1972

Least Distance Estimators: A Geometric Interpretation

Ronald J. Wonnacott

Thomas H. Wonnacott

RESEARCH REPORT 7206

LEAST DISTANCE ESTIMATORS: A GEOMETRIC INTERPRETATION

by

R. J. Wonnacott

and

T. H. Wonnacott

LEAST DISTANCE ESTIMATORS: A GEOMETRIC INTERPRETATION

Least distance is an important concept in economics, because it is the principle underlying any least squares regression; in addition it is the basic concept used in some of the more complicated simultaneous equation techniques such as two stage least squares. Unfortunately, however, least distance can only be fully appreciated with some knowledge of vector geometry. Although this has appeared from time to time in enticing form in various econometric sources, including Malinvaud's text [6], these presentations have typically been addressed to mathematicians; consequently the economist with a good working knowledge of econometrics is often left with his appetite aroused but unsatisfied. This introduction to vector geometry and its applications is designed for such an audience, namely those who have had, say, a graduate course in econometrics, using some matrix algebra. In order to make this argument as brief and simple as possible, it has been necessary to forego a number of proofs; but these can be consulted elsewhere [7],[8]. The methods described geometrically include least squares, instrumental variables, generalized least squares, and two stage least squares--including a new heuristic interpretation of this technique. Finally, for the econometrician *per se*, the argument is extended in the last section to describe a new least distance limited information technique which is then compared with the other least distance estimators suggested by Brown [1], Malinvaud [6], and Zellner [9].

[1] We should like to thank Robin Carter and T. M. Brown for many helpful comments.

Since vector geometry requires breaking away from the traditional frame of reference, and into an entirely new and unfamiliar one, we illustrate with the simplest possible example: the regression

$$y = \beta x + e$$

The usual assumptions[1] are made about the error term e. In addition, to keep the geometry simple, note that we assume prior knowledge that the intercept term in this model is zero; also suppose that to estimate β there is a sample of only two observations of x and y as shown in Figure 1(a). The sample data are typically displayed in vector notation in the form:

$$\vec{y} = \hat{\beta}\,\vec{x} + \vec{\hat{e}} \tag{1}$$

i.e.,

$$\begin{bmatrix} 11 \\ 10 \end{bmatrix} = \underbrace{\hat{\beta}\begin{bmatrix} 24 \\ 13 \end{bmatrix}} + \begin{bmatrix} \hat{e}_1 \\ \hat{e}_2 \end{bmatrix} \tag{2}$$

or

$$\vec{y} = \vec{\hat{y}} + \vec{\hat{e}} \tag{3}$$

where $\vec{\hat{y}}$ represents the fitted value of $\vec{y}$. The problem is to select $\hat{\beta}$ so as to minimize, in some sense, the residuals $\hat{e}_1$ and $\hat{e}_2$. Specifically, if the familiar least squares criterion is used, we

$$\text{minimize} \quad \hat{e}_1^2 + \hat{e}_2^2 \tag{4}$$

Figure 1(b) displays exactly this same information in alternative vector geometry. Whereas in (a) each <u>observation</u> is plotted in a "variable" space (i.e., this space is defined by variables on the axes),

[1] I.e., e has zero mean, constant finite variance, and is independent of x and its own previous values.

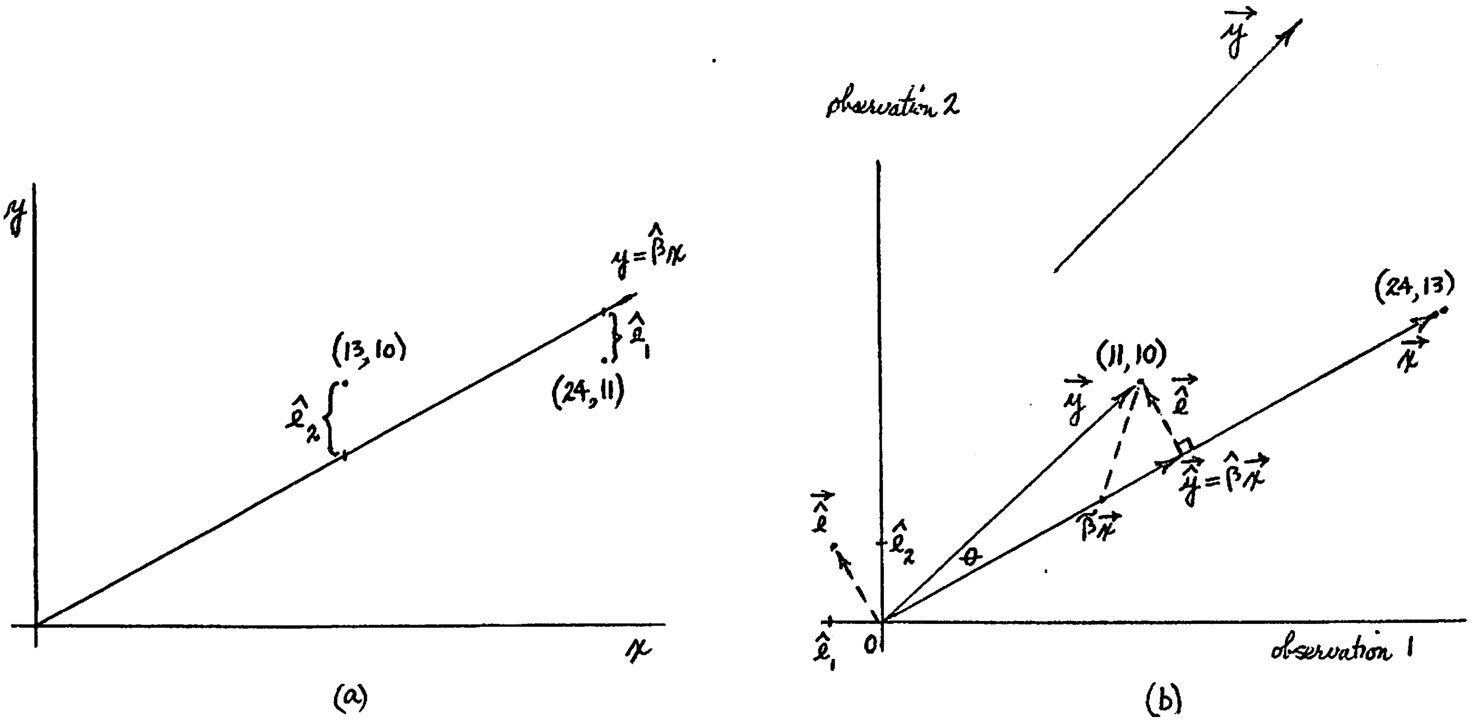


Figure 1
The simplest possible regression, using (a) standard geometry, and (b) vector geometry

in (b) each variable is plotted in an "observation" space. Accordingly, whereas a point in (a) represents a row in (2), a point in (b) represents a column.

An important warning is in order. In any practical statistical problem the number of observations will exceed the number of variables; hence (b) will be a higher-dimensioned space than (a). (Our only motive for keeping (b) a two-dimensional space is to introduce these principles with the simplest possible geometry.[1])

[1] Limiting this analysis to two sample observations in turn requires the assumption that the intercept is known to be zero; otherwise estimation would involve zero degrees of freedom (two sample values to estimate two parameters).

Since any point in Figure (b) is a vector (i.e., a column in equation (2)), it may be represented either as a point in this diagram or as an arrow. An arrow will often be convenient, since it can be shifted, provided of course that its essential characteristics of length and direction are maintained. Thus the vector $\vec{y}$ is shown shifted away from the origin at the top of this figure. On the other hand, if a vector is represented by a point, it cannot of course be shifted.

Comparing (2) and (3) we can see that the problem of selecting $\hat{\beta}$ can be viewed as a problem of selecting the fitted $\hat{\vec{y}}$, where $\hat{\vec{y}} = \beta\hat{\vec{x}}$ is a vector in the line from the origin generated[1] by $\vec{x}$.

<u>What is least squares?</u>

Intuitively, it would be desirable to select a $\hat{\vec{y}}$ in[2] $\vec{x}$ "as close as possible" to the observed $\vec{y}$, i.e., a $\hat{\vec{y}}$ that best "fits" the observed $\vec{y}$. This is shown in (b) as the result of a perpendicular (or othogonal) projection of $\vec{y}$ onto $\vec{x}$. (Note that any other fitted vector in the $\vec{x}$ line (e.g., $\tilde{\beta}\vec{x}$, would not be as close to $\vec{y}$.) The difference between the observed and fitted $\vec{y}$ vectors, $(\vec{y} - \hat{\vec{y}})$, is the residual $\hat{\vec{e}}$, shown in (b) as the

[1] You can confirm geometrically that $2\vec{x} = 2\begin{bmatrix}24\\13\end{bmatrix} = \begin{bmatrix}48\\26\end{bmatrix}$ is a vector lying in the line generated by $\vec{x}$, with twice the length of $\vec{x}$. Similarly $\hat{\beta}\begin{bmatrix}24\\13\end{bmatrix}$ is a vector lying in the line generated by $\vec{x}$, with $\hat{\beta}$ times the length of $\vec{x}$).

[2] More specifically, in the line generated by $\vec{x}$ (hereafter referred to as the "$\vec{x}$ line").

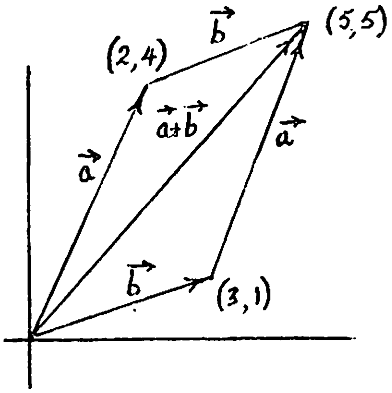
vector joining[1] $\hat{\vec{y}}$ to $\vec{y}$. $\vec{e}$ is also shown shifted to the origin, with its two coordinates confirmed as the two residual terms shown in figure 1(a).

Since the perpendicular projection of $\vec{y}$ onto $\vec{x}$ minimizes the length of $\vec{e}$ (designated as $\|\vec{e}\|$), it will also minimize the squared length $\|\vec{e}\|^2$. But from the left hand side of (b) we confirm that this may be expressed, according to the Pythagorean theorem, as

$$\|\vec{e}\|^2 = \hat{e}_1^2 + \hat{e}_2^2 = \text{sum of squared residuals}$$

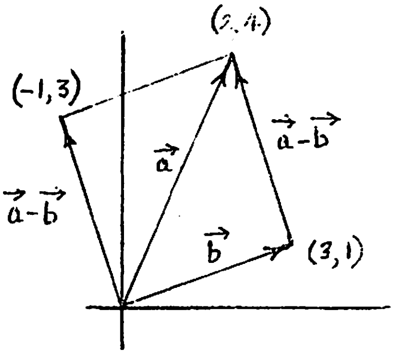
which is recognized in (4) to be what we minimize in applying least squares. Thus the vector geometry interpretation of least squares is the perpendicular

[1] An example will confirm that in vector geometry, addition of two vectors ($\vec{a} + \vec{b}$) is defined by "adding one to the other," thus



Addition $\begin{bmatrix}2\\4\end{bmatrix} + \begin{bmatrix}3\\1\end{bmatrix} = \begin{bmatrix}5\\5\end{bmatrix}$ ✓

while subtraction ($\vec{a} - \vec{b}$) is defined by joining $\vec{b}$ to $\vec{a}$ thus



Subtraction $\begin{bmatrix}2\\4\end{bmatrix} - \begin{bmatrix}3\\1\end{bmatrix} = \begin{bmatrix}-1\\3\end{bmatrix}$ ✓

projection of $\vec{y}$ onto $\vec{x}$; accordingly least squares is the "least distance" estimator.[1]

It is now appropriate to interpret $\hat{\beta}$, which (according to our fourth footnote) is the length of $\vec{\hat{y}}$ relative to the length of $\vec{x}$, or about .5 in our example. The other more familiar interpretation of $\hat{\beta}$ is the slope of the estimated regression line in figure (a); here again its value of approximately .5 is confirmed.

Finally, this vector geometry clarifies a property that is generally much more difficult to grasp using matrix algebra alone: the orthogonality (perpendicularity) of the residual vector $\vec{\hat{e}}$ and the regressor $\vec{x}$.

What is correlation?

The higher the correlation of $\vec{x}$ and $\vec{y}$, the better the regression fit and the smaller the length of $\vec{\hat{e}}$; or the tighter the angle θ between $\vec{x}$ and $\vec{y}$. A standard measure of closeness of two vectors is cos θ, which can be shown to be r, the correlation of the two variables. Thus

$$r = \cos\theta = \frac{\text{length of } \vec{\hat{y}}}{\text{length of } \vec{y}} \qquad (5)$$

In the limit, if $\vec{y}$ lies in the $\vec{x}$ line, then $\vec{\hat{y}}$ and $\vec{y}$ coincide, and $r = \cos\theta = 1$; on the other hand, if $\vec{y}$ is perpendicular to $\vec{x}$, i.e., if

$$\vec{y} \perp \vec{x} \qquad (6)$$

then the $\perp$ projection (read "perpendicular projection") of $\vec{y}$ onto $\vec{x}$ is at the origin, hence the numerator of (5) disappears, and r = 0.

[1] Not to be confused with the "least lines" estimator,(MAD), which involves minimizing the absolute deviations

$$|\hat{e}_1| + |\hat{e}_2|$$

Note that this would have a clear distance interpretation (i.e., minimize the combined lengths of $\hat{e}_1$ and $\hat{e}_2$) in figure (a), but <u>not</u> in (b).

When does least squares work best?

The answer is: "when the linear model holds, with the standard assumptions about the population error term[1] $\vec{e}$." These assumptions can be consulted in any standard text. But for our purposes we concentrate on three: the components e_1, e_2, etc., of the true error vector $\vec{e}$ are assumed to

have an expected value of zero and constant finite variance; (7)

be uncorrelated with each other; and (8)

be statistically independent of $\vec{x}$ (9)

In Figure 1 we displayed the estimation technique; now in Figure 2 it is appropriate to show in vector geometry what the underlying population is assumed to look like. The assumptions above imply that the distribution of $\vec{e}$ can be shown as the left hand sphere in this diagram; this represents a boundless cloud, thick at the center, but thinning out in the distance. This "ellipsoid of concentration" delimits most of the possible values of $\vec{e}$, some of which are illustrated as the dotted vectors. We confirm that the ellipsoid is centered on the origin because its expected value is zero (assumption (7)), and it is a sphere by virtue of the other two assumptions.

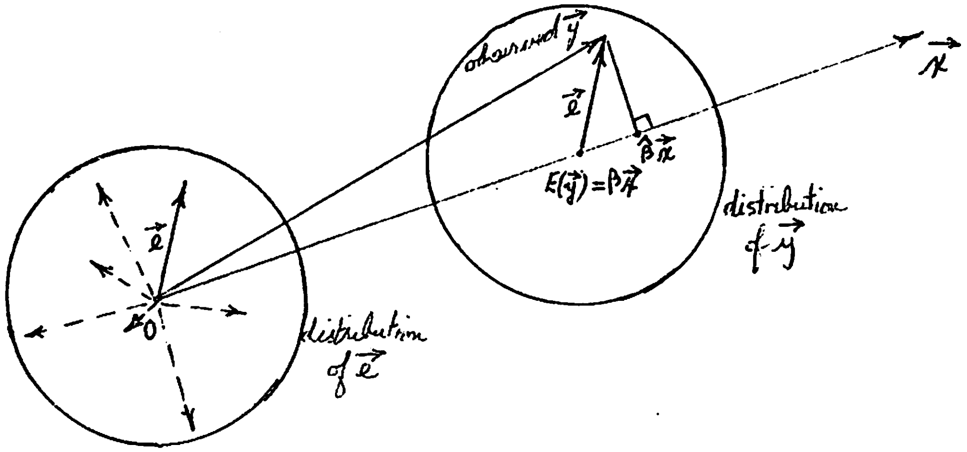

observed $\vec{y}$
$\vec{x}$
$\vec{e}$
$\beta\vec{x}$
$E(\vec{y})=\beta\vec{x}$
distribution of $\vec{y}$
$\vec{e}$
0
distribution of $\vec{e}$

Figure 2

Least Squares as an unbiased, consistent estimator

[1] Not to be confused with the estimated error $\hat{\vec{e}}$, which hereafter we call the "residual".

A typical error vector that we suppose this population has generated is shown as the solid arrow $\vec{e}$ within this sphere of concentration. Its correlation with $\vec{x}$ happens to be slightly positive; but note that some of the other possible (dotted) errors would be negatively correlated with $\vec{x}$, so that averaged over all possible errors, the correlation of $\vec{x}$ and $\vec{e}$ is zero.

Our population model is[1]

$$\vec{y} = \beta\vec{x} + \vec{e} \qquad (10)$$

Taking expectations, and noting (7),

$$E(\vec{y}) = \beta\vec{x} \qquad (11)$$

i.e., E(y) lies in the $\vec{x}$ line, as shown in Figure 2. Moreover, from (10) any observed $\vec{y}$ will be this expected value ($\beta\vec{x}$) plus the error term $\vec{e}$; the specific $\vec{y}$ which we observe in this case is accordingly shown in this figure. More generally, the distribution of all possible $\vec{y}$ values is derived by shifting the spherical distribution of $\vec{e}$ from the origin to $\beta\vec{x}$.

Now consider the statistician who cannot observe $\vec{e}$, but only $\vec{y}$ and $\vec{x}$. Applying least squares he estimates $\hat{\beta}$ by an orthogonal projection of $\vec{y}$ onto $\vec{x}$. In this case (because of the positive correlation of $\vec{e}$ and $\vec{x}$) he would overestimate β; but this procedure is just as likely to underestimate. Hence, on average $\hat{\beta}\vec{x}$ equals $\beta\vec{x}$, that is, the least squares $\hat{\beta}$ is an unbiased estimator of β.

<u>When does least squares not work?</u>

The answer is "when some of the assumptions (7), (8), or (9) do not hold." For example:

[1] recalling our prior knowledge that the usual intercept term (α) is zero.

(a) <u>If $\vec{e}$ and $\vec{x}$ are positively correlated</u>, violating assumption (9); (we continue to assume that (7) and (8) hold). The most common economic example of this occurs in simultaneous equations, where any endogenous variable appearing on the right-hand side of an equation is correlated with the error.[1]

The positive (population) correlation between $\vec{x}$ and $\vec{e}$ is shown in Figure 3; for a given $\vec{x}$, the distribution of $\vec{e}$ will tend to be in the same direction. In other words, as the dotted arrows show, it is now more likely than not that in any specific sample the model will generate a positively correlated $\vec{x}$ and $\vec{e}$.

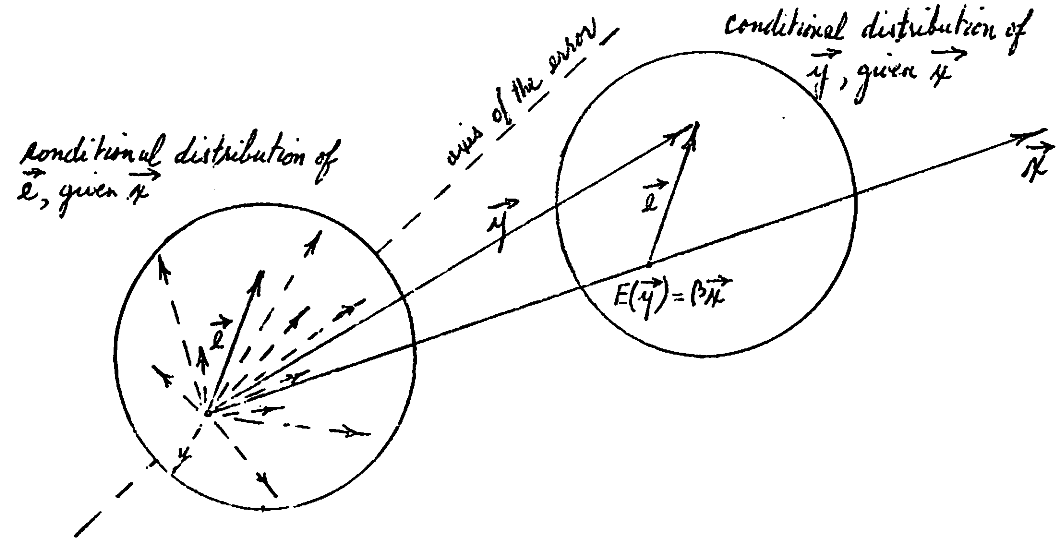

axis of the error
conditional distribution of $\vec{y}$, given $\vec{x}$
conditional distribution of $\vec{e}$, given $\vec{x}$
$\vec{x}$
$\vec{y}$
$\vec{e}$
$E(\vec{y}) = \beta\vec{x}$
$\vec{e}$
y

Figure 3

Least Squares as a biased, inconsistent estimator if $\vec{e}$ and $\vec{x}$ are correlated

[1] While we concentrate on the vector geometry of this problem, those who wish to refer to an equivalent interpretation using standard geometry may consult discussion of the simplest two-equation consumption/income model in [8], pp. 155-159. Although the models are identical, the notation differs: the variables y, x, z and e in Figure 3 correspond, respectively, to C, Y, I and e in [8].

As before, the distribution of $\vec{y}$ is defined by shifting the spherical distribution of $\vec{e}$ from the origin to $\beta\vec{x}$, as shown. Suppose the sample yields the specific $\vec{e}$, hence the observed $\vec{y}$, as shown. A least squares $\perp$ projection of $\vec{y}$ onto $\vec{x}$ would in this case yield an estimated $\hat{\beta}$ which is too large. Moreover, most such estimates will also be too large, because of the "off-center" or skewed disposition of the errors (i.e., due to the population correlation of $\vec{x}$ and $\vec{e}$). Accordingly a least-squares estimator provides a biased estimator. Moreover, increasing sample size does not cure this problem; hence least squares is also inconsistent.

How can a consistent estimator of β be obtained? The answer is the skewed projection of $\vec{y}$ onto $\vec{x}$ <u>in the direction of the axis of the error</u>, as shown in Figure 4. This sounds difficult, because the only projection we have developed so far is a perpendicular one. However this is possible if we can find another variable $\vec{z}$ (shown in Figure 4 and called an "instrumental variable") which <u>is</u> perpendicular to the axis of the error, i.e., which on average has a zero correlation with $\vec{e}$. Then a perpendicular projection onto this variable will achieve the desired result, as shown in Figure 4.

First, both variables $\vec{y}$ and $\vec{x}$ are $\perp$ projected onto $\vec{z}$, yielding the fitted vectors $\hat{\vec{y}}$ and $\hat{\vec{x}}$. Then $\hat{\beta}$ is simply the length of $\hat{\vec{y}}$ relative to $\hat{\vec{x}}$; or, by the similarity of triangles, the length of $\hat{\beta}\vec{x}$ relative to $\vec{x}$. While in this particular sample we underestimate β, we were just as likely to overestimate. Hence this "instrumental variable" (IV) technique has provided an unbiased estimator. Unfortunately, in higher-dimension cases IV does not provide exact unbiasedness, but only asymptotic unbiasedness and consistency.[1]

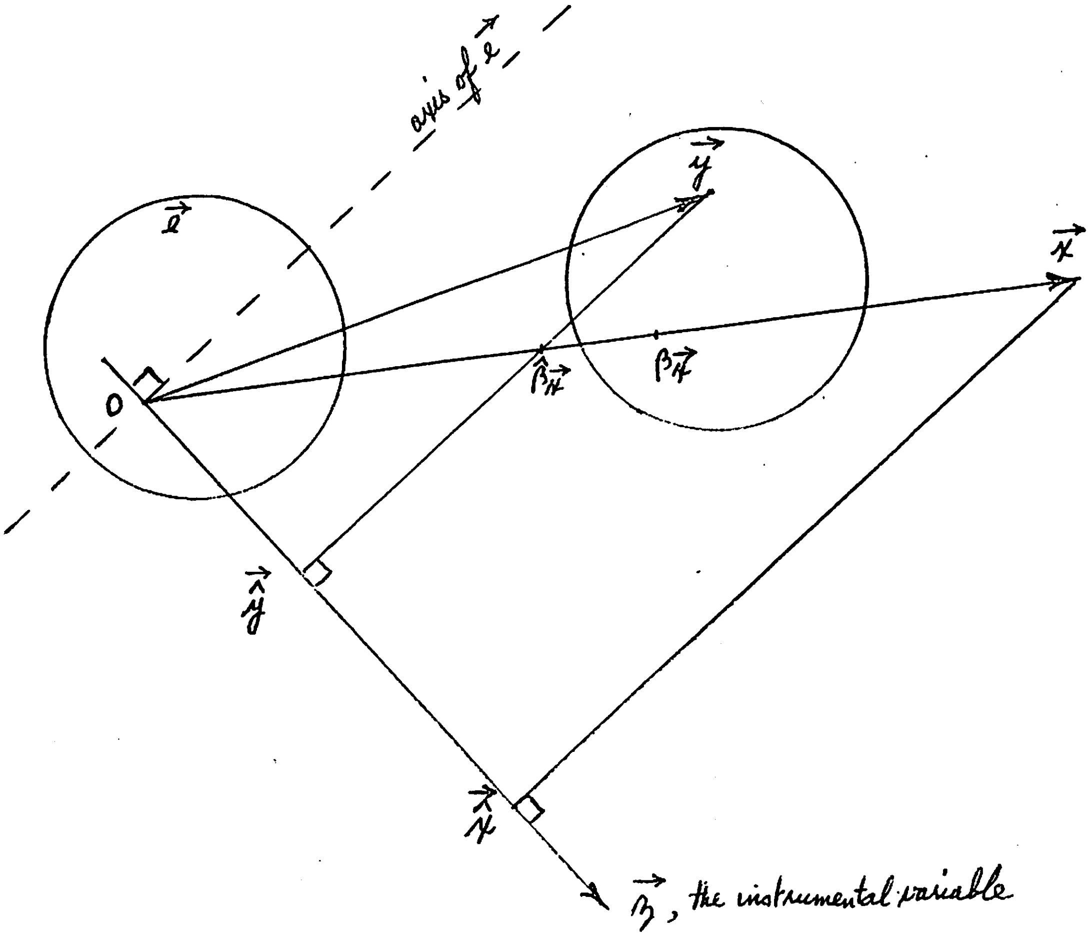


Figure 4

How an instrumental variable gives a consistent estimator when $\vec{e}$ and $\vec{x}$ are correlated

[1] Although the geometry does give a clear intuitive idea of what is going on, it glosses over another, even more subtle complication that occurs in this case. As A. L. Nagar has pointed out, ([3], Chapter 3), in such an exactly-identified case, the two-stage least squares estimator (or equivalently, the IV estimator) in Figure 4 has an infinite-variance Cauchy distribution, with no defined mean. Unbiasedness in this case can therefore not be defined in the usual sense; instead we must talk about median-unbiasedness. (But this raised no great problem, since the (Cauchy) distribution of the estimator is symmetric).

(The equivalence of IV and two-stage least squares in the exactly-identified case, alluded to above, is shown below in Figure 7).

Finally, note that IV is also a least distance projection, but in a very different sense: it is least distance to the instrumental variable, but not to the regressor.

(b) Serial correlation. In this case assumption (8) is violated, but assumptions (7) and (9) hold: in particular the error $\vec{e}$ has zero mean and is independent of $\vec{x}$. Thus its distribution can be shown before we have any information on $\vec{x}$; (in this respect it is similar to the error in Figure 2, but quite different from the error in Figure 3, which could not be shown without first having knowledge of $\vec{x}$).

With serial correlation the error configuration is the tilted ellipse shown around the origin in Figure 5. It is centered on the origin because the expected value of $\vec{e}$ is zero; it is tilted in this way because of the tendency for successive values of e to be alike, i.e., a positive first value of e is likely to be followed by a positive second value; or

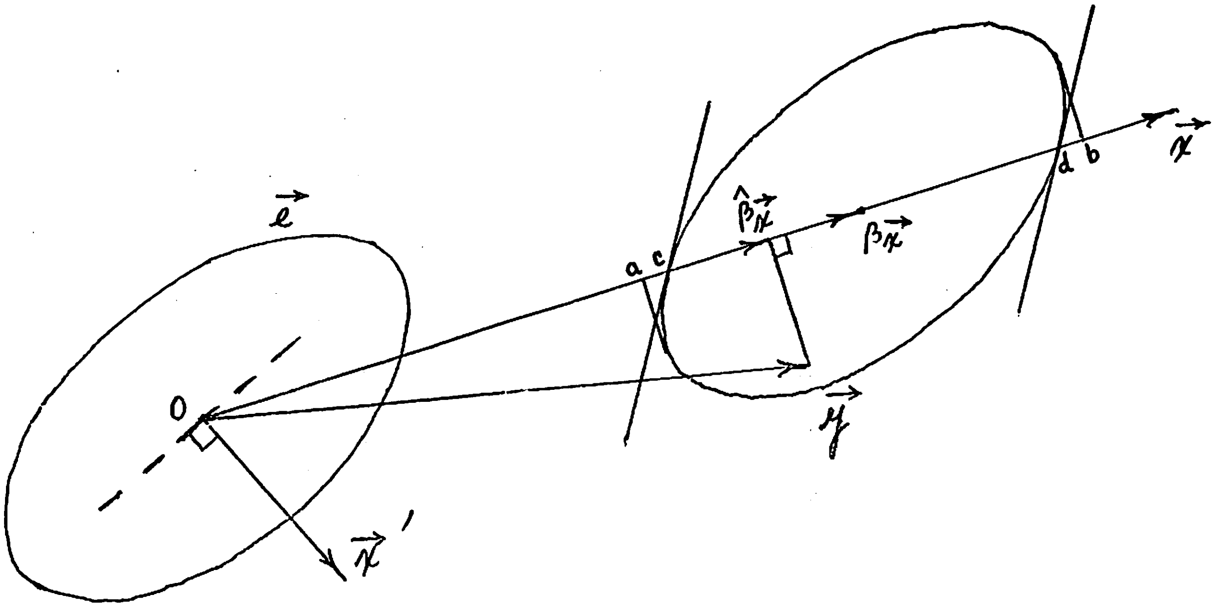


Figure 5
Estimation when there is serial correlation

a negative first value is likely to be followed by another negative value. Hence the probability is concentrated to the northeast and southwest of the origin.[1]

Now suppose $\vec{x}$ is as shown, with the distribution of $\vec{y}$ accordingly centered on $E(\vec{y}) = \beta\vec{x}$. A least squares perpendicular projection will be unbiased (i.e., as likely to yield an overestimate of β as an underestimate). In fact, a projection in any direction will be unbiased; thus the problem is not bias, but efficiency, i.e., the variance of the estimator. It is clear that least squares is not the most efficient estimator, since it yields estimates within the range[2] ab, while projection at a more appropriate angle would yield estimates in the narrower range cd. The latter is the geometric interpretation of generalized least squares (GLS). Or, more precisely, the equivalent Aitken transformation involves "stretching" the vector space so that the distribution of $\vec{y}$ becomes spherical, after which ordinary least squares (OLS) can be applied with maximum efficiency.

What is GLS worth? Noting that cd is almost as wide as ab, we might be skeptical that the Aitken transformation would provide much improvement over OLS. This has been noted in a number of other cases as well; in particular, if $\vec{x}$ were approximately perpendicular to the axis of $\vec{e}$ (like, say, $\vec{x}\,'$), then the perpendicular OLS projection could not be improved upon. Thus the advantage of GLS depends, among other things, on

[1] A negative serial correlation would, of course, result in an ellipse tilted in the other direction (running northwest and southeast).

[2] More precisely, estimates of β generated by the great majority of $\vec{y}$ values (i.e., those $\vec{y}$ values falling within the ellipse) will fall within the range ab.

the configuration of $\vec{x}$. (The importance of the $\vec{x}$ configuration is a general theme that runs through time series analysis; in particular note how the Durbin-Watson test statistic depends on it).

GLS will also provide little improvement over OLS if ρ, the coefficient of serial correlation,[1] is very small. As this approaches zero, the elliptical distribution of $\vec{e}$ approaches the spherical distribution of Figure 2, and OLS becomes a special case of GLS, and as such is the minimum variance estimator. At the other extreme as ρ approaches 1, the distribution of $\vec{e}$ tends to collapse on its major axis; in this case the geometry confirms that GLS will have very small variance relative to OLS, and hence provides a big improvement. To sum up: the smaller the coefficient of serial correlation, the smaller the advantage of GLS.

All this assumes that precise and accurate prior information on ρ exists.[2] If it does not, and ρ must somehow be estimated, then it is no longer clear that GLS will be superior to OLS; in fact, if the estimation of ρ is subject to substantial error, GLS may be inferior. The crucial importance of prior knowledge in good estimation is a general principle that is frequently encountered. As another example, instrumental variable estimation in Figure 4 requires the prior knowledge that $\vec{z}$ is uncorrelated with $\vec{e}$. If this prior knowledge is accurate, exploiting it improves the estimator; if it is false, the estimator may be worse than OLS.

[1] I.e., the correlation of e_t (the error at any time t) with its previous value e_{t-1}.

[2] It also assumes of course that GLS is correctly applied. If it is not, then the result may be worse than OLS; see [4].

Least Squares in Multiple Regression

We now return to the standard assumptions of the general linear model (i.e., spherical and independent errors, etc.), in order to now consider the problem of regressing $\vec{y}$ on two regressors $\vec{x}_1$ and $\vec{x}_2$. This problem is shown in three dimensions in Figure 6. The two vectors $\vec{x}_1$ and $\vec{x}_2$ generate a plane.[1] (Note that any vector not in this plane, such as $\vec{y}$, is designated with a shaded arrowhead, while any vector in the plane, such as $\vec{x}_1$ or $\vec{x}_2$, is designated with an empty arrowhead.) Least squares estimation remains the same as in the earlier simple regression case: the fitted value $\hat{\vec{y}}$ is derived by perpendicular projection of $\vec{y}$ onto the $(\vec{x}_1, \vec{x}_2)$ plane, with OLS again being a "least-distance estimator".

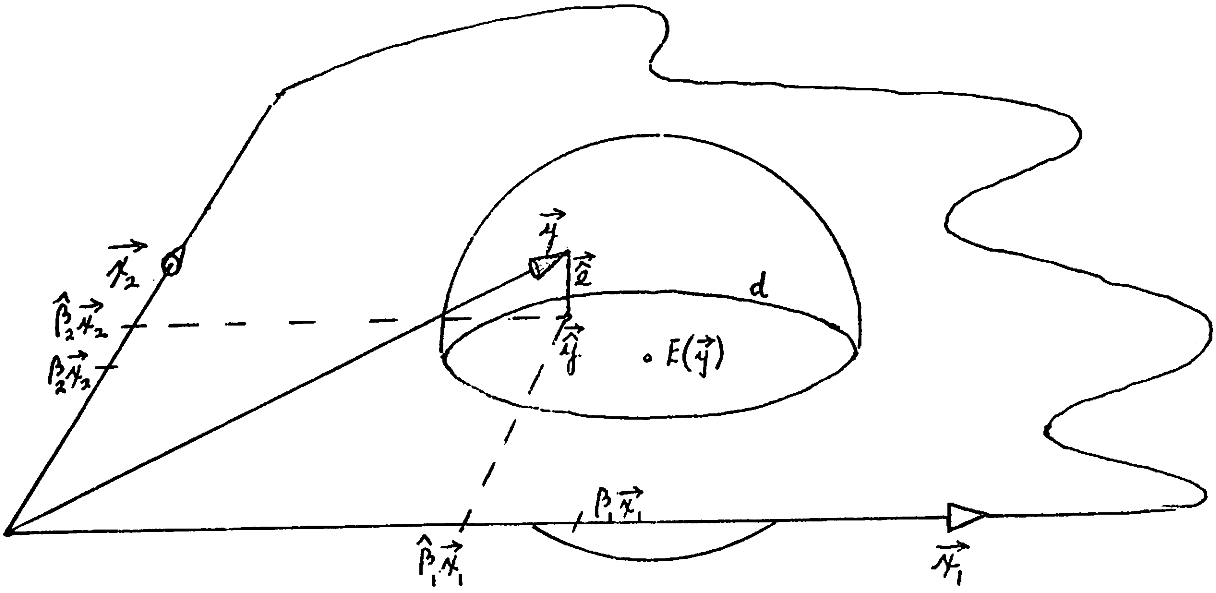


Figure 6
Multiple OLS Regression

[1] unless they are collinear--i.e., unless $\vec{x}_1$ and $\vec{x}_2$ lie on the same line from the origin. In such circumstances, $\vec{y}$ cannot be estimated as a unique function of $\vec{x}_1$ and $\vec{x}_2$, as will become evident geometrically.

Since $\vec{y}$ lies in the $(\vec{x}_1, \vec{x}_2)$ plane, it can be expressed as a linear combination of $\vec{x}_1$ and $\vec{x}_2$ as follows[1]:

$$\vec{y} = \hat{\beta}_1 \vec{x}_1 + \hat{\beta}_2 \vec{x}_2 \qquad (12)$$

In other words, once $\vec{y}$ has been estimated by a least squares $\perp$ projection, it remains only to determine $\hat{\beta}_1$ and $\hat{\beta}_2$, (the $\vec{x}_1$ and $\vec{x}_2$ coordinates of $\vec{y}$). Geometrically, this involves the projection of $\vec{y}$ onto $\vec{x}_1$ at $\hat{\beta}_1\vec{x}_1$; note that this projection within the $(\vec{x}_1, \vec{x}_2)$ plane is <u>not</u> a perpendicular projection[2], but rather a projection in the $\vec{x}_2$ direction, i.e., parallel to $\vec{x}_2$. Finally, of course, (12) requires a projection (in the $\vec{x}_1$ direction) of $\vec{y}$ onto $\vec{x}_2$ at $\hat{\beta}_2\vec{x}_2$.

Also shown in Figure 6 is the parent population, with the spherical distribution of possible $\vec{y}$ values distributed around $E(\vec{y})$, which is also in the $(\vec{x}_1, \vec{x}_2)$ plane,[3] like $\vec{y}$. Note the disc, or circular slice d of possible $\vec{y}$ centered on $E(y)$. Although in this particular case the sample $\vec{y}$ we

[1] Recall in the simple regression case that the intercept term (α) was known to be zero. We continue to assume this; hence $\vec{x}_1$ and $\vec{x}_2$ both represent variable (non-dummy) regressors. If $\vec{x}_1$ were a dummy variable (a column of 1's) then $\hat{\beta}_1$ in (12) would be interpreted as the estimated intercept.

[2] Except in the special case in which the regressors are orthogonal (i.e., $\vec{x}_1 \perp \vec{x}_2$).

[3] because $E(\vec{y})$ can be expressed as a linear combination of $\vec{x}_1$ and $\vec{x}_2$; more precisely,

$$E(\vec{y}) = \beta_1 \vec{x}_1 + \beta_2 \vec{x}_2 \qquad (13)$$

with β_1 and β_2 being the $\vec{x}_1$ and $\vec{x}_2$ coordinates of $E(\vec{y})$, as shown in Figure 6.

happened to observe led us to underestimate β_1 and overestimate β_2, it was just as likely to have been the other way around; this again illustrates the unbiased, consistent nature of OLS.[1]

Two Stage Least Squares (2SLS)

Suppose we wish to estimate the first equation in the simplest possible system of equations:

$$\vec{y}_1 = \beta \vec{y}_2 + \vec{e} \tag{14}$$

$$\vec{y}_2 = f(\vec{y}_1, \vec{z}_1, \vec{z}_2) \tag{15}$$

where $\vec{e}$ is assumed independent of the two exogenous variables $\vec{z}_1$ and $\vec{z}_2$. The problem in estimating (14) is that $\vec{e}$ cannot be assumed independent of the regressor $\vec{y}_2$; (in (14) $\vec{e}$ influences $\vec{y}_1$, while in (15) $\vec{y}_1$ in turn influences $\vec{y}_2$; thus $\vec{y}_2$ is dependent on $\vec{e}$). The problem is identical to the one encountered in the section on instrumental variables above. OLS will not provide a consistent estimate; on the other hand this can be accomplished by using an instrumental variable. Either $\vec{z}_1$ or $\vec{z}_2$ could be used, as shown in Figure 7. On the one hand, using $\vec{z}_1$ involves a $\perp$ projection of $\vec{y}_1$ and $\vec{y}_2$ onto $\vec{z}_1$, at A and B respectively, with β estimated to be OA/OB. Alternatively, using $\vec{z}_2$ involves a $\perp$ projection of $\vec{y}_1$ and $\vec{y}_2$ onto $\vec{z}_2$, at C and D respectively, with β estimated as OC/OD.

[1] We can also easily see one of the advantages of having orthogonal regressors: they produce the same estimate of β_1, regardless of whether $\vec{y}$ is regressed on both $\vec{x}_1$ and $\vec{x}_2$, or just on $\vec{x}_1$ alone. If $\vec{y}$ is regressed on $\vec{x}_1$ alone, then $\hat{\beta}_1$ is obtained by a $\perp$ projection of $\vec{y}$ onto the single vector $\vec{x}_1$. This will coincide with the multiple regression estimate $\hat{\beta}_1$ in Figure 6 only if $\vec{y}$ in that Figure is $\perp$ projected onto $\vec{x}_1$, i.e., only if $\vec{x}_2 \perp \vec{x}_1$. This argument is easily extended to show that in a model misspecification that omits a regressor ($\vec{x}_2$), bias is avoided only if $\vec{x}_1$ and $\vec{x}_2$ are orthogonal (uncorrelated).

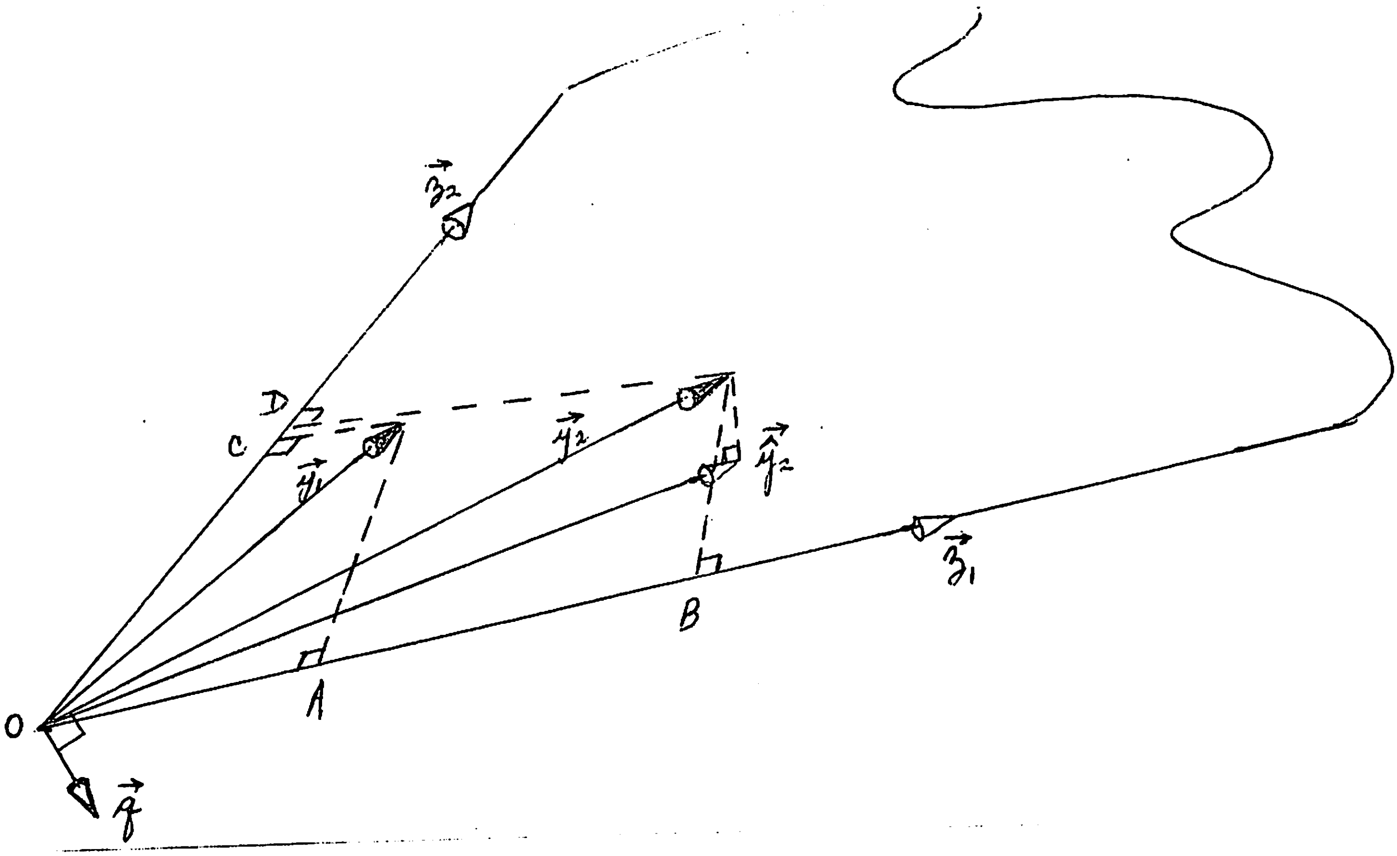


Figure 7

How alternative instrumental variables could be used to estimate β in equation (14)

Now if only one instrumental variable (say $\vec{z}_1$) appeared in (15), there would be no problem;[1] to put it another way, (14) would be exactly identified. But both $\vec{z}_1$ and $\vec{z}_2$ appear in (15), and either instrument[2] may be used,

[1] In fact the problem would then be equivalent (with the appropriate renaming of variables) to the one shown in Figure 4.

[2] In the interests of brevity, the term "instrument" hereafter replaces the more awkward "instrumental variable". Although this is a convenient abbreviation here, it should be emphasized that this conflicts with its ordinary usage in economics, where instrumental variable means any exogenously determined variable (like weather or the interest rate), while instrument refers only to the subset of these (like interest rate) that may be set by the government to achieve certain policy objectives.

with each yielding a different estimate.[1] We are faced with the dilemma of "which one to use". This "oversupply" of instruments is of course just another way of saying that this equation is overidentified.

This dilemma may be resolved in many possible ways: the 2SLS solution is to use not a single $\vec{z}$, but rather a combination of the two, (i.e., another vector in the $(\vec{z}_1,\vec{z}_2)$ plane). Now any such linear combination of two variables, each of which is independent of $\vec{e}$, will yield a new variable also independent of $\vec{e}$; thus <u>any</u> vector in the $(\vec{z}_1,\vec{z}_2)$ plane can be used as an instrument. It must however have some correlation with the regressor $\vec{y}_2$. To see why this is so, consider the vector $\vec{q}$ which is uncorrelated (perpendicular) to $\vec{y}_2$; then the perpendicular projection of $\vec{y}_2$ onto $\vec{q}$ is zero, which means there is no solution for $\hat{\beta}$. Thus we see that there are two requirements for an instrument: first, it must be uncorrelated with the error, and second, it must be correlated with the regressor[2]. In fact, the greater this correlation, i.e., the closer the instrument in the $(\vec{z}_1,\vec{z}_2)$ plane is to $\vec{y}_2$, the better; and this closest vector is $\vec{\hat{y}}$, the $\perp$ projection of $\vec{y}_2$ onto this plane. But of course this is just the fitted value that results from a least-squares $\perp$ projection of $\vec{y}_2$ onto $(\vec{z}_1,\vec{z}_2)$. So the first stage of 2SLS--namely the OLS regression of $\vec{y}_2$ on all the exogenous (instrumental) variables--is seen to be equivalent to the selection of the "best possible" instrument $\vec{\hat{y}}_2$.

[1] In an infinite sample these estimates would be identical, since the IV technique is a consistent one. But this doesn't solve the dilemma in small-sample estimation.

[2] Incidentally, it can now be seen that OLS on a single equation model (as in Figure 2) is just a special case of IV, with the regressor $\vec{x}$ being the instrument. $\vec{x}$ satisfies both the above requirements for an instrument: it is uncorrelated with $\vec{e}$, and it is highly correlated (in fact perfectly correlated) with itself.

The second stage of 2SLS, shown in Figure[1] 8 is to regress $\vec{y}_1$ onto $\hat{\vec{y}}_2$ at $\hat{\hat{\vec{y}}}_1$; then $\hat{\beta}$ is the length of $\hat{\hat{\vec{y}}}_1$ relative to $\hat{\vec{y}}_2$. This is also seen to be equivalent to using the instrument $\hat{\vec{y}}_2$ to estimate $\hat{\beta}$, as follows. The projection of $\vec{y}_2$ onto the instrument $\hat{\vec{y}}_2$ is of course the same $\hat{\vec{y}}_2$, while the projection of $\vec{y}_1$ onto $\hat{\vec{y}}_2$ is $\hat{\hat{\vec{y}}}_1$; thus $\hat{\beta}$ is again the relative lengths of $\hat{\hat{\vec{y}}}_1$ and $\hat{\vec{y}}_2$.

In summary, 2SLS has been interpreted traditionally as a first stage regression of $\vec{y}_2$ on all exogenous variables $\vec{z}_1$ and $\vec{z}_2$, yielding $\hat{\vec{y}}_2$; the second stage is a regression of $\vec{y}_1$ on $\hat{\vec{y}}_2$. Alternatively, as a number of authors (e.g., [5]) were quick to point out, 2SLS may be interpreted as instrumental variable estimation: the first stage involves selecting the best instrument $\hat{\vec{y}}_2$, while the second stage involves applying it.

There is a third useful interpretation. We might view the first stage as a least squares $\perp$ projection of <u>all</u> variables[2] into the instrument plane, (or in general, into the instrument subspace), yielding $\hat{\vec{y}}_1$ and $\hat{\vec{y}}_2$. Once in the instrument subspace, the original problem (of correlated regressor(s) and error(s)) no longer exists, and the second stage can proceed as though this problem had never arisen; thus OLS may be applied, as required by (14), to regress $\hat{\vec{y}}_1$ on $\hat{\vec{y}}_2$. (Confirm in Figure 8 that $\hat{\hat{\vec{y}}}_1$ results, regardless of whether $\vec{y}_1$ or $\hat{\vec{y}}_1$ is projected onto $\hat{\vec{y}}_2$.) This important

[1] In this diagram $\vec{y}_1$ is $\perp$ projected onto (z_1,z_2), just as $\vec{y}_2$ was in Figure 7. We now concentrate on the instrument space (plane) defined by $\hat{\vec{y}}_1$ and $\hat{\vec{y}}_2$, which is of course the same as the plane in Figure 7, since both vectors $\hat{\vec{y}}_1$ and $\hat{\vec{y}}_2$ lie in the $(\vec{z}_1,\vec{z}_2)$ plane.

[2] In our example both $\vec{y}_1$ and $\vec{y}_2$; (note that $\vec{z}_1$ and $\vec{z}_2$ are already in the instrument plane).

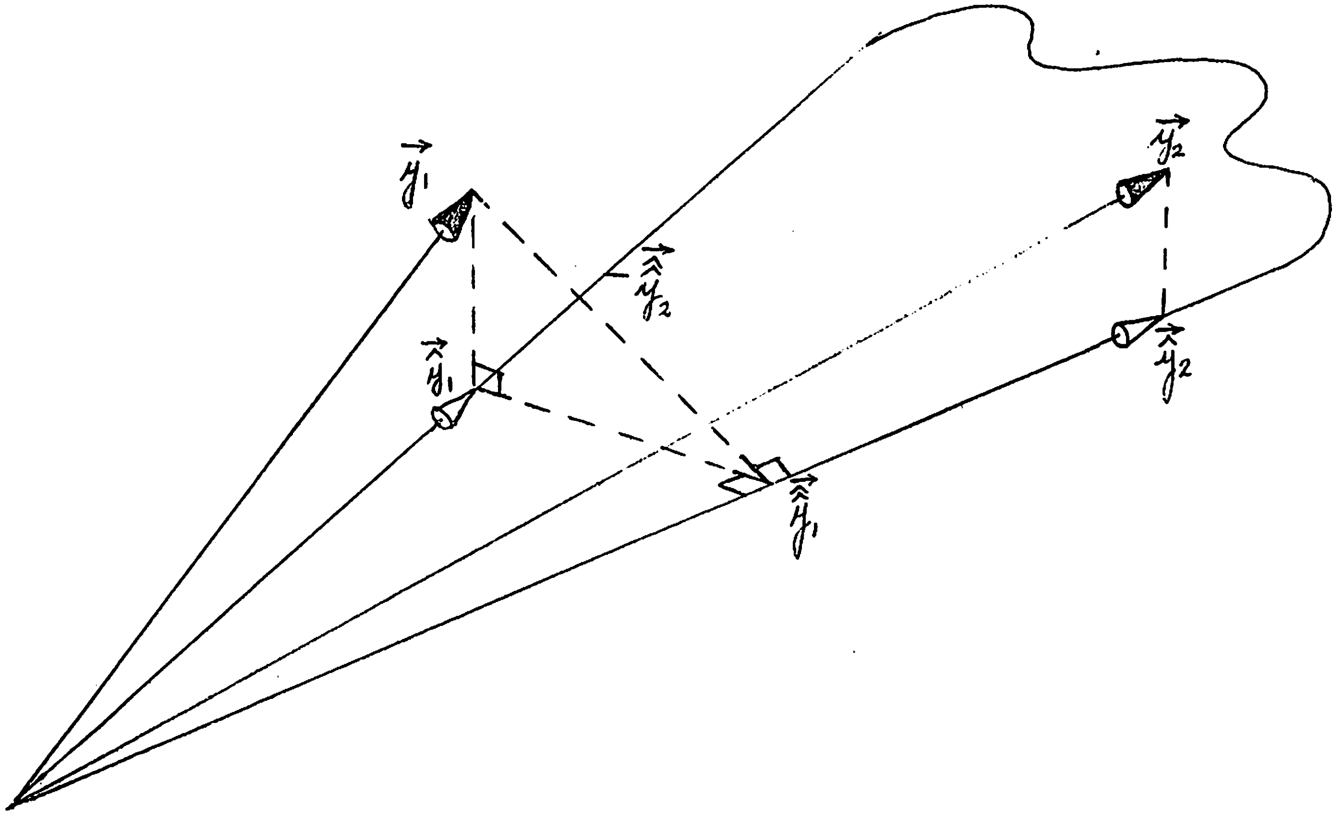


Figure 8

Various interpretations of Two Stage Least Squares

conclusion applies in general in higher-dimension cases: *a sufficient condition for consistent estimation is to first project all variables into the instrument subspace, and then proceed as though the simultaneous equation problem did not exist*. Moreover, a "least distance" interpretation of 2SLS is also clear: the first stage involves a least distance projection of all variables into the instrument subspace, while the second stage involves the appropriate least distance projection within that subspace.

It is well known that 2SLS estimation depends on the way an equation is normalized; in other words, the estimated numerical relationship between $\vec{y}_1$ and $\vec{y}_2$ would have been different had we specified that $\vec{y}_2$ was dependent on $\vec{y}_1$ in (14), rather than vice versa. Note in Figure 8 that this alternative specification would have required in the second stage that $\vec{y}_2$ be $\perp$ projected onto $\vec{y}_1$, which we denote by $\hat{\hat{\vec{y}}}_2$. There is no reason to expect that the relative lengths of $\hat{\vec{y}}_1$ and $\hat{\hat{\vec{y}}}_2$ would correspond to the relative lengths of $\hat{\hat{\vec{y}}}_1$ and $\hat{\vec{y}}_2$.

Various normalization-free procedures have been suggested; we now show one which very explicitly involves the least distance concept.

Least weighted variance (LWV)

The $\perp$ projection of $\vec{y}_1$ and $\vec{y}_2$ into the instrument plane is often referred to as a "free fit", or unrestricted OLS fit, of each $\vec{y}$ onto the reduced form; i.e., fit

$$\vec{y}_1 = \overbrace{\hat{\pi}_{11}\,\vec{z}_1 + \hat{\pi}_{12}\,\vec{z}_2}^{\hat{\vec{y}}_1} + \hat{\vec{v}}_1 \qquad (16)$$

$$\vec{y}_2 = \underset{\hat{\vec{y}}_2}{\hat{\pi}_{21}\,\vec{z}_1 + \hat{\pi}_{22}\,\vec{z}_2} + \hat{\vec{v}}_2 \qquad (17)$$

where the $\hat{\vec{v}}_i$ are the residuals. Because it is OLS, this procedure for estimating the π's involves minimizing the squared vector lengths of the residuals shown in Figure 9, i.e.,

$$\text{minimize } \|\hat{\vec{v}}_1\|^2 \text{ and minimize } \|\hat{\vec{v}}_2\|^2 \qquad (18)$$

This does not represent a final solution, however, since the structural condition (14) has not yet been imposed; taking its expected values yields

$$E(\vec{y}_1) = \beta E(\vec{y}_2) \qquad (19)$$

which means that the expected values of $\vec{y}_1$ and $\vec{y}_2$ lie on the same straight line from the origin. It seems eminently reasonable to restrict our fitted values in the same way,

$$\vec{\hat{\hat{y}}}_1 = \hat{\beta}\, \vec{\hat{\hat{y}}}_2 \qquad (20)$$

with the double hats indicating that these fitted values conform to the structural restrictions (19). This means, as shown in Figure 9, that these two fitted values must lie on a common line from the origin, say $\hat{L}$.

The problem then is to estimate the π's as we did in the free fit (16) and (17), except now we will have to select different "restricted" $\vec{\hat{\hat{y}}}_1$ and $\vec{\hat{\hat{y}}}_2$ to satisfy (20); this of course will give rise to a new restricted set of errors $\vec{\hat{\hat{v}}}_1$ and $\vec{\hat{\hat{v}}}_2$, as shown in Figure 9. Estimation involves minimizing some function of the residuals; in the spirit of least squares we might

$$\text{minimize } \|\vec{\hat{\hat{v}}}_1\|^2 + \|\vec{\hat{\hat{v}}}_2\|^2 \qquad (21)$$

with each of these components representing the squared length of that vector in Figure 9. However this would be appropriate only if we had prior knowledge that the true errors in $\vec{y}_1$ and $\vec{y}_2$ (namely $\vec{v}_1$ and $\vec{v}_2$) had equal (or very similar) variances. If we know that they don't, then following the familiar heteroscedasticity analysis, a form of weighted least squares is preferred, with the less reliable (higher variance) observations accorded less weight than the more reliable (lower variance) ones. In the absence of prior information we might ask what the data tell us; in our example, the vector $\vec{y}_2$ seems to provide more reliable information than $\vec{y}_1$. This follows because each $\vec{y}$, having its expected

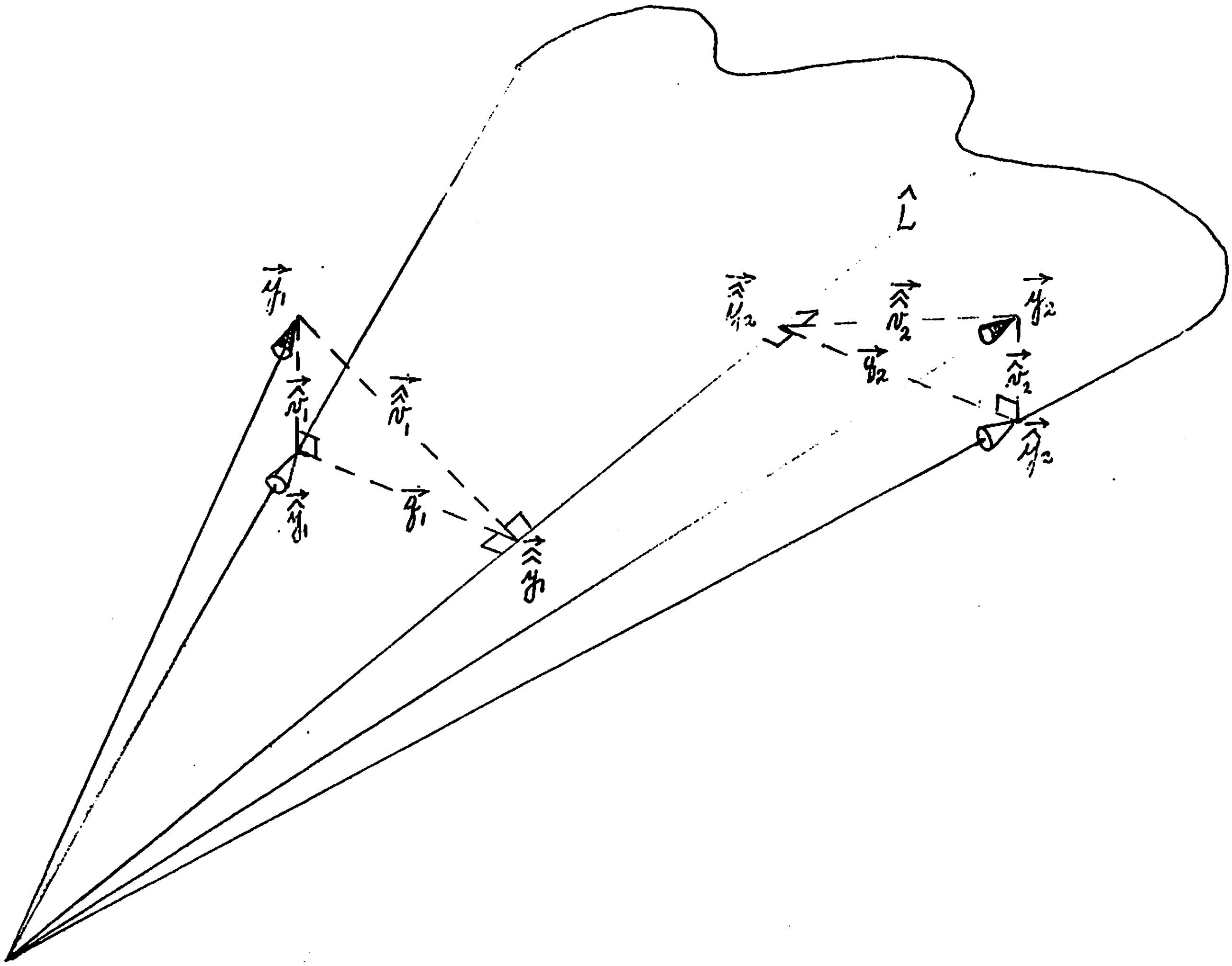


Figure 9

A weighted least distance estimator (LWV) value in the instrument plane,[1] will give some indication of its variance by its distance from the instrument plane. Accordingly, since $\vec{y}_2$ is observed relatively close to the instrument plane (the length of $\vec{\hat{v}}_2$ is less than $\vec{\hat{v}}_1$), $\vec{y}_2$ seems to yield more reliable information on the population than $\vec{y}_1$. This leads us to suggest weighting the components in (21) by their observed reliability, i.e.,

$$\text{minimize} \quad \frac{\|\vec{\hat{\hat{v}}}_1\|^2}{\|\vec{\hat{v}}_1\|^2} + \frac{\|\vec{\hat{\hat{v}}}_2\|^2}{\|\vec{\hat{v}}_2\|^2} \qquad (26)$$

which is the criterion used in Least Weighted Variance (LWV).

Unlike 2SLS, LWV does not depend on any normalization rule, since $\hat{\beta}$ is read off a unique[2] estimated line $\hat{L}$; ($\hat{\beta}$ is the length of $\vec{\hat{\hat{y}}}_1$ relative to $\vec{\hat{\hat{y}}}_2$). But like 2SLS, LWV may be interpreted as a least distance estimator:

[1] The population assumptions about the $\vec{y}$'s include (19), and

$$\vec{y}_1 = \pi_{11}\,\vec{z}_1 + \pi_{12}\,\vec{z}_2 + \vec{v}_1 \qquad (22)$$

$$\vec{y}_1 = \pi_{21}\,\vec{z}_1 + \pi_{22}\,\vec{z}_2 + \vec{v}_2 \qquad (23)$$

which are of course the population statements from which (16) and (17) derive. Taking expected values yields

$$E(\vec{y}_1) = \pi_{11}\,\vec{z}_1 + \pi_{12}\,\vec{z}_2 \qquad (24)$$

$$E(\vec{y}_2) = \pi_{21}\,\vec{z}_1 + \pi_{22}\,\vec{z}_2$$

Thus the expected value of each $\vec{y}$ can be expressed as a linear function of the $\vec{z}$'s, i.e., each lies in the instrument plane.

[2] The same interpretation of course may be made for all normalization—free limited information techniques, like Maximum Likelihood and Least Generalized Variance.

we are minimizing the squared lengths of $\vec{\hat{\hat{v}}}_1$ and $\vec{\hat{\hat{v}}}_2$, with each being a $\perp$ least-distance projection. The only twist is that each is weighted; but once again, this is according to a least-distance set of weights. Moreover, like 2SLS, LWV is a consistent estimator because it may alternatively be viewed as first a projection of all variables ($\vec{y}_1$ and $\vec{y}_2$) into the instrument subspace (at $\vec{\hat{y}}_1$ and $\vec{\hat{y}}_2$), then followed by a procedure within that instrument subspace that would be appropriate had the problem of dependence between the errors and regressors not existed.[1]

Other Least Distance Estimators

Malinvaud's least distance estimator[2] [6], does not have a simple geometric interpretation,[3] but it is a useful point of reference in comparing various other suggested methods. But first, we must clearly distinguish between the estimated covariance matrix of errors, which may be written[4]

$$\vec{W} = \begin{bmatrix}\vec{\hat{\hat{v}}}_1, \vec{\hat{\hat{v}}}_2\end{bmatrix}'\begin{bmatrix}\vec{\hat{\hat{v}}}_1, \vec{\hat{\hat{v}}}_2\end{bmatrix} \qquad (27)$$

[1] This interpretation would involve using the criterion (26) except that $\vec{\hat{\hat{v}}}_1$ and $\vec{\hat{\hat{v}}}_2$ would be replaced by $\vec{q}_1$ and $\vec{q}_2$, with $\vec{q}_1$, for example, representing the $\perp$ projection of $\vec{\hat{y}}_1$ onto $\hat{L}$. Because $\vec{\hat{v}}_1$ and $\vec{\hat{v}}_2$ are constant regardless of which method is used, estimation using this criterion yields the same results as estimation using (26).

[2] He originally formulated this as a full information technique, but we consider its single equation analogue. The following discussion, of course, could easily be cast in a full information context: many of the estimators like LWV, originally suggested in a limited information context, can be applied as full information techniques.

[3] The reason it doesn't is because it is least distance estimation only in a transformed space, the transformation involving the matrix $\vec{\Omega}$ in (29) below.

[4] Except for a common degrees of freedom divisor, which can be ignored in the minimization procedure, and is hereafter dropped in this and other similar contexts.

and the equivalent covariance matrix of <u>true</u> errors

$$\vec{\Omega} = \begin{bmatrix} \mathrm{var}(\vec{v}_1), & \mathrm{cov}(\vec{v}_1,\vec{v}_2) \\ & \\ \mathrm{cov}(\vec{v}_1,\vec{v}_2), & \mathrm{var}(\vec{v}_2) \end{bmatrix} = E\left[\vec{v}_1,\vec{v}_2\right]'\left[\vec{v}_1,\vec{v}_2\right] \qquad (28)$$

Malinvaud's method is to select the π's subject to the structural restriction[1] (19), in such a way as to

$$\text{minimize } \mathrm{tr}(\vec{\Omega}^{-1}\,\vec{W}) \qquad (29)$$

In other words, if $\vec{\Omega}$ were known, then this would provide information on the reliability of the various observed errors in $\vec{W}$, and hence could be used as in (29) to weight the elements of $\vec{W}$ in the minimization procedure.

The problem, of course, is that the true $\vec{\Omega}$ is generally not known, and must be replaced by an estimated $\vec{\hat{\Omega}}$; thus (29) becomes

$$\text{minimize } \mathrm{tr}(\vec{\hat{\Omega}}^{-1}\;W) \qquad (30)$$

with the method of estimating $\vec{\hat{\Omega}}$ being crucial. Zellner's suggestion (first proposed in a somewhat different context—see [9]) was to use the covariance matrix of freely fitted reduced form residuals, i.e., in our notation, let

$$\vec{\hat{\Omega}} = \left[\vec{\hat{v}}_1,\vec{\hat{v}}_2\right]'\left[\vec{\hat{v}}_1,\vec{\hat{v}}_2\right] \qquad (31)$$

On the other hand, LWV in (26) can be interpreted as a special case of

[1] Since this, of course, applies to all methods discussed below, we do not repeat it again.

this, where all off-diagonal terms in (31) are set equal to zero[1] (corresponding to the assumption that the true error terms are uncorrelated), and the diagonal terms in $\vec{\hat{\Omega}}$ are retained as the best available initial estimates of the error variances.[2]

T. M. Brown's simultaneous least squares (SLS), also originally presented as a full information method, (see [1]) similarly has an interesting limited information analogue.[3] Whereas LWV involves minimizing the weighted trace of $\vec{W}$, Brown's method involves minimizing its unweighted trace, (i.e., setting $\vec{\Omega} = \vec{I}$ in (29)). In line with our earlier observations, this is appropriate if the true errors $\vec{v}_1$ and $\vec{v}_2$ are known to have (approximately) the same variance (and their covariances can be assumed to be zero). This, of course, implies a scaling problem,[4] as

[1] i.e., in (31), set

$$\vec{\hat{\Omega}} = \begin{bmatrix} \|\vec{\hat{v}}_1\|^2 & 0 \\ 0 & \|\vec{\hat{v}}_2\|^2 \end{bmatrix}$$

and note that the trace elements of $\vec{W}$ in (27) are, as always $\|\vec{\hat{\hat{v}}}_1\|^2$ and $\|\vec{\hat{\hat{v}}}_2\|^2$. Under these circumstances (30) is equivalent to (26).

[2] R. A. L. Carter has suggested that this might be extended into an iterative procedure with the $\vec{\hat{\hat{v}}}_1$ and $\vec{\hat{\hat{v}}}_2$ falling out of each iteration being used to provide the $\vec{\hat{\Omega}}$ for the next iteration. In such a procedure, it would be clear how fast convergence is occurring, since in the iterative limit, $\vec{\hat{\Omega}}$ should approach $\vec{W}$, hence (30) should approach m (the dimension of this matrix).

[3] which is, of course, free of some of the controversy that has arisen about whether or not SLS is a "full information" method.

[4] Since part of $\vec{y}$ is an error, $\vec{y}$ cannot be rescaled without rescaling the error. But then Brown's assumption of equal error variances would no longer hold.

pointed out by Theil, and recognized by Brown, who also suggested a possible cure, (see [2]). The scaling problem of course does not apply to Zellner's technique or LWV, since both have their own scaling "built in" in the form of an appropriate set of weights $\vec{\Omega}$.

Finally, LGV (Least Generalized Variance, which yields a solution identical to Limited Information Maximum Likelihood) can be interpreted as minimizing

$$|\vec{W}| \tag{32}$$

Although there is no weighting of elements in this determinant this method is invariant to changes in scale: although rescaling a $\vec{y}$ variable does involve rescaling its error, this would result in a proportionate change only throughout one row and column of this determinant, and minimizing a determinant remains invariant to such changes. Goldberger and Olken have shown that this is also equivalent to Zellner's method in (30) and (31).

Geometrically, LGV may be viewed in Figure 9 as a procedure (like LWV) in which the $\vec{\hat{y}}$'s are selected in a line like $\hat{L}$—the difference now being that these fitted $\vec{\hat{\hat{y}}}$'s do <u>not</u> represent a $\perp$ projection of each $\vec{y}$ into $\hat{L}$. Such a free-swinging set of projections suggests (but of course by no means proves) that the resulting estimate of β (i.e., the relative lengths of the $\vec{\hat{\hat{y}}}$'s) may be somewhat unstable — especially in small samples. This view has received some very tentative support in recent preliminary Monte Carlo results [3].

REFERENCES

[1] Brown, T. M. "Simultaneous Least Squares: a Distribution-free Method of Equation System Structure Estimation," International Economic Review, vol. 1, 1960, pp. 173-191.

[2] Brown, T. M. "Simultaneous Least Squares and Invariance under Changes in Units of Measurement," International Economic Review, vol. 8, 1967, pp. 97-102.

[3] Brown, T. M., Nagar, A. L., Carter, R. A. L., et al. Forthcoming volume on the results of a current Monte Carlo study at the University of Western Ontario.

[4] Kadiyala, K. R. "A Transformation Used to Circumvent the Problem of Autocorrelation," Econometrica, vol. 36, No. 1 (Jan. 1968), pp. 93-96.

[5] Klein, L. R. "On the Interpretation of Theil's Method of Estimating Economic Relationships," Metroeconomica, vol. 7, Dec. 1955.

[6] Malinvaud, E. Statistical Methods of Econometrics. Chicago: Rand McNally, 1966, Chapter 9.

[7] Theil, H. Principles of Econometrics. New York: John Wiley, 1971, Chapter 1.

[8] Wonnacott, R. J., and Wonnacott, T. H. Econometrics. New York: Wiley, 1970.

[9] Zellner, A. "An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias," Journal of the American Statistical Association, vol. 57 (June, 1962), pp. 348-368.