a continuum of hypotheses, this principle asserts that the likelihood function contains all the necessary information.

A. W. F. Edwards, *Likelihood*, 1972.

You are living on a Plane. What you style Flatland is the vast level surface of what I may call a fluid, or in, the top of which you and your countrymen move about, without rising above or falling below it.

I am not a plane Figure, but a Solid. You call me a Circle; but in reality I am not a Circle, but an infinite number of Circles, of size varying from a Point to a Circle of thirteen inches in diameter, one placed on the top of the other. When I cut through your plane as I am now doing, I make in your plane a section which you, very rightly, call a Circle. For even a Sphere--which is my proper name in my own country--if he manifest himself at all to an inhabitant of Flatland--must needs manifest himself as a Circle.

E. A. Abbott, *Flatland*, 1884.

It was six men of Indostan
To learning much inclined,
Who went to see the Elephant
(Though all of them were blind),
That each by observation
Might satisfy his mind.
The First approached the Elephant,
And happening to fall
Against his broad and sturdy side,
At once began to bawl:
"God bless me! but the Elephant
Is very like a wall!"
The Second, feeling of the tusk,
Cried, "Ho! what have we here
So very round and smooth and sharp?
To me 'tis mighty clear
This wonder of an Elephant
Is very like a spear!"

The Third approached the animal,
And happening to take
The squirming trunk within his hands,
Thus boldly up and spake:
"I see," quoth he, "the Elephant
Is very like a snake!"
The Fourth reached out an eager hand,
And felt about the knee.
"What most this wondrous beast is like
Is mighty plain," quoth he;
" 'Tis clear enough the Elephant
Is very like a tree!"
The Fifth, who chanced to touch the ear,
Said: "E'en the blindest man
Can tell what this resembles most;
Deny the fact who can
This marvel of an Elephant
Is very like a fan!"

The Sixth no sooner had begun
About the beast to grope,
Than, seizing on the swinging tail
That fell within his scope,
"I see," quoth he, "the Elephant
Is very like a rope!"
And so these men of Indostan
Disputed loud and long,
Each in his own opinion
Exceeding stiff and strong,
Though each was partly in the right,
And all were in the wrong!
Moral:So oft in theologic wars,
The disputants, I ween,
Rail on in utter ignorance
Of what each other mean,
And prate about an Elephant
Not one of them has seen!

John Godfrey Saxe , "The Blind Men and the Elephant: A Hindoo Tale," 1880

Preface

This paper applies the *likelihood principle* of Fisher (1921, 1922, 1925 and 1932), Barnard (1949, 1951, 1966, 1967 and Barnard, Jenkins and Winsten, 1962) and Birnbaum (1962) to the problem of inference in dynamic panel models.[1] Beginning with Chamberlain (1984) an extensive literature on non-likelihood methods for estimation an inference about dynamic panel models has emerged, much of it surveyed in Sevestre and Trognon (1996) and by Baltagi (1996, Chapter 8, pp. 125 - 148). I do not propose to survey this literature here or to compare the alternative estimates suggested with the results of likelihood inference, although such comparison would no doubt be highly useful.

In section 1, I develop the principle that the likelihood function contains "...*all* the information which the data provide concerning the relative merits of..." alternative parametric hypotheses. The usual asymptotic theory of maximum likelihood is shown to be based on a quadratic approximation to the likelihood function in the nearby neighborhood of a local maximum of the function. I argue that one needs to look at the entire function more broadly in order to ascertain the true significance of the data for the hypotheses under consideration, not only because of the possibilities of multiple local maxima and boundary solutions, but also because the data are typically differentially informative with respect to different regions of the parameter space. In order to handle cases in which the likelihood function depends on more than two parameters, I introduce the devices of "concentrating" and of "slicing" or sectioning the function in the direction of a hyperplane or surface reflecting the variation of all but two of the parameters.

In section 2, I derive the likelihood functions for two basic dynamic panel models: (1) a model involving individual-specific effects which reflect the influence of latent time-persistent variables; (2) a model involving individual-specific time trends which reflect the nonstationarity introduced by trending latent variables. In developing the likelihood functions for these two leading cases, I argue for reduction of models of type (2) to stationary models of type (1) by differencing. In this case, however it is necessary to modify the likelihood functions to reflect the effects of differencing on the unobserved residual variation. The differenced model now has a different interpretation from the original models of type (1) in that the individual-specific effects now represent individual-specific trend slopes. I further argue that in stationary cases, which include both levels models and differenced models, the initial observations of the dependent variables contain useful information on the process which must have generated those observations in the past, before the panel was observed, and that this information depends positively on their variance and on the number individuals in the panel, and is thus of particular importance for "shallow" panels.[2]

---

[1] Although *likelihood* and inference from likelihood resembles Laplace's method of *inverse probability* (Laplace, 1774 - 1814), which provides the principal basis for the Bayesian approach to inference, Fisher (1932) was a great pains to distinguish the two, and, indeed, was sharply critical of the use of prior distributions, especially of the use of "non-informative" priors to represent ignorance.

[2] Maddala (1971) discusses a similar problem, pooling cross-section and time-series data, from a Bayesian point of view. The analysis with diffuse priors is similar a number of respects to that presented here based on the likelihood principle. Some of Maddala's results are discussed below. As is the case with much of the literature in this area, however, those about likelihood or maximum likelihood are based on a likelihood function which conditions on the initial observations. Breusch's (1987) remarkable result, for example, about the convergence of iterated Generalized Least Squares to the ML estimates holds only for the case in which the likelihood function is conditional on the initial observations in the dynamic case.

Finally, in section 3, to assess the feasibility and power of likelihood methods for inference about dynamic panel models, I use data on 94 countries for the period 1960 - 1985, and a subsample of 22 OECD countries, from the Penn World Tables 5.6, publicly available from the NBER web site. The 22-country sample consists of primarily European countries, all highly developed and tied together by a network of trading relations; the 94-country sample is much more heterogeneous, consisting of the aforementioned 22 plus 72 additional countries ranging from Mozambique and Haiti to the "Asian Tigers." This is the same data set which has been used in dozens of previous studies.

In an previous paper (Nerlove, 1996), I compared some commonly used methods of estimation in dynamic panel models with one another and contrasted the results obtained from likelihood methods which take account of the information contained in the initial observations about the process which must have generated those observations in the pre-sample period. I showed that many of the earlier findings are probably statistical artifacts arising from biases in the econometric methods employed. Here I focus especially on the need to take advantage of the relatively large amount of information contained in the initial observations and to take account of differing country-specific trends. Using a simple variant of the Solow-Swan growth model widely used in recent studies of the convergence process, I demonstrate here that likelihood methods which take account of individual-specific trends and of the information present in the initial observations leads to acceptance of the convergence hypothesis, with the best-supported value of conditional convergence in the order of about 90% within 13 or 14 years for a broad sample of 94 countries. The analysis demonstrates the power and feasibility of general methods of likelihood inference, especially to reveal problems of inference and areas of ignorance.

1. Introduction: The Likelihood Principle

Although clearly implied in what Fisher wrote in the 1920's (1922,1925), the likelihood principle, which essentially holds that the likelihood function is the sole basis for inference, did not come into prominence until the 1950's and 1960's, principally through the work of Barnard, Birnbaum, and Edwards (see the references cited below, Barndorff-Nielsen, 1988, and Lindsey, 1996) written largely in reaction to both the classical Neyman-Pearson (frequentist) and the Bayesian approaches to inference (Jeffereys, 1934, 1961; see also Press, 1989).

A statistical model consists of a random vector $x \varepsilon X$ of observations having a joint distribution function $F(x;\theta)$, with corresponding density $f(x;\theta)$, depending on the unknown parameters $\theta \varepsilon \Theta$. It is assumed that F is known. The *likelihood function* determined by any given outcome x is defined as the function on Θ equal to $cf(x;\theta)$ where c is an arbitrary positive constant which may depend on x but does not depend on θ. Two likelihood functions defined on the same parameter space Θ, whether arising from the same "experiment" or from different "experiments," E_1 and E_2, are *equivalent* if their ratio is positive and independent of Θ for all $\theta \varepsilon \Theta$ except possibly at points at which both functions are zero (so that the ratio is undefined).

The *likelihood principle* asserts that for a given experiment E, the evidential meaning of any outcome x, for inference regarding θ is contained entirely in the likelihood function determined by x. All other aspects of how the data may have been generated are irrelevant, e.g., the sample space, provided, of course, that the sample space itself doesn't depend on θ. It follows that if two "experiments," E_1 and E_2, have pdf's $f(x,\theta)$ and $g(y,\theta)$, respectively, and if for some particular outcomes, x^* of E_1 and y^* of E_2,

$$f(x^*,\theta) = h(x^*, y^*)g(y^*,\theta),\ h(x^*,y^*) > 0, \text{ for all } \theta \varepsilon \Theta,$$

then these outcomes must result in the same inference about θ.

Birnbaum (1962) derives the likelihood principle from the sufficiency principle and a still more basic assumption, the so-called *conditionality principle*. This principle states that if an "experiment"

involving θ is chosen from a collection of possible experiments, *independently of* θ, then any experiment not chosen is irrelevant to the statistical analysis. The conditionality principle makes clear the implication of the likelihood principle that any inference should depend only on the outcome observed and not on any other outcome we might have observed and thus sharply contrasts the method of likelihood inference from the Neyman-Pearson, or frequentist, approach, in which inference does depend crucially on a hypothetical sequence of experiments, the outcome of but one of which is observed. In particular, questions of unbiasedness, minimum variance, consistency and the like and the whole apparatus of confidence intervals, significance levels, and power of tests, are ruled out of bounds. While maximum-likelihood estimation does satisfy the likelihood principle (and thus sufficiency and conditionality), the frequentist assessment in terms of asymptotic properties is irrelevant. In this paper, I apply the likelihood principle to the problem of inference about the parameters of dynamic panel models and try to make clear the role of the maximum of the likelihood function and its Hessian evaluated at the maximum in approximating the whole of the likelihood function for purposes of inference.

The likelihood principle is clearly incomplete from the standpoint of inference since it nowhere states how the evidential meaning of the likelihood function is to be determined. To the principle, therefore, "likelihoodists" generally append the *method of support* (a term coined by Jeffereys, 1934). The *support function* is defined as the natural logarithm of the likelihood function. Since the likelihood function incorporates an arbitrary constant, the support function is defined only up to the addition of an arbitrary constant. Conventionally, this constant is often taken to be the value which makes support at the maximum equal zero. In multiplicative terms, this is equivalent to normalizing the likelihood function by dividing it by its value at the maximum. Only relative support for a particular parameter value over another can be interpreted in any case, so the constant disappears when looking at the difference between support values of different parameter values. The *method of maximum support* is the *method of maximum likelihood*. But the interpretation of the parameter value which yields this maximum and of the inverse of the negative of the Hessian at the point of maximum is different than in the frequentist interpretation in terms of asymptotic properties. The likelihoodist interpretation of these magnitudes is in terms of a quadratic approximation to the support function in the neighborhood of its maximum..

It is clear that the difference in the value of the support function at two different values of a parameter has the significance that the value for which support is greater is more consistent with the observed data than the value of lesser support. What we have is essentially a likelihood ratio test without the frequentist apparatus of asymptotic chi-square. It is also clear that the values of parameters for which maximum support is obtained (that is, the maximum-likelihood estimates), especially if the maximum is unique, have a special significance in relation to other possible values. Moreover, how sharply defined such a maximum of the likelihood function, if a unique maximum exists, is also clearly relevant to any inference we may wish to draw. On the negative side, a poorly behaved likelihood function, for example, one having ridges of equal likelihood, many local maxima, or a maximum on the boundary of an a priori admissible region of the parameter space, is generally indicative of an incompletely or ill-formulated underlying statistical model.

From a frequentist point of view what matters about the likelihood function is only its maximum and curvature in the neighborhood of the maximum, and all the desirable properties and the assessment of the reliability of the maximum-likelihood estimates are only asymptotic. Greene (1993, pp.111-116) gives a very brief discussion of these matters; Davidson and MacKinnon (1993, Chapter 8, pp.243-287) give a more complete and rigorous discussion; a more intuitive discussion with many econometric examples is given by Cramer (1986). That only the maximum and the Hessian at the maximum are all the matters from a frequentist point of view is perhaps not surprising in view of the fact that for the mean of a normal distribution the quadratic approximation is exact (see the discussion below) and because of the central limit theorem in its many forms many estimators, including ML estimators in regular cases, tend to normality in distribution.

When we are dealing with only one or two parameters looking at the whole of the likelihood or support function is feasible, although some summary measures may be helpful. For three or more parameters, however, it is no longer possible to examine the whole of the support function. In this case, concentrating the likelihood function and corresponding support function may be helpful, and looking at a quadratic approximation to the support function in the neighborhood of the maximum may be revealing.

First, we can section or slice the support function along the plane of all but one or two of the parameters; in the case in which all but one of the parameters has been eliminated in this way, we are back to a two-dimensional plot; when we have done this for all but two parameters we can plot a three-dimensional surface and associated contours of equal support. The latter is particularly useful if we want to examine how two of the parameters interact with one another. It would be natural to choose the values of all but one or two of the parameters equal to the maximizing values. Proceeding in this way amounts to looking at the *concentrated likelihood function* and associated *concentrated support function.*

A second, but not mutually exclusive alternative is to follow the lead of those frequentists who maximize likelihood functions and characterize the entire likelihood function by the point in the parameter space at which the maximum is attained and a quadratic approximation to the entire function at that point. The point of maximum support, particularly if unique, obviously has considerable intuitive appeal. A quadratic approximation at that point is likely to be pretty good if we want to consider only points quite nearby and has the added advantage of being directly interpretable from a frequentist point of view in terms of the information matrix of asymptotic maximum-likelihood theory. The disadvantage is that except for cases, such as the mean or regression function associated with a normal distribution, for which the quadratic approximation is exact, the approximating function may be quite wide of the mark. Moreover, when the likelihood function has two or more local maxima, which may be far apart, the inferential significance of this fact may be lost if one focuses exclusively on the behavior of the function in the vicinity of the highest maximum. Boundary maxima, which are of frequent occurrence in dynamic panel problems, also present a special problem from a frequentist point of view since the asymptotic theory is no longer applicable. But from the standpoint of likelihood inference there is nothing that stops us from comparing the value of support at the boundary values of the parameters with other values in the interior of the permissible region.

What I am suggesting for viewing the support function in a multiparameter case is essentially what one typically does in viewing a three-dimensional surface when we look at a contour map: We take a slice through the surface in the direction parallel to the plane of the two arguments. A slice can, of course, be thought of more generally as any lower dimensional hyperplane, whether parallel to the plane defined by the axes of a subset of arguments or in some other direction. In four dimensions, a slice in any two-dimensional plane, which eliminates all the arguments but two, yields a surface of the functional values in three dimensions. Fixing, or conditioning on, the values of any subset of parameters is obviously a way of defining a particular hyperplane corresponding to the remaining parameters; in this instance, those values which maximize support, given the values corresponding to a point chosen on the hyperplane, on which we want to view the support, assume a special significance. In discussions of maximum likelihood, *concentration of the likelihood function* with respect to a subset of parameters corresponds to selecting a hyperplane for the remaining parameters in just this way. Sometimes we say that we are "maximizing out" the deselected parameters. In the method of maximum likelihood, for example, it frequently turns out that, given the values of one or two of the parameters, it is very easy to maximize with respect to the remaining ones.

2. Likelihood Functions for Two Basic Dynamic Panel Models

In this section, I present a new method of maximum-likelihood estimation based on the density of the observations *unconditional on the initial or starting values of the dependent variable, in which the same process as that under investigation is assumed to generate the data prior to the point at which we*

begin to observe them.[3] I argue more generally for methods of inference which look at more than just the maximum of the likelihood function, on the basis of the *likelihood principle* of Fisher (1922; 1925). This approach fully takes into account what information the initial conditions contain about how the process has operated in the past and is thus of special relevance to short time-dimension ("shallow") panels. I extend this method to the case of country-specific trends. These make the underlying processes being investigated nonstationary, but with simple forms of nonstationarity that can be removed by differencing the data.

A good summary of the current state of knowledge about the properties of various estimators in dynamic panel models is contained in Sevestre and Trognon (1992, 2nd. ed. 1996). Trognon (1978) was the first to show the possible inconsistency of maximum likelihood conditional on the initial individual observations. Nickell (1981) shows the inconsistency of the estimates of the fixed-effects in a dynamic panel model. Kiviet (1995) derives exact results for the bias of leading estimators. I will assume a random effects model for the disturbance for the reasons set forth in Nerlove and Balestra (1996) and because fixed effects can be viewed as a special case from the standpoint of estimation.

2.1. The Model in Levels

For simplicity, I restrict attention to the simple model containing one exogenous variable x_{it} and one lagged value of the dependent variable y_{it-1} as explanatory. Extension to the case in which more than one exogenous explanatory variable is included presents no serious difficulty.

(1) $$y_{it} = \alpha + \beta x_{it} + \gamma y_{it-1} + \mu_i + \varepsilon_{it} \quad , i=1,...N, t=1,...T.$$

Taking deviations from overall means eliminates the constant α. The usual assumptions are made about the properties of the μ_i and the ε_{it}:

(i) $E(\mu_i) = E(\varepsilon_{it}) = 0$, all i and t,

(ii) $E(\mu_i \varepsilon_{jt}) = 0$, all i, j and t,

(iii) $$E(\mu_i \mu_j) = \begin{cases} \sigma_\mu^2 & i = j \\ 0 & i \neq j, \end{cases}$$

(iv) $$E(\varepsilon_{it} \varepsilon_{js}) = \begin{cases} \sigma_\varepsilon^2 & t = s,\ i = j \\ 0 & otherwise \end{cases}.$$

---

[3] Anderson and Hsiao (1981, 1982) have also considered unconditional ML and its relation to conditional ML for a number of different cases. See also Hsiao (1986). In Anderson and Hsiao (1981), they study a simple autoregressive process with additive term specific to the unit under the following alternative assumptions about the initial conditions: (a) initial state fixed; (b) initial state random; (c) the unobserved individual effect independent of the unobserved dynamic process with the initial value fixed; (d) the unobserved individual effect independent of the unobserved dynamic process with the initial value random. The problem is greatly complicated by the presence of exogenous regressors and is studied in Anderson and Hsiao (1982) for panel data both with and without lagged dependent variables. The same four cases are studied as for the simple autoregression considered in the 1981 paper but a number of different assumptions are made about the exogenous explanatory variable. The key distinction is between time-varying and time-invariant exogenous variables. Clearly to examine asymptotic properties some assumptions have to be made about the behavior of the exogenous variables, which is a tricky matter since, being exogenous, we effectively deny knowledge of how they might be generated; however, see my solution below. The important point is that none of the four alternative assumptions about the initial state of the system being observed presupposes that it must have been in operation prior to the initial observation. Such is, I hope, the contribution of this paper.

Both μ_i and ε_{it} are assumed to be uncorrelated with x_{it} for all i and t. While this assumption is far from innocuous, for example, if the independent variable x_{it} is not independent of the dependent variable y_{it} or unobserved factors which affect it, I adopt it here, not only because it is conventional but because one has to cut off somewhere. Clearly, however, y_{it-1} cannot be assumed to be uncorrelated with μ_i.

The intraclass correlation coefficient ρ is defined as $\sigma_\mu^2 / (\sigma_\mu^2 + \sigma_\varepsilon^2)$. This parameter measures the extent of unobserved or latent time-invariant, individual-specific, variation relative to the total unobserved variation in the sample,. It is extremely important in understanding the nature of the variation, both observed and unobserved, in the panel. Also useful are the characteristic roots of Ω = the variance-covariance matrix of the disturbances $u_{it} = \mu_i + \varepsilon_{it}$: $\xi = 1 - \rho + T\rho$ and $\eta = 1 - \rho$.[4] $\lambda = 1 + T\rho/(1-\rho)$ measures the *relative* information contributed over time by the individual specific unobserved effects μ for each individual.

I restrict attention broadly to likelihood inference about the parameters of the model characterized by (1) and (i) - (iv) and its first-difference extension. Finding the maximum of the likelihood function and the Hessian there is also of specific interest.

(a) Maximum Likelihood Conditional on the Initial Value of the Lagged Dependent Variable.

When the likelihood function for the model (1) with $u_{it} = \mu_i + \varepsilon_{it} \sim N(0, \sigma^2\Omega)$ is derived in the usual way from the product of the densities of y_{it} conditional on x_{it} and y_{it-1}, the joint density is conditional on y_{i0}.[5] This likelihood function can be written in terms of the notation introduced above as

(2)
$$\log L(\alpha, \beta, \gamma, \sigma_\mu^2, \sigma_\varepsilon^2 | y_{11}, \ldots, y_{NT}; x_{11}, \ldots x_{NT}; y_{10}, \ldots, y_{N0})$$
$$= -\frac{NT}{2}\log 2\pi - \frac{NT}{2}\log\sigma^2 - \frac{N}{2}\log\xi - \frac{N(T-1)}{2}\log\eta$$
$$-\frac{1}{2\sigma^2}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{it}^* - \alpha\xi^{-1/2} - \beta x_{it}^* - \gamma y_{it-1}^*\right)^2,$$

where y*, x* and y*$_{-1}$ are the transformed variables.[6] Since

$\xi = \dfrac{T}{1+\lambda(T-1)}$ and $\eta = \dfrac{\lambda T}{1+\lambda(T-1)}$, logL can be expressed as a function solely of λ, σ^2, α, β, and

---

[4] The GLS estimates for a regression based on the variance-covariance matrix of the disturbances in (1) has a ready interpretation in terms of these roots: A regression of the means for each cross-sectional observation over time and a pooled regression of the individual observations taken as deviations from these means both contain information about the parameters of the model: The means regression reflects purely cross-sectional variation; whereas the fixed-effects regression reflects the individual variation over time. GLS combines these two types of information with weights which depend on the characteristic roots of $Euu' = \sigma^2\Omega$. The individual means themselves are weighted by the reciprocal of the square root of $\xi = 1 - \rho + T\rho$, while the deviations from these means are weighted by the reciprocal of the square root of $\eta = 1-\rho$. A representative transformed observation is

$$y_{it}^* = \xi^{-1/2}\bar{y}_{i.} + \eta^{-1/2}(y_{it} - \bar{y}_{i.}), \quad i = 1, \ldots, N, \quad t = 1, \ldots, T.$$

The GLS estimates are just the OLS estimates using the transformed observations.

[5] By repeated application of the definition of the joint distribution in terms of the product of the conditional and the marginal, it can easily be seen that the *unconditional joint distribution* is the product of the joint conditional distribution on which (2) is based and the marginal distribution of the initial observations y_{i0}, so that the unconditional likelihood is the sum of (2) and a term which reflects the marginal likelihood of the initial values.

γ. λ is defined above. Trognon (1978) shows that, when the exogenous variable x is generated by a first-order autoregression with white noise input, $w \sim wn(0, \sigma_w^2 I)$, also assumed in the Monte Carlo experiments reported in Nerlove (1971),

$$(3) \qquad x = \delta x_{-1} + w \quad ,$$

maximization of the conditional likelihood function (3) yields boundary solutions $\hat{\rho} = 0$, which, unlike interior maximum likelihood solutions, are inconsistent, for a considerable, and indeed likely, range of parameter values. In particular, there is a value of γ in (1),

$$\gamma^* = \frac{(T-3)^2 - 8}{(T+1)^2} \quad ,$$

such that when $\gamma < \gamma^*$ there exists an interior maximum of (2) which yields consistent ML estimates, but that when $\gamma \geq \gamma^*$ there are values of ρ for which the conditional likelihood function (2) is maximized at the boundary $\rho = 0$, i.e., for the OLS estimates of the pooled regression of untransformed observations , which we know to be inconsistent. The problem is that when T is small the permissible range of γ, the coefficient of the lagged dependent variable is implausible (e.g., negative or very small). For example, for T = 5, $\gamma^* = -0.11$, while for T = 10, $\gamma^* = 0.34$. When $\gamma \geq \gamma^*$, whether or not an interior maximum with consistent ML estimates occurs depends on the value of ρ: For $\rho < \rho^*$ boundary maxima occur where

$$\rho^* = \left(\frac{T-1}{T+1}\right)^2 \frac{\beta^2 \sigma_w^2}{\sigma^2} \frac{1-\gamma}{(\gamma - \gamma^*)(1-\gamma\delta)^2} \quad .$$

For example, when T = 5, $\beta = 1.0$, $\gamma = 0..75$, $\delta = 0.5$, and $\frac{\sigma_w^2}{\sigma^2} = 1.0$, $\gamma^* = -0.11$ and the critical value of ρ is $\rho^* = 0.31$. That means that any true value of the intraclass correlation less than 0.31 is liable to produce a boundary solution to (2) $\rho = 0$ and inconsistent estimates of all the parameters. Using these results, Trognon (1978) is able to replicate the Monte Carlo results reported in Nerlove (1971).[7]

Even though ML may yield inconsistent estimates when the nonnegligible probability of a boundary solution is taken into account, it is nonetheless true that the likelihood function summarizes the information contained in the data about the parameters. From a conventional, Neyman-Pearson point of view what matters about the likelihood function is only its maximum and curvature in the neighborhood of the maximum, and all the desirable properties and the assessment of the reliability of the maximum-likelihood estimates are only asymptotic. That only the maximum and the Hessian at the maximum are all that matters from a conventional point of view is perhaps not surprising in view of the fact that for the mean of a normal distribution the quadratic approximation is exact and because of the central limit theorem in its many forms many estimators, including ML estimators in regular cases, tend to normality

---

[6] See footnote 1.

[7] Maddala (1971, pp. 346 - 347) gives a condition for the gradient of the concentrated likelihood function to be positive at a boundary $\rho = 0$ (OLS on the pooled data) for the conditional likelihood function. so if ρ is constrained to the interval [0,1) this implies a local maximum at the boundary 0. Breusch (1987) shows that this condition can be easily checked at the start of his iterative GLS procedure by beginning with the pooled OLS estimates and $\rho = 0$. Unfortunately these results apply only to the likelihood function when no lagged value of the dependent variable is included or when those initial values are conditioned upon. I have not been able to derive a similar result for the unconditional likelihood function below.

in distribution. So the problem of possible inconsistency of the ML estimates should not concern us unduly from the standpoint of likelihood inference. It is the whole shape of the likelihood function which expresses what the data have to say about the model and its parameters which matters.[8] For this reason, "slices" or sections of some of the multidimensional likelihood functions are also presented in the empirical example of the next section. In subsection 2.2 I consider a model in which first differences are taken to eliminate a linear deterministic trend; in this case, the individual-specific time invariant effects become differences in the trend slopes. This makes the interpretation of the model in first-difference form different from that in levels. Moreover, the time- and individual varying disturbance is now likely to be serially correlated, a fact which needs to be taken into account in the formulation of the unconditional likelihood function. A parallel set of results for the individual-specific trends model is presented in 2.2 below.

(b) Unconditional Likelihood and Unconditional Maximum Likelihood.

While it is not guaranteed that a boundary solution to the likelihood equations is obtained, which would yield ML estimates which are inconsistent, it is apparent, as suggested above, that in panels with a short time dimension the initial values provide important information about the parameters of the model, and to condition on them is to neglect this information.

It is not, in fact difficult to obtain the unconditional likelihood function once the marginal distribution of the initial values is specified.[9] The problem is a correct specification of this distribution. If $|\gamma| \geq 1$ or the processes generating the x_{it} are not stationary, it will not, in general be possible to specify the marginal distribution of the initial observations. I will assume that, possibly after some differencing, both the y_{it} and the x_{it} are stationary. The derivation of the unconditional likelihood function in the case in which deterministic or stochastic trends are included is contained in the next subsection.

Under this assumption, the dynamic relationship to be estimated is stationary and $|\gamma| < 1$. Consider equation (2) can be rewritten with the intercept eliminated, for y_{i0} and the infinite past as:

$$(4) \qquad y_{i0} = \sum_{j=1}^{\infty} \gamma^j \beta x_{i,-j} + \frac{1}{1-\gamma}\mu_i + \nu_{i0} \text{ , where } \nu_{it} = \gamma \nu_{it-1} + \varepsilon_{it} \text{ .}^{10,11}$$

---

[8] Maddala (1971, p. 346) shows that the likelihood function may have at most two local maxima. In work not reported here, I have obtained likelihood functions with two local maxima, one for large values of ρ and small γ, the other for large γ and ρ close to 0. When these yield a similar value of the likelihood function, I would argue that the data are telling us that it's difficult to distinguish. A method which does must therefore be misleading. For the cross-country data considered below in this paper, I do not find evidence of two local maxima but rather that the likelihood function is rather flat in the γ-ρ plane, which leads to a similar conclusion.

[9] See footnote 5.

[10] For a particular time period T and the infinite past

$y_{iT} = \gamma^{\infty} y_{i-\infty} + \sum_{j=0}^{\infty} \gamma^j \beta x_{i-j} + \frac{1-\gamma^{\infty}}{1-\gamma}\mu_i + \nu_{iT}$, where $\nu_{iT} = \sum_{j=0}^{\infty} \gamma^j \varepsilon_{iT-j}$. Since $1 \geq |\gamma|$ and $\nu_{iT} = \sum_{j=0}^{\infty} \gamma^j \varepsilon_{iT-j}$ is the MA form of a first-order autoregression with white noise input, equation (24) follows.

[11] If all variables are expressed as deviations of from their overall means, there is no need to include an intercept; if not, μ_i should be replaced by $\alpha + \mu_i$.

If $\beta = 0$, so that the relationship to be estimated is a pure autoregression for each y_{it}, the vector of initial values $y_0 = (y_{10}, \ldots y_{N0})'$ has a joint normal distribution with means 0 and variance-covariance matrix $[\frac{\sigma_\mu^2}{(1-\gamma)^2} + \sigma_\nu^2]I_N = (\frac{\sigma_\mu^2}{(1-\gamma)^2} + \frac{\sigma_\varepsilon^2}{1-\gamma^2})I_N$. The unconditional likelihood is therefore

(5)
$$\log L(\gamma, \rho, \sigma_\mu^2, \sigma_\varepsilon^2 | y_{11}, \ldots, y_{NT}; \ldots; y_{10}, \ldots, y_{N0})$$
$$= -\frac{NT}{2}\log 2\pi - \frac{NT}{2}\log\sigma^2 - \frac{N}{2}\log\xi - \frac{N(T-1)}{2}\log\eta$$
$$- \frac{1}{2\sigma^2}\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it}^* - \gamma y_{it-1}^*)^2$$
$$- \frac{N}{2}\log(\frac{\sigma_\mu^2}{(1-\gamma)^2} + \frac{\sigma_\varepsilon^2}{1-\gamma^2}) - \left[\frac{1}{2(\frac{\sigma_\mu^2}{(1-\gamma)^2} + \frac{\sigma_\varepsilon^2}{1-\gamma^2})}\right]\sum_{i=1}^{N} y_{i0}^2$$

This likelihood function can easily be concentrated: To maximize, express σ_μ^2, σ_ε^2, ξ and η in terms of ρ and γ. For given ρ and γ in the interval [0,1), concentrate the likelihood function with respect to σ^2. It follows that

$$\hat{\sigma}^2(\gamma,\rho) = \frac{RSS^*(\gamma,\rho)}{N(T+1)} \quad where\ RSS^*(\gamma,\rho) = \sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it}^* - \gamma y_{it-1}^*)^2 + (\sum_{i=1}^{N} y_{i0}^2 \Big/ \left[\frac{\rho}{(1-\gamma)^2} + \frac{1-\rho}{1-\gamma^2}\right].$$

Thus, the concentrated LF is

$$\log L^*(\gamma,\rho) = -\frac{N(T+1)}{2}\log 2\pi - \frac{N}{2}\log\xi - \frac{N(T-1)}{2}\log\eta$$
$$- \frac{N(T-1)}{2}\log\left\{\frac{RSS^*(\gamma,\rho)}{N(T-1)}\right\} - \frac{N}{2}\left\{\frac{\rho}{(1-\gamma)^2} + \frac{1-\rho}{1-\gamma^2}\right\}$$
$$- (1/2RSS^*/N(T+1))\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it}^* - \gamma y_{it-1}^*)^2 - \sum_{i=1}^{N} y_{i0}^2 / \{(2/N(T+1))\left[\frac{\rho}{(1-\gamma)^2} + \frac{1-\rho}{1-\gamma^2}\right]RSS^*\}$$

Maximizing L* is quite a bit more complicated than the usual minimization of the sum of squares in the penultimate term because RSS*, in that term, depends on $\sum_{i=1}^{N} y_{i0}^2 = N \operatorname{var} y_0$, as well as on ρ and γ, which enter the final terms as well. var y_0 is the observed sample variance of the initial observations. This magnitude relative to the theoretical unconditional variance of the y_0's is crucial in the relation between the conditional and the unconditional likelihood functions. When $\beta \neq 0$, things are more complicated still. But more important than finding the maximum of L* is its shape above the γ-ρ plane. It is apparent from the results presented below that there may be significant trade-offs between γ and ρ without large effects on the value of the likelihood.

Various alternative specifications of the likelihood function are considered in the literature are reported and analyzed in Sevestre and Trognon (1996, pp. 136-138).[12] Considerable simplification, however, can be

[12] One interesting possibility discussed by Sevestre and Trognon (1996, p. 136-138) is to choose y_{i0} a linear function of some *observed* individual-specific time-invariant exogenous variables and a disturbance which is decomposed as the sum of the individual-specific disturbances μ_i and a remainder. The first-

obtained if, following Nerlove (1971), we are willing to assume that x_{it} follows a well-specified common stationary time-series model for all individuals i. The first term in (13) is

$\varphi_{i0} = \beta\sum_{j=0}^{\infty}\gamma^j x_{i,-j}$. Hence, for any stationary processes x_{it}, which may be serially correlated,

$$\frac{\varphi_{it}}{\beta} = \gamma\frac{\varphi_{it-1}}{\beta} + x_{it}$$

with variances

(6) $$\sigma^2_{\varphi_t} = \frac{\beta^2\sigma^2_{x_t}}{1-\gamma^2} \quad .$$

If we suppose that the variance of the x_{it} is the same for all i, then the random variable

$$\phi_{it} = \sum_{j=0}^{\infty}\gamma^j\beta x_{it-j}$$

has a well defined variance which is the same for all i and a function of β, γ, and $\sigma^2_{x_t}$. This then enters the final term in the unconditional likelihood (5), which now becomes:

(7) $$\log L(\beta,\gamma,\sigma^2_\mu,\sigma^2_\varepsilon \mid y_{11},\ldots,y_{NT};x_{11},\ldots x_{NT};y_{10},\ldots,y_{N0})$$

$$= -\frac{N(T+1)}{2}\log 2\pi - \frac{NT}{2}\log\sigma^2 - \frac{N}{2}\log\xi - \frac{N(T-1)}{2}\log\eta$$

$$-\frac{1}{2\sigma^2}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(y^*_{it} - \beta x^*_{it} - \gamma y^*_{it-1}\right)^2$$

$$-\frac{N}{2}\log\left(\frac{\beta^2\sigma^2_x}{1-\gamma^2} + \frac{\sigma^2_\mu}{(1-\gamma)^2} + \frac{\sigma^2_\varepsilon}{1-\gamma^2}\right) - \left[\frac{1}{2\left(\frac{\beta^2\sigma^2_x}{1-\gamma^2} + \frac{\sigma^2_\mu}{(1-\gamma)^2} + \frac{\sigma^2_\varepsilon}{1-\gamma^2}\right)}\right]\sum_{i=1}^{N} y^2_{i0}.$$

Concentrating the likelihood function to permit a one- or two-dimensional grid search is no longer possible. Nor is it possible to graph the likelihood surface with respect to variations in all of the parameters. Although "slicing" the likelihood function along any hyperplane in the parameter space can reveal the trade-offs between any pair of parameters. If gradient or search procedures yield an interior maximum, the ML estimates obtained are consistent as long as the random variables $\phi_{it} = \sum_{j=0}^{\infty}\gamma^j\beta x_{i,t-j}$ have well-defined variances and covariances.

---

order equations for maximizing the likelihood then take on a simple recursive form when β = 0, and permit other simplification when β ≠ 0. But if we knew some individual-specific time-invariant observed variables influenced behavior why not incorporate them directly in (2) , the equation to be estimated?

which they will if the x_{it} are generated by a stationary process. It doesn't really matter what this process is as long as it is stationary. Besides, since the x_{it} are assumed to be exogenous, we really have no basis on which to model their determination and are likely to misspecify this part of the model. In this sense we ought to prefer this kind of "almost full-information" maximum likelihood. Still we have to assume something about the variance of the x process in order to proceed. I suggest estimating σ_x^2 from the sample data.

To generalize these results to the case in which their are several explanatory variables in addition to the lagged value of the dependent variable, assume that X_{it} follows a stationary VAR process and replace βx^*_{it} by $X^*_{it}\beta$ and $\beta^2\sigma_x^2$ by $\beta'\Sigma_{XX}\beta$ in the above formula.

The expression

$$\varphi^2 = \frac{\beta^2\sigma_x^2}{1-\gamma^2} + \frac{\sigma_\mu^2}{(1-\gamma)^2} + \frac{\sigma_\varepsilon^2}{1-\gamma^2} = \frac{1}{1-\gamma^2}\left[\beta^2\sigma_x^2 + \sigma^2(1+\frac{2\gamma\rho}{1-\gamma})\right]$$

is the unconditional variance of the initial observations y_{i0}. The absolute value of the difference between the log of the unconditional likelihood function and the log of the conditional likelihood function is

$$(8) \qquad f(\varphi^2) = \frac{N}{2}\left[\log 2\pi + \log\varphi^2 + \frac{\text{var } y_0}{\varphi^2}\right].$$

$f(\varphi^2)$ is an increasing function of N and var y_0, but given N and var y_0, reaches a minimum for φ^2 = var y_0, i.e. when the sample value is close to the true value of the unconditional variance of the initial observations. So the larger the number of cross-section observations and the larger the sample variance of the initial observations the greater the information contained in them about the prior operation of the process which generated the data. But the closer are φ^2 and var y_0 the less informative are the initial observations on the dependent variable. In Figure 1, I have plotted the function $f(\varphi^2)$ for the values of N and var y_0 for two samples of countries in a panel study of growth used to illustrate these methods in the next section.

2.2. The Model in First Differences[13]

Adding an individual-specific trend, t, to (2)

$$(2') \qquad y_{it} = \alpha + \beta x_{it} + \gamma y_{it-1} + \tau_i t + \mu_i + \varepsilon_{it}\ ,\ i=1,...N,\ t=1,...T,$$

and differencing,

$$(2'') \qquad \Delta y_{it} = \beta\Delta x_{it} + \gamma\Delta y_{it-1} + \tau_i + \omega_{it}\ ,\ \omega_{it} = \Delta\varepsilon_{it}\ ,\ i=1,...N,\ t=1,...T,$$

where Δ denotes the first-difference operator and τ_i is the individual-specific trend coefficient, assumed to have mean zero (enforced by eliminating any overall constant in the differences by deducting the sample means). Thus, not only is the meaning of ρ altered, but if ε_{it} did not contain a unit root to start with ω_{it} will now; in particular, if ε_{it} is not serially correlated to start with, ω_{it} will follow a first-order moving average process with

---

[13] I am indebted to Pietro Balestra for his suggestions on how to work out the likelihood functions in the first-difference case and to Jinkyo Suh for his help in the details. Baltagi and Li (1994), in a paper which later came to may attention, also give a transformation which would permit such a derivation.

unit root. The variance -covariance matrix of the new disturbances $v_{it} = \tau_i + \Delta\varepsilon_{it}$ is now block diagonal with blocks:

$$A = \tilde{\sigma}^2 \begin{bmatrix} 1 & a & b & \dots b \\ a & 1 & a & b.. \\ b & a & 1 & a... \\ \vdots & \vdots & \vdots & \dots 1 \end{bmatrix} \text{ where } \tilde{\sigma}^2 = \sigma_\tau^2 + 2\sigma_\varepsilon^2, a = \frac{\sigma_\tau^2 - \sigma_\varepsilon^2}{\sigma^2}, \text{ and } b = \frac{\sigma_\tau^2}{\tilde{\sigma}^2}.$$

Let $z_{it} = [\Delta y_{i,t-1} \Delta x_{it}]$ and $\delta = [\gamma \ \beta]'$, $v_{i,t} = \tau_i + \omega_{i,t}$, $\omega_{i,t} = \Delta\varepsilon_{i,t}$. Assume

(i) $E(\tau_i) = E(\omega_{i,t}) = 0, \ \forall \ i,t$

(ii) $E(\tau_i \omega_{j,t}) = 0, \qquad \forall i,j,t$

Stacking observations over time we can write the model (2") as:

(2''') $\Delta y_{it} = z_{it} + v_{it}, i = 1,2,\dots,N, t = 2,\dots,T$, or

(9) $\Delta y_i = z_i\delta + v_i = z_i\delta + (\tau_i \underset{\sim(T-1)}{1} + \omega_i), \ i = 1,2,\dots,N,$

where $\underset{\sim(T-1)}{1}$ is $(T-1)\times 1$ column vector of ones. Now, consider mean vector and the variance-covariance matrix of v_i:

$$E(v_i) = E(\tau_i \underset{\sim(T-1)}{1} + \omega_i) = E(\tau_i) \underset{\sim(T-1)}{1} + E(\omega_i) = 0, \qquad \text{by assumption (i).}$$

$$E(v_i v_i') = E[(\tau_i \underset{\sim(T-1)}{1} + \omega_i)(\tau_i \underset{\sim(T-1)}{1} + \omega_i)'] = E(\tau_i^2) \underset{\sim(T-1)}{1} \underset{\sim(T-1)}{1'} + E(\omega_i \omega_i')$$

$$= \sigma_\tau^2 \underset{\sim(T-1)}{1} \underset{\sim(T-1)}{1'} + \sigma_\varepsilon^2 B, \qquad \text{by assumption (ii),}$$

where B is the $(T-1)\times(T-1)$ variance-covariance matrix of ω_i. B is a tri-diagonal matrix having 2 on the main diagonal and (-1) on the adjacent diagonals. Note

$$E(\omega_{i,t}^2) = E[(\varepsilon_{i,t} - \varepsilon_{i,t-1})^2] = 2\sigma_\varepsilon^2$$

$$E(\omega_{i,t}\omega_{i,t-1}) = E[(\varepsilon_{i,t} - \varepsilon_{i,t-1})(\varepsilon_{i,t-1} - \varepsilon_{i,t-2})] = -\sigma_\varepsilon^2.$$

We transform the first-differenced data so as to reduce the model in first-differences to the previous case.

Since matrix B is a positive definite and symmetric, there exits a non-singular matrix, S, such that $SBS' = I_{T-1}$ or $SS' = B^{-1}$. Pre-multiplying (9) by S yields

(10) $S\Delta y_i = Sz_i\delta + Sv_i = Sz_i\delta + S(\tau_i \underset{\sim(T-1)}{1} + \omega_i)$ or

(10') $Y_i = Z_i\delta + V_i$, where $Y_i = S\Delta y_i$, $Z_i = Sz_i$, and $V_i = Sv_i$.

Then, $E(V_i) = E(Sv_i) = SE(v_i) = 0$, and

$$E(V_i V_i') = E[(Sv_i)(Sv_i)'] = SE(v_i v_i')S' = S[\sigma_\tau^2 \underset{\sim(T-1)}{1}\underset{\sim(T-1)}{1'} + \sigma_\varepsilon^2 B]S'$$

$$= \sigma_\tau^2 (S \underset{\sim(T-1)}{1})(S \underset{\sim(T-1)}{1'}) + \sigma_\varepsilon^2 SBS' = \sigma_\tau^2 l_T l_T' + \sigma_\varepsilon^2 I_{T-1}$$

where $l_T = S \underset{\sim(T-1)}{1}$, which is same as the usual error component structure assumed in the previous case. Thus, the conditional and unconditional likelihood functions derived above apply to the transformed data with appropriate re-interpretation of the parameters ρ and σ^2.

We proceed as follows to determine the matrix S and the vector l_T explicitly in order to obtain the required transformation of the data: In stacked form, the variance-covariance matrix of the disturbances in (10') is $\sigma^2\Omega$ where $\sigma^2 = \sigma_\tau^2 + \sigma_\varepsilon^2$,

$$\Omega = \rho(I_N \otimes l_T l_T') + (1-\rho) I_{N(T-1)} = \rho(I_N \otimes l_T l_T') + (1-\rho)(I_N \otimes I_{T-1})$$

$$= I_N \otimes \{\rho(l_T l_T') + (1-\rho) I_{T-1}\} = I_N \otimes A$$

and A is $\rho(l_T l_T') + (1-\rho) I_{N(T-1)}$, $\rho = \dfrac{\sigma_\tau^2}{\sigma^2}$.

Since A is a symmetric matrix, there exists an orthogonal matrix C_{T-1} such that $C_{T-1}' C_{T-1} = I_{T-1}$ and $C_{T-1}' A C_{T-1} = \Lambda_{T-1}$, where Λ_{T-1} is a diagonal matrix with the characteristic roots of A on the diagonal. Note that since C_{T-1} is an orthogonal matrix $C_{T-1}' C_{T-1} = I_{T-1}$, $C_{T-1} C_{T-1}' = I_{T-1}$, and $C_{T-1}' = C_{T-1}^{-1}$. This matrix C_{T-1} is same as the following:

$$C_{T-1} = \left[l_T (l_T' l_T)^{-1/2} \quad \underset{(T-1)\times(T-2)}{C_1} \right] \text{ such that}$$

(a) $C_1' l_T = 0$,

(b) $C_1' C_1 = I_{T-2}$,

(c) $C_1 C_1' = I_{T-1} - l_T (l_T' l_T)^{-1} l_T'$

Given C_1 satisfying (a), (b), and (c), C_{T-1} is an orthogonal, in as much as

$$C_{T-1}' C_{T-1} = \begin{bmatrix} l_T' (l_T' l_T)^{-1/2} \\ \underset{(T-1)\times(T-2)}{C_1'} \end{bmatrix} \left[l_T (l_T' l_T)^{-1/2} \quad \underset{(T-1)\times(T-2)}{C_1} \right] = \begin{bmatrix} l_T' l_T (l_T' l_T)^{-1} & l_T' (l_T' l_T)^{-1/2} C_1 \\ C_1' l_T (l_T' l_T)^{-1/2} & C_1' C_1 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 \\ 0 & I_{T-2} \end{bmatrix} = I_{T-1}$$

and

$$C_T C_T' = \left[l_T (l_T' l_T)^{-1/2} \underset{(T-1)\times(T-2)}{C_1} \right] \begin{bmatrix} l_T' (l_T' l_T)^{-1/2} \\ C_1' \\ (T-1)\times(T-2) \end{bmatrix} = l_T (l_T' l_T)^{-1} l_T' + C_1' C_1$$

$$= l_T (l_T' l_T)^{-1} l_T' + I_{T-1} - l_T (l_T' l_T)^{-1} l' = I_{T-1}$$

Therefore, since $A = \{\rho(l_T l_T') + (1-\rho) I_{T-1}\}$,

$$C_{T-1}' A C_{T-1} = \rho C_{T-1}' (l_T l_T') C_{T-1} + (1-\rho) C_{T-1}' C_{T-1}$$

$$= \rho \begin{bmatrix} l_T' (l_T' l_T)^{-1/2} \\ C_1' \end{bmatrix} (l_T l_T') \left[l_T (l_T' l_T)^{-1/2} \quad C_1 \right] + (1-\rho) I_{T-1}$$

$$= \rho \begin{bmatrix} (l_T' l_T) & 0 \\ 0 & 0 \end{bmatrix} + (1-\rho) I_{T-1} = \begin{bmatrix} 1-\rho+(l_T' l_T)\rho & 0 & 0 \\ 0 & 1-\rho & 0 \\ 0 & 0 & 1-\rho \end{bmatrix} = \Lambda_{T-1}$$

Thus, the $(T-1)$ characteristic roots of A are:

$$\xi = 1 - \rho + (l_T' l_T)\rho$$

with multiplicity one (note the subtle difference between this and the previous case), and

$$\eta = 1 - \rho \text{ , with multiplicity } (T-2).$$

Define $C = I_N \otimes C_{T-1}$ and $\Lambda = I_N \otimes \Lambda_{T-1}$, then $C'\Omega C = \Lambda$, since $C'\Omega C = (I_N \otimes C_{T-1})'(I_N \otimes A)(I_N \otimes C_{T-1}) = I_N \otimes (C_{T-1}' A C_{T-1}) = I_N \otimes \Lambda_{T-1}$. Therefore, the $N(T-1)$ characteristic roots of Ω are

$$\xi = 1 - \rho + (l_T' l_T)\rho \text{ ,}$$

with multiplicity N, and

$$\eta = 1 - \rho \text{ , with multiplicity } N(T-2).$$

It remains to determine the structure of column vector l_T. Since l_T is defined as $l_T = S. \underset{\sim(T-1)}{1}$, we have to know the structure of transformation matrix S. Here, we know the variance-covariance matrix B and so the transformation matrix S is easily obtained from the following formula[14]

[14] See Balestra (1980).

$S = DL$, where $D = diag[\{t(t+1)\}^{-1/2}]$ and $L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 2 & 0 & 0 \\ .. & .. & .. & 0 \\ 1 & 2 & \dots & (T-1) \end{bmatrix}$.

Then,

$$l_T = S \underset{\sim(T-1)}{1} = \frac{1}{2}\begin{bmatrix} (1\times 2)^{1/2} \\ (2\times 3)^{1/2} \\ \dots \\ [(T-1)\times T]^{1/2} \end{bmatrix} \quad \text{and} \quad l_T' l_T = \frac{(T-1)T(T+1)}{12}.$$

Note the difference between this and the previous case.

To transform the data, write the model as

(11) $Y_i = Z_i\delta + V_i$, where $V_i = (\tau_i l_T + w_i)$, i = 1,...,N.

Pre-multiply the i-th equation by $A^{-1/2}$, where

$$A^{1/2} = C_{T-1}\Lambda_{T-1}^{-1/2}C_{T-1}' = \begin{bmatrix} l_T(l_T'l_T)^{-1/2} & C_1 \end{bmatrix}\begin{bmatrix} \xi^{-1/2} & & 0 \\ & \eta^{-1/2} & \\ & . & \\ 0 & & \eta^{-1/2} \end{bmatrix}\begin{bmatrix} l_T'(l_T'l_T)^{-1/2} \\ C_1' \end{bmatrix}$$

$$= \xi^{-1/2} l_T(l_T'l_T)^{-1}l_T' + \eta^{-1/2}C_1C_1' = \xi^{-1/2} l_T(l_T'l_T)^{-1}l_T' + \eta^{-1/2}\{I_{T-1} - l_T(l_T'l_T)^{-1}l_T'\}$$

Thus, $\underset{(T-1)\times 1}{Y_i^*} = \underset{(T-1)\times(T-1)}{A^{-1/2}}\underset{(T-1)\times 1}{Y_i} = \underset{(T-1)\times(T-1)}{A^{-1/2}}\underset{(T-1)\times(T-1)}{S}\underset{(T-1)\times 1}{\Delta y_i}$, and similarly $Z_i^* = A^{-1/2}Z_i$.

To obtain the results for the conditional and the unconditional likelihood functions which parallel (2) and (5) above, simply replace y_{it}, $y_{i,t-1}$, and z_{it} by the transformed values, σ_μ^2 by σ_τ^2, and note

$\xi = 1 - \rho + (l_T'l_T)\rho$, where $\tilde{T} = l_T'l_T = \frac{(T-1)T(T+1)}{12}$ enters where T did before. Thus,

$$\ln L(\gamma,\beta,\sigma_\tau^2,\sigma_\varepsilon^2) = -\frac{N(T-1)}{2}\ln 2\pi - \frac{N(T-1)}{2}\ln\sigma^2 - \frac{N}{2}\ln(1-\rho+\tilde{T}\rho)$$

$$-\frac{N(T-2)}{2}\ln(1-\rho) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\sum_{t=2}^{T}(y_{it}^{*} - \gamma y_{i,t-1}^{*} - x_{it}^{*}\beta)^2 \qquad (12)$$

$$-\frac{N}{2}\ln 2\pi - \frac{N}{2}\ln\varphi^2 - \frac{1}{2\varphi^2}\sum_{i=1}^{N}(y_{i0})^2$$

and

$$\varphi^2 = \frac{\beta^2\sigma_x^2}{1-\gamma^2} + \frac{\sigma_\tau^2}{(1-\gamma)^2} + \frac{\sigma_\varepsilon^2}{1-\gamma^2} = \frac{1}{1-\gamma^2}\left[\beta^2\sigma_x^2 + \sigma^2\left(1+\frac{2\gamma\rho}{1-\gamma}\right)\right],$$

$$\rho = \frac{\sigma_\tau^2}{\sigma^2} = \frac{\sigma_\tau^2}{\sigma_\tau^2+\sigma_\varepsilon^2}.$$

The changes required for the conditional likelihood function are exactly parallel.

3. An Empirical Example: A Study of Cross-Country Economic Growth Using Panel Data

To assess the feasibility and power of likelihood methods for inference about dynamic panel models, I used data on 94 countries for the period 1960 - 1985, and a subsample of 22 OECD countries, from the Penn World Tables 5.6, publicly available from the NBER web site at **ftp://nber.harvard.edu/pub/**. The 22-country sample consists of primarily European countries, all highly developed and tied together by a network of trading relations; the 94-country sample is much more heterogeneous, consisting of the aforementioned 22 plus 72 additional countries ranging from Mozambique and Haiti to the "Asian Tigers." This is the same data set which has been used in dozens of previous studies. Following Islam (1995), s and n were computed as quinquennial means over the preceding 5-year span for the 5 years 1965, 1970, 1975, 1980, 1985; y was taken as the value reported in that year and in 1960 for the lagged value applicable to 1965. In the Table, I present the results both for the levels model, in which country-specific effects affect the intercepts of the growth equation, and for the country-specific trends model, which requires differencing to reduce the process to stationarity. In this case, what I call the first difference model, the conditional and unconditional likelihood functions are defined for the first differences of the original data and the likelihood functions modified from those for the levels model as described in the preceding section.

The Solow-Swan model is the basis for almost all previous investigations and for mine as well: Let y_t = per capita output, k_t = the capital-output ratio, s = the savings rate, δ = the depreciation rate of capital, and n = the exogenous rate of population growth and labor force. All of these variables may differ over time as indicated by their subscript t, but also, in a cross-country context, they are certain to differ from one country to another in a fashion which persists over time. An additional subscript is introduced in the sections which follow this one to indicate that fact. If the production function is Cobb-Douglas, $y_t = A_t k_t^{\alpha}$, where A_t reflects other than conventional factors of production affecting growth and where α, the elasticity of per capita output with respect to the capital-labor ratio, is often interpreted in terms of capital's share as implied by payment of capital at its marginal product. Under these circumstances it can easily be shown, using a simple partial adjustment model, that

$$(13) \qquad \log y_t = \frac{\alpha(1-\gamma)}{1-\alpha}[\log s - \log(n+\delta)] + \frac{1-\gamma}{1-\alpha}\log A_t + \gamma \log y_{t-1}.$$

The speed of convergence to equilibrium is inversely proportional to γ. With growth convergence $0 < \gamma < 1$. In equilibrium, per capita GDP depends only on the parameters n, s, and the time path of A. In an empirical context, these differ from time to time and country to country. Clearly the extent of convergence is conditional on s, n, δ and the time path of A_t. In empirical investigations, changing n and s and sometimes a measure of changing A have been introduced. Below I examine both models in which A is assumed to be constant although differing from one country to another and models in which A_t can be represented by a simple linear trend which plausibly also differs from country to country.

Recent Empirical Investigations

Equation (13) has been widely used to examine the hypothesis of growth convergence (Mankiw, et al , 1992, p.410; Barro and Sala-i-Martin, 1995, Chapter 12; Islam, 1995, p. 1133; Lee, et al. 1996, Casseli, et al. 1996). In empirical work, y_t is replaced by real per capita GDP; when varying s and n are taken into account, s is replaced by an average savings rate over the period t-1 to t, and n is replaced by the growth rate of population over the period t-1 to t. It is usual to use rates averaged over several years; following Islam (1995) and others, I have used quinquennial averages. The restriction on the coefficients of ln(s) and ln(n+δ), which arises from the constant-returns-to-scale assumption implies that ln(s) and ln(n+δ) can be collapsed into a single variable. Testing the growth convergence hypothesis, in this context, revolves largely around the coefficient γ of the initial level of per capita real GDP. If this is positive but much less than one, the implication is that on average countries with low initial values are growing faster than those with high initial values and is therefore evidence of convergence. Whereas if this coefficient is close to one, perhaps even slightly larger than one, the implication is that initial values have little or no effect or even a perverse one on subsequent growth; such a finding is therefore evidence against the neoclassical theory which implies convergence. For example, if $\gamma = 0.9$, convergence to within 90% of final equilibrium occurs only in 22 periods, which, given quinquennial data, implies 110 years! Similarly, 0.8 requires 53 years, 0.7 32 years, while 0.2 requires only 7 years and 0.1 is within 90% in 5 years. Using unconditional ML below, I obtain a value of about 0.45 which yields convergence to within 90% of equilibrium within 15 years.

The estimates of γ obtained heretofore using cross-country quinquennial data are generally in excess of 0.7 no matter what econometric procedure is employed, but vary over a wide range depending on the method, 0.7 to 0.98. But for the differenced model, many estimates of γ are much smaller, in the vicinity of 0.5.[15] (See Nerlove, 1996, for a summary and comparison of many of the standard methods of analysis with one another and with the likelihood methods proposed here.) It is apparent that, for all practical purposes, coefficients in excess of 0.7 represent negligible convergence, since, with unchanging s, n, and A, it would take more than a generation to achieve 90% of equilibrium real per capita GDP.

---

[15] Using a GMM estimator derived from a modified Chamberlain approach (Chamberlain, 1984; Crépon and Mairesse, 1996), Caselli, et al. (1996) obtain an estimate of about 0.51 - 0.53, i.e., much more rapid convergence and close to the estimates obtained for the 94-country sample using either conditional or unconditional ML. My estimates for the 22-country sample are much higher, however.

It is interesting to note that these methods are basically instrumental variable methods which use lagged values of the explanatory variables as instruments, an approach which was employed in Balestra and Nerlove (1966) to obtain initial consistent estimates of the residual variance-covariance matric on which to base feasible GLS.

Most recent work attempts to test whether $\gamma = 1$; however, this is a test for unit root in log y_{it}. Even under the best of circumstances testing for a unit root is problematic.[16]

Tests based on a single cross-section (which can be viewed as a panel of time dimension 1) or on pooled cross-section time series (panel) data generally have yielded contradictory results: Pooled panel data studies tend to reject the hypothesis of convergence (relatively high γ's), even after controlling for population growth rates, savings rates and other variables. Dynamic fixed-effects models are of course not possible for a single cross-section, but recent work (Islam, 1995) using a dynamic fixed-effects panel model yields results supporting convergence. There are serious problems with tests such as these which rely on the estimated coefficients of the initial, or lagged value, of the dependent variable in dynamic panel models, or in the special case of a single cross-section, which arise from two sources of bias. In this paper, I contrast these findings with results obtained from likelihood methods which take account of the information contained in the initial observations about the process which must have generated those observations in the pre-sample period. In Nerlove (1996), I showed that many of the earlier findings are probably statistical artifacts arising from biases in the econometric methods employed, and, now here, especially failure to take advantage of the relatively large amount of information contained in the initial observations and failure to take account of differing country-specific trends. This demonstrates the sensitivity of the conclusions drawn about γ to the econometric method employed, irrespective of the validity of the relationship estimated.

The first source of bias are omitted variables, especially infrastructure and investments over time in infrastructure, and the natural resource base available to each country in cross-sectional or panel studies. Systematic differences in these across countries or regions will systematically bias the conclusions. Because such variables are likely to be correlated with savings or investment rates in conventional or in human capital and with population growth rates it is not altogether clear what the net effect of omitting them on the coefficient of the initial value will be in a single cross-section.[17] But in a pooled model it is clear that, to the extent such differences are persistent, they will be highly correlated with the initial value and therefore omitting them will bias the coefficient of that variable upwards towards one and thus towards rejecting convergence. This source of bias has been well-known since the early paper by Balestra and Nerlove (1966) and is well-supported by the Monte Carlo studies reported in Nerlove (1971). In this light, it is not surprising that pooled panel data, or single cross-sections, which are a special case of panels with $T = 1$, even with inclusion of additional variables, often reject convergence.

Second, since there are likely to be many sources of cross country or cross region differences, many of which cannot be observed or directly accounted for, it is natural to try to represent these by fixed effects in a panel context. But, as is well-known from the Monte Carlo investigations reported in Nerlove (1971) and demonstrated analytically by Nickell (1981), inclusion of fixed effects in a dynamic model biases the coefficient of the initial value of the dependent variable included as an explanatory variable downwards, towards zero and therefore towards support for the convergence hypothesis. This may account for Islam's (1995a) recent findings.[18]

Alternative estimates based on more appropriate random-effects models, such as two-stage feasible Generalized Least Squares or maximum likelihood conditional on the initial observations are also biased in small samples and inconsistent in large, or in the case of Instrumental Variable estimates have poor sampling properties or are difficult to implement. For example, the papers by Knight, Loayza

---

[16] Bernard and Durlauf (1995) use cointegration techniques on rather longer time series for 15 OECD countries to test alternative time-series definitions of convergence and contrast the results with the standard formulation.

[17] Caselli, et al. (1996) attempt to control for the endogeneity of these explanatory variables.

[18] As pointed out to me by Hashem Pesaran, these country-specific effects may also be trends, since many of the latent variables specific to each individual in the cross section may themselves be trending. Taking such country-specific trends into account has been one of the great challenges of doing this paper.

and Villanueva(1993), Loayza (1994), and Islam (1995a) employ a method, among others, proposed by Chamberlain (1984), generally referred to as the Π-matrix approach.[19] The alternative of unconditional maximum likelihood suggested in Nerlove and Balestra (1996) is implemented here. In the case of country-specific trends, this method requires differencing the data in order to achieve stationarity, which, in turn, requires a reformulation of both the conditional and the unconditional lilkelihood functions.[20]

Even if one has little interest in the question of convergence, or its rate, per se, the question of whether the coefficient of the state variable, lagged dependent or initial value, is biased in the sense of being inconsistent is an important one since biases in this coefficient will affect the estimates of the coefficients of other variables correlated with it and their levels of significance. To the extent such estimates are important in the formulation of policies to promote growth, the matter is indeed a serious one.[21]

New Estimates Based on Maximizing the Conditional and Unconditional Likelihood Functions

Consider first the levels model: The estimates of ρ, the ratio of the unobserved country-specific variation relative to the total unobserved residual variation, are much higher for the 22-country sample than for the 94-country, 0.48 vs. 0.11 for conditional estimates, but 0.77 vs. 0.13 for unconditional estimates. The estimate of the overall residual variation is about the same for the 94-country sample whether one conditions on the initial observations or not, while that for the 22-country sample is reduced dramatically by conditioning. The values of the other coefficients differ significantly for the two samples but are quite similar for the conditional and the unconditional ML methods. The variance of the initial values of real per capita GDP is only 0.256 for the 22-country sample vs. 0.799 for the 94-country sample. On the other hand, the value of

$$\varphi^2 = \frac{1}{1-\gamma^2}\left[\beta^2\sigma_x^2 + \sigma^2(1+\frac{2\gamma\rho}{1-\gamma})\right],$$

which together with the variance of the initial values, determines the "distance" between the conditional and the unconditional likelihood functions is in both cases, at the unconditional ML estimates, very close to these variance values, 0.25 for the 22-country sample and 0.91 for the 94-country sample. The difference or similarity of the estimates obtained from the two likelihood functions reflects the off-setting effects of the larger sample and greater heterogeneity for the 94 countries vs. the opposite for the 22 countries. At the same time the estimate of ρ behaves in such a way as to minimize the "distance" between the two functions. See Figure 1. The estimates of γ, the adjustment coefficient, are very high, irrespective of method and the estimates of α, which should approximate capital's share of GDP, are also considerably higher than one would expect.

Turning now to the estimates of the first-difference model: The relation between the estimates of ρ for the 22-country sample and the 94-country are now reversed, being much smaller and by conventional measures insignificantly different form zero, for the 22-country sample than for the 94-country sample. This is just what one would expect if the 22 countries had more-or-less common trends, while in the larger more hetrerogeneous 94-country sample trends were more diverse. The difference between the levels and the first-difference models is

---

[19] See also Crépon and Mairesse (1996).

[20] Lee, et al. (1996) also estimate from what they maintain is an unconditional likelihood function, but inasmuch as they do not transform to stationarity (their relationship includes both a constant and a linear trend), I do not think their formulation of the likelihood function based on the unconditional density of the dependent variable is correct. They use annual observations to obtain sufficient degrees of freedom to estimate individual country-specific trends, but I think they are only fooling themselves if they think that much of the information contained in the annual observations is real, as opposed to interpolated.

[21] For example in (13) the parameter α could be derived from the coefficient of the variable log s - log (n+δ) as coefficient/(coefficient +1-γ), so there is a double source of bias. Indeed, a number of authors accept or reject statistical formulations based on the estimated value of α which should approximate capital's share.

striking. With the exception of β and the implied α, the estimates of the other coefficients are similar for the conditional and unconditional methods, although they differ greatly as between the 94-country and the 22-country samples. The estimates of α are now of reasonable magnitude, about 20%, except higher for the unconditional estimates for 22 countries. Perhaps the most dramatic change is the great reduction in both ρ and γ as compared with the levels model. The estimates of γ are still large for the 22-country sample, but now for the 94-country sample imply convergence to equilibrium in only 13 or 14 years. On balance, the estimates for the first-difference model are much more reasonable than for the levels model, while except for rather different values of ρ, which of course has a rather different interpretation for the two models, there is not much basis for choosing between conditional vs. unconditional ML.

Lest we accept these conclusions too uncritically, however, remember that the value of the likelihood function at the maximum and its quadratic approximation there is only a partial and imperfect basis for inference. In the models considered here, likelihood is a function of four parameters, ρ, γ_0, β and σ^2, assuming that the overall intercept has been removed by taking deviations from the overall sample means. Unfortunately, in terms of assessing the support for various combinations of parameter values, we are in the position of the poor hexagon in his dealings with the sphere in *Flatland.* Our likelihood functions are five-dimensional creatures while we, poor souls, are only three-dimensional. Taking two parameters at a time and "slicing" or sectioning the likelihood function along a plane defined by some values of the other two, say, for example, the values which maximize the likelihood function, is potentially a way out. But, inasmuch as there are $\binom{4}{2} = 6$ possible combinations of the 4 parameters taken 2 at a time, we are left in a situation like that of the six blind men of Indostan asked to describe an elephant, each one by feeling only one part of the animal. For reasons of limited space and time, I choose to look only at one pair, ρ and γ, here. "Slices" of the four unconditional likelihood functions, levels and differences, 94-country and 22-country samples, are presented in Figures 2 - 5. Each figure consists of two parts, a close-up near the maximum value and a panoramic view of nearly the whole surface of the "slice." The "slices" are all defined by the plane of β and σ^2 which maximize the joint unconditional likelihood of the observations. Both surface and contour plots are presented. Each Figure, therefore, consists of four graphs.

The "slices," ρ vs. γ, suggest that the conventional asymptotic standard errors, which are a reflection of the quadratic approximation to the likelihood around its maximum, are of some help but not a sure guide to assessing support for the possible values of the parameters. While γ is well-determined in most instances, ρ is not. Although the algorithms for maximizing the likelihood functions converged without difficulties in every case and no boundary solutions were encountered, it is clear, especially from the panoramic views that the likelihood functions of the ρ-γ "slices" are rather flat over considerable ranges. Rather than offering much support for the ML estimates, these partial views suggest rather clearly what is *not* supported by the data. The first-difference model has the greatest support for a reasonably small value of γ in the 94-country case, but there is a clear trade-off between ρ and γ not revealed by the conventional asymptotic standard errors. Perhaps the greatest benefit of looking at the likelihood function at points in the parameter space away from the likelihood maximizing values is to reveal the fragility of inferences based on the ML and other estimates which may appear to be rather precise.

REFERENCES

Abbott, E. A., *Flatland.* New York: Dover, 1884, reprinted 1946.

Anderson, T. W., and C. Hsiao, "Estimation of Dynamic Models with Error Components." *Journal of the American Statistical Association, 76*: 598 - 606, 1981.

Anderson, T. W., and C. Hsiao, "Formulation and Estimation of Dynamic Models Using Panel Data," *Journal of Econometrics, 18*: 47 - 82, 1982.

Balestra, P., " A Note on the Exact Transformation Associated with the First-Order Moving Average Process," *Journal of Econometrics, 14:* 381-94, 1980.

Balestra, P., and M. Nerlove, "Pooling Cross-Section and Time-Series Data in the Estimation of a Dynamic Economic Model: The Demand for Natural Gas." *Econometrica, 34*:585-612, 1966.

Barnard, G. A., "Statistical Inference," *Jour. Royal Statistical Society, Ser. B 11:* 115-149, 1949.

Barnard, G. A., "The Theory of Information," *Jour. Royal Statistical Society, Ser. B 13:* 46-64, 1951.

Barnard, G. A., "The Use of the Likelihood Function in Statistical Practice," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1:* 27-40, 1966.

Barnard, G. A., "The Bayesian Controversy in Statistical Inference," *Journal of the Institute of Actuaries, 93:* 229-69, 1967.

Barnard, G. A., G. M. Jenkins and C. B. Winsten "Likelihood Inference and Time Series," *Jour. Royal Statistical Society, Ser. A 125:*321-72, 1962.

Barndorff-Nielsen, O. E., *Parametric Statistical Models and Likelihood,* Lecture Notes in Statistics, No. 50. Berlin: Springer-Verlag, 1988.

Barro, R. J., and X. Sala-I-Martin, *Economic Growth.* New York: McGraw-Hill: 1995.

Baltagi, B. H., *Econometric Analysis of Panel Data,* New York: Wiley, 1995.

Baltagi, B. H. and Q. Li, "Estimating Error Component Models with General MA(q) Disturbances," *Econometric-Theory; 10:* 396-408, 1994.

Baumol, W., "Productivity Growth, Convergence, and Welfare: What the Long-Run Data Show." *American Economic Review, 76:* 1072-1085, 1986.

Bernard, A. B., and S. N. Durlauf, "Convergence in International Output." *Journal of Applied Econometrics, 10*:97-108, 1995.

Birnbaum, A., "On the Foundations of Statistical Inference," *Jour. American Statistical Association, 57:*269-306, 1962.

Breusch, T. S., "Maximum Likelihood Estimation of Random Effects Models," *Journal of Econometrics, 36*: 383 - 389, 1987.

Caselli, F., G. Esquivel and F. Lefort, "Reopening the Convergence Debate: A New Look at Cross-Country Growth Empirics," *Journal of Economic Growth, 1*: 363-389, 1996.

Chamberlain, G., "Panel Data." Pp. 1247-1313 in Z. Griliches and M. Intriligator, eds., *Handbook of Econometrics, II.* Amsterdam: Elsevier: 1984 .

Crépon, B., and J. Mairesse, "The Chamberlain Approach." Pp. 323-391 in L. Mátyás and P. Sevestre, *op.cit.*, 1996.

Cramer, J. S., *Econometric Applications of Maximum Likelihood Methods.* Cambridge: University Press, 1986.

de la Fuente, A., "The Empirics of Growth and Convergence: A Selective Review," *Journal of Economic Dynamics and Control, 21*: 23-73, 1997.

Davidson, R. and J. G. MacKinnon, *Estimation and Inference in Econometrics,* New York: Oxford University Press, 1993.

Edwards, A. W. F., *Likelihood.* Cambridge: University Press, 1972.

Fisher, R. A., "On the Probable Error of a Coefficient of Correlation Deduced from a Small Sample Mean," *Metron, 1(4)*: 3 - 32, 1921.

Fisher, R. A., "On the Mathematical Foundations of Theoretical Statistics," *Philosophical Transactions of the Royal Society of London, Series A, 222:*309-368, 1922.

Fisher, R. A., "Theory of Statistical Estimation," *Proceedings of the Cambridge Philosophical Society, 22*:700-725, 1925.

Fisher, R. A., "Inverse Probability and the Use of Likelihood, *Proceedings of the Cambridge Philosophical Society, 28*: 257 - 261, 1932.

Fisher, R. A., *Statistical Methods, Experimental Design, and Scientific Inference,* being a reprint of *Statistical Methods for Research Workers* (1925), *The Design of Experiments* (1935), and *Statistical Methods and Scientific Inference* (1956). Oxford: University Press, 1990.

Greene, W. H., *Econometric Analysis, 2nd Edition,* New York: Macmillan, 1993.

Hsiao, C., *Analysis of Panel Data,* Cambridge: University Press, 1986.

Islam, N., "Growth Empirics: A Panel Data Approach." *Quarterly Journal of Economics, 110*:1127-1170, 1995.

Knight, M., N. Loayza and D. Villanueva, "Testing the Neoclassical Growth Model," *IMF Staff Papers, 40*: 512-41, 1993.

Jeffereys, H., "Probability and Scientific Method," *Proceedings of the Royal Society, Ser. A, 146:* 9-16, 1934.

Jeffereys, H., *Theory of Probability.* Oxford: University Press, 1961.

Laplace, Pierre Simon, 1774 - 1814, especially *Théorie Analytique des Probabilités*, 1812. Cited in Stephen M. Stigler, *The History of Statistics*, Chapter 3, "Inverse Probability," pp. 99 - 138. Cambridge: Harvard University Press, 1986.

Lee, K., M. H. Pesaran, and R. Smith, "Growth and Convergence: A Multicountry Empirical Analysis of the Solow Growth Model," unpublished, 1996.

Lindsey, J. K., *Parametric Statistical Inference.* Oxford: Clarendon Press, 1996.

Loayza, N., "A Test of the International Convergence Hypothesis Using Panel Data." *Policy Research Working Paper No. 1333.* The World Bank, 1994 .

Mátyás, L. and P. Sevestre, 1996. *The Econometrics of Panel Data: Handbook of Theory and Applications.* 2nd. ed.1996. Boston: Kluwer Academic Publishers, 1992.

Maddala, G. S., "The Use of Variance Components Models in Pooling Cross-Section and Time Series Data," *Econometrica, 39:*341-358, 1971.

Maddala, G. S., "The Likelihood Approach to Pooling Cross-Section and Time-Series Data," *Econometrica, 39:* 939 – 953, 1971.

Mankiw, N. G., D. Romer and D. N. Weil, "A Contribution to the Empirics of Economic Growth." *Quarterly Journal of Economics, 108*:407-437, 1992.

Mundlak, Y., "On the Pooling of Cross-Section and Time-Series Data," *Econometrica, 46:* 69-86, 1978.

Nerlove, M., "Further Evidence on the Estimation of Dynamic Economic Relations from a Time Series of Cross-Sections." *Econometrica, 39*:359-382 (1971).

Nerlove, M., "Growth Rate Convergence, Fact or Artifact? An Essay in Panel Data Econometrics." Paper presented to the Sixth Conference on Panel Data Econometrics, Amsterdam, 28-29 June 1996.

Nerlove, M., and P. Balestra, "Formulation and Estimation of Econometric Models for Panel Data." Pp. 3-22 in L. Mátyás and P. Sevestre, *op.cit.*, 1996.

Nickell, S., "Biases in Dynamic Models with Fixed Effects." *Econometrica, 49*:1417-1426, 1981.

Press, S. J., *Bayesian Statistics: Principles, Models and Applications.* New York: Wiley, 1989.

Saxe, John Godfrey, *The Poems: Complete Edition,* Boston: Houghton, Mifflin & Co., 1880.

Sevestre, Patrick, and Alain Trognon, "Propriétés de grands échantillons d'une classe d'estimateurs des modéles autoregréssives à erreurs compossées," *Annales de l'INSEE, 50:*25-49, 1983.

Sevestre, Patrick, and Alain Trognon, "Dynamic Linear Models," pp. 120-144 in Mátyás and Sevestre, *op. cit.*, 1996.

Trognon, Alain, "Miscellaneous Asymptotic Properties of Ordinary Least Squares and Maximum Likelihood Methods in Dynamic Error Components Models," *Annales de l'INSEE,30-31:*631-657, 1978.

TABLE: CONDITIONAL AND UNCONDITIONAL MAXIMUM-LIKELIHOOD ESTIMATES FOR THE LEVELS MODEL AND THE FIRST-DIFFERENCE MODEL, 94-COUNTREY AND 22-COUNTRY SAMPLES

LEVELS MODEL

	94-Country Sample	22-Country Sample
Conditional ML		
ρ	0.1133 (0.0497)	0.4796 (0.1584)
γ	0.9339 (0.0122)	0.8189 (0.0245)
β	0.1370 (0.0131)	0.1908 (0.0438)
Implied α	0.6744 (0.0289)	0.5131 (0.0664)
Residual Variance	0.0194 (0.0013)	0.0052 (0.0012)
Unconditional ML		
Estimates of σ_x^2 used	0.0826	0.0069
ρ	0.1288 (0.0456)	0.7700 (0.0731)
γ	0.9385 (0.0105)	0.8085 (0.0228)
β	0.1334 (0.0124)	0.1815 (0.0521)
Implied α	0.6846 (0.0277)	0.4865 (0.0791)
Residual Variance	0.0197 (0.0013)	0.0113 (0.0028)

FIRST-DIFFERENCE MODEL

	94-Country Sample	22-Country Sample
Conditional ML		
ρ	0.2267 (0.0664)	0.0126 (0.0405)
γ	0.4540 (0.0651)	0.6187 (0.0490)
β	0.1368 (0.0208)	0.0815 (0.0601)
Implied α	0.2004 (0.0358)	0.1762 (0.1159)
Residual Variance	0.0122 (0.0009)	0.0021 (0.0003)
Unconditional ML		
Estimate of σ_x^2 used	0.0597	0.0058
ρ	0.2335 (0.0632)	0.0936 (0.0696)
γ	0.4364 (0.0578)	0.7254 (0.0512)
β	0.1340 (0.0201)	0.1478 (0.0727)
Implied α	0.1921 (0.0317)	0.3500 (0.1326)
Residual Variance	0.0120 (0.0008)	0.0027 (0.0004)

Figures in parentheses are asymptotic standard errors.

Notes to the Table

Data on 94 countries for the period 1960 - 1985 from the Penn World Tables 5.6, publicly available from the NBER web site at ftp://nber.harvard.edu/pub/.

22-Country Sample:

Japan
Austria
Belgium
Denmark
Finland
France
Germany (FRG)
Greece
Ireland
Italy
Netherlands
Norway
Portugal
Spain
Sweden
Switzerland
Turkey
U. K.
Canada
U. S.
Australia
New Zealand

94-country Sample = 22-Country Sample + the Following:

Algeria
Botswana
Cameroon
Ethiopia
Ivory Coast
Kenya
Madagascar
Malawi
Mali
Morocco
Nigeria
Senegal
South Africa
Tanzania
Tunisia
Zambia
Zimbabwe
Costa Rica
Dominican Rep.
El Salvador
Guatemala
Haiti
Honduras
Jamaica
Mexico
Nicaragua
Panama
Trinidad & Tobago
Argentina
Bolivia
Brazil
Chile
Colombia
Ecuador
Paraguay
Peru
Uruguay
Venezuela
Bangladesh
Hong Kong
India
Israel
Jordan
Korea
Malaysia
Burma
Pakistan
Philippines
Singapore
Sri Lanka
Syria
Thailand
Angola
Benin
Burundi
Central African Republic
Chad
Congo
Egypt
Ghana
Liberia
Mauritania
Mauritius
Mozambique
Niger
Rwanda
Somalia
Togo
Uganda
Zaire
Nepal
Papua New Guinea

FIGURES

Figure 1: Graphs of log(ratio conditional/unconditional lf) against phi2.

Figure 2: Unconditional likelihood. Levels. 94 countries. Rho vs. gamma. Sigma2 and beta fixed. Closeup and panoramic views.

Figure 3: Unconditional likelihood. Levels. 22 countries. Rho vs. gamma. Sigma2 and beta fixed. Closeup and panoramic views.

Figure 4: Unconditional likelihood. First differences. 94 countries. Rho vs. gamma. Sigma2 and beta fixed. Closeup and panoramic views.

Figure 5: Unconditional likelihood. First differences. 94 countries. Rho vs. gamma. Sigma2 and beta fixed. Closeup and panoramic views.

FIGURE 1.1: GRAPH OF log(RATIO CONDITIONAL/UNCONDITIONAL LF) AGAINST PHI2:
94 COUNTRIES, ESTIMATED PHI2 = 0.91.

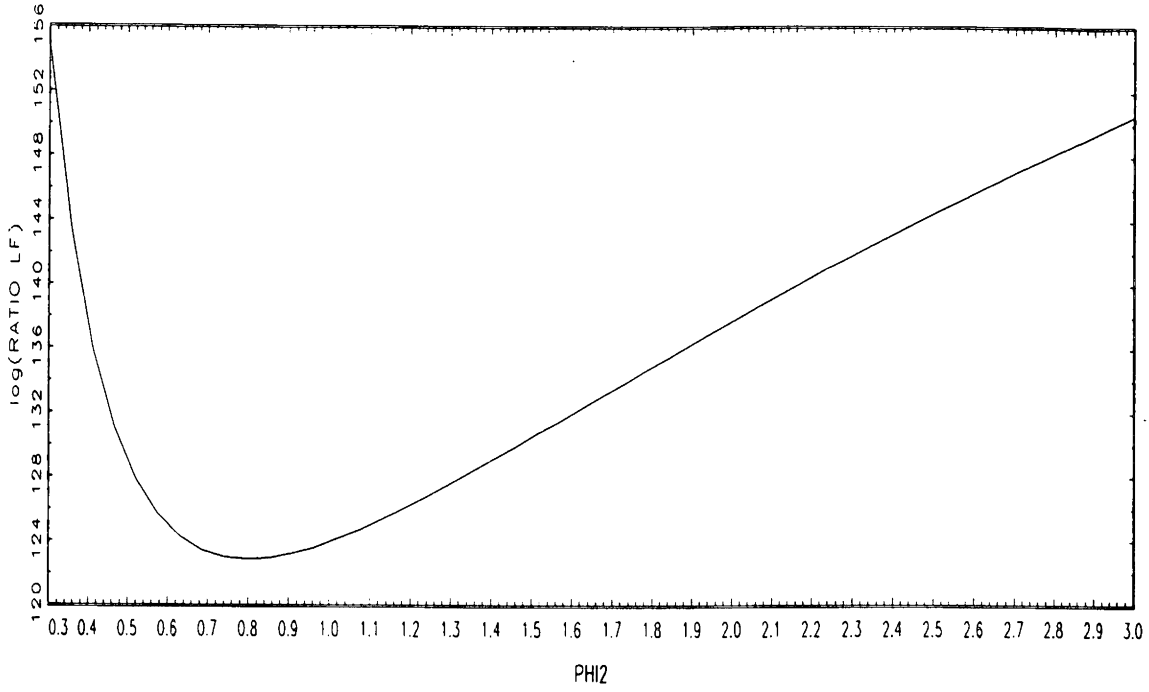

FIGURE 1.2: GRAPH OF log(RATIO CONDITIONAL/UNCONDITIONAL LF) AGAINST PHI2:
22 COUNTRIES. ESTIMATED PHI2 = 0.25.

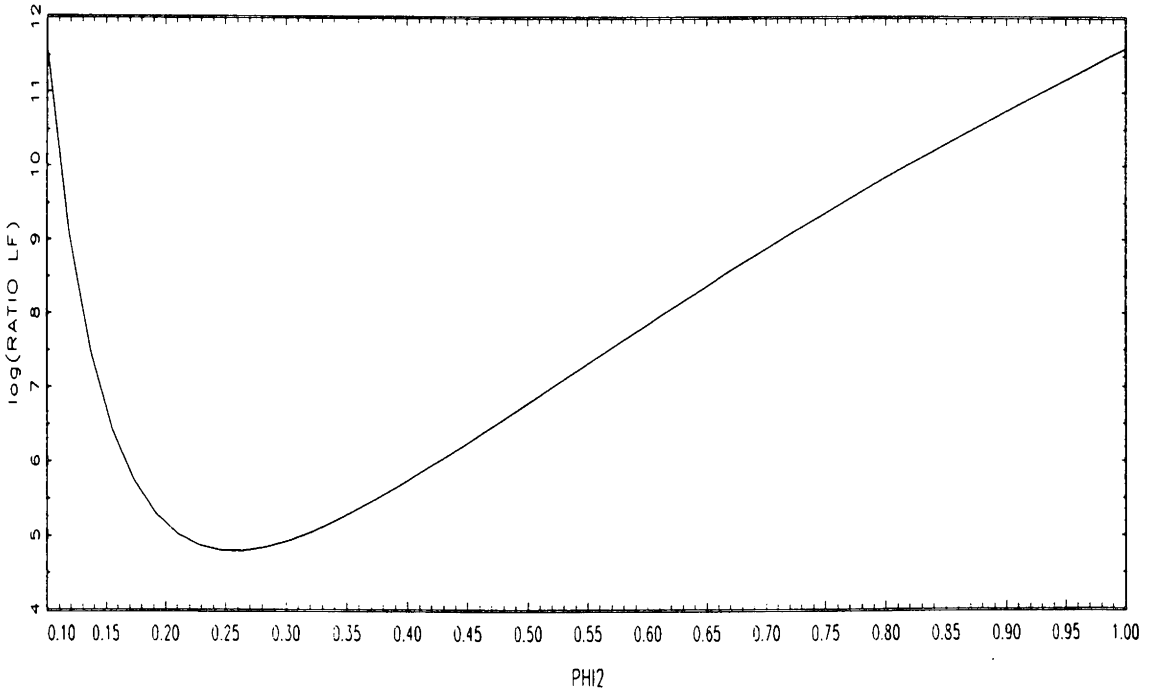

FIGURE 2.1: SURFACE. UNCONDITIONAL LIKELIHOOD, LEVELS, 94 COUNTRIES, RHO VS. GAMMA. SIGMA2 AND BETA FIXED. CLOSEUP VIEW.

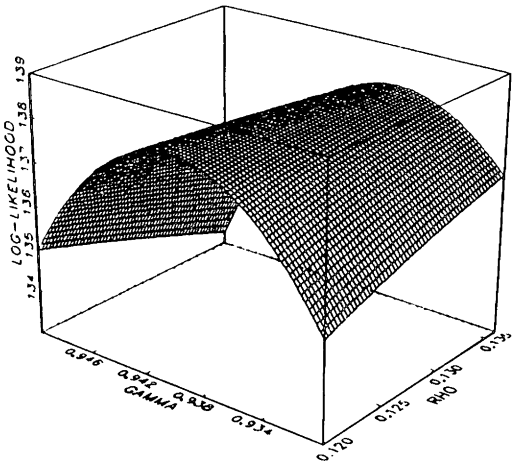

FIGURE2.2: CONTOURS. UNCONDITIONAL LIKELIHOOD, LEVELS, 94 COUNTRIES, RHO VS. GAMMA. SIGMA2 AND BETA FIXED. CLOSEUP VIEW.

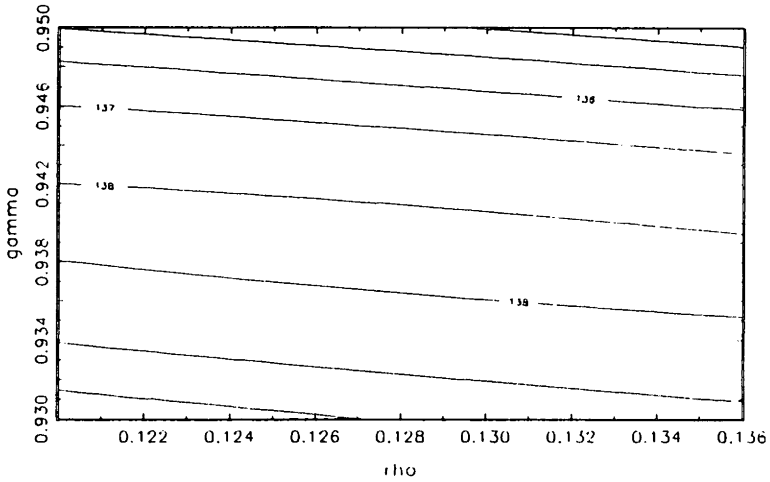

FIGURE 2.3: SURFACE. UNCONDITIONAL LIKELIHOOD, LEVELS, 94 COUNTRIES, RHO VS. GAMMA. SIGMA2 AND BETA FIXED. PANORAMIC VIEW.

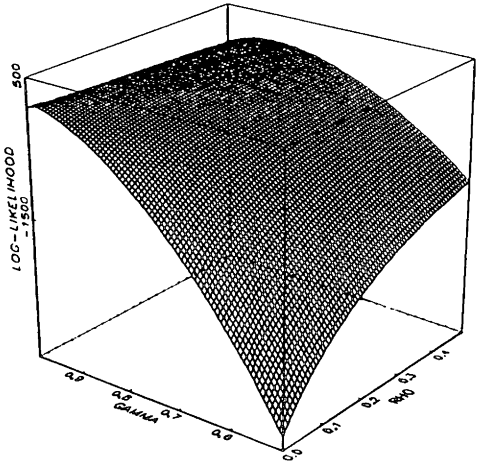

FIGURE 2.4: CONTOURS. UNCONDITIONAL LIKELIHOOD, LEVELS, 94 COUNTRIES, RHO VS. GAMMA. SIGMA2 AND BETA FIXED. PANORAMIC VIEW

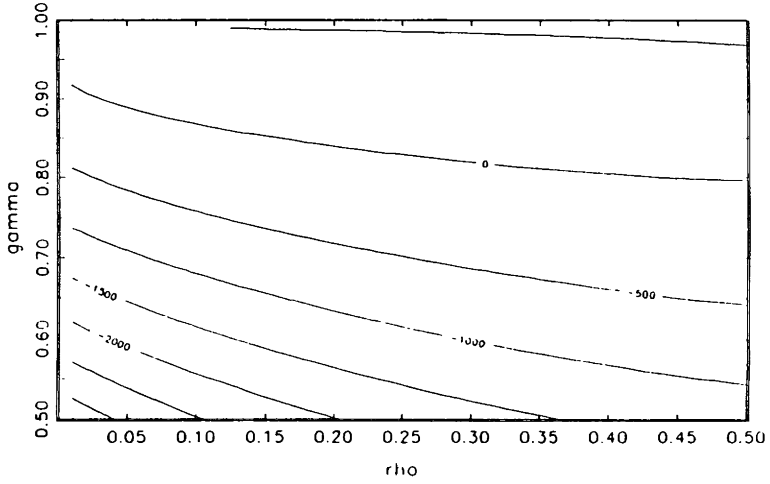

FIGURE 3.1: SURFACE. UNCONDITIONAL LIKELIHOOD. LEVELS. 22 COUNTRIES. RHO VS. GAMMA. SIGMA2 AND BETA FIXED. CLOSEUP VIEW.

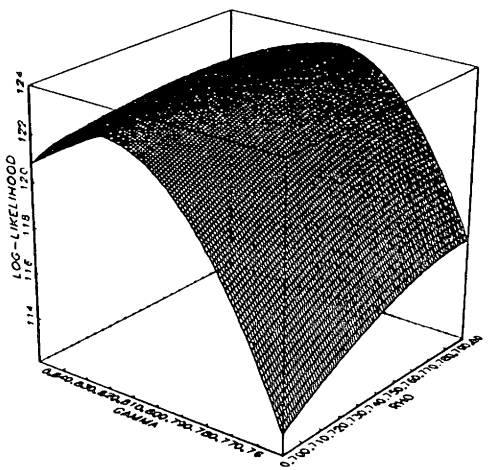

FIGURE3.2: CONTOURS. UNCONDITIONAL LIKELIHOOD. LEVELS. 22 COUNTRIES. RHO VS. GAMMA. SIGMA2 AND BETA FIXED. CLOSEUP VIEW.

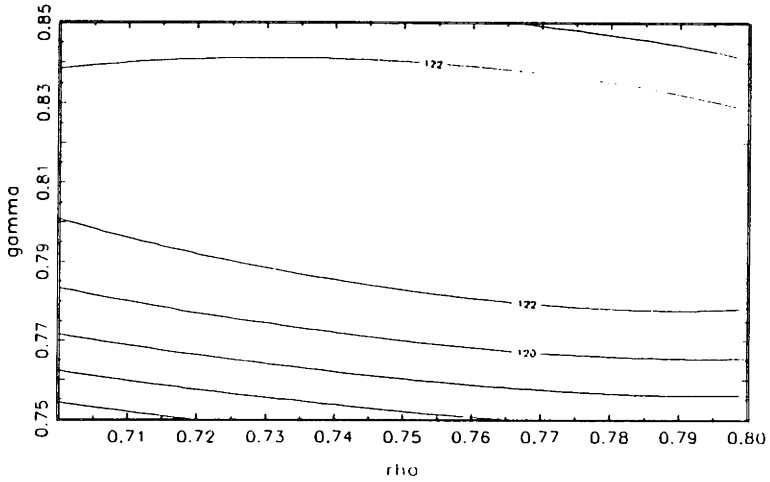

FIGURE 3.3: SURFACE. UNCONDITIONAL LIKELIHOOD. LEVELS. 22 COUNTRIES. RHO VS. GAMMA. SIGMA2 AND BETA FIXED. PANORAMIC VIEW.

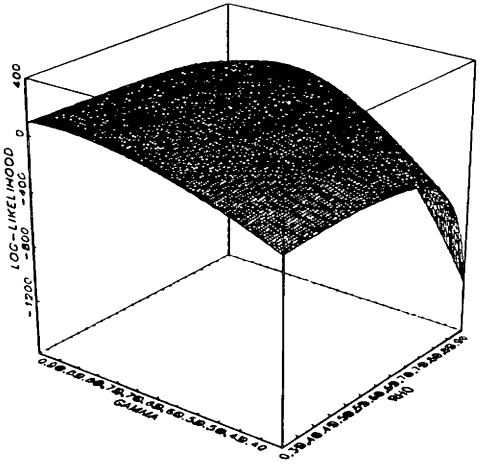

FIGURE 3.4: CONTOURS. UNCONDITIONAL LIKELIHOOD. LEVELS. 22 COUNTRIES. RHO VS. GAMMA. SIGMA2 AND BETA FIXED. PANORAMIC VIEW

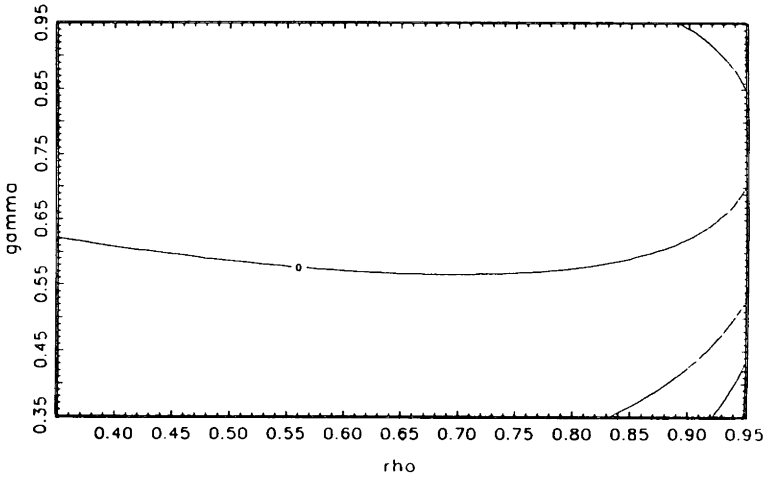

FIGURE 4.1: SURFACE. UNCONDITIONAL LIKELIHOOD. FIRST DIFFERENCES. 94 COUNTRIES. RHO VS. GAMMA. SIGMA2 AND BETA FIXED. CLOSEUP VIEW.

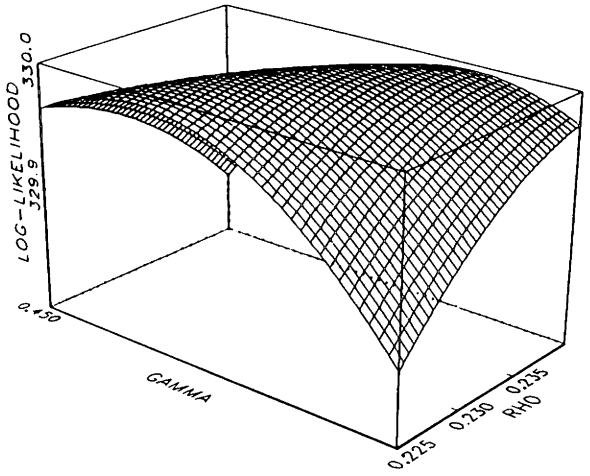

FIGURE 4.2: CONTOURS. UNCONDITIONAL LIKELIHOOD. FIRST DIFFERENCES. 94 COUNTRIES. RHO VS. GAMMA. SIGMA2 AND BETA FIXED. CLOSEUP VIEW

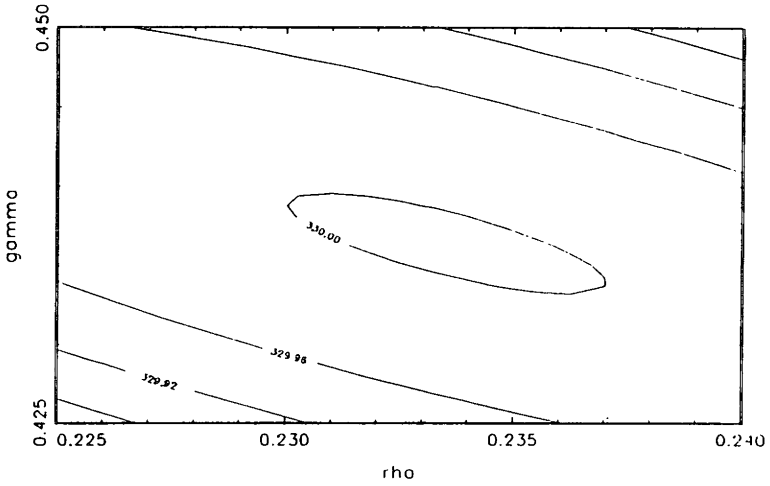

FIGURE 4.3: SURFACE. UNCONDITIONAL LIKELIHOOD. FIRST DIFFERENCES 94 COUNTRIES, RHO VS. GAMMA. SIGMA2 AND BETA FIXED. PANORAMIC VIEW.

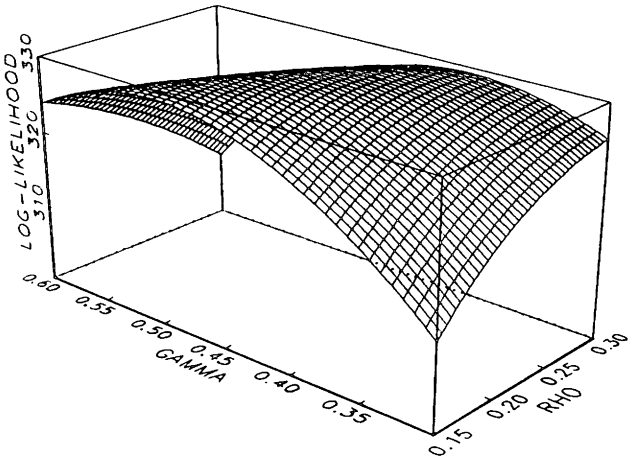

FIGURE 4.4: CONTOURS. UNCONDITIONAL LIKELIHOOD. FIRST DIFFERENCES. 94 COUNTRIES, RHO VS. GAMMA. SIGMA2 AND BETA FIXED. PANORAMIC VIEW.

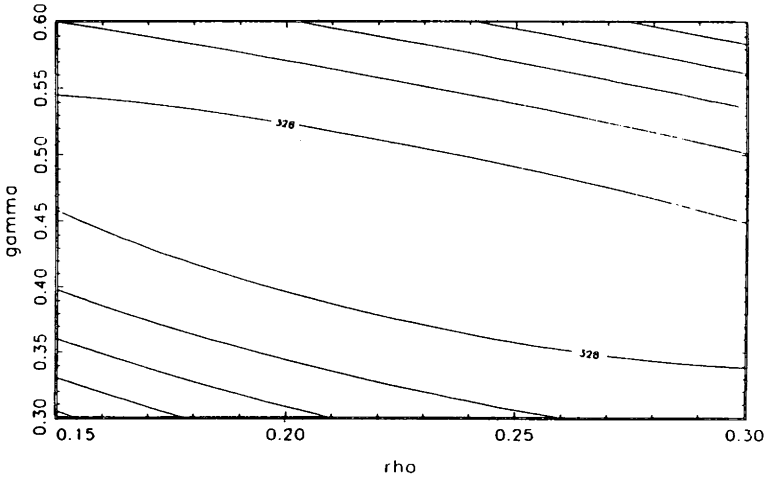

FIGURE 5.1: SURFACE. UNCONDITIONAL LIKELIHOOD. FIRST DIFFERENCES. 22 COUNTRIES. RHO VS. GAMMA. SIGMA2 AND BETA FIXED. CLOSEUP VIEW.

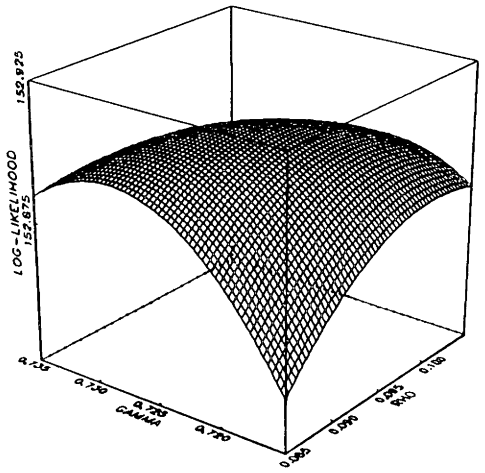

FIGURE 5.2: CONTOURS. UNCONDITIONAL LIKELIHOOD. FIRST DIFFERENCES. 22 COUNTRIES, RHO VS. GAMMA. SIGMA2 AND BETA FIXED. CLOSEUP VIEW

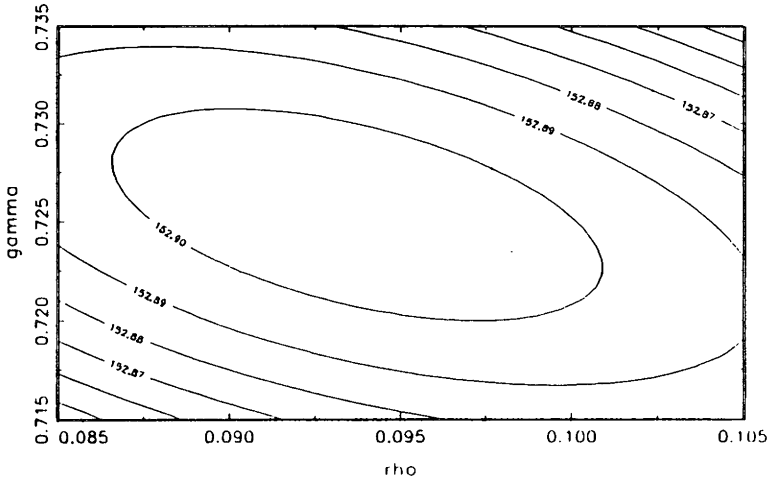

FIGURE 5.3: SURFACE. UNCONDITIONAL LIKELIHOOD. FIRST DIFFERENCES 22 COUNTRIES, RHO VS. GAMMA. SIGMA2 AND BETA FIXED. PANORAMIC VIEW.

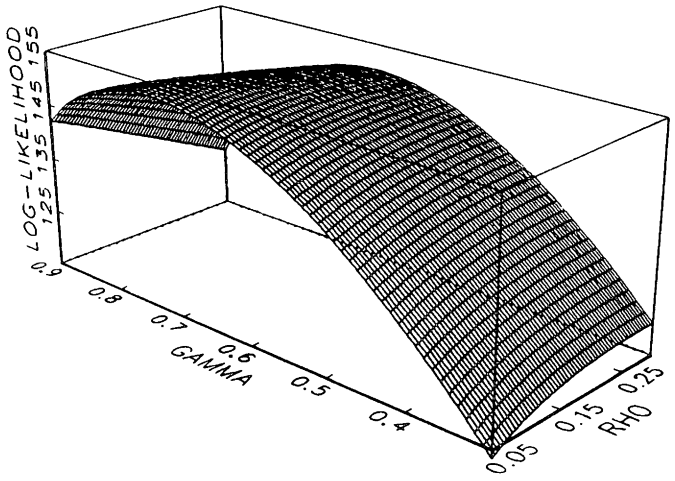

FIGURE 5.4: CONTOURS. UNCONDITIONAL LIKELIHOOD. FIRST DIFFERENCES. 22 COUNTRIES, RHO VS. GAMMA. SIGMA2 AND BETA FIXED. PANORAMIC VIEW.