TEL-AVIV

Working paper 26-82

# THE FOERDER INSTITUTE FOR ECONOMIC RESEARCH

TEL-AVIV UNIVERSITY

RAMAT AVIV ISRAEL

מכון למחקר כלכלי ע"ש ד"ר ישעיהו פורדר ז"ל

ע"י אוניברסיטת תל-אביב

# Model-Free Statistical Inference with Lorenz Curves, Income Shares, and Gini Coefficients

By:

Charles M. Beach
Queen's University
September 1979

Discussion Paper No. 360

The author wishes to thank A.B. Atkinson, F.A. Cowell, A.S. Goldberger, and A.F. Shorrocks for helpful discussion; the author, however, retains full responsibility for all shortcomings of the paper. The paper was written during visits to the London School of Economics and Political Science and the Institute for Research on Poverty (University of Wisconsin). Partial support was provided by the Social Sciences and Humanities Research Council of Canada.

## I. Introduction

One of the most frequently used devices to describe and compare distributional inequality in economics is the Lorenz curve. It has intuitive appeal and can be easily estimated. It is generally defined and not dependent on any prior specification of an underlying distribution function. It is the basis of a necessary and sufficient condition for ranking two distributions independent of utility functions (Atkinson (1970)).[1] It is also the basis for several summary measures of income (or wealth) inequality such as Gini concentration coefficient, perhaps the most frequently used single measure of inequality. Finally, the Lorenz curve also provides a disaggregated overview of the share structure of inequality in a distribution, so that one can see over which regions of a distribution inequality is relatively marked.

So far, however, Lorenz curves and income shares have been used essentially as descriptive devices and not as tools for rigorous statistical inference. This is at least in part due to the complexities of the sampling distributions associated with these devices, but is also partly due to a surprising lack of inquiry into the problem of formalizing statistical inference with Lorenz curves. Such a state of affairs is particularly troublesome in light of the massive outflow of recent empirical work using micro data to compare income and wealth inequality in different distributions, and of the current general interest in distributional considerations. This paper offers

[1] Other than that the utility functions be increasing and concave.

a solution to this problem by forwarding a new approach to distributional inference based on quantile analysis and the asymptotic distribution of sample income quantiles. Indeed, it will be shown that statistical inferences with Lorenz curves, income shares, and Gini coefficients are (asymptotically) _distribution-free_ or _model-free_ in the sense of not requiring knowledge of the underlying distribution model or parent distribution of the sample.

So far, statistical inference and confidence intervals have been worked out only for a few summary inequality measures (Gastwirth (1974) and Kakwani (1974)). But such measures frequently hide much interesting distributional detail, and contain implicit value norms that may not be adequately recognized or generally acceptable. The present paper is written in the spirit of these studies, but extends the analysis to _disaggregated_ inequality levels so as to permit a much richer and more detailed understanding of the structure of inequality in a distribution. As a useful corallary, the analysis also provides for inferences and standard errors of the Gini coefficient as well.

This paper focuses on the problem of disaggregated statistical inference, and for convenience and clarity we will assume to be working with samples of micro data. The approach thus contrasts with that of Gastwirth (1972) and Gastwirth and Glauberman (1976) who focus on interpolation methods for estimation of Lorenz curves and thus on "interpolation error" as opposed to "sampling error". In contrast to Gastwirth (1974) and Kakwani (1974), the present approach is disaggregative in orientation and leads to model-free inferences -- unlike maximum likelihood procedures, for example. And, in contrast to Kakwani and Podder (1973, 1976) and Thurow (1970), the current approach does not require any curve-fitting or iterative nonlinear estimation techniques in order to carry out inferences on Lorenz curves and income shares. The approach also avoids the need to fit specific distribution models or density function to empirical distributions in order to extract the relevant inequality information from the data -- again in contrast to analyses, for example, by Aigner and Goldberger (1970) and Kloek and van Dijk (1977, 1978). The present work,however, can be seen as an extension of the model-free approach of Beach (1977) of basing distributional analysis on a set of income quantiles, so that the overall structure of inequality in a distribution can be studied without the need of fitting specific functional forms.

The objectives of the paper are thus (i) to draw economists' attention to a body of statistical theory on sample quantiles that can be usefully exploited in distributional analysis; and (ii) to provide model-free inference techniques to Lorenz curves, income shares, and Gini coefficients.

The paper proceeds as follows. The next section introduces income quantiles and reviews some of the basic sampling theory to be used. Sections III and IV apply the theory to derive asymptotic distributions of Lorenz curve ordinates, income shares, and Gini coefficients. Sections V and VI then illustrate various inference procedures, and a few general comments are provided in the brief concluding section.

## II. Review of Sampling Distributions of Income Quantiles

### II.1) Lorenz Curves and Quantiles

In order to define a Lorenz curve conveniently, let f(y) be the (continuous) parent density function of income recipients. Then the proportion of recipients with incomes up to $y$ is the (cumulative) distribution function (or c.d.f.)

$$F(y) = \int_{-\infty}^{y} f(u)\, du \qquad (2.1)$$

and the proportion of total income receipts in the distribution by recipients with incomes up to $y$ is the incomplete (first) moment function

$$\Phi(y) = \frac{1}{\mu} \int_{-\infty}^{y} u\, f(u) du \qquad (2.2)$$

where the mean income level, $\mu$, is assumed to exist. Then just as the Lorenz curve abscissa $F(y)$ varies from 0 to 1, the Lorenz curve ordinate $\Phi(y)$ also varies from 0 to 1 monotonically where we assume for convenience that all incomes are positive. The so-called curve of concentration or Lorenz curve is the function $\Phi(F)$ defined parametrically in terms of income levels $y$ by (2.1) and (2.2).[2]

An income quantile $\xi_p$ corresponding to an abscissa value $p(0 \leq p \leq 1)$ on a Lorenz curve is defined implicitly by $p = \int_0^{\xi_p} f(u)du$ or $F(\xi_p) = p$ where $F(y)$ is assumed to be strictly monotonic. For example, the first decile level is $\xi_{.1}$ such that $.1 = \int_0^{\xi_{.1}} f(u)du$, and the median income level is $\xi_{.5}$ such that $.5 = \int_0^{\xi_{.5}} f(u)du$, so that half the recipients have incomes less than or equal to $\xi_{.5}$ and half have more.[3] Thus, corresponding to a set of K abscissas

[2] For an explicit definition of $\Phi$ in terms of F, see Gastwirth (1971) and Dorfman (1979).

[3] It may be of interest to remark that concern with income quantiles has also recently developed in the theoretical literature on measuring economic inequality as well (Sen (1973), p. 31; Donaldson and Weymark (1979)).

$p_1 < p_2 < \ldots < p_K$, we have a set of K population income quantiles $\xi_{p_1} < \xi_{p_2} < \ldots < \xi_{p_K}$. Note that the $\xi_{p_i}$'s are not in general parameters of a distribution, but simply distribution characteristics which we seek to estimate by sample statistics. Consequently, while quantile procedures are "nonparametric", they are not necessarily "distribution-free" (Bradley (1968) p. 15). Note also that the quantile abscissas, $p_i$, need not necessarily be equally spaced. We shall assume for convenience in this paper that they are (e.g., that the $\xi_{p_i}$'s are all deciles, centiles, or quartiles, say). But if one were particularly interested in upper and lower shares, for example, one might choose closer quantiles over those regions than elsewhere in the distribution.

### II.2) Exact Distributions of Order Statistics

Consider a random sample of N observations drawn from the probability density model f(y) with corresponding c.d.f. F(y), and order the observations from the smallest to the largest. Then $Y_\ell$ in the ordered sample represents the $\ell$'th smallest observation where $1 \leq \ell \leq N$. The probability that $(\ell-1)$ of the sample observations fell below a value $y_\ell$, one falls in the range $y_\ell \pm \frac{1}{2} dy_\ell$, and the remaining $(N-\ell)$ fall above $y_\ell$ is then given by (Kendall and Stuart (1969), pp. 236,252; Wilks (1962), p. 236) the probability element

$$dG(y_\ell) = \frac{N!}{(N-\ell)!(\ell-1)!} [F(y_\ell)]^{\ell-1}[1-F(y_\ell)]^{N-\ell} f(y_\ell)\, dy_\ell. \qquad (2.3)$$

The corresponding mean and variance of the $\ell$'th order-statistic, $Y_\ell$, are thus given by (Sarhan and Greenberg (1962), p. 13)

$$E(Y_\ell) = \frac{N!}{(N-\ell)!(\ell-1)!} \int_0^\infty u[F(u)]^{\ell-1} [1-F(u)]^{N-\ell} f(u)du$$

and

$$V(Y_\ell) = E(Y_\ell^2) - E(Y_\ell)^2$$

$$= \frac{N!}{(N-\ell)!(\ell-1)!} \; [\int_0^\infty u^2 \, [F(u)]^{\ell-1}[1-F(u)]^{N-\ell}f(u) \; du$$

$$- \{\int_0^\infty u[F(u)]^{\ell-1} \, [1-F(u)]^{N-\ell}f(u)du\}^2].$$

From these expressions it can be readily seen that exact sampling distributions for order-statistics have two important characteristics. First, the observations in an <u>ordered</u> sample will no longer be independent[4] or identically distributed even when the original sample observations were. Second, the exact sampling distributionsof order statistics are relatively complicated to handle analytically and depend very directly upon the underlying parent density model f(y), so that exact inferences about the parent quantiles $\xi_{p_i}$ based on such order-statistics are not distribution-free or "model-free".[5]

[4] Corresponding joint distributions and covariances for any two order statistics $Y_\ell$ and $Y_k$ can also be found in Sarhan and Greenberg (1962), p. 13; Wilks (1962), p. 236; and Kendall and Stuart (1969), pp. 270,325.

[5] It is worth noting, however, that pairs of order-statistics can be used to set distribution-free <u>confidence intervals</u> for population quantiles. In particular, it can be easily shown that, if $F(y_\ell) \leq p \leq F(y_k)$,

$$\text{Prob}(Y_\ell \leq \xi_p \leq Y_k) = \Sigma_{j=\ell}^{k-1} \binom{N}{j} p^j \; (1-p)^{N-j}$$

for order-statistics $Y_\ell$, $Y_k$ (Wilks (1962), pp. 330-331; Kendall and Stuart (1969), pp. 517-'8). However, as we shall want to work with functions or transforms of sample quantiles and obtain smooth confidence bands for the set of transformed quantiles, we shall deal directly with their sampling distribution functions and not just with confidence intervals for conveniently selected order-statistics.

### II.3 Asymptotic Distributions of Sample Quantiles

An asymptotic approximation to the distribution of sample quantiles, however, does provide the basis for distribution-free inference for sample shares and Lorenz curve ordinates. Given a random sample of N observations,[6] define an estimate of the population quantile $\xi_p$ to be

$$\begin{aligned} \hat{\xi}_p &= Y_{Np} \text{ if Np is an integer} \\ &= Y_{[Np]+1} \text{ if Np is not an integer} \end{aligned} \qquad (2.4)$$

where [Np] denotes the greatest integer not exceeding Np. These corresponding sample quantiles are known to have several useful statistical properties.

In particular, it can be shown that, if F(y) is strictly monotonic, $\hat{\xi}_p$ defined in (2.4) has the property of strong or almost sure consistency (Rao (1965), p. 355); that is, $\lim_{N\to\infty} \hat{\xi}_p = \xi_p$ with probability one, so that a fortiori it is weakly consistent as well. In addition, the $\hat{\xi}_{p_i}$'s are also asymptotically normal with a relatively simple covariance structure. More formally, we state this result (without proof) as the basic corner stone of this paper.

[6] Since this paper is concerned essentially with statistical inference and not estimation, it is assumed throughout that the analyst has access to actual micro data. If, however, he does not and the distribution data are available only in interval or histogram form, then standard interpolation procedures must be employed to obtain estimates of quantile income levels and income shares (e.g., Gastwirth (1972)). In this case, interpolation errors occur in addition to sampling errors in estimating the $\xi_{p_i}$ and in computing asymptotic standard errors.

Theorem 1:

Suppose that, for the set of proportions $\{p_i\}$ such that $0 < p_1 < p_2 < \ldots < p_K < 1$, $\hat{\xi} = (\hat{\xi}_{p_1}, \hat{\xi}_{p_2}, \ldots, \hat{\xi}_{p_K})'$ is a vector of K sample quantiles from a random sample of size N drawn from a continuous population density f(y) such that the $\xi_{p_i}$'s are uniquely defined and $f_i \equiv f(\xi_{p_i}) > 0$ for all $i = 1, \ldots, K$. Then the vector $\sqrt{N}\ (\hat{\xi}-\xi)$ converges in distribution to a K-variate normal distribution with mean zero and covariance matrix $\Lambda$. That is, $\hat{\xi}$ is asymptotically normal with mean vector $\xi = (\xi_{p_1}, \xi_{p_2}, \ldots, \xi_{p_K})'$ and asymptotic covariance matrix $(1/N)\ \Lambda$ where

$$\Lambda = \begin{bmatrix} \frac{p_1(1-p_1)}{f_1^2} & \cdots & \frac{p_1(1-p_K)}{f_1 f_K} \\ \vdots & \ddots & \vdots \\ \frac{p_1(1-p_K)}{f_1 f_K} & \cdots & \frac{p_K(1-p_K)}{f_K^2} \end{bmatrix} \qquad (2.5a)$$

If P denotes the matrix

$$P = \begin{bmatrix} p_1 & p_1 & \cdots & p_1 \\ p_1 & p_2 & \cdots & p_2 \\ \vdots & \vdots & \ddots & \vdots \\ p_1 & p_2 & \cdots & p_K \end{bmatrix},$$

$p' = (p_1, \ldots, p_K)$, and $F = \text{Diag}\ [f_1, \ldots, f_K]$, $\Lambda$ can be expressed in matrix form as

$$\Lambda = F^{-1}\ [P - pp']F^{-1} \qquad (2.5b)$$

Proofs of Theorem 1 can be found, for example, in Wilks (1962), pp. 273-'4, and Kendall and Stuart (1969), pp. 237-'9.[7] Since $\hat{\xi}_{p_i}$ is a consistent estimate of $\xi_{p_i}$, one can of course calculate a consistent asymptotic standard error of $\hat{\xi}_{p_i}$ as $[p_i(1-p_i)/Nf(\hat{\xi}_{p_i})^2]^{\frac{1}{2}}$.

It is important to note, however, that asymptotic inference on quantile income levels still requires knowledge of the underlying distribution model $f(\cdot)$ in computing the standard errors. It is thus desirable to work with transforms of these quantile variables which will allow model-free inferences. We now make use of Theorem 1 in deriving asymptotic distributions of sample share statistics and Lorenz curve ordinates.

## III. Income Shares and Lorenz Curves

### III.1) Asymptotic Distribution of Lorenz Curve Ordinates

To estimate Lorenz curve ordinates, recall first of all from (2.2) that

$$\Phi(\xi_{p_i}) = \frac{1}{\mu}\int_0^{\xi_{p_i}} uf(u)du \quad = \quad \frac{F(\xi_{p_i})}{\mu}\int_0^{\xi_{p_i}} \frac{uf(u)du}{F(\xi_{p_i})}$$

$$= p_i \cdot \frac{E(Y|Y \leq \xi_{p_i})}{E(Y)} \quad = \quad \frac{\tau_i}{\mu} .$$

Consequently, the sample estimate of $\Phi(\xi_i)$ may be computed as

$$\hat{\Phi}_i = \sum_{Y_j \leq \hat{\xi}_{p_i}} Y_j \Big/ \sum_{j=1}^{N} Y_j \doteq p_i \left(\frac{\bar{Y}_{\hat{\xi}_{p_i}}}{\bar{Y}}\right), \; i = 1, \ldots, K \qquad (3.1)$$

where $\bar{Y}_{\hat{\xi}_{p_i}} = \Sigma_{Y_j \leq \hat{\xi}_{p_i}} Y_j/n_i$ and $n_i = [Np_i]$. This will be referred to as the feasible or sample estimator of $\Phi(\xi_{p_i})$.

[7] Stronger and broader results than Theorem 1 can also be found in Chernoff, Gastwirth, and Johns (1967), pp. 56,58; and Bahadur (1966).

It will also be convenient to define the population income share function evaluated at the sample quantile estimate as

$$\Phi(\hat{\xi}_{p_i}) = \frac{1}{\mu} \int_0^{\xi_{p_i}} uf(u)du. \qquad (3.2)$$

While a random variable since it depends upon $\hat{\xi}_{p_i}$, it is also clearly dependent on the (unknown) population distribution function. This will be referred to as the infeasible estimator of $\Phi(\xi_{p_i})$. A Lorenz curve in this paper is represented by a set of K ordinates $\{\Phi(\xi_{p_i})\}$ which are to be estimated from the sample. The line of argument of this section involves, first, establishing the asymptotic distribution of the infeasible estimators $\Phi(\hat{\xi}_{p_i})$ for i=1,...,K as transforms of the sample quantiles (Lemma 1); then arguing that $\hat{\Phi}_i$ and $\Phi(\hat{\xi}_{p_i})$ have the same limiting distribution (Lemma 2); and thence concluding that the asymptotic distribution of the feasible estimators $\hat{\Phi}_i$, i=1,...,K, is exactly that derived for the $\Phi(\hat{\xi}_{p_i})$'s.

In order to derive the asymptotic distribution of a set of Lorenz curve ordinates $\{\Phi(\hat{\xi}_{p_i})\}$, it is useful first of all to recall the following result (Rao(1965), p. 321)) on limiting distributions of continuous functions of random variables. Suppose that $T_N$ is a K-dimensional statistic $(t_{1N}, t_{2N}, \ldots, t_{KN})'$ and $\theta = (\theta_1, \ldots, \theta_K)'$ a corresponding vector of constants such that the limiting distribution of the scaled vector $\sqrt{N}(T_N - \theta)$ is a K-variate normal with mean zero and covariance matrix $\Sigma$. Suppose also that a scalar function of the statistic vector $T_n$, $g(T_N)$, is totally differentiable. Then it follows that the limiting distribution of $\sqrt{N}(g(T_N) - g(\theta))$ is also normal with mean zero and variance $v = j' \Sigma j$ where

$$j = \left(\frac{\partial g(T_N)}{\partial t_{1N}}, \ldots, \frac{\partial g(T_N)}{\partial t_{KN}}\right)'\Bigg|_\theta$$

is the gradient vector of $g(\cdot)$ evaluated at $\theta$. More generally, if $g = (g_1(T_N), \ldots, g_M(T_N))'$ is an M-dimensional vector-valued function with each $g_i$ a function of the statistic vector $T_N$ and each $g_i$ is again totally differentiable, the M-dimensional vector $\sqrt{N}(g(T_N) - g(\theta))$ has an M-variate normal limiting distribution with zero mean and (M x M) covariance matrix $V = J \Sigma J'$ where

$$J = [J_{ij}] = \left[\frac{\partial g_i(T_N)}{\partial t_{jN}}\right]\Bigg|_\theta$$

is now an (M x K) matrix in which the i'th row contains the gradient of $g_i$ again evaluated at $\theta$.

In order to apply these results to the present situation, let $g_i$, $i = 1, \ldots, K$, be the incomplete (first) moment function $\Phi(y)$ defined in (2.2). The gradient of the function (2.2) evaluated at the population value $\xi_{p_i}$ can be seen to be simply $\xi_{p_i} f(\xi_{p_i})/\mu = (1/\mu)\xi_{p_i} f_i$. Consequently, setting $T_N = (\hat{\xi}_{p_i}, \ldots, \hat{\xi}_{p_K})'$, $\theta = (\xi_{p_1}, \ldots, \xi_{p_K})'$, $g(T_N) = (\Phi(\hat{\xi}_{p_1}), \ldots, \Phi(\hat{\xi}_{p_K}))'$, and $\Sigma = \Lambda$, we note that $J_L = \text{Diag}\,[(1/\mu)\xi_{p_1} f_1, \ldots, (1/\mu)\xi_{p_K} f_K]$, so that the variance of the limiting distribution corresponding to V in the case of Lorenz curve ordinates is

$$V_L = \begin{bmatrix} \left(\frac{\xi_{p_1}}{\mu}\right)^2 p_1(1-p_1) & \cdot & \cdot & \cdot & \left(\frac{\xi_{p_1}\xi_{p_K}}{\mu^2}\right) p_1(1-p_K) \\ \vdots & \cdot & \cdot & \cdot & \vdots \\ \left(\frac{\xi_{p_1}\xi_{p_K}}{\mu^2}\right) p_1(1-p_K) & \cdot & \cdot & \cdot & \left(\frac{\xi_{p_K}}{\mu}\right)^2 p_K(1-p_K) \end{bmatrix} \quad (3.3a)$$

$$= R[P - pp']\,R \quad (3.3b)$$

where R = Diag $[\xi_{p_1}/\mu, \ldots, \xi_{p_K}/\mu]$. We thus have the result

Lemma 1: Under the conditions of Theorem 1, the (scaled) vector of infeasible Lorenz curve ordinate estimates with elements $\sqrt{N}(\Phi(\hat{\xi}_{p_i}) - \Phi_i)$ calculated from (3.2) is asymptotically K-variate normal with mean zero and covariance matrix $V_L$ given in (3.3). Consequently, the (infeasible) Lorenz curve ordinates $\Phi(\hat{\xi}_{p_i})$ are asymptotically joint normal with mean $\Phi_i = \Phi(\xi_{p_i})$ and asymptotic covariance matrix $(1/N)V_L$.

So far, however, we have established the asymptotic distribution only of an infeasible set of estimators $\{\Phi(\hat{\xi}_{p_i})\}$ of the Lorenz curve ordinates. What are calculated from the sample are the feasible or sample estimates $\{\hat{\Phi}_i\}$ defined in (3.1). However, analogous to the results for Aitken generalized - least - squares estimators in econometrics, the feasible and infeasible estimators can be shown to be asymptotically equivalently distributed.

Lemma 2: Under the conditions of Theorem 1, if the population density has finite mean and variance, $\sqrt{N}(\hat{\Phi}_i - \Phi_i)$ and $\sqrt{N}(\Phi(\hat{\xi}_{p_i}) - \Phi_i)$ have the same limiting distributions. Proof of this result is based on a modification of Theorem 1 in Gastwirth (1974) and is provided in the Appendix. Basically, the argument involves showing that the conditional and unconditional means, $\bar{Y}_{\hat{\xi}_{p_i}}$ and $\bar{Y}$, in (3.1) are both asymptotically normal with appropriate means and variances inspite of the fact that $\bar{Y}_{\hat{\xi}_{p_i}}$ is stochastically conditioned.

Combining Lemmas 1 and 2, one now has the principal result of this paper.

Theorem 2: Under the conditions of Lemma 2, the vector of sample estimators $\hat{\Phi} = (\hat{\Phi}_i, \ldots, \hat{\Phi}_K)$ of Lorenz curve ordinates is asymptotically normal in that $\sqrt{N}(\hat{\Phi} - \Phi)$ has a limiting K-variate normal distribution with mean zero and covariance matrix $V_L$ specified in (3.3).

Consequently, asymptotic standard errors for the sample estimates $\hat{\Phi}_i$ are given by

$$\sqrt{\frac{v_{ii}^L}{N}} = \left(\frac{\hat{\xi}_{p_i}}{\hat{\mu}}\right) \sqrt{\frac{p_i(1-p_i)}{N}} \quad \text{for } i = 1, \ldots, K. \tag{3.4}$$

The important thing to note about $V_L$, of course, is that, in contrast to $\Lambda$, it does not require knowledge of the underlying model density function $f(\cdot)$. It depends solely upon the chosen proportions $p_i$, the population mean $\mu$, and the population quantile income levels $\xi_{p_i}$ which can be estimated consistently from the sample. Thus statistical inferences about the Lorenz curve ordinates can be carried out without having to know or estimate the underlying model or parent density function. It is in this sense that we say that Lorenz curve inferences are model-free. It is perhaps interesting to remark that this distribution-free aspect of Lorenz curve inference in the statistical field usefully complements Atkinson's (1970) Lorenz curve criterion in the field of welfare economics for making distributional inferences independent of the exact form of underlying utility functions as well. Consequently, one has further reason to be interested in using Lorenz curve analysis in applied distribution work.

It is worth noting that the present result implies that it is unnecessary for Lorenz curve inference to fit functional forms to empirical Lorenz curves as suggested,for example, by Kakwani and Podder (1973, 1976) and Thurow (1970). It also implies that, to make Lorenz curve inferences, it is unnecessary as well to fit various density functions to empirical distributions such as done in Aigner and Goldberger (1970) and in Kloek and Dijk (1977,1978).

In addition, it suggests that, along with (cumulative) income shares and means, it is useful in applied work and published data also to provide estimates of income quantiles. Indeed, the only new information that will be required to compute standard errors and various test statistics for Lorenz curves is a set of income quantiles, $\{\hat{\xi}_{p_i}\}$.

Furthermore, note from (2.1) and (2.2) that the derivative of the population Lorenz curve,

$$\frac{d\Phi}{dF} = \frac{d\Phi(y)/dy}{dF/dy} = \frac{(y/\mu)f(y)}{f(y)}$$

$$= y/\mu, \qquad (3.4)$$

is the so-called relative-mean-income curve (Kendall and Stuart (1969), p. 49; Levine and Singer (1970)) which has a number of useful inequality properties in its own right. Corresponding to the abscissa points $p_1$, $p_2$, ..., $p_K$, the relative-mean-income curve ordinates are thus $\xi_{p_1}/\mu$, $\xi_{p_2}/\mu$, ..., $\xi_{p_K}/\mu$.[8] It

[8] As an illustration of a relative-mean-income curve, consider the Pareto distribution with $F(y) = 1 - y^{-\alpha}$ and $\alpha > 1$. Then $\mu = \alpha/(\alpha-1)$, and $\xi_{p_i} = (1-p_i)^{-1/\alpha}$, so that the relative-mean-income-curve ordinates are $\xi_{p_i}/\mu = (\alpha-1/\alpha)(1-p_i)^{-1/\alpha}$. Thus for selected upper-tail values of $p_i$ and alternative values of $\alpha$, the corresponding relative-mean-income ordinates are easily computed.

| | $p_i$ = | .7 | .8 | .9 | .95 |
|---|---|---|---|---|---|
| $\alpha$ = 1.5 | - | .7438 | .9746 | 1.5474 | 2.4562 |
| 2.0 | - | .9129 | 1.1181 | 1.5813 | 2.2364 |
| 2.5 | - | .9712 | 1.1422 | 1.5072 | 1.9887 |
| 3.0 | - | .9958 | 1.1400 | 1.4362 | 1.8097 |
| 4.0 | - | 1.0134 | 1.1216 | 1.3338 | 1.5860 |

can be seen, then, that the elements of covariance matrix (3.3) are simply the products of selected proportions and their corresponding Lorenz curve derivatives.[9] Consequently, an alternative way of saying that it is useful for an applied distribution analyst to provide a set of income quantiles to go with an estimated Lorenz curve is that he should provide an estimated relative-mean-income curve as well, as done, for example, in some work of Beach _et_. _al_. (1980). A relative-mean-income curve thus has an important inferencial role in applied work as well as a useful descriptive role in distribution analysis.

Remark may also be made of the relatively simple structure of the asymptotic covariance matrix in (33). For positive incomes, $V_L$ has all positive elements; that is, between _cumulative_ income shares, covariances are quite reasonably positive. As one moves down the principal diagonal of terms $(\xi_{p_i}/\mu)^2$ $p_i(1-p_i)$, the component $p_i(1-p_i)$ increases to a maximum at the median value $p_i$ = .5 and then decreases, while the square of the relative-mean-income value increases steadily from $(\xi_{p_i}/\mu)^2$ to $(\xi_{p_K}/\mu)^2$. Thus the variances increase over the range $p_i$ to beyond the median and then may either increase or decrease depending on which effect dominates.[10] Typically, for skewed distributions of

[9] This should not be at all surprising since we know that (i) the proportions $F(\xi_{p_i})$ and $F(\hat{\xi}_{p_j})$ for $i < j$ are asymptotically normal with asymptotic covariance $p_i(1-p_j)/N$ (Wilks (1962), p.271), and that (ii) the derivative of the function $\Phi(F(\xi_{p_i}))$ is $d\Phi(\xi_{p_i})/dF = \xi_{p_i}/\mu$. Consequently, the income share functions $\Phi(F(\hat{\xi}_{p_i}))$ and $\Phi(F(\hat{\xi}_{p_j}))$ are also asymptotically normal with asymptotic covariance $(\xi_{p_i}/\mu)$ $(\xi_{p_j}/\mu)$ $p_i$ $(1-p_j)/N$.

[10] In the case of the Pareto distribution with $F(y) = 1 - y^{-\alpha}$ for $\alpha > 1$, the asymptotic variance is

income or wealth, the estimated variances have been found to reach a maximum in the interval between p = .70 and p = .85 and thereafter decline. Also note that the asymptotic squared correlation coefficient between cumulative shares corresponding to $p_i$ and $p_j$ ($p_i < p_j$) is $p_i(1-p_j)/p_j(1-p_i)$. That is, the correlations are independent even of the quantile levels and depend solely on the (known) abscissa proportions $p_i$, $p_j$. As one moves along the minor diagonal of $V_L$ where $p_i + p_j = 1$, the correlation is maximized at the median when i = j and minimized at the two ends of the diagonal where asy. $\mathrm{cor}^2(\hat{\Phi}_i, \hat{\Phi}_j) = p_1^2/p_K^2$.

### III.2 Asymptotic Distribution of Income Shares

The line of argument to derive the asymptotic distribution of Lorenz curve ordinates holds also for a set of income shares. If the Lorenz curve ordinates represent cumulative income shares, the differences between successive ordinates corresponding to adjacent quantiles represent income shares between different quantiles. If there are K quantiles (e.g., K = 9 in the case of deciles), then there are K + 1 (population) quantile shares

$$\psi_i = \Phi(\xi_{p_i}) - \Phi(\xi_{p_{i-1}}) \qquad i = 1,2, \ldots, K+1 \qquad (3.5)$$

where we set $\Phi(\xi_{p_0}) = 0$ and $\Phi(\xi_{p_{K+1}}) = 1$. Since $\hat{\psi}_i = \hat{\Phi}_i - \hat{\Phi}_{i-1}$ is just a

$$\frac{v_{ii}}{N} = (\frac{1}{N}) (\frac{\alpha-1}{\alpha})^2 p_i(1-p_i)^{\frac{\alpha-2}{\alpha}}.$$

For given N and $\alpha$, this is maximized at

$$p^* = \frac{1}{2}(\frac{\alpha}{\alpha-1}).$$

Consequently, when $\alpha$ = 2, 2.5, and 3, p* = 1.0, .8333, and .75 respectively.

difference in sample Lorenz curve ordinates which are asymptotically normal with asymptotic covariance matrix $(1/N)V_L$, it is clear that the sample income share statistics are also asymptotically (K+1) - variate normal with asymptotic mean $\psi = (\psi_1, \psi_2, \ldots, \psi_{K+1})'$ and asymptotic covariance matrix $(1/N)V_s$ where $V_s = J_s\, V_L\, J_s'$ and the (K+1) x K gradient matrix

$$J_s = [\frac{\partial\psi_i}{\partial\Phi_j}] = \begin{bmatrix} 1 & & & & \\ -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & \ddots & \\ & & & -1 & 1 \\ & & & & -1 \end{bmatrix} \qquad (3.6)$$

Thus combining (3.3) and (3.6), one can check that the ij'th element of the symmetric matrix $V_s$ where $1 \leq i \leq j \leq K+1$ is equal to

$$\begin{aligned} v_{ij}^s = (1/\mu^2)[&\xi_{p_{i-1}}\xi_{p_{j-1}}(1-p_{j-1}) - \xi_{p_i}\,\xi_{p_{j-1}}\,p_i\,(1-p_{j-1}) \\ &- \xi_{p_{i-1}}\,\xi_{p_j}\,p_{i-1}\,(1-p_j) + \xi_{p_i}\,\xi_{p_j}\,p_i(1-p_j)] \qquad (3.7) \end{aligned}$$

where $p_0 = 0$, $p_{K+1} = 1$, $\xi_{p_0} = 0$, and $\xi_{p_{K+1}}$ is assumed finite.

Again, it is evident that $V_s$ does not depend upon the underlying population density function $f(\cdot)$, so that model-free inferences concerning income shares are again feasible. Note also that, in contrast to $V_L$, $V_s$ is of dimension (K+1) x (K+1) and singular since the sum of the K+1 income shares is identically one.

In order to compute (asymptotic) standard errors for income shares, one simplifies (3.7) by setting i = j to

$$v_{ii}^s = (1/\mu^2)\ [\xi_{p_{i-1}}^2\ p_{i-1}(1-p_{i-1}) - 2\xi_{p_i}\xi_{p_{i-1}}\ p_{i-1}(1-p_i) + \xi_{p_i}^2\ p_i(1-p_i)] \quad (3.8$$

It is then immediately evident from (3.8) that, to compute standard errors for income shares, one need compute only 2K-1 elements -- the K diagonal elements and K-1 first-superdiagonal elements -- of the $V_L$ matrix and not the full set of K(K-1)/2 different elements in $V_L$. The (asymptotic) standard error for the i'th income share $\hat{\psi}_i$ can thus be computed as

$$[(1/N\hat{\mu}^2)\ [\hat{\xi}^2_{p_{i-1}} p_{i-1}(1-p_{i-1}) - 2\ \hat{\xi}_{p_i}\ \hat{\xi}_{p_{i-1}}\ p_{i-1}(1-p_i) + \hat{\xi}^2_{p_i}\ p_i(1-p_i)]^{\frac{1}{2}} \qquad (3.9)$$

The asymptotic variances of bottom and top income shares are particularly easy to compute. The share statistic for the lowest 100 $p_i$% of the sample is simply $\hat{\psi}_i = \hat{\Phi}_i$ which has the (asymptotic) standard error $(\frac{\hat{\xi}_{p_i}}{\bar{Y}})\sqrt{\frac{p_i(1-p_i)}{N}}$. The share statistic for the top 100(1-$p_i$)% is $\hat{\psi}_i = 1 - \hat{\Phi}_i$, so that the corresponding (asymptotic) standard error is also $(\frac{\hat{\xi}_{p_i}}{\bar{Y}})\sqrt{\frac{p_i(1-p_i)}{N}}$.

## IV. Standard Errors for Gini Coefficients

A corollary of deriving the asymptotic distribution of sample Lorenz curve ordinates is that one can also do so for an interpolated approximation to the Gini coefficient, perhaps the single most frequently used summary measure of income inequality in a distribution. While Gastwirth (1974) and Kakwani (1974) have derived asymptotic distributions for estimates of various other summary inequality measures, this appears to be the first such derivation for the Gini coefficient. The approach again is model-free, and does not require *a prior* specification of the underlying parent distribution such as involved in maximum likelihood methods used by Kakwani (1974). The geometric approach used here also avoids the rather substantial difficulties of the perhaps more natural approach (Kendall and Stuart (1969), p. 241) of first examining the distribution of the mean absolute difference

$$\Delta = \int_0^\infty \int_0^\infty | y_1 - y_2 | dF(y_1) dF(y_2)$$

which appears in the numerator of the Gini coefficient.

The (population)Gini coefficient of concentration,$\Gamma$, lying in the interval (0,1) for positive incomes, is geometrically equal to twice the area between the Lorenz curve and the absolute equality diagonal (Kendall and Stuart (1969), p. 49). If one interpolates linearly along the Lorenz curve between adjacent quantile ordinates and uses a trapezoidal integration formula, the Gini coefficient[11] may be estimated as

$$\hat{\Gamma} = G = (1/K+1) \Sigma_{i=1}^{K+1} (p_i - \hat{\Phi}_i + p_{i-1} - \hat{\Phi}_{i-1}) \qquad (4.1)$$

[11] Note that this is the only point at which interpolation has been used in this paper. The expression for the estimated variance of G is thus approximate in that it reflects both sampling errors as well as interpolation errors. One could if one wished also use an alternative interpolation formula such as Gastwirth's (1972) "upper-bound" interpolation rule or some rule-of-thumb combination of the two.

if the $p_i$'s are equally spaced. Therefore the (Kx1) gradient vector for the linear transformation (4.1) is $j = (-2/(K+1), \ldots, -2/(K+1))'$, and one obtains from the results of Section III.1 that $\sqrt{N}(G-\Gamma)$ also has a limiting normal distribution with mean zero and variance

$$j'V_L\, j = (4/(K+1)^2)\Sigma_{i=1}^{K}\ \Sigma_{j=1}^{K}\ v_{ij}^{L}$$

where the summation is over all elements of the $V_L$ covariance matrix. The corresponding (asymptotic) standard error of G is thus

$$\text{S.E.}(G) = \frac{2}{(K+1)}\left[\frac{\Sigma_i\Sigma_j\hat{v}_{ij}^{L}}{N}\right]^{\frac{1}{2}} \qquad (4.2)$$

where $\hat{v}_{ij}^{L} = (\hat{\xi}_{p_i}/\bar{Y})(\hat{\xi}_{p_j}/\bar{Y})\ p_i\ (1-p_j)$ for $i < j$.

Since the Gini coefficient is expressed as a function of the Lorenz curve ordinates for given $p_i$'s, it too has the property of allowing model-free statistical inference. The relative mean deviation inequality statistic, in contrast, does not (Beach (1979)). However, the estimated coefficient and its standard error do depend on the coarseness of the interpolation intervals $[p_i, p_{i-1}]$, so that it is advisable when reporting inference results based on (4.1) and (4.2) to indicate also the interval size (e.g., deciles or quintiles) used in the interpolation.

## V. Hypothesis Testing with Quantile Results

### V.1) Hypothesis Tests on Income Shares

Given the asymptotic distribution results on estimated income shares derived in the last section, one is now able to consider directly the problem of hypothesis testing with income shares.

### i) Tests on Single Share Statistics

First of all, consider the case where there is some hypothesized value $\psi_i^0$ to which the sample share statistic, $\hat{\psi}_i$ is being compared (for example, that the bottom 10% of recipients get only 5% of total income). From the results of Section III.2, it is clear that the appropriate test statistic under $H_o$: $\psi_i = \psi_i^0$ is $z_i = (\hat{\psi}_i - \psi_i^0)/(\hat{v}^s_{ii}/N)^{\frac{1}{2}}$ which is to be compared to the critical values on a standard normal table for a specified level of significance $\alpha$.

More typically, however, the distribution analyst is more interested in comparing income shares between two alternative distributions (for example, between two time periods or two regions). Specifically, suppose one has two corresponding income share statistics $\hat{\psi}_{1i}$ and $\hat{\psi}_{2i}$ based respectively on two independent samples of sizes $N_1$ and $N_2$. According to a null hypothesis, $H_o$: $\psi_{1i} = \psi_{2i}$ against, say, $H_1$: $\psi_{1i} \neq \psi_{2i}$ for a given particular quantile share. Under the independence assumption, the appropriate standard normal test statistic becomes $z_2 = (\hat{\psi}_{1i} - \hat{\psi}_{2i})/[(\hat{v}^{s_1}_{ii}/N_1) + (\hat{v}^{s_2}_{ii}/N_2)]^{\frac{1}{2}}$ where $\hat{v}^{s_1}_{ii}$ and $v^{s_2}_{ii}$ are the estimated variances based on (3.8) for samples 1 and 2 respectively.

Tests on single share statistics such as just considered are most likely to be appropriate when looking at either top or bottom shares in a distribution.[12]

[12] It may be remarked that standard "t-ratios" typically reported for individual regression coefficients are not so interesting for estimated share statistics. Perhaps the more appropriate "standard" on which to base individual test statistics is the null hypothesis of absolute equality. Consequently, instead of reporting individual "t-ratios", $t = \hat{\psi}_i/\sqrt{\hat{v}^s_{ii}/N}$, it may be more appropriate to report individual "z-ratios", $z = (\hat{\psi}_i - p_i)/\sqrt{\hat{v}^s_{ii}/N}$.

### ii) Joint Test on a Set of Income Shares

When evaluating an overall distribution of income, one may be more concerned with a set of income shares. For purposes of exposition, suppose one is interested in the full set of K quantile share statistics (one share statistic, say the last, is omitted as being linearly dependent on the others). For example, one may have a model of income generating behaviour as in Fair (1971) and wish to compare an actual distribution of income shares, say $\hat{\psi} = (\hat{\psi}_1, \hat{\psi}_2, \ldots, \hat{\psi}_K)'$, to an hypothesized set of income shares $\psi^0 = (\psi_1^0, \psi_2^0, \ldots, \psi_K^0)'$ specified by the theoretical model. In this case, one wishes to test $H_0: \psi = \psi^0$ against the uninformative alternative $H_1: \psi \neq \psi^0$. From the results of Section III.2, under the null hypothesis, $\sqrt{N}(\hat{\psi} - \psi^0)$ is asymptotically distributed as a K-variate normal with mean zero and covariance matrix $\bar{V}_s$, where the bar notation on $V_s$ indicates that the last row and column of the $V_s$ matrix have been deleted. Consequently, the test statistic

$$c_1 = N(\hat{\psi} - \psi^0)' \, \hat{\bar{V}}_s^{-1}(\hat{\psi}-\psi^0) \qquad (5.1)$$

is asymptotical distributed under $H_0$ as a (central) chi-squared variate with K degrees of freedom.

It should be remarked, however, that the actual computations involved in the income share test (5.1) (and in subsequent tests as well) are much simpler than may first appear as there is no need to invert the matrix $\hat{\bar{V}}_s$ numerically. If the (KxK) nonsingular matrix $\bar{J}_s$ is defined as

$$\bar{J}_s = -\begin{bmatrix} 1 & & & & \\ -1 & 1 & & \text{\Huge 0} & \\ & -1 & 1 & & \\ & & \ddots & \ddots & \\ & \text{\Huge 0} & & & \\ & & & -1 & 1 \end{bmatrix} ,$$

it can be seen that the share covariance matrix

$$\bar{V}_s = \bar{J}_s V_L \bar{J}_s',$$

so that

$$(\bar{V}_s)^{-1} = (\bar{J}_s')^{-1} V_L^{-1} (\bar{J}_s)^{-1}$$

where it can also be checked that

$$(\bar{J}_s)^{-1} = \begin{bmatrix} 1 & & & 0 \\ 1 & 1 & & \\ \vdots & \vdots & \ddots & \\ 1 & 1 & \cdots & 1 \end{bmatrix}$$

Thus any arbitrary quadratic form in the matrix $(\bar{V}_s)^{-1}$ can be written as

$$a'(\bar{V}_s)^{-1}a = b'V_L^{-1}b \qquad (5.2)$$

where

$$b = (\bar{J}_s)^{-1}a = \begin{bmatrix} a_1 \\ a_1+a_2 \\ a_1+a_2+a_3 \\ \vdots \\ a_1+a_2+\cdots+a_K \end{bmatrix}, \qquad (5.3)$$

so that it becomes now a quadratic form in the matrix $V_L^{-1}$, the inverse of the Lorenz curve (asymptotic) covariance matrix.

$V_L$, however, can be shown to have a simple analytic inverse. Specifically, it will be recalled that $V_L$ = RAR where R is a diagonal matrix and A = P - pp' from (3.3). Now the matrix $A^{-1}$ can be seen to have a very simple structure, with elements

$$a^{ii} = \frac{p_{i+1} - p_{i-1}}{(p_{i+1}-p_i)(p_i-p_{i-1})} \qquad \text{for } i = 1, \ldots, K, \tag{5.4a}$$

$$a^{i,i+1} = a^{i+1,i} = \frac{-1}{(p_{i+1}-p_i)} \qquad \text{for } i = 1, \ldots, K-1, \tag{5.4b}$$

and zeros elsewhere (Mosteller (1946), p. 385). Again, for convenience, set $p_0 = 0$ and $p_{K+1} = 1$. Consequently, any quadratic form in the matrix $V_L^{-1}$ can be written as

$$b'V_L^{-1}b = \Sigma_{i=1}^K \frac{(p_{i+1}-p_{i-1})}{(p_{i+1}-p_i)(p_i-p_{i-1})} b_i^2 - 2\,\Sigma_{i=2}^K \frac{b_i b_{i-1}}{(p_i-p_{i-1})} \tag{5.5}$$

Thus one needs to compute only 2K-1 terms in (5.5) instead of inverting a (KxK) matrix numerically. When one is working with deciles or vigintiles, for examples, this is a substantial computational reduction. The test statistic in (5.1) can thus be re-expressed as

$$c_1 = N[\Sigma_{i=1}^K \frac{(p_{i+1}-p_{i-1})}{(p_{i+1}-p_i)(p_i-p_{i-1})} \left(\frac{\hat{\xi}_{p_i}}{\hat{\mu}}\right)^{-2} b_i^2$$

$$- 2\,\Sigma_{i=2}^K \frac{b_i\ b_{i-1}}{(p_i-p_{i-1})} \left(\frac{\hat{\xi}_{p_i}}{\hat{\mu}}\right)^{-1} \left(\frac{\hat{\xi}_{p_{i-1}}}{\hat{\mu}}\right)^{-1}] \tag{5.6a}$$

where $b_i = \Sigma_{j=1}^i (\hat{\psi}_j - \psi_j^0)$ . (5.6b)

Clearly, one could also work out an intermediate case where a test is performed on a set of only L quantile shares where $1 \leq L \leq K$ based on an asymptotic chi-squared statistic with L degrees of freedom.

### iii) Joint Test of a Difference of Two Independent Sets of Income Shares

When one is comparing alternative distributions, however, one may be more concerned with testing for differences in sets of share statistics between two sample distributions corresponding, for example, to different periods or different regions. Specificially, suppose one distribution is characterized by a set of K quantile shares $\hat{\psi}_1 = (\hat{\psi}_{11}, \hat{\psi}_{12}, \ldots, \hat{\psi}_{1K})'$ and the second by $\hat{\psi}_2 = (\hat{\psi}_{21}, \hat{\psi}_{22}, \ldots, \hat{\psi}_{2K})'$ and the samples are drawn independently of size $N_1$ and $N_2$ respectively. The null hypothesis one may wish to test then is

$$H_0: \ \psi_1 = \psi_2 \text{ against } H_1: \ \psi_1 \neq \psi_2.$$

Now the two share covariance matrices $V_{s1}$ and $V_{s2}$ can be seen to be equal if and only if $(\xi_{1p_i}/\mu_1) = (\xi_{2p_i}/\mu_2)$ for all i; that is, if the relative mean income curves are the same for the two distributions. But if the relative mean income curves are the same, so also are the corresponding Lorenz curves, and the corresponding sets of quantile share statistics. Consequently, under the null hypothesis that $\psi_1 = \psi_2$, we shall also assume that the two covariance matrices are equal, $V_{s1} = V_{s2} = V_s$.

Under the null hypothesis, then, one can see that the vector difference $(\hat{\psi}_1 - \hat{\psi}_2)$ is asymptotically K-variate normal with mean zero and covariance matrix $(1/N_1 + 1/N_2)\ \bar{V}_s$. Consequently, an appropriate test statistic for $H_0$ is

$$c_2 = \left(\frac{N_1 N_2}{N_1+N_2}\right) (\hat{\psi}_1 - \hat{\psi}_2)'\ \hat{\bar{V}}^{-1}\ (\hat{\psi}_1 - \hat{\psi}_2) \qquad (5.7)$$

which will also be asymptotically chi-squared with K degrees of freedom.[13]

[13] Since covariance matrices are assumed the same in the two samples, estimates of the elements of $\bar{V}_s$ should be based on a combined sample. A convenient approximation to the combined relative mean income ordinates, however, may be provided simply by the weighted average

$$\frac{\hat{\xi}_i}{\hat{\mu}} = \left(\frac{N_1}{N_1+N_2}\right) \left(\frac{\hat{\xi}_{1p_i}}{\hat{\mu}_1}\right) + \left(\frac{N_2}{N_1+N_2}\right) \left(\frac{\hat{\xi}_{2p_i}}{\hat{\mu}_2}\right).$$

Following the same argument presented for $c_1$, one can alternatively and more simply compute $c_2$ by the formula (5.6a) where now

$$b_i = \Sigma_{j=1}^{i} (\hat{\psi}_{1j} - \hat{\psi}_{2j}) \tag{5.7b}$$

Again one can also formulate joint tests for differences in subsets of quantile shares as well.

## V.2) Hypothesis Tests and Confidence Bands on Lorenz Curves

In the case of Lorenz curves, tests of individual ordinates are not typically of much concern, so that we consider only joint tests on the full set of K Lorenz curve ordinates analogous to those just discussed for income shares.

### i) Joint Tests on Lorenz Curve Ordinates

Since much of the framework for hypothesis testing of Lorenz curve ordinates has already been laid out, the present discussion can be fairly brief. To compare a hypothetical or theoretical Lorenz curve $\Phi^0 = (\Phi_1^0, \Phi_2^0, \ldots, \Phi_K^0)'$ against an empirically estimated curve $\hat{\Phi} = (\hat{\Phi}_1, \hat{\Phi}_2, \ldots, \hat{\Phi}_K)'$ in order to test $H_o$: $\Phi = \Phi^0$ vs $H_1$: $\Phi \neq \Phi^0$, one can again use an asymptotic chi-squared test statistic

$$c_3 = N(\hat{\Phi} - \Phi^0)' \hat{V}_L^{-1}(\hat{\Phi} - \Phi^0) \tag{5.8}$$

with K degrees of freedom. To compare two separate Lorenz curve estimates $\hat{\Phi}_1$ and $\hat{\Phi}_2$ from independent samples, in order to test $H_0$: $\Phi_1 = \Phi_2$ vs $H_1$: $\Phi_1 \neq \Phi_2$, one can use the statistic

$$c_4 = (\frac{N_1 N_2}{N_1 + N_2}) (\hat{\phi}_1 - \hat{\phi}_2)' \hat{V}_L^{-1} (\hat{\phi}_1 - \hat{\phi}_2) \tag{5.9}$$

which is also asymptotically chi-squared with K degrees of freedom under the null hypothesis and accompanying assumption of equal variances.

Just as the share test statistics can be computed without having to invert numerically a (KxK) covariance matrix, so also can $c_3$ and $c_4$. Specifically, using the result in (5.5), one can re-express (5.8) as

$$\begin{aligned} c_3 = N[\Sigma_{i=1}^{K} & \frac{(p_{i+1}-p_{i-1})}{(p_{i+1}-p_i)(p_i-p_{i-1})} \left(\frac{\hat{\xi}_{p_i}}{\hat{\mu}}\right)^{-2} (\hat{\Phi}_i - \Phi_i^0)^2 \\ & -2\Sigma_{i=2}^{K} \frac{1}{(p_i-p_{i-1})} \left(\frac{\hat{\xi}_{p_i}}{\hat{\mu}}\right)^{-1} \left(\frac{\hat{\xi}_{p_{i-1}}}{\hat{\mu}}\right)^{-1} (\hat{\Phi}_i - \Phi_i^0)(\hat{\Phi}_{i-1}-\Phi_{i-1}^0)] \end{aligned} \qquad (5.10)$$

and (5.9) as

$$\begin{aligned} c_4 = \left(\frac{N_1N_2}{N_1+N_2}\right)[\Sigma_{i=1}^{K} & \frac{(p_{i+1}-p_{i-1})}{(p_{i+1}-p_i)(p_i-p_{i-1})} \left(\frac{\hat{\xi}_{p_i}}{\hat{\mu}}\right)^{-2} (\hat{\Phi}_{1i} - \hat{\Phi}_{2i})^2 \qquad (5.11) \\ & -2\Sigma_{i=2}^{K} \frac{1}{(p_i-p_{i-1})} \left(\frac{\hat{\xi}_{p_i}}{\hat{\mu}}\right)^{-1} \left(\frac{\hat{\xi}_{p_{i-1}}}{\hat{\mu}}\right)^{-1} (\hat{\Phi}_{1i}-\hat{\Phi}_{2i})(\hat{\Phi}_{1i-1}-\hat{\Phi}_{2i-1})]. \end{aligned}$$

One particularly interesting problem where one may wish to apply the above inference procedures is that of statistically testing Atkinson's (1970) distributional ranking criterion involving Lorenz curves. Specifically, one may wish to use the criterion of noninteresting Lorenz curves to define a ranking or comparison of inequality between two distributions (as opposed to defining a ranking of distributions <u>per se</u>), as applied for example in Beach (1980). To test empirically the hypothesis of one Lorenz curve lying statistically significantly inside another, one may start from a situation of one estimated Lorenz curve $\hat{\Phi}_1$ indeed lying uniformly above another $\hat{\Phi}_2$ (i.e.: $\hat{\Phi}_{1i} > \hat{\Phi}_{2i}$ for all i = 1, ..., K), and then use statistic $c_4$ to test $H_0$: $\Phi_1 = \Phi_2$ against the one-sided alternative $H_1$: $\Phi_1 > \Phi_2$.

### ii) Confidence Band for Lorenz Curves

Along with the hypothesis tests so far described, it would be desirable from a graphical point of view to supplement an illustrated Lorenz curve with some kind of confidence band about it over its full length. One could then immediately see graphically how accurately the illustrated Lorenz curve has been estimated, and particularly over some regions more tightly than others.

Perhaps an initial approach to this problem might simply be to construct a band of, say, two standard errors of $\hat{\Phi}_i$ on both sides of the estimated Lorenz curve ordinates. While such a band may have some descriptive interest in illustrating the relative widths of individual ordinate confidence intervals, it is not a very useful analytical device because it treats individual ordinate estimates as separate and unrelated. What is wanted instead is a joint confidence band or set of simultaneous confidence intervals that incorporate the market interdependence of the individual ordinate estimates for Lorenz curves. As is well known in the statistical literature this is the classical problem of determining a set of simultaneous confidence intervals or multiple comparisons for a given joint level of confidence, and there is no unique way of handling the problem. Perhaps the best known approach for our purposes is Scheffé's (1959 , pp. 68-70) projection method. See the last reference or Wilks (1962, pp. 291-'3) for details. If $d_\alpha = \sqrt{\chi^2_K}$ is the square root of the 100(1-$\alpha$)% critical value on a chi-square distribution with K degrees of freedom, then the probability is at least 100(1-$\alpha$)% that the K intervals $(\hat{\phi}_i - d_\alpha\sqrt{v^L_{ii}}\ , \hat{\Phi}_i + d_\alpha\sqrt{v^L_{ii}})$

jointly contain the K population ordinates $\Phi_i$, $\Phi_2$, ..., $\Phi_K$. Consequently, an approximate set of simultaneous confidence intervals is provided by a band of $d_\alpha$ standard errors in width on both sides of the estimated Lorenz curve ordinates. In the case of decile ordinates (K=9) with $\alpha = .05$, the corresponding value of $d_\alpha$ is $d_\alpha = \sqrt{16.919} = 4.11$. This compares with the two-standard-errors rule that corresponds to treating the ordinates as separate and unrelated.

Alternative approaches to the simultaneous confidence interval problem are also available (Seber (1977), pp. 126-132). Bonferroni t-intervals, for example, are based on the critical value of $t_\upsilon^{\alpha/2K}$ for $d_\alpha$ from the t-distribution with $\upsilon$ degrees of freedom. Asymptotically, one may simply use $z^{\alpha/2K}$ from the standard normal distribution for large micro-data samples. In the above case where $\alpha = .05$ and $K = 9$, the Bonferroni critical value is $d_\alpha = 2.78$ which is substantially smaller than that obtained from the Scheffé procedure, and consequently in this case to be preferred.

## VI. Illustrative Empirical Results

Several of the tests of Section V are now illustrated with two sources of micro data, one for the United States from Danziger and Taussig (1978),[14] and one for Canada from Beach, et. al. (1980).

Table 1 provides the background data on decile income levels, decile shares, and Lorenz curve ordinates for United States census unit households (reporting positive income) from the CPS for the two years 1967 and 1976.

[14] The author would like to thank Prof. Sheldon Danziger for providing the data in Table 1.

These estimates are based on very large data sets ($N_1$ = 48, 191 for 1967 and $N_2$ = 58,063), and appear roughly similar except for the inflation of income values over the period resulting in the sample mean increasing from $7692 in 1967 to $14,087 in 1976.

Table II provides (asymptotic) standard errors on the decile income shares as computed by (3.9) (given in percents) for the two years, and also z-statistics on the difference of individual shares, $\hat{\psi}_{1i} - \hat{\psi}_{2i}$. Judging the shares separately, one can see the differences are individually significantly different from zero in the first, third, fourth, fifth, and eighth decile shares on convential significance levels. Note also how the standard errors are consistently slightly smaller for 1976 because of the larger sample size.

Table III provides more summary test statistics for differences in overall inequality between the two years. A joint test of the difference between the two Lorenz curves is computed from (5.11) to be $C_4$ = 54.46 which is seen to be highly significant at any conventional levels of significance. The Gini coefficient standard errors are also computed (based on deciles) and yield test statistics for significant difference from zero (i.e., absolute equality) of 163. and 183. for 1967 and 1976 respectively. However, the difference between the two Ginis has a z-ratio of only -2.091 which lies between a 95% and 99% confidence-level cut-off on the normal table with a two-tailed test. Thus it is quite clear that a test on Gini coefficients is not at all equivalent to a test on significant differences in the overall Lorenz curve. In the first place, one is a single test, while the other a joint test. Secondly, one has an assumed aggregation structure and implicit social welfare function built into it while the other does not. In the case of two

## TABLE I

### Decile Incomes, Shares, and Lorenz Curves: 1967,1976

1) United States CPS Households, 1967

| Decile | Decile Level | Decile Share | Lorenz C. Ord. |
|---|---|---|---|
| 1 | $1,441 | 1.00% | 1.00% |
| 2 | 2,700 | 2.66 | 3.66 |
| 3 | 4,056 | 4.38 | 8.04 |
| 4 | 5,457 | 6.24 | 14.28 |
| 5 | 6,750 | 7.95 | 22.23 |
| 6 | 8,000 | 9.59 | 31.82 |
| 7 | 9,504 | 11.34 | 43.16 |
| 8 | 11,390 | 13.55 | 56.71 |
| 9 | 14,500 | 16.57 | 73.28 |
| 10 | | 26.72 | 100.00 |
| | $G_1$ = .3992 | $\hat{\mu}_1$ = $7,692 | $N_1$ = 48,191 |

2) United States CPS Households, 1976

| Decile | Decile Level | Decile Share | Lorenz C. Ord. |
|---|---|---|---|
| 1 | $2,935 | 1.16% | 1.16% |
| 2 | 4,875 | 2.73 | 3.89 |
| 3 | 7,000 | 4.18 | 8.07 |
| 4 | 9,285 | 5.78 | 13.85 |
| 5 | 11,870 | 7.50 | 21.35 |
| 6 | 14,580 | 9.36 | 30.71 |
| 7 | 17,540 | 11.37 | 42.08 |
| 8 | 21,350 | 14.10 | 56.18 |
| 9 | 27,450 | 17.01 | 73.19 |
| 10 | | 26.80 | 100.00 |
| | $G_2$ = .4061 | $\hat{\mu}_2$ = $14,087 | $N_2$ = 58,063 |

Source: See footnote 14, and Danziger and Taussig (1978).

## TABLE II

### Decile Shares and Standard Errors: U.S. 1967 and 1976

| | 1967 | 1976 | z dif. |
|---|---|---|---|
| 1) | 1.00%<br>(0.026) | 1.16%<br>(0.026) | -4.40* |
| 2) | 2.66<br>(0.051) | 2.73<br>(0.045) | -1.04 |
| 3) | 4.38<br>(0.074) | 4.18<br>(0.036) | 2.06* |
| 4) | 6.24<br>(0.096) | 5.78<br>(0.081) | 3.66* |
| 5) | 7.95<br>(0.116) | 7.50<br>(0.101) | 2.93* |
| 6) | 9.59<br>(0.134) | 9.36<br>(0.121) | 1.27 |
| 7) | 11.34<br>(0.156) | 11.37<br>(0.143) | -0.14 |
| 8) | 13.55<br>(0.182) | 14.10<br>(0.168) | -2.20* |
| 9) | 16.57<br>(0.215) | 17.01<br>(0.202) | -1.49 |
| 10) | 26.72<br>(0.258) | 26.80<br>(0.243) | -0.23 |

*denotes significantly different from zero on the basis of atwo-tailed test of a standard normal variate with $\alpha = .05$.

Source: See Table I.

## TABLE III

### Summary Test Statistics:  US 1967 and 1976

| | 1967 | 1976 |
|---|---|---|
| Gini Coef. - | .3992<br>(.00245)z=162.9 | .4061<br>(.00222)z=182.9 |
| Lorenz Curve Difference: | $c_4 = 54.46 > \chi_9^2 = 23.59$ at $\alpha .005$ | |
| Gini Coef. Difference: | $d = G_{67} - G_{76} = -.0069$ | |
| | S.E.(d) = .00330 | |
| | zd = -2.091 | |
| | $(z(\alpha=.05) = 1.960,\ z(\alpha=.01) = 2.326)$. | |

Source:  See Table I.

intersecting Lorenz curves, for example, the corresponding Gini coefficients can be the same while the Lorenz curves are quite different. In general, the Lorenz curve joint test is to be preferred to that on the Gini coefficient as a less restrictive test.

It can also be seen that with such large sample sizes, even rather similar looking distributions can be quite sharply distinguished as to their relative structure of inequality. At the same time, the size of "sampling error" is on the low side relative to "interpolation error" as found, for example, by Gastwirth (1972) who computed interpolation error bounds on the Gini coefficients for the 1967 CPS data with 10 income groups. The width of the interval between upper and lower interpolation bounds for three different interpolation procedures was calculated as .020, .019, and .009. These may be compared to an approximate 95% confidence interval on G for 1967 of ±2 standard errors or an interval width of .009.

Finally, Table IV provides Lorenz curve data on family total income for all (census) family units in the province of Ontario, Canada, for 1973 taken from a recent empirical study by the author and others (Beach et al (1980)) and computed based on a vigintile (K+1=20) income disaggregation and a sample size of 7624 family units. This finer level of disaggregation shows the Lorenz curve standard errors increasing up until the sixteenth vigintile and then decreasing in size. The third column provides joint confidence intervals for the nineteen vigintile ordinates based on "Bonferroni-z" intervals. At a 95% level of confidence, the asymptotic Bonferroni-z value for $d_\alpha$ is 3.01 (Seber (1977), p. 131) compared to the corresponding asymptotic Scheffé value for $d_\alpha$ which would be $d_\alpha = \sqrt{\chi^2_{19}} = 5.49$. Consequently, the narrower Bonferroni intervals have been used in the table.

## TABLE IV

## Lorenz Curve Vigintile Ordinates<br>Family Total Income for All Family Units<br>Ontario, 1973

| Vig. | Pt. Est. | Est.±3.01 S.E. | S.E. |
|---|---|---|---|
| 1 | 0.39% | 0.29 - 0.49% | 0.034% |
| 2 | 1.23 | 1.03 - 1.43 | 0.067 |
| 3 | 2.37 | 2.05 - 2.70 | 0.108 |
| 4 | 3.95 | 3.45 - 4.45 | 0.166 |
| 5 | 5.98 | 5.31 - 6.65 | 0.223 |
| 6 | 8.47 | 7.61 - 9.33 | 0.287 |
| 7 | 11.47 | 10.39 - 12.55 | 0.358 |
| 8 | 14.97 | 13.71 - 16.23 | 0.417 |
| 9 | 18.92 | 17.49 - 20.35 | 0.475 |
| 10 | 23.32 | 21.73 - 24.91 | 0.529 |
| 11 | 28.14 | 26.42 - 29.86 | 0.572 |
| 12 | 33.38 | 31.53 - 35.23 | 0.614 |
| 13 | 39.07 | 37.13 - 41.01 | 0.644 |
| 14 | 45.21 | 43.20 - 47.22 | 0.668 |
| 15 | 51.83 | 49.78 - 53.88 | 0.682 |
| 16 | 59.00 | 56.94 - 61.06 | 0.684 |
| 17 | 66.83 | 64.81 - 68.85 | 0.670 |
| 18 | 75.51 | 73.59 - 77.43 | 0.638 |
| 19(K) | 85.63 | 83.95 - 87.31 | 0.558 |

$\hat{\mu}$ = \$11,091    G = 3.74<br>
N = 7624.    (.00639) z = 58.5

Source: Beach *et al* (1980), Tables 9.1, 9.4, and 9.5.

Finally, one may remark on the substantially larger standard error for the estimated Gini coefficient in Table IV compared to Table III due to the smaller sample size on which it is based. It has also been computed from vigintile values, whereas the earlier figures were based on decile values. However, if one recomputed the standard error in Table IV in more aggregated fashion, one would obtain values of .006335 from decile figures and .006160 from quintile figures compared to .006391 from the reported vigintile figures. That is, the Gini standard errors appear quite insensitive to the level of aggregation used and differ less than 4% between using quintile and vigintile levels of disaggregation.

## VII. Review and Conclusions

The general objective of this paper has been to extend the standard techniques of statistical inference to applied income distribution work at a disaggregated level of analysis. Sections II-IV of the paper introduced the essential background material on the asymptotic distributions of income quantiles, and then used them to derive model-free standard errors and confidence intervals for income share statistics, Lorenz curve estimates, and estimated Gini coefficients. The only additional information required to estimate the asymptotic covariance matrices involved is that of a relative mean income curve. Sections V and VI then provided several hypothesis tests on income shares and Lorenz curves which are typically of most interest to applied distribution analysts.

Three general conclusions emerge from this paper. First, it clearly follows that model-free statistical inference on Lorenz curves, income shares, and Gini coefficients is both feasible and remarkedly simple to carry out. Consequently, it is hoped that henceforth applied distribution analysis will be carried on in the framework of standard statistical inference. Second, when an analyst is reporting his empirical results in terms of Lorenz curves, he should also report estimated relative-mean-income ordinates so as to allow a reader to carry out inferences on the Lorenz curve figures. Third, statistical agencies providing published distribution data should also include, along with income share and histogram data, quantile income level estimates such as decile levels which researchers can then use for statistical inference purposes.

## Appendix

Lemma 2: Under the conditions of Theorem 1, if the population density has finite mean and variance, $\sqrt{N}(\hat{\Phi}_i - \Phi_i)$ and $\sqrt{N}(\Phi(\hat{\xi}_{p_i}) - \Phi_i)$ have the same limiting distribution .

Proof: The first part of the proof is a modification of the arguments in Gastwirth's (1974) Theorem 1.

Recall, first of all, that by the Central Limit Theorem $z = N^{\frac{1}{2}}(\bar{Y} - \mu)/\sigma$ has an asymptotic standard normal distribution if the Y's are drawn (as assumed) from a random sample. Also by Theorem 1 of the text,

$$\varepsilon = N^{\frac{1}{2}}(\hat{\xi}_{p_i} - \xi_{p_i})f(\xi_{p_i})/[p_i(1-p_i)]^{\frac{1}{2}} \qquad (A1)$$

is asymptotically standard normal as well.

Now in order to transform a conditional mean problem into an unconditional mean problem, introduce the random variable

$$\begin{aligned} I_j^i &= 1 \text{ if } Y_j < \xi_{p_i} \\ &= 0 \text{ otherwise} \end{aligned} \qquad (A2)$$

where $Y_j$ denotes the j'th observation in the random sample drawn from the continuous density $f(\cdot)$ with finite mean and variance. The number of observations less than $\xi_{p_i}$ is a binomial random variable with parameters N and $p_i$, and

$$\tau_i = E(I_j^i Y_j) = \int_0^{\xi_{p_i}} y \, dF(y) = p_i \, E(Y_j | Y_j \leq \xi_{p_i}).$$

Consider then the asymptotic distribution of the conditional mean estimator $\bar{Y}_{\hat{\xi}_{p_i}} = (1/n_i) \Sigma_{Y_j \leq \hat{\xi}_{p_i}} Y_j$ where $n_i = [Np_i0]$. Let

$$S_i = \sqrt{N}\, p_i\, [\bar{Y}_{\hat{\xi}_{p_i}} - E(Y_j | Y_j \le \xi_{p_i})]$$

$$\doteq \sqrt{N}\, [N^{-1}\, \Sigma_{Y_j \le \hat{\xi}_{p_i}} Y_j - \tau_i] \qquad (A3)$$

Then consider the first term in (A3).

$$\Sigma_{Y_j \le \hat{\xi}_{p_i}} Y_j = \Sigma_1^N I_j Y_j + \Sigma_{Y_j \varepsilon (\xi_{p_i}, \hat{\xi}_{p_i})} Y_j \qquad (A4)$$

$$\doteq \Sigma_1^N I_j Y_j + \xi_{p_i} R + O_p(1) \qquad (A4)$$

where it is assumed for convenience that $\xi_{p_i} < \hat{\xi}_{p_i}$, and where R represents the (signed) number of observations between $\hat{\xi}_{p_i}$ and $\xi_{p_i}$. Since the number of observations in a small interval of length $\Delta$ about $\xi_{p_i}$ is approximately $Nf(\xi_{p_i})\Delta$, and since the (signed) length of the interval between $\hat{\xi}_{p_i}$ and $\xi_{p_i}$ is approximately

$$N^{-\frac{1}{2}}[p_i(1-p_i)]^{\frac{1}{2}}\, \varepsilon / f(\xi_{\mu_i}) \text{ from (A1),}$$

$$R \doteq N^{\frac{1}{2}}[p_i(1-p_i)]^{\frac{1}{2}}\varepsilon. \qquad (A5)$$

Thus, from (A4) and (A5),

$$S_i = N^{-\frac{1}{2}}\, \Sigma_1^N\, (I_j Y_j - \tau_i) + \xi_{p_i}[P_i(1-p_i)]^{\frac{1}{2}}\varepsilon + O_p(1)$$

where the first term is asymptotically normal with zero mean by the Central Limit Theorem and the second has also been shown to be asymptotically normal with mean zero in Theorem 1 of the text. Consequently, $S_i$ is also asymptotically normal with mean zero, and $p_i \bar{Y}_{\hat{\xi}_{p_i}}$ is asymptotically normal with mean $p_i\; E(Y_j | Y_j \le \xi_{p_i}) = \tau_i$.

Now, by the argument in Section III.1, the limiting distribution of a continuous function of asymptotically normal random variables is also asymptotically normal. In particular, consider the ratio $p_i \bar{Y}_{\hat{\xi}_{p_i}}/\bar{Y}$ both of whose arguments have been shown to be asymptotically normal with means $\tau_i$ and $\mu$ respectively. Then it follows that

$$\sqrt{N}\,[p_i \frac{\bar{Y}_{\hat{\xi}_{p_i}}}{\bar{Y}} - \frac{\tau_i}{\mu}] = \sqrt{N}(\hat{\Phi}_i - \Phi_i)$$

is also asymptotically normal with mean zero and a constant variance for $i = 1, \ldots, K$. That is, $\sqrt{N}(\hat{\Phi}_i - \Phi_i)$ and $\sqrt{N}(\Phi(\hat{\xi}_{p_i}) - \Phi_i)$ have the same probability limit of zero, so that the feasible estimator $\hat{\Phi}_i$ and the infeasible estimator $\Phi(\hat{\xi}_{p_i})$ are asymptotically equivalently distributed for all $i = 1, \ldots, K$ (Rao (1965), p. 101(ix)).

## References

[1] Aigner, D.J., and A.S. Goldberger, "Estimation of Pareto's Law from Grouped Observations", *Journal of the American Statistical Association,* vol. 65 (1970), pp. 712-723.

[2] Atkinson, A.B., "On the Measurement of Inequality", *Journal of Economic Theory,* vol. 2 (1970), pp. 244-263.

[3] Bahadur, R.R., "A Note on Quantiles in Large Samples", *Annals of Mathematical Statistics* vol. 37 (1966), pp. 577-580.

[4] Beach, C.M., "Cyclical Sensitivity of Aggregate Income Inequality", *Review of Economics and Statistics,* vol. 59 (1977), pp. 56-66.

[5] __________, "Inference with the Relative Mean Income Curve and Associated Inequality Measures", Discussion Paper, Dept. of Economics, Queen's University (1979), forthcoming.

[6] __________, with the assistance of D.E. Card and F. Flatters, *Distribution of Income and Wealth: Theory and Evidence for Ontario* (Toronto: Ontario Economic Council, 1980).

[7] Bradley, J.V., *Distribution-Free Statistical Tests* (Englewood Cliffs, N.J.: Prentice-Hall, 1968).

[8] Chernoff, Herman, J.L. Gastwirth, and M.V. Johns Jr., "Asymptotic Distribution of Linear Combinations of Functions of Order Statistics with Applications to Estimation", *Annals of Mathematical Statistics,* vol. 38 (1967), pp. 52-72.

[9] Danziger, Sheldon, and M.K. Taussig, "The Income Unit and the Anatomy of Income Distribution", Institute for Research on Poverty Discussion Paper No. 516-78, University of Wisconsin (Madison)(1978).

[10] Donaldson, David, and J.A. Weymark, "A Single-Parameter Generalization of the Gini Indices of Inequality", Discussion Paper 79-11, Dept. of Economics, University of British Columbia (1979).

[11] Dorfman, Robert, "A Formula for the Gini Coefficient", *Review of Economics and Statistics,* vol. 61 (1979), pp. 146-149.

[12] Eltetö, O., and E. Frigyes, "New Income Inequality Measures as Efficient Tools for Causal Analysis and Planning", *Econometrica,* vol. 36 (1968), pp. 383-396.

[13] Fair, R.C., "The Optimal Distribution of Income", *Quarterly Journal of Economics,* vol. 85 (1971), pp. 551-579.

[14] Gastwirth, J.L., "A General Definition of the Lorenz Curve", *Econometrica,* vol. 39 (1971), pp. 1037-1039.

[15] __________, "The Estimation of the Lorenz Curve and Gini Index", *Review of Economics and Statistics,* vol. 54 (1972), pp. 306-316.

[16] __________, "Large Sample Theory of Some Measures of Income Inequality", *Econometrica,* vol. 42 (1974), pp. 191-196.

[17] Gastwirth, J.L., and M. Glauberman, "The Interpolation of the Lorenz Curve and Gini Index from Grouped Data", *Econometrica,* vol. 44 (1976), pp. 479-484.

[18] Kakwani, N.C., "A Note on the Efficient Estimation of the New Measures of Income Inequality", *Econometrica,* vol. 42 (1974), pp. 597-600.

[19] Kakwani, N.C. and N. Podder, "On the Estimation of Lorenz Curves from Grouped Observations", *International Economic Review* vol. 14 (1973) pp. 278-292.

[20] _________, "Efficient Estimation of the Lorenz Curve and Associated Inequality Measures from Grouped Observations", *Econometrica,* vol. 44 (1976), pp. 137-148.

[21] Kendall, M.G. and A. Stuart, *The Advanced Theory of Statistics,* 3rd Ed., vol. I (London: Charles Griffen & Co., 1969).

[22] Kloek, T., and H.K. van Dijk, "Further Results on Efficient Estimation of Income Distribution Parameters", *Economie Appliquée* vol. 30 (1977) pp. 1-21.

[23] _________, "Efficient Estimation of Income Distribution Parameters", *Journal of Econometrics,* vol. 8 (1978), pp. 61-74.

[24] Levine, D.B., and N.M. Singer, "The Mathematical Relation Between the Income Density Function and the Measurement of Income Inequality", *Econometrica,* vol. 38 (1970), pp. 324-330.

[25] Martić, Ljubomir, "A Geometrical Note on New Income Inequality Measures", *Econometrica,* vol. 38 (1970), pp. 936-937.

[26] Mosteller, Frederick, "On Some Useful Inefficient Statistics", *Annals of Mathematical Statistics,* vol. 17 (1946), pp. 377-407.

[27] Rao, C.R., *Linear Statistical Inference and Its Applications* (N.Y.: John Wiley & Sons, 1965).

[28] Sarhan, A.E., and B.G. Greenberg, ed., *Contributions to Order Statistics* (N.Y.: John Wiley & Sons, 1962).

[29] Scheffé, Henry, *The Analysis of Variance* (N.Y.: John Wiley & Sons, 1959).

[30] Seber, G.A.F., *Linear Regression Analysis* (N.Y.: John Wiley & Sons, 1977).

[31] Sen, Amartya, *On Economic Inequality* (Oxford: Clarendon Press, 1973).

[32] Thurow, L.C., "Analyzing the American Income Distribution", *American Economic Review,* vol. 60 (1970), pp. 261-270.

[33] Wilks, S.S., *Mathematical Statistics* (N.Y.: John Wiley & Sons, 1962).

[34] Wold, Herman, "A Study on the Mean Difference, Concentration Curves, and Concentration Ratio", *Metron,* vol. 12 (1935), pp. 39-58.

Irvine

ISSN 0316-5078

# INSTITUTE FOR ECONOMIC RESEARCH

QUEEN'S UNIVERSITY

Kingston, Ontario, Canada K7L 3N6

# Power Grid Economics in a Peak Load Pricing Framework: Comment*

by

John Rowse

Queen's University

Discussion Paper No. 352

July, 1979

Preliminary Draft: Please do not quote without the author's consent.

## Abstract

In a recent paper focussing on the analysis of electric power supply from a power grid, Professor Michael Berkowitz has set forth a model of an electric power grid and determined, under three sets of conditions, rules for optimal pricing, transmission and capacity expansion. In light of the rising real cost of electric power his research addresses an important issue, but the modelling approach he has adopted ignores several major considerations for power grid supply which, if taken into account, give rise to substantial differences in these optimal decision rules. It is the purpose of this note to discuss several of these considerations, to analyze a power grid model modified to allow for them, and to compare the analytical findings with those of Berkowitz in order to determine their implications. The principal conclusion is that, for his results to be made applicable to power grid economics, they must be established using a modified modelling approach.

## Introduction

In a recent paper, Professor Michael Berkowitz has set forth a model of an electric power grid and determined, under three sets of conditions, rules for optimal pricing, transmission and capacity expansion.[1] In light of the rising real cost of electric power his research addresses an important issue, but the modelling approach he has adopted ignores several major considerations for power grid supply which, if taken into account, give rise to substantial differences in these optimal decision rules. It is the purpose of this note to discuss several of these considerations, to analyze a power grid model modified to allow for them, and to compare the analytical findings with those of Berkowitz in order to determine their implications. The principal conclusion is that, for his results to be made applicable to power grid economics, they must be established using a modified modelling approach.

This paper is organized as follows. The next section offers several observations on power grid analysis and the model of Professor Berkowitz and indicates why his model requires modification. The following section sets out a modified model allowing for several of these observations and presents a brief analysis of the Kuhn-Tucker conditions necessary for model optimization, from which come the rules for optimal pricing, transmission and capacity expansion. A brief summary of the differences in analytical results determined by this modified model and those of Professor Berkowitz is then recorded, and implications of these differences are drawn. The final section offers concluding remarks.

## Observations on Power Grid Analysis and the Berkowitz Model

### a) Time Orientation

Professor Berkowitz makes no mention of the time frame he has in mind for his analysis. It is possible, therefore, that he intends his model to be without a time dimension; that is, his analysis compares the steady-state behaviour of a group of autarkic power systems with that of a group of fully integrated systems. Such a comparison is not adequate for analyzing power grid economics, however, because power grid analysis requires a temporal orientation and specifically, a long-term perspective. Most large generating stations either built or under construction have long expected lifetimes[2] and the utilization of prospective stations for a protracted period of time must be considered in any model assembled to examine power grid supply, in order to prevent near term demands from exerting undue influence on the types and extent of generating capacity selected for expansion in the model. Furthermore, it is not possible (and would not likely prove optimal even if possible) to redesign existing generating systems from scratch to realize the benefits arising from power supply integration; hence a gradual approach over time toward an optimal spatial generating and transmission grid configuration, taking into account existing supply capabilities, is the proper objective of power grid analysis. Finally, principal genefits from integrating capacity expansion and utilization accrue over a long time period, not a short one.

On the other hand, perhaps the author does have a particular time frame in mind, but has simply neglected to specify it. If so, then his time horizon cannot be more than four or five years at most -- much too short a time frame for power grid analysis -- since his capacity expansion decision variables (the $X_{ij}$) are to be optimized only once for all time periods. If the time frame intended is in fact much longer than this time span, then the model is flawed because its once-only selection of capacity increments implies that capacity for use at the beginning of the planning period is identical to capacity installed at the horizon; assuming a planning horizon two or three decades distant, this implication translates into enormous overcapacity during the early years.[3] If a long term orientation is intended, then the obvious solution is to define capacity expansion variables for successive time periods and allow the model to select which regions or firms are to expand capacity over time to meet evolving grid demands and which are not. A major advantage of such an approach is that it would allow an optimal intertemporal grid system to exhibit early exploitation of particularly favourable region or firm-specific expansion alternatives for grid supply, to be followed subsequently by expansion using less attractive alternatives, precisely what should be expected from a joint supply approach.

b) Member Alternatives for Capacity Expansion

All electric utilities possess multiple plant expansion alternatives for meeting electrical loads, with each alternative exhibiting unique construction and operating costs and load-carrying characteristics. This range in generating alternatives possible mitigates to some extent the cost consequences of operating in isolation, since each supplier will be able to select a mix of capacity types -- peak, intermediate and base -- which best fits the pattern of loads experienced. Berkowitz assumes a single type of capacity for each member and does not rule out the case of complete self-generation; thus *all* capacity he defines must be base load in nature because at least some fraction of it must be able to operate continuously. Hence he implicitly rules out intermediate and peak load capacities and thereby any role that each may play in solving the peak load problem. In his statement of Conclusions, the author speculates that inclusion of "technique choice" could well modify decisions of grid members, but the point deserves more than mere mention since for each grid member several generating types will compete with possible grid supply in the decision as to which source is to supply power at any given time. For instance, for the meeting of peak demands for a utility it is certainly conceivable that domestic peak load capacity could compete successfully in cost with supply from the grid.

c) The Nexus Between Generating Capacity and Support of Transmission

Save for rare exceptions, no electric utility commits specific generating capacity for a long period of time to supporting transmission to another utility, even though long term supply contracts between utilities certainly are possible and do exist. The difference is that power supplied under contract will be generated by whatever capacity it is most convenient to operate, and as time passes this supporting capacity most likely will change. Moreover, in the context of a power grid, where transmission flows among members will change in direction and intensity as demands evolve and progressively more costly supply alternatives are exploited, major benefits will be lost if specific generating capacities of one member are tied to supporting transmission to other members. For example, for a period of time it may be optimal for member A to utilize capacity X to generate and transmit power to member B. Subsequently, as circumstances evolve, member B may construct and use its own capacity Y or import power from member C, thus releasing capacity X for domestic use by A or for power generation to some other member D. If capacity X were always committed to generating power for member B by A, this substitution of uses for X could not occur and hence one principal source of grid benefits would disappear. The point of this discussion is that in a long term model framework generating capacity expansion must be defined differently than it is by Berkowitz since his decision variables $X_{ij}$ exhibit precisely this dedication of generating capacity of one grid member to the support of transmission to another member.

### d) Interregional Power Losses

Power losses normally are associated with long distance transmission and the extent of such losses may well differ from one region to another. Although the Berkowitz model makes no specific allowance for power losses in transmission, it can readily be modified in interpretation to allow for power losses if such losses are taken to be identical *everywhere* in the power grid, for then transmission losses will have principal consequences only for the cost of delivered power. On the other hand, if power losses differ among pairs of grid members, then such losses must be represented as physical characteristics of the power grid modelled since they will have distinctly different consequences for the amount of generating capacity needed to support transmission to a power importer depending upon the losses experienced over different transmission links, as well as implications for the cost of delivered power. For instance, whether member A opts for power supply from member B or from member C will have different consequences for the amount of generating capacity required to support transmission in the grid if the AB link experiences one percent transmission losses and the AC link experiences ten percent losses.

e) Power Grid Financial Integrity

Emerging from the particular case of stochastic demand in the Berkowitz model is the result that any firm purchasing generating capacity must realize a producer's deficit. To remedy this problem, the author raises the possibility of subsidization of producers from a public authority, which to cover the subsidy costs might sell -- at the marginal cost of capacity construction -- future rights to purchase electric power; this subsidy and futures-market scheme for grid financial solvency subsequently is used in the derivation of numerous results in the paper. Disregarding the utter implausibility of such a scheme, it is important to examine the assumptions upon which the solvency result is predicated, in order to form an impression of its generality. The author in fact offers no proof that the futures market will work well enough to allow costs of subsidizing producers to be covered, but he does refer the reader to two earlier papers by Brown and Johnson (1969, 1973).[4] I must assume, therefore, that the result concerning grid financial solvency is intended to rest upon the proof developed in their papers. Now Brown and Johnson (1969) derived their result on the basis of a futures market without transactions costs, availability of only a single type of capacity, purchase of production capacity at a constant per-unit cost, absence of a risk premium on resale of future rights by speculators, a perfectly operating spot market for the commodity bought and sold, as well as complete faith that the public authority would indeed be able to provide the commodity exactly when desired, not to mention an unstated prohibition of temporal adjustment in capacity or operating costs. Presuming the author implicitly is adopting *all* these same restrictive assumptions (and for *each* location in his power grid) in order to obtain the desired solvency, the relevance for power grid economics of the analytical results derived under such a scheme must surely be called into question.

One alternative for securing grid financial solvency might simply be to insist upon it directly: since the author is dealing with a second-best world in any event (he has, for instance, quite reasonably neglected income distributional considerations and environmental externalities associated with power supply, and omitted regional employment consequences of specific capacity expansion projects) why could grid financial integrity not be made part of the model by imposing a probabilistic "solvency constraint" insisting that revenues exceed costs with some exogenously specified high probability and the consequences for supply and pricing explored under this restriction?

## A Modified Power Grid Model

The first four observations of the previous section have focussed on considerations in modelling electric power systems either neglected or given only scant attention in the Berkowitz power grid model. These observations are relevant, however, only to the extent that they give rise to theoretical and policy results that differ significantly from those determined by Berkowitz. It is the purpose of this section, therefore, to construct a power grid model modified to take into account the above considerations and analyze it to demonstrate how conclusions regarding pricing, transmission and capacity expansion rules change. Only the deterministic case will be treated, and for brevity the deterministic model of Berkowitz will henceforth be referred to simply as the B model.

Formulation of the modified power grid model has several elements that contrast sharply with the B model: (i) there are four generic types of intertemporal decision varibles, namely variables for generating capacity utilization, generating capacity expansion, transmission capacity utilization and transmission capacity expansion; (ii) there are several different types of capacity which may be chosen for expansion by each member; (iii) there is allowance for existing capacity; and (iv) the planning orientation is intended to be long term, from twenty to thirty years or longer. Concerning (i), the generating capacity - transmission support nexus of the B model is broken, since there is no reason to insist that particular generating capacity of one grid member be committed to the transmission of power to a second grid member over the entire planning period selected, as the B model requires.

a) Model Formulation[5]

For convenience, Table 1 provides definitions of indexes, variables and parameters for the modified model. There are $m$ grid members, indexed by both $i$ and $j$; there are $R$ types of capacity, indexed by $r$ and assumed available to *each* grid member; there are $T$ time periods, indexed by $t$, each of which may be one or several years in duration; and there are $K$ subperiods of each time period, indexed by $k$, which together account for all the different time intervals of each time period during which demand varies cyclically. These subperiods may be thought of as corresponding to certain blocks of the approximate load duration curve,[6] in which case they are not sequential in time (although this interpretation involves considerable difficulty), or they may be thought of as days, weeks, months or seasons of each time period, in which case they are sequential in time. Electric power, measured in megawatts (MW), is the commodity generated and transmitted in the grid network,[7] and is sold at a price $p^i_{tk}$ during period $t$, subperiod $k$, to customers of grid member $i$. Power demand $D^i_{tk}$, the quantity of electric power demanded by customers of member $i$ during $t,k$ is responsive to power price, however, so that demand can be written functionally as $D^i_{tk} = D^i_{tk}(p^i_{tk})$. Adopting the assumptions of the $B$ model, demands of one subperiod are not responsive to prices of other subperiods, and the demand functions are all monotonically decreasing and differentiable. The inverse demand functions will be denoted by $p^i_{tk}(D^i_{tk})$.

Primal decision variables consist of: generating capacity utilization variables $U^i_{rtk}$, denoting the quantity of power generated by grid member $i$ using capacity type $r$ during $t,k$; generating capacity expansion

Table 1

Indexes, Variables and Parameters

Indexes:

$i,j$ - index the $m$ power grid members

$r$ - indexes the $R$ types of generating capacity

$t$ - indexes the $T$ time periods

$k$ - indexes the $K$ subperiods of each time period

Primal Decision Variables:

$D^i_{tk}$ - power consumption by the customers of grid member $i$ during subperiod $k$ of time $t$; a function only of $p^i_{tk}$, the prevailing price of electric power

$U^i_{rtk}$ - power generation by grid member $i$ using capacity type $r$ during subperiod $k$ of time $t$

$V^i_{jtk}$ - power transmission from grid member $j$ to member $i$ during subperiod $k$ of time $t$

$X^i_{rt}$ - generating capacity of type $r$ constructed by member $i$ to be ready for first use during time $t$

$Y^i_{jt}$ - transmission capacity linking grid members $i$ and $j$ constructed for first use during time $t$

Parameters:

$\alpha_{ij}$ - percentage of power transmitted from grid member $j$ to grid member $i$ which is received successfully by member $i$; $0 < \alpha_{ij} \leq 1$

$X^o_{ri}$ - existing generating capacity of type $r$ owned by grid member $i$ at the start of the planning period

$Y^o_{ij}$ - existing transmission capacity linking grid members $i$ and $j$ at the start of the planning period

$\overline{X}^i_r$ - exogenously specified maximum amount of generating capacity type $j$ which may be constructed by grid member $i$ over the planning period

$a^i_{rtk}$ - discounted per-megawatt operating cost of power generation by grid member i from capacity type r during subperiod k of time t

$b^i_{rt}$ - discounted per-megawatt cost of constructing capacity type r by grid member i for first use during time t

$c^i_{jtk}$ - discounted per-megawatt cost of transmitting power from grid member j to member i during subperiod k of time t; consists essentially of maintenance costs

$d^i_{jt}$ - discounted per-megawatt cost of constructing transmission capacity linking grid members i and j for first use during time t

Dual Decision Variables:

$\lambda^i_{tk}$ - imputed value of a unit increment in the demand for electric power by the customers of grid member i during subperiod k of time t

$\eta^i_{rtk}$ - imputed value to grid member i of an incremental megawatt of generating capacity type r during subperiod k of time t

$\mu^i_{jtk}$ - imputed value to the power grid of an incremental megawatt of transmission capacity linking grid members i and j during subperiod k of time t

$\omega^i_r$ - imputed value to grid member i of a unit increment in the availability of generating capacity type r over the full planning period variables $X^i_{rt}$, denoting the quantity of capacity of type $r$ to be ready for first use during $t$ for grid member $i$; transmission capacity utilization variables $V^i_{jtk}$, denoting the quantity of electric power transmitted to grid member $i$ from grid member $j$ during $t,k$; and transmission capacity expansion variables $Y^i_{jt}$, denoting the quantity of transmission capacity linking grid members $i$ and $j$ to be ready for first use during time $t$. Units of all decision variables and hence of all constraints listed below are megawatts.[8]

With these definitions as background, the model constraint set may be set forth as follows, where the variable in parentheses beside each constraint group is the Lagrangian multiplier or dual variable.

$$(1) \qquad \sum_{r=1}^{R} U^i_{rtk} + \sum_{\substack{j=1 \\ j\neq i}}^{m} \alpha_{ij} V^i_{jtk} - \sum_{\substack{j=1 \\ j\neq i}}^{m} V^j_{itk} \geq D^i_{tk}(p^i_{tk}) \qquad (\lambda^i_{tk}) \quad \left\{ \begin{array}{l} i=1,\ldots,m \\ t=1,\ldots,T \\ k=1,\ldots,K \end{array} \right.$$

These constraints insist that whatever power demands are made during $t,k$ by grid member $i$ must be satisfied, either from generation using $i$'s own generating capacity, namely the first sum on the left hand side of (1), or from power import from other grid members, namely the second (weighted) sum on the left hand side. Because of power losses in the network, however, only a certain fraction of power transmitted from member $j$ to $i$ -- $\alpha_{ij}$ -- is successfully received by $i$ for the meeting of demands at $i$.[9] The third sum on the left hand side represents power transmitted to all other grid members from $i$ and has a negative sign because it represents demands placed upon the supply system of $i$. Because it does not make sense to transmit power from $i$ to itself, variables such as $V^i_{itk}$ have not been defined and have therefore been excluded from two of the sums in (1). Finally, at the optimum it must be true that (1) will be satisfied as an equality since it will never be optimal to supply more power than is demanded.[10]

$$(2) \qquad U^i_{rtk} \leq X^o_{ri} + \sum_{s=1}^{t} X^i_{rs} \qquad (\eta^i_{rtk}) \qquad \left\{ \begin{array}{l} i=1,\ldots,m \\ r=1,\ldots,R \\ t=1,\ldots,T \\ k=1,\ldots,K \end{array} \right.$$

These constraints require that power generation from capacity type $r$ during $t,k$ for grid member $i$ be no greater than initial capacity $X^o_{ri}$ -- which may be zero -- plus whatever capacity has been added for use by time $t$, this latter capacity being given by the sum on the right hand side of (2). As specified in the inequality, generating capacity is cumulative and remains fully intact over time, exhibiting no downtime on a short term basis for maintenance or breakdown and undergoing no retirement as time passes.

$$(3) \qquad V^i_{jtk} + V^j_{itk} \leq Y^o_{ij} + \sum_{s=1}^{t} Y^i_{js} \qquad (\mu^i_{jtk}) \qquad \left\{ \begin{array}{l} \left.\begin{array}{l} i=1,\ldots,m-1 \\ j=2,\ldots,m \end{array}\right] j>i \\ t=1,\ldots,T \\ k=1,\ldots,K \end{array} \right.$$

These constraints ensure that power transmission during $t,k$ between grid members $i$ and $j$ is no greater than initial transmission capacity $Y^o_{ij}$ -- which may be zero -- plus whatever transmission capacity is added over time for use by $t$, as specified by the second term on the right hand side. As with generating capacity, transmission capacity is cumulative and remains fully intact over time once constructed. Only a single type of transmission capacity is allowed in the model, an assumption also implicit in the B model. It should be noted that variables representing transmission of power in both directions at once can be included in the same constraint because at most one of the power transmission variables will be positive in the optimal solution; the facts that power generation and transmission costs are linear and that maximization of social welfare requires production cost minimization together are responsible for this property of optimality.[11]

$$(4) \qquad \sum_{t=1}^{T} X_{rt}^{i} \leq \overline{X}_{r}^{i} \qquad (\omega_{r}^{i}) \qquad \begin{cases} r=1,\ldots,R \\ i=1,\ldots,m \end{cases}$$

These constraints ensure that for each grid member $i$ no more than an exogenously specified maximum amount of capacity type $r$, $\overline{X}_{r}^{i}$, is constructed over the planning period. Such constraints are intended to allow for (i) restrictions deriving from physical constraints, exemplified by unequal spatial endowments of potential capacity (some members may have *none* of the capacity types allowed), inherent scarcity of low cost generating sources such as choice hydro sites (although little emphasis here is placed upon hydro because of its specialized nature), or restricted availability of cooling water for conventional thermal generation; or (ii) restrictions deriving from "policy" constraints, such as the ceiling on allowable nuclear expansion by any grid member over the horizon. Clearly, if there are no physical or policy constraints, the $\overline{X}$'s may be taken to be arbitrarily large, implying that these constraints in no way bear upon the optimal solution and hence requiring the associated dual variables $\omega_{r}^{i}$ to be zero.[12]

$$(5) \qquad U_{rtk}^{i} \geq 0, \quad V_{jtk}^{i} \geq 0, \quad X_{rt}^{i} \geq 0, \quad Y_{jt}^{i} \geq 0$$

for all variables for which $i$, $j$, $r$, $t$ and $k$ are defined

These constraints require that all primal decision variables be non-negative.

The objective to be maximized is the discounted sum of consumers' plus producers' surplus that is generated over the grid planning horizon. The objective is therefore to

$$\text{Max } \Omega = \sum_{t=1}^{T} \sum_{i=1}^{m} \sum_{k=1}^{K} \int_{0}^{D^i_{tk}} p^i_{tk}(w)\, dw$$

$$- \sum_{t=1}^{T} \sum_{i=1}^{m} \left[ \sum_{r=1}^{R} \left( \sum_{k=1}^{K} a^i_{rtk}\, U^i_{rtk} \quad + \quad b^i_{rt}\, X^i_{rt} \right) \right]$$

$$- \sum_{t=1}^{T} \left[ \sum_{i=1}^{m} \sum_{\substack{j=1 \\ j\neq i}}^{m} \sum_{k=1}^{K} c^i_{jtk}\, V^i_{jtk} \; + \; \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} d^i_{jt}\, Y^i_{jt} \right]$$

subject to the above constraints, where the first term defining $\Omega$, namely the triple sum of integrals, represents consumers' surplus plus revenue received from sales of power over the planning horizon (the p's denoting *discounted prices*); the second and composite term represents costs of power generation (the coefficient $a^i_{rtk}$ representing the discounted per MW cost of power generation for grid member $i$ of operating capacity type $r$ during subperiod $k$ of time $t$) and capacity expansion (the coefficient $b^i_{rt}$ representing the discounted per MW cost of expanding capacity type $r$ to be ready for first use during time $t$); and the third and composite term represents the discounted costs of power transmission (the coefficient $c^i_{jtk}$ representing the discounted per MW cost of transmitting power from grid member $i$ to member $j$ during $t,k$) which are apt to be very small because they embrace only costs of maintaining transmission capacity, and costs to the grid of expanding transmission capacity (the coefficient $d^i_{jt}$ representing the discounted per MW cost of expanding transmission capacity between grid members $i$ and $j$ to be ready for first use during the time period $t$). If desired, it is possible to assume cost symmetry, namely $c^i_{jtk} = c^j_{itk}$, but it is not necessary.[13] Adding the second and third terms to the first yields consumers' surplus plus revenue minus production, transmission and expansion costs, or the discounted sum of consumers' plus producers' surplus.

* * * * * * * * * * *

Such is the modified model for grid planning, and although it constitutes an elaboration of the deterministic B model, it is still a stark simplification of reality. For instance, the demand functions are assumed known with certainty; no reserves are required because there are no unforeseen demands, forced outages or downtimes for maintenance; all existing and constructed capacity never wears out; each grid member has access to exactly the same range of capacity types; each type of capacity never has restricted availability, such as would be the case with hydro capacity during the years of low flows; discreteness in capacity construction, a phenomenon known to be important for electric power supply, is not allowed; there are no scale economies available in generating and transmission capacity construction; each grid member has all generating capacity and demands located at a single point, so that there are no transmission and distribution losses internal to each member; and all grid members are linked directly to each other. All of these simplifications could be relaxed, but at some cost in terms of complexity in the model and results, though undoubtedly with further insight. One example is the case of scale economies in generating capacity allowed in the B model[14], which modifies that model slightly and forces slight reinterpretation of one set of the dual constraints.

b) Interpretation of the Dual Constraints

Consisting of the Kuhn-Tucker conditions necessary for maximizing $\Omega$, the dual constraints are set out next, where the primal variable in parentheses beside each inequality is the Lagrangian multiplier associated with the dual constraint.[15] At the optimum, the complementary slackness conditions require that if the primal variable is positive, then the slack in the associated dual constraint must be zero; that is, the associated dual constraint must be an equality. Conversely, if the dual slack is positive, then the associated primal variable must be zero. Similar relationships hold for the primal constraints (1) through (4) and their associated dual variables. The dual constraints are as follows, the index ranges being those of the primal decision variables:

$$\lambda^i_{tk} - \eta^i_{rtk} \leq a^i_{rtk} \qquad (U^i_{rtk}) \tag{6}$$

$$\alpha_{ij}\,\lambda^i_{tk} - \lambda^j_{tk} - \mu^i_{jtk} \leq c^i_{jtk} \qquad (V^i_{jtk}) \tag{7}$$

$$\sum_{s=t}^{T}\sum_{k=1}^{K} \eta^i_{rtk} - \omega^i_r \leq b^i_{rt} \qquad (X^i_{rt}) \tag{8}$$

$$\sum_{s=t}^{T}\sum_{k=1}^{K} \mu^i_{jtk} \leq d^i_{jt} \qquad (Y^i_{jt}) \tag{9}$$

$$\lambda^i_{tk} \geq p^i_{tk} \qquad (D^i_{tk}) \tag{10}$$

All dual variables are restricted to be non-negative.

If it is assumed that the quantity of power $D^i_{tk}$ sold to the customers of grid member $i$ during subperiod $k$ of time $t$ is positive in the optimal solution (an assumption difficult *not* to make since it is hard to conceive of an *optimal price* high enough to force any power demand $D^i_{tk}$ to zero), then constraint (10) becomes an equality, and substitution can be made into constraints (6) and (7) to derive:

$$\text{(6a)} \qquad p^i_{tk} - \eta^i_{rtk} \leq a^i_{rtk} \qquad (U^i_{rtk})$$

$$\text{(7a)} \qquad \alpha_{ij}\, p^i_{tk} - p^j_{tk} - \mu^i_{jtk} \leq c^i_{jtk} \qquad (V^i_{jtk})$$

The dual constraints may now be examined for their economic content. If $U^i_{rtk} > 0$, that is, if capacity $r$ is operated to provide power to grid member $i$ during $t,k$ then such power generation is carried to the point at which price is equal to operating cost plus the imputed value of an incremental megawatt expansion of that particular capacity:

$$\text{(11)} \qquad p^i_{tk} = a^i_{rtk} + \eta^i_{rtk} \qquad \text{for} \qquad U^i_{rtk} > 0$$

This equality is derived from (6a) and complementary slackness. Clearly, there may be many types of capacity $r$ for which this equality holds, each with different operating costs and consequently different imputed values $\eta$. This equality, it might be noted, is analogous to but not identical with equation (3a) in Berkowitz (1977, p. 624), the difference being that (11) can hold for *many types of capacity for each member, not just one*. Further, it might be noted that if the marginal plant supplying power (i.e. that for which operating cost is highest) is not operated to capacity, then $\eta^{i}_{rtk} = 0$ by complementary slackness between the dual variable and primal slack of equation (2), and hence $p^{i}_{tk} = a^{i}_{rtk}$; that is, price during the k'th subperiod of time t is equal to the operating cost of the marginal plant, as might be expected. This result holds, however, *only* when either there are no power imports, *or* when all power import capacity is exhausted.[16]

If grid member i imports power during t,k from member j, then (7a) can be rewritten as:

$$(12) \qquad \alpha_{ij}\, p^{i}_{tk} = p^{j}_{tk} + c^{i}_{jtk} + \mu^{i}_{jtk} \qquad \text{for} \quad V^{i}_{jtk} > 0$$

Consequently, power import from member j is carried to the point at which the prices of electric power sold to customers of i and j are related as in equation (12). If the fraction of power successfully transmitted from j to i is less than unity, then price at i must exceed price at j, the excess also depending on the marginal cost of power transmission (if there is any) and the imputed value of an incremental MW of transmission capacity. *If* region j does not import power, *if* its marginal plant is not operated to capacity, and *if* the fraction of power successfully transmitted from j to i is unity, then equation (12) can be rewritten as:

$$(13) \qquad p^{i}_{tk} = a^{j}_{rtk} + c^{i}_{jtk} + \mu^{i}_{jtk}$$

This equation is analogous to, but once again not identical with, equation (3b) in Berkowitz (1977, p. 624). Main differences in interpretation stem from the way in which the three right hand side elements of (13) are defined.

What is interesting about (11) and (12), taken together, is the range of price and generating configurations that they cover. Generation configurations for each grid member range from complete self-sufficiency with power export, through autarkic self-sufficiency and partial self-sufficiency supplemented with some power import, to complete dependence upon power import from other grid members. Perhaps the most important choice allowed each member is the *intermediate choice,* namely utilization of some domestic generating capacity supplemented with purchase of power from one or more grid members, undoubtedly a position in which most if not all grid members in reality will at some time find themselves when cyclical variations in demand are taken into account. For grid member $i$ this case obtains when both (11) and (12) hold simultaneously for some $r$ and $j$ and thus $p_{tk}^{i}$ satisfies these two equalities. The B model does not allow for the possibility that this intermediate position may be optimal because of the restrictiveness of the options allowed grid members in that model; these options give rise to the "knife-edge" character of the optimal solution: *either* import all power supply *or* self-generate completely. (Only in the highly improbable case when the cost of power import exactly equals the cost of self-generation will import and self-generation by member $i$ be observed simultaneously in the B model, but even in such a circumstance no advantage is to be had over complete self-generation.) Further for the modified model, with regard to the decision to import or self-generate, the procedure is more complex than that of the B model. Because of restrictions on generating and transmission capacity, the modified model requires examination of the question of *how much power to import from each member, not just whether to import or not.* Member i must consider its own range of generating costs $a^i_{rtk}$ of capacities r not fully utilized, the ranges of generating costs $a^j_{rtk}$ of all other members j likewise not fully utilized, all marginal import costs, and the fractions of power successfully transmitted $\alpha_{ij}$ from all other members. Hence the set of considerations influencing the decision to import is far broader and more complete than that arising from analysis of the B model and displayed in Table 1 of Berkowitz (1977, p. 625).

Forming a second part of the dual constraint set are the rules for expanding different types of generating capacity. If capacity of type r is to be added for grid member i for first use during time t, then such construction must be carried to the point at which:

$$(14) \qquad \sum_{s=t}^{T} \sum_{k=1}^{K} \eta^i_{rtk} - \omega^i_r = b^i_{rt} \quad \text{for} \quad X^i_{rt} > 0$$

that is, the sum, over all subperiods and time periods extending from t to the planning horizon, of imputed values of an additional megawatt of generating capacity, less the imputed scarcity value of capacity r -- which may be zero if development is unrestricted by physical or "policy" constraints --is exactly equal to its per MW discounted cost. This decision rule is analogous to the rule enunciated verbally in Berkowitz (1977, p. 624): "Plant i capacity is employed to the point at which the sum of the marginal contributions *over all periods* exactly equals its cost."[17] Principal differences here lie in the facts that (i) the sum of "marginal contributions" or imputed values is over all subperiods and times for which capacity is available for use prior to the horizon of the long term modified model and not just to the (uncertain) time horizon of the B model, and (ii) more than a single generation type for each member is involved in the capacity choice. Given the long expected lifetimes of actual generating capacity and the range of capacity choices typically available, this modified decision rule makes much more sense than that determined from the B model. Secondary differences stem from (iii) the presence in (14) of the imputed scarcity value of capacity type j, which may indeed be positive if prospective capacity is both limited in extent and sufficiently attractive that from the grid standpoint it should be developed fully for grid supply, and (iv) the allowance implicit in (14) for *intertemporal development sequences.* Capacity alternatives for which construction and/or operating costs are highly attractive to the grid will be developed first and the associated $\overline{X}$'s will be attained early in the planning period, less attractive alternatives will be developed (and their $\overline{X}$'s attained) subsequently, and the least attractive alternatives will be deferred for development until the latter part of the planning period. Prospective capacity increments will be "evaluated" through the $\eta$'s and $\omega$'s such that this expansion pattern obtains over time. This phasing of expansion in an electric power grid is precisely what economic intuition would predict and in the optimum what the modified model specified numerically would exhibit; such phasing, however, is denied the B model because capacity expansion is allowed only once for each grid member for the planning period.

Forming the final part of the dual constraint set are the rules for expanding transmission capacity:

$$(15) \qquad \sum_{s=t}^{T} \sum_{k=1}^{K} \mu^{i}_{jtk} = d^{i}_{jt} \quad \text{for} \quad Y^{i}_{jt} > 0$$

If any capacity linking grid members i and j is to be constructed for first use during time t, then such construction is to be carried to the point at which the sum, over all subperiods and time periods from t to the planning horizon, of all imputed values of an additional megawatt of transmission capacity, is just equal to its per MW discounted cost. Because the decision on transmission capacity construction is separated from that concerning generating capacity construction, there is no counterpart in the B model interpretation for this decision rule. Two facts in particular should be noted. First, power transmission is tied to no particular capacity type, as it is in the B model; power transmission from member i is simply deducted from the aggregate of power generated by i as indicated in equation (1). Second, transmission flows in both directions along the transmission lines linking i and j can give rise to positive imputed values which in aggregate dictate transmission capacity expansion. These two facts together allow for the following possibilities in the model: *at different times,* (i) different capacity types may support power transmission from each member, and (ii) transmission flows may move in different directions along any single transmission link if there are, for instance, noncoincident peaks, differential demand growth rates, or different comparative advantages in peak, intermediate and base load generation experienced by different members. Intuitively, allowance for bi-directional transmission flows along the same transmission link and the expectation that their occurrence may well influence construction of transmission capacity are important for modelling a power grid; nowhere, however, do such allowance and expectation appear in the B model.

To this point there has been no discussion of the importance of including existing capacity in the modified model and what advantage such inclusion offers over the B model. Admittedly, the formal pricing and capacity decision rules remain unchanged by the introduction of existing capacity. What will be modified, however, is the pattern of time phasing of grid development: unless all existing capacity is mothballed or decommissioned when the grid is formed (a drastic step to contemplate in a developed country!), inclusion of existing capacity must have implications for expansion of the power grid. Some part of the existing capacity will undoubtedly be used, particularly early in the planning period, implying less need for construction of new capacity; technically, this reduced need will manifest itself in smaller $\omega$'s and sums of $\eta$'s in (8) for specific capacity expansion variables and in less capacity construction in aggregate. Most important, perhaps, is the point that, because so much capacity of an hypothetical power grid is already fixed prior to formation, analysis of the advantages stemming from power grid formation requires a long term perspective in order to allow for the necessary temporal transformation to a spatial generating grid structure which best utilizes the most attractive existing and potential capacity usually available only to a selected few favoured members.

Several extensions to the dual constraint interpretations which contribute to an understanding of power grid economics are easily made. As one illustration it may be instructive to examine what the introduction of environmental costs of generation - heretofore excluded from consideration - holds for grid decision making. Since the analysis is not central to the arguments presented here, however, it will be confined to the Appendix.

### c) Summary of Differences in Analytical Results

In a nutshell, what differences in results between the B model and the modified model have been identified? Highly visible differences in the pricing rules stem from the allowance of the modified model for multiple generation types for grid members and thus for potential optimality of the "intermediate case" of simultaneous self-generation and power import described above, and the fact that in deciding on power import each member must answer at the same time the question of *how much* to accept from other members. The import decision procedure must entail not just the restricted set of conditions discussed by Berkowitz, but consideration of the range of operating costs of capacity types not fully utilized (both for the member contemplating import and all other members), marginal import costs, power losses in transmission from other members, and capacity restrictions on generation and transmission. Regarding capacity expansion rules, differences stem from the evaluation in the modified model of the performance of generating capacity -- particularly near term capacity increments -- over a much longer horizon than that of Berkowitz (however distant his might be) in deciding on capacity to be constructed, the fact that peak and intermediate load generating capacity expansion may well prove optimal even though not allowed in the B model, the fact that there is no allowance in the B model for the temporal phased expansion of generating capacity in the grid toward an optimal spatial generating and transmission capacity configuration although such phased expansion is precisely the logical outcome of "high-order interconnectedness" among grid members, and -- because Berkowitz does not allow transmission capacity expansion variables into his model -- the fact that there is no allowance for bi-directional transmission flows between any pair of grid members to influence intertemporal transmission capacity expansion linking them. Less visible but nonetheless important differences stem from the omission of existing capacity in the B model, since such capacity strengthens the argument for a long term perspective, and the omission of bounds on the extent of capacity development -- either limiting low cost expansion of particular generation, or giving expression to physical or policy restrictions on development -- since such restrictions are inherently a part of any grid development and will account to a great extent for the phasing of capacity expansion over time and space.

d) Implications

Although the modified model represents an elaboration of the deterministic model of Berkowitz (exclusive of his allowance for scale economies in generating capacity construction), it remains no more than a considerable simplification of reality. Thus the analytical results of the modified model cannot claim to add significantly to an understanding of power grid economics. It should be apparent, however, that this model does allow for several factors, either known or expected to be of crucial importance in analyzing the economics of power grid formation, which are omitted from the B model. In view of the contrasting results from the modified model and the B model, the latter model must be considered highly incomplete, offering misleading rules for optimal pricing and capacity cexpansion. But such a conclusion must therefore imply that the two stochastic models discussed subsequently by Professor Berkowitz suffer from the same criticism, since they build directly upon his deterministic model. A glance at equations (7), (10), (17a) and (17b), Table 2, and the pricing and production theorems in Berkowitz (1977) should suffice to convince the reader that the incomplete structure of the deterministic B model is perpetuated in the analytical results of the stochastic models.

## Concluding Remarks

Modelling electric power systems is a highly complex task, requiring consideration of numerous technical factors associated with power supply, a careful treatment of the timing and location of power consumption, generation and capacity expansion, and utilization of the familiar economic concepts of investment analysis. The modified power grid model of this paper represents only one step in the direction of capturing reality in an economic model for examining optimal pricing and capacity expansion rules, and it too suffers from the criticism that it fails to take account of numerous and important factors that bear upon the problem of system planning, as mentioned briefly at the end of the subsection on model formulation. Yet it does allow for several considerations known to be important for the expansion of electric power systems or the formation of a power grid, and therefore has the ability to demonstrate why the simpler deterministic model of Berkowitz must be regarded as incomplete and hence must yield results that cannot be considered applicable for decisions concerning pricing, transmission and capacity expansion in an electric power grid. By inference the results derived in a stochastic setting likewise cannot be considered applicable. This judgment concerning the results derived by Berkowitz rests primarily on the absence of a long term perspective from his model and his lack of allowance for: (i) multiple capacity expansion alternatives for each producer, (ii) differing power losses among pairs of grid members, (iii) existing generating and transmission capacity, (iv) separation of the decisions to construct generating and transmission capacity, and

(v) representation of the inherent scarcity of low cost capacity expansion alternatives. But might this view be regarded as true yet without substance, because the approach of Professor Berkowitz is really intended only as a high-level abstraction of an electric power grid from which several propositions can be be deduced? Since the author discusses important real problems and issues facing electric utilities at the outset of his paper, his work "represents an effort to integrate the peak-load properties of electricity with the economic studies of time-space equilibrium" and his model and analysis throughout emphasize "high levels of co-ordination" among grid members, it is clear that his approach is believed to constitute more than a high-level abstraction.

Should it be necessary to examine the issues of peak loads and uncertainty within the context of a highly integrated power grid, I would suggest that it be done in a framework similar but not necessarily identical to that of the modified model, recognizing the simplifications inherent even in this model. Acquaintance with observations, methods and analysis by Turvey and Anderson (1977, Chapters 13 and 14) and thoughts by Kleindorfer (1977) should also prove helpful in this task. Admittedly, given the complexity of the algebraic derivation and manipulation necessary, such analysis certainly will not constitute a trivial exercise.

Conversely, I do believe that Berkowitz is right in asserting that rate-of-return regulation for private electric utilities may well act to impede interutility capacity sharing and the consequent exploitation of potential economies. Indeed he could add that public ownership and operation need not by itself lead to co-operation in power supply. In the Canadian context, legislation providing the mandates for electric power supply from publicly operated electric utilities apparently has failed to emphasize adequately the importance of interutility co-operation, possibly because major stress in the mandates inevitably has been placed upon supply to those specific regions falling within the compass of the utilities mandated. For instance, public utilities with attractive expansion alternatives evidently have been loath to exploit those alternatives rapidly in an effort to benefit neighbouring utilities as well as themselves, preferring instead to hold the alternatives in reserve for their own supply.[18] From a slightly different perspective, the activities of (i) co-ordinating supply planning, (ii) determining that supply economies do exist (with a high degree of certainty) and that they can be attained, and (iii) negotiating the distribution of benefits and costs of joint supply, and the partial surrender of sovereignty in decision-making, all involve substantial costs to prospective grid members and hence may constitute other formidable barriers to co-operation in supply. Whatever the reasons, I doubt that much progress will be made in overcoming the hurdles blocking co-operative expansion planning between two or more neighbouring electric utilities in Canada until persuasive numerical evidence is forthcoming which confirms that joint planning is not only desirable in theory but truly rewarding in practice.

# APPENDIX

## Environmental Costs in the Modified Model

One of several ways in which the dual of the modified model can be extended simply for improved understanding of factors influencing grid decisions is to include environmental costs, heretofore neglected, in the model. Since essentially all of the formal analysis for this task has already been carried out, exploration of the consequences of incorporating environmental costs is particularly easy to do.

Suppose first that all environmental damage associated with power supply can be quantified solely in terms of costs that are directly proportional to power generation by each grid member. "Real" operating costs to society are therefore equal to the initial actual operating costs augmented on a per MW basis by environmental premiums $e^i_{rtk}$, which may be thought of as costs of pollution abatement. With this change, pricing rule (12) remains intact, but (11) is amended to:

$$(16) \qquad p^i_{tk} = a^i_{rtk} + e^i_{rtk} + \eta^i_{rtk} \qquad \text{for} \qquad U^i_{rtk} > 0$$

What consequences does this hold for the grid? If all generating capacity employed is utilized fully and the environmental premiums are "small enough", there *may* be no change in optimal supply, price and grid expansion. In this polar case the only consequence is that at the optimum, producers' surplus declines. Almost certainly, however, some or all of the grid members will reduce the optimal quantity of power sold to their customers during most or all subperiods of each time period, and for the grid as a whole, aggregate capacity expansion will shrink over the planning period. Although this result is not easily derived formally because of the complexity of the grid network structure and the possibilities for substitution among different producers that it allows, the result can be seen heuristically as follows. If grid member $j$ is enormously efficient in power generation relative to all other members and consequently never requires import of power, then its price for power consumption will initially be set by the operating cost of the marginal plant $z$; i.e., $p^j_{tk} = a^j_{ztk}$, as shown earlier. With the addition of the environmental premium $e^j_{ztk}$, price will rise to $a^j_{ztk} + e^j_{ztk}$ provided that $z$ remains the marginal plant, in turn implying because of the monotonicity of $p^j_{tk}(D^j_{tk})$ that the optimal quantity of power supplied to the customers of grid member $j$ must decrease. If member $j$ is an exporter of power to other members, as may well be the case because of its dominant position with regard to generation, this price increase *may* fan outward to other members through equation (12), in turn implying a rise in prices for those members and consequent decreases in the optimal quantities of power to provide. The difficulty, of course, lies in specifying precisely the price and supply consequences for these other grid members, since it is possible that the power price increase may be absorbed in the decline in imputed values $\mu^i_{jtk}$ of transmission capacity linking members $i$ with $j$ (if transmission capacity is fully utilized), or the price increase may largely be offset if alternative sources of supply (with small or negligible environmental costs) from other grid members can be found. In any event, consequences can easily be conjectured: grid members penalized relatively less by environmental costs will be called upon to export more than before or certainly no less, whereas the opposite will be true for those members penalized relatively more heavily than before; furthermore, the optimal capacity expansion programs for grid members and the phasing of capacity expansion will likely be differentially affected by the relative sizes of environmental costs associated with generation of different members and different types.

It should be clear that similar kinds of consequences will be determined if the environmental premium is placed sequentially and separately upon costs of capacity expansion $b^i_{rt}$ (to reduce sulphur oxide emissions at coal-fired plants through introduction of stack scrubbers, for instance), costs of transmission construction $d^i_{jt}$ (reflecting possibly the burial of transmission cables or relocation of transmission corridors to accommodate public pressure groups) or greater costs for transmission utilization $c^i_{jtk}$ (to allow greater preventive maintenance for reduction of potential hazard for example), or whether some or all of these premiums are imposed together. Somewhat the same, though clearly not identical consequences for the grid, can be derived through adjustment in the allowable extent of development of particular generating capacity by reduction of one or more of the $\overline{X}$'s entering into equation (4), according as opposition emerges publicly to construction of plants which have perceived undesirable externalities. If the alternatives so restricted are precisely those which are highly desirable from a cost standpoint, then the consequences undoubtedly will be what economic intuition would dictate: higher prices for power, decreased supply in many or all time periods to the grid, minor or major changes in the spatial phasing of electric power capacity expansion, and a reduction in aggregate capacity construction over the planning period. The scarcity values $\omega^i_r$, in addition, will measure on a per MW basis the opportunity cost to the power grid of imposing policy-mandated restrictions on the availability of these alternatives.

The above qualitative discussion of the implications of incorporating environmental costs in the power grid structure has been made not because it is particularly enlightening for power grid economics but because it forms a straightforward example of the types of analysis possible using primal and dual programs of the modified model. Without question, the consequences of including environmental costs associated with power generation in a grid framework would better be analyzed in the context of a model which allowed relaxation of some of the more stringent assumptions incorporated into the modified model (as listed at the end of the subsection on model formulation).

## Footnotes

[*] For helpful suggestions on a preliminary draft of this note I extend thanks to my colleagues John Baldwin and John Hartwick.

[1] See Berkowitz (1977).

[2] For example, accounting lifetimes for hydro installations are commonly taken to be from fifty to seventy-five years, while such lifetimes for coal-fired and nuclear thermal stations similarly are taken to be anywhere from twenty-five to forty years.

[3] For instance, even at a modest compound growth rate of two percent annually, power demand a decade in future is twenty-two percent larger than current demand.

[4] In a comment upon the paper by Brown and Johnson (1969), Ralph Turvey (1970) asserted: "As for their idea about covering capacity costs via transferable future rights in telephone calls or therms of gas, it was, I assume, meant to be funny". In their reply, Brown and Johnson (1970) were not disposed to disagree. But in a later reply, Brown and Johnson (1973) reaffirmed the seriousness of their notion of sales of future rights, apparently undaunted.

[5] The modified model here described draws heavily upon the general linear model formulated in the survey article by Anderson (1972). Every effort has been made to maintain comparability in the notation with the B model but unfortunately numerous alterations and extensions have been necessary.

[6] See Anderson (1972, pp. 282, 295) for discussion of the load duration curve and its approximation.

7. Berkowitz does not specify what the units of his commodity are, megawatts or megawatt-hours. But I suspect they are megawatts also, since his demands are measured in the same units as capacity, which in the electric power industry is measured in kilowatts, megawatts or gigawatts.

8. All power demands and supplies in this paper are measured in terms of megawatts, despite the fact that demand and price normally are for megawatt-hours. Given a specified time interval, however, demand for megawatt-hours during that interval translates into demand for megawatts, the units adopted. Of course, conversion to megawatt-hours is easily made through multiplication of demand or supply by $\theta_k$, the duration in hours of subperiod $k$; for instance, demand for electrical energy at $i$ during $t,k$ when power has price $p^i_{tk}$ is given by $\theta_k \cdot D^i_{tk}(p^i_{tk})$. This approach is similar to the one taken by Turvey and Anderson (1977, p. 302) in their treatment of demand uncertainty and optimal electricity pricing.

9. Power loss is not proportional to power transmitted, but linearity is retained in the model by assuming it is; this approach is also taken in Anderson (1972, p. 292). Utilization of a nonlinear "successful power transmission" function is only one of numerous refinements that might be introduced into the model to render it more realistic.

10. The dual variable associated with (1) is really intended to be associated with this constraint rewritten as a "$\leq$" inequality, in order for the variable to be non-negative. In view of the interpretation of $\lambda^i_{tk}$ later as an electric power price, non-negativity is a highly desirable property for this Lagrangian multiplier.

11. In consequence there is no need to impose nonlinear constraints of the form $V^i_{jtk} \cdot V^j_{itk} = 0$.

12. In reality, constraints stemming from policy considerations may bear strongly upon feasible capacity expansion patterns within a grid network. For example, one or more members may be favourably endowed

with nuclear or coal-fired potential, so much so that optimal grid decisions will entail rapid and complete development of such potential for grid supply. It should be easy to see how policy considerations might rule out such grid configurations as infeasible, however, if there are uncompensated externalities associated with the planned capacity construction and utilization.

13. The sum over $j$ for the transmission capacity expansion variables in the objective function must be restricted to $j > i$ since construction of capacity between $j$ and $i$ in this model automatically implies usable capacity between $i$ and $j$.

14. See Berkowitz (1977, p. 625).

15. For definition of the dual constraint set in the case of the general nonlinear program, see Balinski and Baumol (1968).

16. Cases in which the marginal cost of generation is *exactly equal* to the cost of power import are implicitly eliminated in making this assertion.

17. This decision rule is "less analogous" than the pricing rules, however, because of the way Professor Berkowitz defines generating capacity (as dedicated to supporting generation to other grid members).

18. Particular cases of international and interprovincial co-operation, such as development of the Columbia River in British Columbia and hydro construction on the Churchill River in Labrador, might best be thought of as aberrant cases rather than examples of continuing "high level" co-operation for mutual advantage between neighbouring electric utilities.

## References

Anderson, D. (1972) "Models for Determining Least-Cost Investments in Electricity Supply". *The Bell Journal of Economics and Management Science* 3, 267-99; reprinted as Chapter 13 in Turvey and Anderson (1977).

Balinski, M. and Baumol, W. (1968) "The Dual in Nonlinear Programming and its Economic Interpretation". *Review of Economic Studies* 35, 237-56.

Berkowitz, M. (1977) "Power Grid Economics in a Peak Load Pricing Framework". *Canadian Journal of Economics* 10, 621-36.

Brown, G. and Johnson, M. (1969) "Public Utility Pricing and Output Under Risk". *American Economic Review* 59, 119-28.

Brown, G. and Johnson, M. (1970) "Public Utility Pricing and Output Under Risk: Reply". *American Economic Review* 60, 489-90.

Brown, G. and Johnson, M. (1973) "Welfare-Maximizing Price and Output with Stochastic Demand: Reply". *American Economic Review* 63, 230-1.

Kleindorfer, P. (1977) "Turvey and Anderson's *Electricity Economics*". *The Bell Journal of Economics* 8, 624-6.

Turvey, R. (1970) "Public Utility Pricing and Output Under Risk: Comment". *American Economic Review* 60, 485-6.

Turvey, R. and Anderson, D. (1977) *Electricity Economics: Essays and Case Studies* (Baltimore: The Johns Hopkins University Press).

TEL-AVIV

Working paper 26-82

# THE FOERDER INSTITUTE FOR ECONOMIC RESEARCH

TEL-AVIV UNIVERSITY<br>
RAMAT AVIV ISRAEL

מכון למחקר כלכלי ע״ש ד״ר ישעיהו פורדר ז״ל

ע״י אוניברסיטת תל־אביב

Econ.

ISSN 0316-5078

# INSTITUTE FOR ECONOMIC RESEARCH

QUEEN'S UNIVERSITY

Kingston, Ontario, Canada K7L 3N6

# INCREASING RETURNS, MONOPOLISTIC COMPETITION, AND FACTOR MOVEMENTS: A WELFARE ANALYSIS*

by

Elhanan Helpman and Assaf Razin

Working Paper No. 26-82

May, 1982

FOERDER INSTITUTE FOR ECONOMIC RESEARCH,
Faculty of Social Sciences, Tel Aviv University,
Ramat Aviv, Israel.

# INCREASING RETURNS, MONOPOLISTIC COMPETITION, AND FACTOR MOVEMENTS: A WELFARE ANALYSIS *

by

Elhanan Helpman and Assaf Razin
Tel Aviv University

## 1. INTRODUCTION

The welfare economics of international factor movements have been widely discussed in the literature. In private ownership economies factor owners choose the location of employment of their factors of production according to the highest reward and when permitted this includes locations in different countries. In competitive economies with convex technologies these private considerations coincide with social welfare (except for monopolistic considerations of large countries). As one might suspect, this coincidence of goals does not necessarily hold in economies which are characterized by industries which operate with increasing returns to scale and in which firms engage in monopolistic competition. The aim of this study is to identify the channels of influence of factor movements on social welfare which are special to such market structures, and in view of their existence to evaluate in welfare terms the performance of the private sector's decisions about the international allocation of productive resources. In the main analysis we will use foreign investment as a case study, but it should be clear that our findings apply to all factor movements except for labor migration. An analysis of labor migration requires as an input the results reported below, but since migration is guided by utility differentials rather than wage differentials it requires a separate treatment (see, for example, the discussion in Helpman and Razin (1980)).

Our main concern is with welfare aspects. For this reason we will assume that reward differentials exist (thereby inducing factor movements) without specifying the factors that generate these reward differentials. In the case discussed in this paper primary inputs can be differently priced in different countries for the same reasons that are advanced in the standard trade models (see, for example, Jones (1967)). It should only be pointed out that in the present framework they can also be differentially priced due to pure size differences among countries (see Helpman and Razin (1980)).

In order to have a benchmark for our main findings, we present in Section 2 a standard analysis of the welfare effects of capital mobility. In Section 3 we provide a detailed analysis of the effects of changes in the capital stock on a country's gross domestic product for economies with an increasing returns to scale sector in which firms engage in average cost pricing. The results of this analysis are then used in Section 4 to perform a cost-benefit analysis of international capital movements for an economy which produces differentiated products under increasing returns to scale. Concluding comments are provided in Section 5.

## 2. THE STANDARD WELFARE ANALYSIS OF FACTOR MOVEMENTS

As a prelude to our main discussion, we present in this section an analysis of welfare gains from factor movements for a competitive economy with a convex technology. For simplicity, we aggregate all traded goods into a single commodity Y and choose $p_Y = 1$ as its price. The aggregation is based on the assumption that relative prices of traded goods do not change as a result of factor movements (the small country assumption in commodity markets) in order to avoid welfare changes that result from adjustments in the terms of trade, because our main analysis sheds no new light on this issue. We also assume that there is a single nontraded good X whose price in terms of Y is p (an extension to many nontraded goods is straightforward).

Assuming the existence of a representative consumer, or a social welfare function which is maximized with costless income redistribution, our country's welfare level can be represented by an indirect utility function $v(p,GNP)$, where GNP stands for gross national product measured in units of Y, which equals net national product due to the lack of depreciation of the capital stock. Assuming that all foreign source income stems from international mobility of capital, GNP equals GDP minus rental payments on domestically employed foreign capital. Hence,

$$GNP = GDP(p,L,K+\Delta) - \rho\Delta \qquad (1)$$

where GDP(·) is the gross domestic product function (which has the standard properties of a restricted profit function as discussed, for example, in Varian (1978)), L and K stand for domestically owned labor and capital (assumed to be inelastically supplied), $\Delta$ stands for foreign capital employed in the home country when $\Delta > 0$ and domestic capital employed abroad when $\Delta < 0$. Finally, $\rho$ represents the rental rate on $\Delta$.

When foreign capital is employed in the home country $\rho = r$, where r is the domestic rental rate on capital and it equals the domestic marginal product value of capital $\partial GDP(\cdot)/\partial K$. Here the assumption is that foreignly owned capital commands the same rental rate as domestically owned capital. On the other hand, when domestic capital is employed abroad its rental rate in the foreign country is $\rho$ , which may be a function of the size of foreign investment.

Choosing a transformation of the utility function such that in equilibrium the marginal utility of income (i.e., $\partial v/\partial GNP$) equals one, differentiation of $v(\cdot)$, using (1) and the properties of the indirect utility and GDP functions yield:

$$dU = (r - \rho)d\Delta - \Delta d\rho + (X - D_X)dp$$

where X is the output level of sector X and $D_X$ is consumption of X. Since X is not traded, in equilibrium $X = D_X$, and we obtain:

(2) $$dU = (r - \rho)d\Delta - \Delta d\rho$$

Suppose that $r$ is smaller than the rental rate that domestic capital can obtain abroad. Then owners of domestic capital will shift part of it into foreign operations thereby increasing domestic welfare due to the first term on the right-hand-side of (2) (since $r < \rho$ and $d\Delta < 0$). If the foreign rental rate is unaffected by the home country's investment abroad, the second term on the right-hand-side of (2) equals zero. If, on the other hand, the foreign rental rate on domestic capital invested abroad declines with the size of the investment and we start with a positive investment level $(\Delta < 0)$, the second term generates a negative welfare effect, but this negative welfare effect is negligible for small investment levels. In the case under discussion $dU$ evaluated at $\Delta = 0$ is positive, so that it pays to invest abroad at least a little. The negative welfare effect (which doesn't exist at $\Delta = 0$) stems from monopoly power in foreign investment and we will disregard it in what follows because our main analysis sheds no new light on this particular aspect of international capital mobility.

Now suppose that $r$ exceeds the rental rate that foreign capital receives abroad. Then foreigners will invest in the home country. In this case $r \equiv \rho$ and (2) reduces to $dU = -\Delta dr$. However, due to the concavity of the GDP function in the employed levels of factors of production the rental rate on capital declines with capital inflows so that for positive investment levels $(\Delta > 0)$ welfare increases. This shows that private considerations about the location of capital coincide with social benefits in the sense that social welfare increases as a result of private decisions to shift capital to the high return location.

## 3. INCREASING RETURNS AND INCOME EFFECTS OF CAPITAL MOVEMENTS

We have seen in the previous section that in a competitive economy with convex technologies private decisions about the location of capital coincide with the goal of social welfare maximization. An important ingredient in that analysis was the effect of capital movements on GDP. In particular, an inflow of one unit of capital increases GDP by exactly the market rental rate on capital ($r = \partial GDP/\partial K$), thus making the private and social returns on capital coincide. This is achieved in a competitive economy due to marginal cost pricing.

In sectors with increasing returns to scale marginal cost pricing is incompatible with profitable production. In such cases free entry drives firms to engage in average cost pricing and indeed this assumption is common in much of the recent literature on international trade in the presence of economies of scale (see the literature surveyed in Helpman (1982)). If this is the case, an inflow of one unit of capital (or an increase in the employed capital stock due to, say, investment) will not increase GDP by the market rental rate on capital. A similar argument also applies to other factors of production. However, for every welfare analysis of factor movements their effect on GDP will be of major importance. For this reason we provide in this section the relevant analysis (which we believe to be of interest in its own right) which will be used in the next section for welfare evaluations.

The following analysis applies to models in which sectors with economies of scale are populated by firms which have identical technologies. They charge the same price and, due to free entry, engage in average cost pricing. For example, recent models of monopolistic competition in differentiated products which confine attention to symmetric equilibria satisfy these requirements (see Helpman (1982)). Assuming again that there are two goods, $X$ and $Y$, which are produced with labor and capital, where this time $Y$ is produced with constant returns to scale and $X$ is produced with increasing returns to scale, the equilibrium conditions in production can be represented as follows:

$$1 = c_Y(w,r) \tag{3}$$

$$p = C_X(w,r,x)/x \tag{4}$$

$$a_{LY}(w,r)Y + \ell_X(w,r,x)N = L \tag{5}$$

$$a_{KY}(w,r)Y + k_X(w,r,x)N = K + \Delta \tag{6}$$

where $c_Y(\cdot)$ is the marginal cost function of $Y$, $w$ and $r$ are the wage rate and the rental rate on capital, $C_X(\cdot)$ is a single firm's cost function in industry $X$, $x$ is the output level of a single firm in industry $X$, $a_{LY}(\cdot)$ $(= \partial c_Y/\partial w)$ is the employed labor-output ratio in the production of $Y$, $a_{KY}(\cdot)$ $(=\partial c_Y/\partial r)$ is the employed capital-output ratio in the production of $Y$, $\ell_X(\cdot)$ $(=\partial C_X/\partial w)$ is the employment of labor by a single firm in industry $X$, $k_X(\cdot)$ $(=\partial C_X/\partial r)$ is the employment of capital by a single firm in industry $X$, $Y$ is the output of product $Y$ and $N$ is the number of firms in industry $X$.

Equation (3) represents the condition of marginal cost pricing in the production of $Y$ (the price of $Y$ equals one) while equation (4) represents the condition of average cost pricing in the production of $X$. Equations (5) and (6) represent equilibrium conditions in factor markets. The demand for labor and capital by a firm in sector X and its cost function are not proportional to its output level $x$ due to economies of scale. In fact, the elasticity of $C_X(\cdot)$ with respect to $x$ is smaller than one, because due to scale economies the standard measure of economies of scale:

$$\theta(w,r,x) \equiv \frac{C_X(w,r,x)/x}{\partial C_X(w,r,x)/\partial x} \tag{7}$$

is larger than one.

Given a single firm's output level $x$, the price of output in the X sector $p$, and the employment of factors of production $L$ and $K+\Delta$, equations (3)-(6) provide a solution to factor prices $w$ and $r$, the output level $Y$ and the number of firms $N$ in the industry producing with increasing returns to scale. We can use equations (3)-(6) to derive a GDP function for the economy under analysis, which is an analogue of the GDP function used in the previous section. For this purpose we transform these equations as follows. Let:

$c_X(w,r;x) \equiv C_X(w,r,x)/x$ = average cost function of a firm in sector $X$

$a_{LX}(w,r;x) \equiv \ell_X(w,r,x)/x$ = labor-output ratio in sector X

$a_{KX}(w,r;x) \equiv k_X(w,r,x)/x$ = capital-output ratio in sector X

$X \equiv Nx$ = output level in sector X .

Using the new variables, equations (3)-(6) can be rewritten as:

(3') $$1 = c_Y(w,r)$$

(4') $$p = c_X(w,r;x)$$

(5') $$a_{LY}(w,r)Y + a_{LX}(w,r;x)X = L$$

(6') $$a_{KY}(w,r)Y + a_{KX}(w,r;x)X = K + \Delta$$

Equations (3')-(6') have the standard form of the production equilibrium conditions in a competitive constant returns to scale economy as long as $x$ is given. In particular, $c_X(\cdot)$ has the usual properties of a unit cost function as far as its dependence on factor prices is concerned. Moreover, $a_{LX}(\cdot) = \partial c_X(\cdot)/\partial w$ and $a_{KX}(\cdot) = \partial c_X(\cdot)/\partial r$, so that by duality there exists a sectoral constant returns to scale production function of $X$ from which $c_X(\cdot)$, $a_{LX}(\cdot)$ and $a_{KX}(\cdot)$ are derivable.[1] Hence, system (3')-(6') implies the existence of a GDP function, $GDP(p,L,K+\Delta;x)$, such that it has the usual properties with respect to

$(p,L,K+\Delta)$. In particular, $\partial GDP/\partial p = X = Nx$ $\partial GDP/\partial L = w$, $\partial GDP/\partial K = r$ and GDP is convex in $p$ and concave in $(L,K+\Delta)$. The difference between this GDP function and that used in the previous section is the dependence of the present one on $x$, the individual firm's output level. It is obvious from (5')-(6') that $x$ operates like technical progress an increase in $x$ reduces unit output costs $c_X(\cdot)$ -- because due to (7) the elasticity of $c_X(\cdot)$ with respect to $x$ is $-1 + 1/\theta(.) < 0$. Let $b = 1 - 1/\theta$ be the absolute value of the elasticity of $c_X(\cdot)$ with respect to $x$, then following the analysis of technical progress in Jones (1965) $b = \theta_{LX}b_L + \theta_{KX}b_K$ where $b_L$ is minus the elasticity of $a_{LX}(\cdot)$ with respect to $x$, $b_K$ is minus the elasticity of $a_{KX}(\cdot)$ with respect to $x$, and $\theta_{jX}$ is the share of factor $j$ in costs of production; $j = L,K$. As Jones (1965) has shown, a one percentage point increase in $x$ has the same effect on <u>output</u> levels as a b percent increase in the price $p$ plus a $\lambda_{LX}b_L$ percent increase in the labor force plus a $\lambda_{KX}b_K$ percent increase in the capital stock, where $\lambda_{LX}$ is the share of labor employed in the production of $X$ and $\lambda_{KX}$ is the share of the capital stock employed in the production of $X$.[2] This can be explained as follows. Suppose $x$ is increased by a one percentage point and the number of firms $N$ is reduced by a one percentage point so that aggregate output in sector X does not change. As a result of the increase in $x$ each firm will increase its employment of labor by $\varepsilon_{LX}$ percent, where $\varepsilon_{LX}$ is its elasticity of labor demand with respect to output, so that the sector's demand for labor will increase by $\varepsilon_{LX}$ percent. On the other hand, due to the decline in the number of firms in the industry, the industry's labor demand will fall by one percent, so that $b_L \equiv 1 - \varepsilon_{LX}$ is the proportion of the industry's labor force that is being released as a result of these changes. Since the industry employs the proportion $\lambda_{LX}$ of the total labor force, $\lambda_{LX}b_L$ is the industry's saving of labor as a proportion of the total labor force. Similarly, $\lambda_{KX}b_K$ is the proportion of total capital saved by industry X as a result of a one percent increase in $x$, holding aggregate output X constant (with the adjustment being made by means of an increase in the number of firms in the industry). In addition to these factor supply effects, a one percentage point increase in $x$ reduces unit production costs by $b$ percent.

Using the above described relationship between the effects on output levels of a one percentage point increase in $x$ and a $b$ percent increase in the price of $p$ plus $\lambda_{jX}b_j$, $j = L,K$, percent increases in the supply of factors of production, one can calculate the change in GDP as a result of a one percentage point increase in $x$ as follows:

$$\frac{\partial GDP}{\partial x}x = (p\frac{\partial X}{\partial p} + \frac{\partial Y}{\partial p})pb + (p\frac{\partial X}{\partial L} + \frac{\partial Y}{\partial L})L\lambda_{LX}b_L + (p\frac{\partial X}{\partial K} + \frac{\partial Y}{\partial K})K\lambda_{KX}b_K$$

The term in the first bracket on the right hand side equals zero (due to the standard tangency condition between the GDP line and the transformation curve), the term in the second bracket is the wage rate $w$ and the term in the third bracket is the rental rate on capital $r$. Hence, using the definition of $\lambda_{jX}$, $j = L,K$, we obtain:

$$\frac{\partial GDP}{\partial x} x = w\, a_{LX} X\, b_L + r\, a_{KX} X\, b_K = pX(\theta_{LX} b_L + \theta_{KX} b_K) = pXb$$

and

$$\frac{\partial}{\partial x} GDP(p,L,K+\Delta;x) = pN(1-\theta^{-1})$$

where use has been made of the relationships $X = Nx$ and $b = (1 - \theta^{-1})$.

Now define $r^*$ to be the increase in GDP that results from an increase in $\Delta$ holding $p$ constant. In the competitive case with convex technologies this was shown to equal $r$ -- the market rental rate on capital. In the case considered here it is:

$$r^* = \frac{\partial}{\partial \Delta} GDP(\cdot) + \frac{\partial}{\partial x} GDP(\cdot) \frac{dx}{d\Delta}$$

where $dx/d\Delta$ is a total derivative. Using the previous result this can be written as:

$$(8) \qquad r^* = r + pN(1 - \theta^{-1}) \frac{dx}{d\Delta}$$

Since $\theta > 1$ (economies of scale), (8) tells us that an inflow of one unit of capital will increase GDP by more than the market rental rate on capital if it brings about an expansion of every firm's output level in sector $X$ and it will increase GDP by less than the market rental rate on capital or even reduce GDP (as we will show) if it brings about a contraction of every firm's output level in sector X. This means that the private sector may undervalue or overvalue the marginal productivity of capital (and labor) as far as GDP is concerned, depending on its marginal effect on the size of operation of firms in the sector with economies of scale (with constant returns to scale $\theta = 1$ and $r^* = r$). However, this is but one consideration in the cost-benefit analysis of international capital movements, although it is an important one. A complete welfare analysis for an economy that produces differentiated products is presented in the next section.

## 4. DIFFERENTIATED PRODUCTS AND THE WELFARE ECONOMICS OF CAPITAL MOVEMENTS

A complete analysis of the welfare effects of international movements of factors of production in the presence of economies of scale and monopolistic competition requires a complete specification of the economy's structure. We chose to analyze an economy in which sector X produces differentiated products and we model it along the lines suggested in Lancaster (1980) and Helpman (1981). However, here we assume that Y is a composite traded good while the differentiated products are nontraded goods. The assumption of nontradedness of the differentiated products simplifies the analysis by enabling us to employ the small country assumption without having to deal explicitly with the effects of factor movements on the number of varieties supplied on world markets. Moreover, it is an assumption of interest in its own right because many services (such as restaurant meals) are nontraded differentiated products.

Following Lancaster (1979) we assume that every consumer has a utility function $u(\cdot)$ defined on the consumption level of good Y, $\alpha_Y$, and the consumption level of his most preferred differentiated product X, $\alpha_X$. We assume that these preferences can be represented by a Cobb-Douglas utility function:

$$u = s^{-s}(1-s)^{s-1} A\alpha_X{}^s\alpha_Y{}^{1-s} \quad , 0 < s < 1, \quad A > 0 \tag{9}$$

If an individual has to consume a variety which is at distance $\delta$ from his ideal product then $\alpha_X(\delta)$ units of this variety provide him with the same level of utility as $\alpha_X(\delta)/h(\delta)$ units of the ideal product, where $h(.)$ is Lancaster's compensation function. This means that the effective price a consumer pays for a unit of his ideal product is $p(\delta)h(\delta)$ if he buys for the price $p(\delta)$ a variety which is at distance $\delta$ from his ideal product. Given his income level I in terms of Y and measuring $p(\delta)$ in units of Y his demand functions are:

$$\alpha_X = \frac{sI}{p(\delta)h(\delta)}$$

$$\alpha_Y = (1-s)I$$

and his indirect utility function is:

$$v = AI[p(\delta)h(\delta)]^{-s} \tag{10}$$

All consumers are assumed to be identical except for their most preferred variety. They are, however, uniformly distributed over the set of varieties in terms of their preferences, where this set is assumed to consist of a circumference of a circle whose length is one (see Helpman (1981)).

Assuming that Y is produced with constant returns to scale while every variety in sector X is produced with the same increasing returns to scale technology, and assuming that firms in industry X engage in monopolistic competition with free entry which enforces average cost pricing, we can describe a symmetric equilibrium of this economy (in a symmetric equilibrium all varieties are equally priced and produced in equal quantities) which translates in the present case into equations (3)-(6) plus the following two conditions (see Helpman (1981)):

(11) $$R(N) = \theta(w,r,x)$$

(12) $$s(pxN + Y - \rho\Delta) = pxN$$

The production conditions (3)-(6) were discussed already. It should only be pointed out that due to the economies of scale every firm in sector X produces a different variety so that $N$ stands for both the number of firms and the number of varieties supplied by local firms. Since X-goods are not traded, $N$ is also the number of varieties that are consumed. Condition (11) stems from monopolistic competition which leads every firm in sector X to equate marginal costs to marginal revenue, and from average cost pricing. These two imply the equality of the degree of monopoly power represented by $R(\cdot)$ to the degree of economies of scale $\theta(\cdot)$. The degree of monopoly power is measured by the ratio of average revenue to marginal revenue which equals in the case of a Cobb-Douglas utility function (and a unit length of the circumference of the circle) to one plus twice the elasticity of $h(\cdot)$ evaluated at $\delta = 1/N$ (see equation (49) in Helpman (1981)). Finally, equation (12) describes the equilibrium condition in the market for nontraded goods -- proportion $s$ of GNP is spent on X-products. From the system of equations (3)-(6) and (11)-(12) we can calculate the effect of capital movements on all endogeneous variables, and in particular on $x$ which is required in order to find out whether the market rental rate on capital $r$ undervalues or overvalues the GDP effect of capital movements.

Producers in sector X supply in equilibrium N varieties. Since every product is sold for the same price, consumers whose ideal product is one of the N that are being produced are better off than other consumers. Using the indirect utility function (10), the fact that all varieties are equally priced and the fact that a proportion $1/N$ of consumers is served by a single firm in sector X, the average utility level is calculated to be:

$$AIp^{-s}N\int_0^{1/N}[h(\delta)]^{-s}d\delta$$

If the produced varieties are drawn from a uniform distribution this represents the ex-ante expected utility level of <u>every</u> consumer. Multiplying the average welfare level by $L$ and taking advantage of the accounting equation $LI = GDP - \rho\Delta$, we obtain the following measure of the economy's aggregate welfare level:

$$U = Ap^{-s}[GDP(p,L,K+\Delta\,;x) - \rho\Delta]\phi(N) \tag{13}$$

where $GDP(\cdot)$ is a function with the properties discussed in the previous section and $\phi(N) \equiv N\int_0^{1/N}[h(\delta)]^{-s}d\delta$ is an increasing function of $N$.

It is seen from (13) that the welfare effects of capital movements (a change in $\Delta$) can be decomposed into four parts; two traditional effects and two new ones. The traditional effects are the direct effect of $\Delta$

on GNP both through its effect on GDP and on repatriation payments and the indirect effect through an induced change in the price p of X-goods. These were discussed in Section 2 in which we presented the traditional analysis and we showed that the price effect is nil due to the nontradedness of X-goods. This will be shown to be true also in the present case. The new channels of influence that appear in (13) are an induced change in the scale of operation of firms in the differentiated product industry, that was discussed in detail in Section 3, and an induced change in the number of varieties that are available to consumers, whose welfare implications are similar to those of public goods.

Total differentiation of (13), using properties of the GDP(p,L,K+Δ;x) function that were derived in the previous section and the definition of r* in (8), we obtain:

$$dU = \frac{U}{\phi}\phi' dN + \frac{U}{GNP}[(r^* - \rho)d\Delta - d\rho\Delta] + \frac{U}{p}(\frac{pxN}{GNP} - s)dp$$

where $\phi' > 0$ is the derivative of $\phi$ with respect to N. Due to the equilibrium condition in the market for nontraded goods the last term -- which captures the induced price effect -- equals zero, just as in the standard analysis. Choosing the constant A so that U = GNP at the initial equilibrium point (which means that the marginal utility of income equals one), the change in welfare is:

$$dU = \frac{U}{\phi}\phi' dN + (r^* - \rho)d\Delta - d\rho\Delta \tag{14}$$

where (from (8))

$$r^* = r + pN(1 - \theta^{-1})\frac{dx}{d\Delta}$$

Comparing this equation to (2) we see immediately the two novel elements in the present welfare analysis of capital flows; the effect on the number of varieties and the difference between the social value of capital as a contributor to GDP, $r^*$, and the private value $r$, which do not coincide unless the scale of operation of firms in sector X does not change.

The above described considerations suggest that private decision to locate capital in the highest private return location may have negative social welfare effects. This is demonstrated by the following two cases.

Case 1. Suppose that a capital outflow reduces the number of varieties supplied in the investing country and it reduces the scale of operation of a representative firm in sector X (i.e., $dN/d\Delta > 0$ and $dx/d\Delta > 0$). In this case $r^* > r$. Suppose also that foreigners offer a rental rate on domestic capital $\rho$ which exceeds $r$ but falls short of the social productivity of domestically employed capital $r^*$. Disregarding the effect of foreign investment on $\rho$, it is seen that in this case private owners of capital will invest abroad ($d\Delta < 0$) bringing about a reduction of domestic welfare ($dU < 0$). The reduction of welfare stems from the fact that the rental rate on capital offered by foreigners falls short of the domestic social productivity of capital in terms of GDP and that a capital outflow makes less varieties available to consumers. Nevertheless, atomistic individuals will invest abroad because they maximize their own income.

Case 2. Suppose that a capital inflow reduces the number of varieties produced in the home country and the scale of operation of a representative firm in industry X. In this case $r^* < r$, which means that the domestic market rental rate on capital overstates its marginal product value. Suppose also that $r^* < \rho < r$. Then foreigners will find it profitable to invest in the home country (because $\rho < r$), but domestic welfare will decline because the capital inflow will reduce GNP and the number of varieties available to domestic consumers.

The two cases discussed above show that an investing country as well as a recipient country may lose from foreign investment, provided the number of varieties and the scale of operation of firms producing these varieties can respond to capital flows as indicated in the suppositions of these cases. Generally, the response of $x$ and $N$ to changes in $\Delta$ can be calculated from the general equilibrium system described by equations (3)-(6) and (11)-(12). For present purposes it is sufficient to bring examples to the effects discussed in Cases 1 and 2, which we do below.

Case 1. Let $Y$ be produced only by means of labor and let $X$ be produced only by means of capital. Let the production function of $X$ be:

$$x = \begin{cases} 0 & \text{for } k_X < \bar{\beta}_X \\ & \qquad\qquad \bar{\beta}_X, \beta_X > 0 \\ (K_X - \bar{\beta}_X)/\beta_X & \text{for } k_X \geq \bar{\beta}_X \end{cases}$$

This is a production function with increasing returns to scale which has associated with it the linear cost function:

$$C_X = r(\bar{\beta}_X + \beta_X x) \quad \text{for} \quad x > 0$$

and the measure of economies of scale:

$$\theta = 1 + \bar{\beta}_X/(\beta_X x) \quad \text{for} \quad x > 0$$

In this case equilibrium conditions (6) and (11) become:

(6a) $(\bar{\beta}_X + \beta_X x)N = K + \Delta$

(11a) $R(N) = 1 + \bar{\beta}_X/(\beta_X x)$

Choosing a compensation function $h(\delta)$ whose elasticity is increasing in $\delta$ at $\delta = 1/N$ assures that $R(N)$ declines in $N$. In this case (6.a) and (11.a) imply $dN/d\Delta > 0$ and $dx/d\Delta > 0$; i.e., a capital outflow reduces the number of varieties and the scale of operation of a representative firm, and $r^* > r$.

Case 2. Suppose that $Y$ is produced with a Lentief technology in which the input-output coefficients $a_{LY}$ and $a_{KY}$ are fixed and $X$ is produced only with labor according to the following production function:

$$x = \begin{cases} 0 & \text{for } \ell_X < \overline{\gamma}_X \\ (\ell_X - \overline{\gamma}_X)/\gamma_X & \text{for } \ell_X \geq \overline{\gamma}_X \end{cases} \qquad \overline{\gamma}_X, \gamma_X > 0$$

In this case the equilibrium conditions (5), (6) and (11) can be written as follows:

(5b) $a_{LY}Y + (\overline{\gamma}_X + \gamma_X x)N = L$

(6b) $a_{KY}Y = K + \Delta$

(11b) $R(N) = 1 + \overline{\gamma}_X/(\gamma_X x)$

It is straightforward to see that in this case $dN/d\Delta < 0$ and $dx/d\Delta < 0$, provided $R(\cdot)$ is declining in $N$, which happens when the elasticity of $h(\delta)$ is increasing in $\delta$ at $\delta = 1/N$.

Our examples show that indeed the social productivity of a factor of production can be understated or overstated by its market reward and that an expansion in the quantity of a factor of production may increase or reduce the number of varieties available to consumers. With a suitable reinterpretation of the equilibrium conditions, taking $\rho \equiv 0$, the last example can be used to produce $r^* < 0$ which shows that in a closed economy with differentiated products capital accumulation may be welfare reducing -- an immiserizing growth result. Finally, the reader should not be left with the impression that changes in the capital stock always affect $N$ and $x$ in the same direction; this is a special feature of our examples in which X-goods are produced with a homothetic production function. In Helpman and Razin (1980) there is an example with a nonhomothetic production function in which they can be affected in opposite directions.

## 5. CONCLUDING REMARKS

We have shown in this paper that in economies with sectors which produce differentiated products under increasing returns to scale, foreign investment may flow in the wrong directions thereby harming the recipient as well as the investing country. This was demonstrated by identifying two channels of influence which are special to such economies and which are not taken into account by private capital owners; the contribution of capital flows to GDP through its inducement of changes in the scale of operation of individual firms and its contribution to welfare through an inducement of changes in the number of varieties supplied to consumers. This finding has a clear policy implication -- it calls for an intervention to prevent harmful capital flows by bringing the private return on foreign investment in line with the social return, with the social return being the one derived in our cost-benefit analysis.

Although this paper deals with capital movements, the issue that is raised in it is much broader; the issue is really that in economies with increasing returns and a monopolistic market structure -- even if it is perfect competition according to Lancaster's (1979) terminology -- private valuations of productive resources do not coincide with social valuations. We have, for example, already indicated in the main text that in such economies the contribution of a factor of production to GDP may be negative and that capital accumulation may bring about a decline in welfare. However, given the market structure, one can use our techniques to compute appropriate shadow prices for policy evaluation purposes.

## FOOTNOTES

* This paper is related to our Seminar Paper No.155, Institute for International Economic Studies, University of Stockholm, but it is not merely a revision of that paper.

1. This function is implicitly defined by $F(xL_X/X, xK_X/X) = x$, where $F(\cdot)$ is the single firm's production function and $(L_X, K_X)$ are employment levels in industry X.

2. This can be easily verified by logarithmic differentiation of (3')-(6').