THE EXPECTED ACCURACY OF A PRICE INDEX FOR DISCONTINUOUS MARKETS

by

Cary Webb

First draft: Feb 1981
Second draft: Oct 1981
Third draft: May 1991

THE EXPECTED ACCURACY OF A PRICE INDEX FOR DISCONTINUOUS MARKETS

Abstract. The primary purpose of this paper is to derive a priori estimates for the standard errors and autocorrelation coefficients of market price levels produced by a random walk model. The model applies to markets such as real estate where goods are traded only irregularly. The data utilized in the model is of the goods that have traded at least twice. The theory is tested with real estate data from the Arizona housing market. The size of the sample needed to achieve a specified accuracy in the price levels is also estimated.

THE EXPECTED ACCURACY OF A PRICE INDEX FOR DISCONTINUOUS MARKETS

I. Introduction

In [W1] we proposed and tested a linear model that was used to estimate market returns in real estate or any other market where the price of a given good is not determined regularly. There it was argued that under certain assumptions the estimates are unbiased and of minimal variance. The assumptions that guaranteed these conclusions are that prices of individual properties follow random walks and that these random walks are independent and identically distributed. The data analyzed in [W1] gave a certain confirmation of these assumptions.

The data to which this model applied are so-called "repeat sales" data in which the transaction prices of each good (e.g., a house) are known for at least two distinct times. Now that a great number of real estate transactions are recorded on computer, large data sets with this characteristic have become available and have materially aided in the development of price index models of the sort discussed in this paper.

The model used in [W1] was developed after studying the Bailey-Muth-Nourse model [BMN]. This is true even though it is possible to derive the model in [BMN], published in 1963, as a special case of Court's model which was published in 1938 [C]. This can be done by simply omitting the "hedonic" variables in Court's model, "suggestion number two," and retaining the time variables. In his paper Court acknowledges that Sidney W. Wilcox, Chief Statistician of the U.S. Bureau of Labor Statistics, was the person who made the original suggestion to construct such models.

Court's work notwithstanding, historically, the use of repeat sales data in linear models whose parameters are estimated by the method of least squares was initiated to the best of our knowledge by [BMN]. The model developed in this work suffers, however, from serious heteroscedasticity. Removing this heteroscedasticity is the primary practical improvement that was accomplished in [W1]. It was done, roughly speaking, by dividing each observation by a quantity that is proportional to the standard deviation of the market return over the period between sales.

Although there is abundant theoretical reason to believe in the presence of heteroscedasticity in the Bailey-Muth-Nourse model and the consequent relative unreliability of the index numbers that are estimated with it, in [W1] index numbers were estimated with the old Bailey-Muth-Nourse model as well as with the new model. The diagnostics described in [W1] suggest that the new model produces more reliable index numbers.

A disadvantage to any model is the number of assumptions that must be made in measuring its effectiveness. In [W1] there is only one assumption, the assumption that percentage changes in the prices of individual goods follow independent and identically distributed random walks in discrete time. In the past, however, many observers of real estate markets have doubted the truthfulness of this. (It is another matter, the robustness of the methods in [W1]. This latter question has not been settled.)

In [CS1] Case and Shiller acknowledge the importance of heteroscedaticity but they approach the problem of differently. Instead of an assumption like a random walk, they perform a 3-stage regression. Stage one is the same as Bailey-Muth-Nourse. Stage two is the regression of the squared residuals from stage one against the length of the holding period. This produces an estimate of the variance of the distribution from which the residual of a given observation is chosen. Stage three is the division of the observation by the square root of this estimate for variance. It is the last stage that removes the heteroscedasticity.

There are two major points to be made. Firstly, the Case and Shiller method assumes that variance is a linear function of the length of the holding period. Secondly, their method reduces to the method in [W1] if it turns out that the constant term in this function is zero.

As a matter of fact, Case and Shiller discover, at least for the data that they analyzed, that their estimate for the constant term is bounded well away from zero. Thus the two methods, in [W1] and [CS1], are, in fact, different. It has not been studied, however, to what extent the two methods produce different estimates for the price levels.

Whereas the Case and Shiller method avoids the random walk assumption, it is exactly this assumption in the method developed in [W1] which allows us in the present work to derive a priori estimates for the standard errors and autocorrelation coefficients of the index numbers and, consequently, to derive also an estimate of how large a sample is needed to produce index numbers of a specified reliability. This is accomplished, basically, by being able to construct the expected structure of the matrix X'X where X is the data matrix (called a "portfolio" matrix in Section III).

II. The Price Index Model

In this section we will recapitulate the construction of the price index model. For a more leisurely and detailed version of this construction, see [W1].

Denote by P_i the (unobserved) price level in a given real estate market in period i, i = 1,2,...,n+1. Let w_i be the logarithm of the associated price relative. That is,

$$(1) \quad w_i = \log(P_{i+1}/P_i).$$

Define w_{ij} to be the sum,

$$(2) \quad w_{ij} = \sum_{k=i}^{j-1} w_k.$$

This latter equality is just the logarithmic form of the identity,

$$(3) \quad P_j/P_i = \prod_{k=i}^{j-1} P_{k+1}/P_k.$$

The random walk assumption implies that the distributions from which the W are drawn are independent and identically distributed. As i a consequence the distributions from which the numbers $Wi/\sqrt{j-i}$ are drawn are homoscedastic, having the same variance as w_{ij}.

Suppose property q was bought in period i for price p_i and next sold in period j for price p_j. Let y_{ijq} be the logarithm of p_j/p_i divided by the square root of j-i. That is,

$$(4) \quad y_{ijq} = \log(p_j/p_i)/\sqrt{j-i}\ .$$

Then y_{ijq} is an estimate for $w_{ij}/\sqrt{j - i}$. In other words, from Equation (2),

$$(5) \quad y_{ijq} = 1/\sqrt{j - i} \sum_{k=i}^{j-1} w_k + e_q,$$

where e_q is the disturbance term. When a regression is performed it will be "weighted least squares" because of the factor, $1/\sqrt{j-i}$.

Equation (5) can be rewritten in more conventional form as,

$$(6) \quad y = \sum_{k=1}^{n} b_k x_k + e.$$

In Equation (6), y is of the form $y_{ij}/\sqrt{j-i}$, $b_k = w_k$, and x_k is the "dummy" variable defined by,

$$x_k = \begin{cases} 1/\sqrt{j-i} & , i \leq k \leq j-1 \\ 0 & , \text{otherwise.} \end{cases} \quad .$$

It is important to notice in Equation (6) that there is no constant term. Hence we cannot expect the residuals from the regression to sum to zero. There is evidence in the data that we have studied, however, that the actual mean is not significantly different from zero.

The model is not affected by inflation as long as the rate of inflation remains constant. And if the rate r changes, the effect on the variance is mitigated because it depends on the change in the square of 1 + r. For example, if the inflation rate increases from 6% to 8% from period i to period i + 1, the variance of the distribution from which the associated index number is chosen increases by about 4%. The only way for this to have a material effect on the index numbers is that this increase (or decrease) in the rate of inflation occurs in most of the periods under study (i.e., inflation--or deflation--is occurring at an exponential rate throughout most of the entire data set). Under such circumstances the correction to heteroscedasticity would be somewhat different, although not radically so, than what is described here.

Suppose there are N buy-sell pairs in the data. Then there are N equations of the form of Equation (6). We represent this system of equations with matrices:

$$(7) \quad Y = Xb + e.$$

In Equation (7), Y is an N-dimensional column vector, X is an N by n matrix, b is the n-dimensional vector $(b_1,\ldots,b_n)$, and e is the vector of disturbances.

Unique least squares estimates of the b_k's exist if and only if X is of full column rank or, equivalently, X'X is a non-singular matrix. (Here, X' represents the transpose of X.) As it happens there are two necessary and sufficient conditions on the data that guarantee the non-singularity of X'X. First of all, there must be a transaction in every period for which we want to estimate a return. Secondly, in every period at least one of the goods must be held and not sold (i.e., in every period we must be strictly between the buy-date and the sell-date of one of the observed goods). We call data that satisfy these two conditions "connected". That these two conditions are necessary and sufficient for the non-singularity of X’X is called the Connectivity Theorem. This theorem was proved in [W1].

It is simple enough to modify the model to accommodate either kind of singularity. First of all, if there is no transaction in a certain period, one merely omits the corresponding variable from the model. Of course, the regression procedure produces no estimate for the return and price level for the missing period. The estimate for the return of the succeeding period includes the return of the "missing" period.

Secondly, suppose all goods purchased prior to a certain period have either been sold or are sold in that period. Then the regression procedure must be broken into two parts. The first part estimates the returns and prices up to the pivotal period. The second part estimates the other returns and prices.

III. Basic Definitions

In order to avoid any possible ambiguity, a period will be one of the standard calendar periods (e.g., month, quarter, year) and a holding period is the interval, measured in periods, between two successive trades of a given good. For example, the holding period of a house bought in 1972 and sold in 1976 is 4 years. We denote by $n+1$ the number of periods in the interval of study.

An elementary holding period has duration one period and there are n elementary holding periods in the interval of study. It is the market returns during the elementary holding periods that our model estimates.

Occasionally it might be important to remember that the actual holding period may differ from our description of it. For example, the holding period of a house bought in 1972 and sold in 1976 is between 5 years (less one day) and 3 years (plus one day) in duration, depending upon the exact days of the transactions. Our model assumes, effectively, that all of these holding periods are precisely 4 years long. Section IX should be consulted for a discussion of how this assumption might affect the standard errors of the coefficient estimates.

Suppose the qth holding period began in period i and ended in period j. Then the i through $j-1$ entries in the qth row of the matrix X as defined in the preceding section are $1/\sqrt{j-i}$. The remaining entries are zero. We will call this matrix the portfolio matrix. It is a matrix that not only describes the composition of the collection of goods under observation at all times during the interval of study but, also, the weight (i.e., $1/\sqrt{j-i}$) appropriate to each observation that is necessary to optimize the least squares estimates.

For the qth row of X, the positive integer $d = j-i$ is called the duration (of that holding period). We denote by m, the mean duration over all holding periods.

For duration d, $f_n(d)$ will denote the probability that a randomly selected holding period has duration d. (The subscript n is necessary because the probability is also a function of the length n of the interval of study.) If $f_n(d)$ represents the sample probabilities, then m is just the mean of d over all holding periods.

IV. Model of Trading Activity

In the next section where probabilities are calculated, some of the results will depend on how the portfolio of properties whose prices are being observed changes through time. One of the purposes of this section is to model trading activity in such a way that these probabilities are as tractable as possible, yet robust enough to be useful.

Denote by M the number of all transactions in a given period. M counts all trades in the market, including those that do not end up in the two-sales data (i.e., in the portfolio matrix). The basic assumption here is that M is constant over all periods in the interval of study.

We may assume with little error that each trade in a given period will be followed at sometime by another trade of the same good. In other words, each trade can be thought of a the "buy" end of a holding period.

Of the M trades in a given period, let M_d be that number (assumed constant over all periods) which are commencing a holding period of duration d. By assumption, M is the sum of the M_d's. Hence the probability that a holding period is of duration d is M_d/M.

We have denoted by n the number of elementary holding periods in the interval of study. Thus N, the total number of observations in the two-sales data set, is the sum of n terms describing, in sequence, the number of "buys" that make it into the two-sales data from each of the first n periods. (No "buys" in the last period qualify.) More precisely,

$$N = (M_n+\ldots+M_1)+(M_{n-1}+\ldots+M_1)+\ldots+(M_2+M_1)+M_1$$

$$N = M_n+2M_{n-1}+\ldots+(n-1)M_2+M_1$$

$$(8) \quad N = \sum_{d=1}^{n} (n-d+1)M_d$$

This means that the probabilities are known. They are,

$$f_n(d) = (n-d+1)M_d/N.$$

The partition (8) of N can also be used to calculate the number of trades $t_n(k)$ in period k that wind up in the two-sales data.

$$t_n(k) = \begin{cases} \sum_{d=1}^{n} M_d & , k=1 \text{ or } n+1 \\ \sum_{d=1}^{k-1} M_d + \sum_{d=1}^{n-k+1} M_d, & 1 < k < n+1. \end{cases}$$

Note that $t_n(k) = t_n(n+2-k)$. Thus $t_n(k)$ is a symmetric function. (It will be, roughly speaking, bell-shaped if and only if $M_d > M_{d+1}$.)

The number of holding periods $p_n(k)$ actually being observed in period k (i.e., the size of the "portfolio" in period k) is closely related to $t_n(k)$:

$$p_n(k) = t_n(k+1).$$

Thus $p_n(k) = p_n(n-k)$. So $p_n(k)$ is also a symmetric function.

This model is not a particularly accurate approximation of the data studied in [W1]. For one thing, in [W1] trading activity increased throughout the interval of study. It is another question, however, the extent to which the results in this paper are robust enough to apply to that kind of data. This question is taken up later.

V. Probabilities

Suppose s is an entry in the portfolio matrix X and t is the next entry in the same row. (The only restriction is that s cannot be in the last column.) In this section we will be concerned primarily with the relation between s and t.

Let the duration of the row containing s and t be d. Then the value of s is either 0 or $1/\sqrt{d}$. The probability that s is 0 is p(0) = (n-m)/n. The probability that s is $1/\sqrt{d}$ is p(d) = $df_n(d)/n$.

The probabilities p(0) and p(d) actually are conditional upon the column of X in which s resides. The values we have given are averages over all columns. More precise values would assign greater probabilities to p(0) when the column number is close to either 1 or n. The exact probabilities conditional upon the column of s depend upon the distribution of portfolio size through time or, equivalently, the distribution of transactions contained in the two-sales data. (See the discussion of $p_n(k)$ and $t_n(k)$ in Section IV.) At the present time we are developing the theory independent of a detailed specification of these distributions. In the future it may prove desirable to include such specifications in our calculations.

The probability of t as a function of s is given in the following table.

	t=0	t#0
s=0	1-n/(n-1)(n-m)	n/(n-1)(n-m)
s#0	n/m(n-1)	1-n/m(n-1)

Comments similar to the ones in the preceding paragraph can be made. For example, p(t=0|s=0) is larger for an end column and smaller for a beginning column. This is especially true if m is large. (The difference is small when m is small.) On the other hand, p(t#0|s#0) is smaller for an end column and larger for a beginning column. This effect is also made greater if m is large and is only minimal if m is small. It should also be pointed out that these probabilities break down for very large values of m (i.e., values close to n). In any practical situation, however, m will be bounded significantly away from n.

The crucial probability for us is p(s=t). It can be derived in the following way using the preceding results.

$$\begin{aligned} p(s=t) &= p(s=t=0) + p(s=t\#0) \\ &= p(t=0|s=0)p(s=0) + p(t\#0|s\#0)p(s\#0) \\ &= (1-n/(n-1)(n-m))((n-m)/n)+(1-n/(m(n-1))m/n \\ &= (n-m)/n - 1/(n-1) + m/n - 1/(n-1) \\ &= 1 - 2/(n-2) \\ &= (n-3)/(n-1). \end{aligned}$$

This is a probability that only makes sense if n>3. Therefore this result and all subsequent ones apply only to the case n>3. It is also true that the quantity (n-3)/(n-1) can itself be derived more simply by noting that there are only two chances in n-1 that s#t.

The particular derivation given of p(s=t) is useful because it allows us to relate our comments on the conditional probabilities p(t|s) to the number (n-3)/(n-1). They imply the amount that the first term in the first line of the derivation exceeds (falls short of) the indicated value is at least partially offset by a corresponding decrease (increase) in the value of the second term. Therefore the expectation is that p(s=t) is quite closely approximated by (n-3)/(n-1) as long as n is sufficiently greater than 3.

VI. Expectations

In this section a column of the portfolio matrix X is represented by x. Our first objective is to compute the expected value of the ordinary "dot product" x•x. This product is the sum of all the squared components of x and appears as a diagonal element of X'X. Geometrically it is the square of the distance of x from the origin in N-dimensional euclidean space.

For d>0, let N_d be the number of components of x whose value is $1/\sqrt{d}$. N_d is defined for all positive integers for 1 to n. Using E to represent the expectation operator,

$$E(x\cdot x) = E\sum_{d=1}^{n}(N_d/d)$$

$$= \sum_{d=1}^{n} E(N_d)/d$$

$$= \sum_{d=1}^{n} (1/d)Ndf(d)/n$$

$$= N/n \quad \sum_{d=1}^{n} f(d)$$

$$= N/n\ (1)$$

$$= N/n.$$

We note that this value of N/n is not an approximation but is exact, given that $E(N_d)$ is independent of column. (This is, strictly speaking, not true for reasons given in Section V.) In ordinary terms, N/n is the mean size of the portfolio represented by the two-sales data set. That is, it is the mean number of properties that are between their buy-date and sell-date at any one time.

The foregoing calculation has determined the expected value of a diagonal element in the matrix X'X. As we have said, because of remarks in Section V, it is known that $E(N_d)$ is not, strictly speaking, independent of the column number of x. Therefore N/n is not the expectation of every diagonal element. It is, however, the mean expectation over all diagonal elements.

In order to approximate the off-diagonal elements, let y be a column different from x. We first assume x and y are adjacent columns. Thus x•y is an element just above (or below) the main diagonal. From Section V we know that the probability is (n-3)/(n-1) that corresponding entries of x and y are equal. Therefore, using the immediately preceding result, we can approximate,

$$E(x \cdot y) = (n-3)/(n-1)\ E(x \cdot x)$$

$$= (n-3)/(n-1)\ \ N/n.$$

The first line of this derivation is only an approximation because the same components of x and y cannot be 1's. Therefore the right hand side of this equation tends to overstate the left hand side.

Now we consider the case when x and y are not adjacent columns of X. Let s, t, and u be three successive elements in a row of X. We know that p(s=t) = p(t=u) = (n-3)/(n-1). Were it true that the events s=t and t=u are independent, it would follow that p(s=t=u) = p(s=t)p(t=u) = $[(n-3)/(n-1)]^2$. By induction it would follow,

$$(9)\quad E(x \cdot y) = N/n\ [(n-3)/(n-1)]^{|i-j|}$$

where x is the ith column of X and y is the jth column of X. Note that this expression is consistent with the earlier result for i-j = -1, 0, or 1.

Although it has not been possible to derive a mathematically precise description of the difference between the two sides of (9), we shall see that results based on this assumption are robust enough to justify it.

X'X is the variance-covariance matrix of the variables $X_1,\ldots,X_n$. (In the next section, the variance-covariance matrix of the estimates $b_1,\ldots,b_n$ is studied.) If N/n is factored out, we obtain the correlation matrix. It is of the form, $A = (a^{|i-j|})$ where a = (n-3)/(n-1).

One must be careful in interpreting the correlation coefficients given by the matrix A. This is because the value of X_j for a given observation (i.e., holding period) is determined by the value of X_i, i<j, and the length of the holding period remaining after period i. Therefore, in an important sense, the stochastic relationship that one usually expects between the variables of a (linear) model is missing.

More properly, what the correlation matrix does measure is the overlap between the portfolio being observed during the ith elementary holding period and the portfolio being observed in the jth elementary holding period. (The portfolio being "observed" during the ith elementary holding period is merely those individual holding periods that include the ith elementary period.) A high value for this correlation coefficient signifies a lot of overlap. A low value signifies little overlap. In this regard, a value of 1 implies the portfolios are identical. A value of zero implies the portfolios are disjoint. (For the particular matrix A, of course, 1's only occur on the diagonal and 0's are entirely absent.) Simple examples show, however, that in general there is no precise relationship between the correlation coefficient and the fraction of overlap.

For the model of correlation represented by the matrix A, the correlation coefficient of lag d is a^{d-1} where $0<a<1$. Therefore the composition of the portfolio is changing and tending toward being completely renewed as time passes. This is exactly our intuitive evaluation of a market.

VII. The Variance-Covariance Matrix

The variance-covariance matrix for the b_i's is $\sigma^2(X'X)^{-1}$ where σ^2 is the variance of the y variable in (6) and X is the portfolio matrix. From Section VI we know that X'X can be approximated by N/n A where the entry a_{ij} in A is $a^{|i-j|}$ and a = (n-3)/(n-1).

The inverse of A has a particularly simple form:

$$A^{-1} = 1/(1-a^2) \begin{bmatrix} 1 & -a & & & & & \\ -a & 1+a^2 & -a & & & & \\ & -a & 1+a^2 & -a & & & \\ & & \cdot & \cdot & \cdot & & \\ & & & \cdot & \cdot & \cdot & \\ & & & & \cdot & \cdot & \cdot \\ & & & & -a & 1+a^2 & -a \\ & & & & & -a & 1 \end{bmatrix}$$

where the other entries of the matrix are zeros. It is not difficult to see that the determinant of A is $(1-a^2)^{n-1}$. [K], [T] and [HJ] can be consulted for a discussion about such matrices.

The information provided by A^{-1} is summarized below. v is used as the covariance operator.

$$v(b_i,b_j) = \begin{cases} \sigma^2 n/N(1-a^2) & , i=j=1 \text{ or } n \\ \sigma^2 n(1+a^2)/N(1-a^2), & 1<i=j<n \\ -\sigma^2 an/N(1-a^2) & , |i-j| = 1 \\ 0 & , |i-j| > 1 \end{cases}$$

The most interesting feature of this result is the negative sign when $|i-j| = 1$. In other words, with a lag of 1, the estimates are negatively correlated. As n increases, it can be shown that the mean of all of the n-1 correlation coefficients of lag 1 approaches the value -0.5 as a limit. Sample values are -0.4236 (n=5), -0.5134 (n=10), and -0.4977 (n=20). (This is studied in more detail in Section X.)

With a lag greater than 1, the model predicts that the estimates b_i are uncorrelated.

VIII. Empirical Results

In this section we compare the sample variances and covariances of the estimated parameters as reported in [W1] with the variances and covariances which are predicted with the theory that has been developed in the preceding sections. In order to formulate a more comprehensive evaluation of this theory, a substantially larger data set is also analyzed. This larger data set is from Pima County, Arizona and contains 3675 buy-sell observations. On the other hand, the original data set, from Cochise County, Arizona, contained only 672 such observations (631 observations in the "edited" version). Particularly for Pima County, a very satisfactory agreement between theory and observation will obtain.

This is probably a good indication for future analysis because in the work of other investigators, data sets notably larger than our data set for Pima County have been collected and analyzed. For example in [CS1] four data sets are analyzed. The smallest contains 6669 observations, whereas the largest contains 15,530 observations. In [AS] one of the data sets analyzed contained 174,000 observations. With the continuing computerization of real estate data, there is reasonable expectation that sufficiently large data sets will be routinely available to guarantee an accuracy that will make regression methods of price level determination applicable to a wide-range of problems in the area of real estate.

The theory predicts variances and convariances as multiples of σ^2, the variance of the y variable (i.e., the log of the price relative divided by the square root of the holding period's duration). Therefore, the comparison between theory and empirical results should be between the matrix $(X'X)^{-1}$ calculated in Section VII and the matrix $(X'X)^{-1}$ arrived at in the analysis of actual data.

TABLES 1 and 2 present the pairs of matrices which correspond to the annual data sets analyzed in [W1] (i.e., Cochise County and the edited Cochise County). TABLE 3 presents the pairs of matrices which correspond to the annual data from Pima County over the same interval of time. The reader is referred to [W1] for a more complete discussion (in the case of Cochise County) of how the empirical results were developed and what the actual return estimates and price levels are.

THE VARIANCE COVARIANCE MATRIX $(X'X)^{-1}$

Cochise County, Arizona, 1971-76

Empirical

.0180	-.0050 (-.346)	-.0004	.0000	.0000
	.0116	-.0044 (-.390)	.0000	-.0000
		.0110	-.0028 (-.315)	-.0003
			.0072	-.0025 (-.346)
				.0072

Theoretical

.0099	-.0050 (-.447)	0	0	0
	.0124	-.0050 (-.400)	0	0
		.0124	-.0050 (-.400)	0
			.0124	-.0050 (-.447)
				.0099

Note. The numbers in parenthesis are correlation coefficients.

Summary	Empirical	Theoretical	Rel. Error
Mean variance	.0110	.0114	4%
Mean covariance (lag 1)	-.0037	-.0050	26%
Mean corr. coef. (lag 1)	-.350	-.424	17%
Mean covariance (lag > 1)	-.0001	0	-

TABLE 1

THE VARIANCE-COVARIANCE MATRIX $(X'X)^{-1}$

Cochise County, Arizona, 1971-76 (edited)

Empirical

.0185	-.0052 (-.350)	-.0005	.0000	.0000
	.0119	-.0046 (-.392)	.0000	-.0002
		.0116	-.0031 (-.322)	-.0003
			.0080	-.0031 (-.378)
				.0084

Theoretical

.0106	**-.0053** **(-.448)**	**0**	**0**	**0**
	.0132	**-.0053** **(-.402)**	**0**	**0**
		.0132	**-.0053** **(-.448)**	**0**
			.0132	**-.0053** **(-.448)**
				.0106

Note. The numbers in parentheses are correlation coefficients.

Summary	Empirical	Theoretical	Rel. Error
Mean variance	.0117	.0122	4%
Mean covariance (lag 1)	-.0040	-.0053	25%
Mean corr. coef. (lag 1)	-.361	-.424	15%
Mean covariance (lag > 1)	-.0002	0	-

TABLE 2

THE VARIANCE-COVARIANCE MATRIX $(X'X)^{-1}$

Pima County, Arizona, 1971-76

Empirical

.0024	-.0011 (-.417)	.0000	.0000	.0000
	.0029	-.0011 (-.457)	.0000	.0000
		.0020	-.0007 (-.391)	.0000
			.0016	-.0006 (-.387)
				.0015

Theoretical

.0018	-.0009 (-.442)	0	0	0
	.0023	-.0009 (-.391)	0	0
		.0023	-.0009 (-.391)	0
			.0023	-.0009 (-.422)
				.0018

Note. The numbers in parentheses are correlation coefficients.

Summary	Empirical	Theoretical	Rel. Error
Mean variance	.0021	.0021	0%
Mean covariance (lag 1)	-.0009	-.0009	0%
Mean corr. coef. (lag 1)	-.413	-.424	3%
Mean covariance (lag > 1)	.0000	0	-

TABLE 3

The single statistic of greatest interest is the variance of the coefficient estimate. This is measured by the appropriate diagonal entry of the $(X'X)^{-1}$ matrix. (Actually the variance is this entry multiplied by the variance of the y variable. But in computing relative differences, which is the appropriate method of measuring error in this situation, this common factor can be disregarded.)

In Tables 1-3, all of the diagonal entries for both the empirical matrices and the theoretical matrices are given for the three data sets that were analyzed. The means of these diagonal entries are also given. The means can be used as an overall measure of the reliability of the coefficient (i.e., price level) estimates.

For the two small data sets from Cochise County, one observes that there is a 4% relative difference between the mean diagonal entry of the empirical matrix and the mean diagonal entry of the theoretical matrix. (This translates into a 2% relative difference between standard errors.) We rate this as being at least modest evidence that the model is working.

For the larger data set represented by Pima County, the mean diagonal entry of the empirical matrix and the mean diagonal entry of the theoretical matrix are identical (to 2 significant digits). We rate this as stronger evidence that the model is working.

With the smaller data sets, there is a degradation of these results for the covariances between the variables. Here the relative differences between means increase to 26% and 25%, respectively. Not a good fit. But with the larger data set, this relative difference is again zero (to the precision of the computation). On balance, at least modest evidence that the model is perfoming well.

The second most interesting statistic is the correlation coefficient of lag 1 between the coefficient estimates. This is because our model leads us to expect a negative correlation coefficient that approaches -0.5 as the number of index numbers to be estimated approaches infinity (see Sections VII and X). For $n = 5$, the case for the annual index series for the Arizona data, the value predicted by the theoretical model is -0.4236. This differs from the empirical results by 15-17% for the small data sets from Cochise County both of which exhibit less than the expected negative correlation. (This accuracy is line with what we have been led to expect from this particular data.) On the other hand, it differs from the empirical results by only 3% for the larger Pima county data set. This is usually a satisfactory result for a measurement of serial correlation.

The other results produced by the variance-covariance matrix is the covariance between the variables of lag greater than 1. The theory predicts a value of zero which corresponds to the (precisely) zero values above the superdiagonal in the theoretical variance-covariance matrix. In fact, except for one value of -0.0005 for the Pima edited data, all of these entries in all three of the empirical $(X'X)^{-1}$ matrices round to zero thousandths.

One notes there are only minor differences between the empirical $(X'X)^{-1}$ matrices for the two sets of Cochise County data. The differences in the reliability of the price level estimates derives almost completely from the substantially larger standard deviation of the y variable. For the unedited data this value is 0.468, whereas for the edited data it is only 0.184. In other words, there is significantly greater variability in the prices of the unedited data. When the observations that contribute to this variability are deleted, the reliability of our computations is of course correspondingly increased.

It is interesting that there is not much difference in the standard deviations of the y variable between Pima County and the edited Cochise County. Pima has a value of 0.221 whereas, as already reported, the edited Cochise is 0.184. Thus the greater reliability of the index numbers for Pima derives from the greater number of observations in the Pima data set. Our theory predicts that the standard errors of the coefficient estimates are inversely proportional to the square root of the size of the data set. (This prediction, by itself, can be derived without relying on our model of the variance-covariance matrix.) One predicts, therefore, that the diagonal entries of the empirical $(X'X)^{-1}$ matrix for the edited Cochise data should be 3675/631 = 5.8 times as large as the corresponding entries for the Pima data. In fact, they are not far from this value, being, on average, 117/21 = 5.6 times as large (see Tables 2 and 3). And much of the discrepancy between these two ratios, only 4% in relative terms, could be explained by a rounding error in the denominator of the second ratio.

Although the results in this Section, especially for Pima County, provide good confirmation of our theory, it should be noted that the theory suggests that for larger values of n the theory should perform even better. (See Section V.) This is only to say that the primary objective in this work is theoretical and that the theory must be applied to a variety of data sets in order to have a complete evaluation of it.

IX. Analysis of Variance

In [W1] it was noted that the model produces non-zero standard errors even if the data is continuous (i.e., there is a price for every good in every period). This "is merely a measure of the dispersion of the yields of the individual properties. Hence it is a measure of how much we can expect the market return to vary over time" [W1].

Let us call the variance that would exist even if a continuous set of date were available, the "continuous" component. The remaining component is called the "discontinuous" component.

Roughly speaking, the size of the discontinuous component is measured by m, the mean duration of a holding period (Section III). That m is greater than 1 is the simplest expression of discontinuity in the data.

The purpose of this section is to measure the relative sizes of the continuous and discontinuous components more precisely. We will show that usually the discontinuous component is greater than the continuous component and, hence, most of the standard error in the coefficient estimates is due to our imperfect knowledge and not to dispersion in the yields of the individual properties.

Denote by P the number of distinct properties represented in the 2-sales data. If the data is continuous, then by what we have observed in [W1] and in Section VII, the variance of each coefficient is σ^2/P, where σ^2 is the variance of the y variable.

In Section VII the variances in the general case were computed. The mean of all of them is,

$$\sigma^2/2N\ (n^2 - 3n + 5 + 1/(n-2)).$$

Thus the fraction f_c of the total variance that is the continuous component is,

$$\begin{aligned} f_C &= (1/P)/((n^2 - 3n + 5 + 1/(n-2))/2N) \\ &= 2N/P(n^2 - 3n + 5 + 1/(n-2)). \end{aligned}$$

If $m_1 = N/P$ denotes the mean number of holding periods per property, then,

$$f_C = 2m_1/(n^2 -3n + 5 + 1/(n-2)).$$

The fraction f_D of the total variance that is the discontinuous component is simply one less f_C:

$$f_D = 1 - f_C.$$

The following Table 4 contain the values of f_D for both the annual Cochise (unedited) data and the annual Pima data.

	Annual Cochise	Annual Pima
N	672	3675
P	586	3130
m_1	1.1468	1.1741
f_C	.1496	.1531
f_D	.8504	.8469

TABLE 4

We see that with both of the data sets, 85% of the variance is the discontinuous component. Thus most of the uncertainty that we report for price levels is a result of the discontinuity of the data and not as a result of the dispersion of the returns on the individual properties that were observed.

This result becomes more dramatic when n increases because m_1 stays relatively constant whereas $n^2 - 3n + 5$ grows rapidly with n. For example, with the Cochise monthly data [W1], n = 76 and $m_1 = 1.1992$. These values produce, $f_C = 0.0004$ and $f_D = 0.9996$. In other words, virtually all of the variance is due to discontinuity of the data.

If we impute $m_1 = 1.2$, then $f_C < 0.05$ for $n > 8$ and $f_C < 0.01$ for $n > 16$. In [CS1], n = 66, and in [AS], n = 17 or n = 64. Thus we see that for at least some of the other data that has been analyzed in the literature, it is true that 99% of the variance is accounted for by the discontinuity of the data and not by the dispersion of the returns of the individual properties.

X. Negative Serial Correlation

Primary assumption has been that market returns are drawings from distributions that are independent and identically distributed. Therefore, by assumption, there is zero correlation between the true returns from different elementary holding periods.

The situation is quite different with the least-squares estimates for these returns, however. In Section VII it was shown that the model leads us to expect negative serial correlation between the estimate of a return and the estimate for the succeeding return. In this section we want to investigate more fully this negative correlation. We rely on results developed in Section VII.

Using these results, we can compute the correlation coefficient between b_i and b_{i+1}. If i = 1 or n-1, this correlation coefficient is,

$$r_i = -(n-3)/ \ 2(n^2 - 4n + 5) \ .$$

For other i, the correlation coefficient is,

$$r_i = -(n-1)(n-3)/2(n^2 - 4n + 5).$$

The mean of these n-1 correlation coefficients is,

$$\text{mean} = - \frac{(n-1)(n-3)^2 + (n-3)\sqrt{8(n^2 -4n + 5)}}{2(n-1)(n^2 - 4n + 5)} \ .$$

As n increases, this last expression (for the mean) approaches -0.5. Examples of the values it assumes are -0.4236 (n=5), -0.5134 (n=10), and -0.4977 (n=20). Note that these values do not approach -0.5 monotonically. There are, however, only a finite number of oscillations about -0.5 before the approach does become monotonic. If n > 6, the mean is within 0.02 of -0.5.

Algebraically, the explanation of negative correlation is the presence (noted in Section VII) of negative entries in the off-diagonal elements (i, i + 1) and (i, i - 1) of the matrix A^{-1} The reason that the correlation coefficient only assumes two values for a given value of n is due to the fact that all of the non-zero off - diagonal entries are equal and that the diagonal entries are equal except for the first and last which are equal to each other. All correlation coefficients with lags greater than 1 are zero because all entries not adjacent to the main diagonal of A^{-1} are zero. These comments summarize all that need be said algebraically about the correlation between coefficient estimates.

It is useful also to portray negative correlation geometrically. The concept that unifies the algebraic and geometric aspects of correlation is the concept of collinearity.

As in previous sections, let x and y denote adjacent columns in the portfolio matrix **X**. The geometric measure of collinearity is the angle θ between x and y in N - dimensional Euclidean space. The cosine of this angle is simply the non - mean - adjusted correlation coefficient, $\frac{x \cdot y}{\|x\| \quad \|y\|}$ between the two variables x and y. This correlation coefficient was estimated in Section VI and is, approximately, $a = \frac{n - 3}{n - 1}$. The following table contains the value of this fraction and the corresponding angle for selected values of n.

n	a	θ
5	.500	60^o
10	.778	39^o
15	.857	31^o
20	.895	27^o
25	.917	24^o
30	.931	21^o
100	.980	12^o
∞	1.000	0^o

TABLE 6

In this paper and (Wl) three different data sets for annual returns were analyzed. The value of n in each of these cases was n = 5. The following table reports the angles between adjacent variables in the market return model for these three cases.

Variables	$\frac{x \cdot y}{\|x\| \quad \|y\|}$		
	Cochise	Cochise (edited)	Pima
X_1 and X_2	.3982	.4021	.4780
X_2 and X_3	.4622	.4610	.5223
X_3 and X_4	.3898	.4027	.4878
X_4 and X_5	.3893	.4168	.4207
Mean	.4099	.4206	.4773
Arccosine of Mean	66°	65°	61°
Predicted Mean	.5000	.5000	.50000
Arccosine of Predicted Mean	60°	60°	60°

TABLE 7

In Table 7, the numerators, x.y, are the entries (i, i + 1) in the matrix X'X. The values of $\|x\|$ and $\|y\|$ are the square roots of the ith and (i +1) th diagonal entries of the same matrix. In other words, the geometry of the variables is determined by the X'X matrix. (Dually, the geometry of the coefficient estimates is determined by the inverse matrix $(X'X)^{-1}$.)

As one can see from Table 7, the agreement between pre-dicted and actual is quite good for Pima,where the theoretical angle between adjacent variables is 60° , whereas the actual is 61° . This corresponds closely with the results reported in Section VIII where the predicted and actual values of the correlation between return estimates were presented. (Good results concerning the matrix $(X'X)^{-1}$ of course go hand in glove with good results concerning the matrix $X'X$.)

Just as in Section VIII, the results are not as good for Cochise. Here the actual values are 66° and 65° whereas the predicted value is again 60° . The reasons for the model's decreased accuracy were discussed at the end of Section VIII. As we have observed there is not significant negative correlation in the annual series (Table 6 of (W1)). This is consistent with the large angles 60° - 70° between adjacent variables in the annual data.

In order to study negative correlation we must look at the monthly, quarterly and semi-annual return estimates. These estimates have already been judged relatively inaccurate (W1).

Table 6 in (W1) reports that the negative correlation between return estimates are - .3132, -.4882, and -.5290 for the monthly (n=76), quarterly (n=25) and semi-annual (n=12) series, respectively. The first of these numbers varies considerably from the predicted value but the latter two correspond quite closely with those predicted in Table 5. On the whole, the actual results conform fairly well with our expectations.

The expected angles between adjacent variables for these three cases are 13° , 24° and 35° , respectively. (These values were computed just as those in Table 6.)

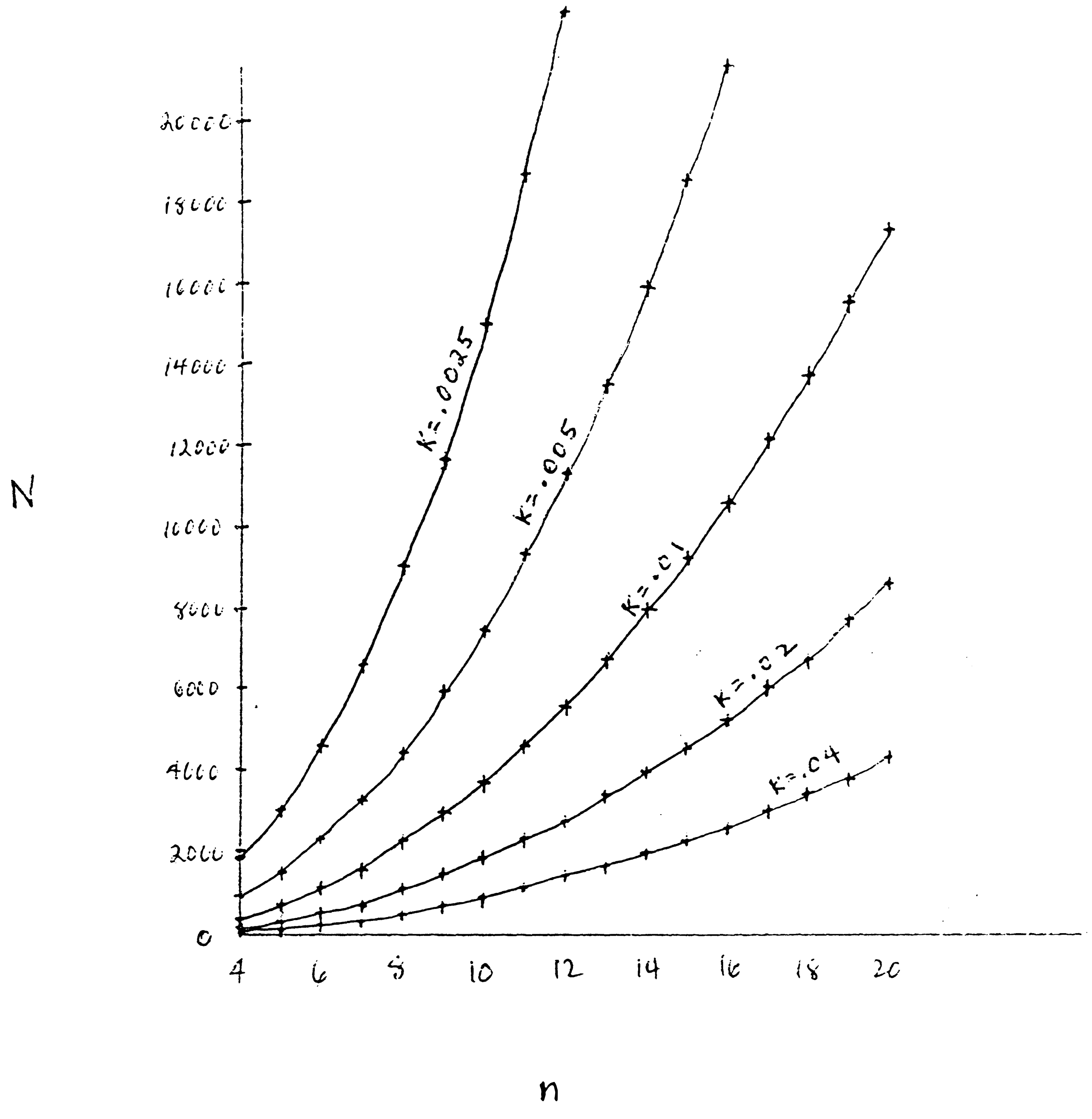
Geometrically, there are two observations that can be made about coefficient estimates made when the angle between the variables is small.

The first observation is the decreased accuracy which is, geometrically, the relatively large interval that is the projection of $Xb= X(X'X)^{-1}X'y$ onto each of the vectors x and y.

The second observation is the relatively small area in the plane generated by x and y that is between the vectors x and y. (It is this area between x and y that corresponds to positive estimate for both coefficients of x and y.) In other words, for larger values of n it is relatively improbable that coefficient estimates for adjacent variables are of the same sign. Hence the expectation that coefficient estimates are negatively correlated. (Negative correlation is not equivalent, of course, to alternation in sign. Alternation explains to some extent, however, the negative correlation.)

With knowledge of the negative correlation between estimates, one might at first suspect more accurate coefficient estimates could be derived perhaps by using an averaging strategy. Common averaging strategies are, however, linear transformations. Therefore, given our assumptions about the model, the Gauss -Markov theorem assures us that the minimum variance unbiased linear estimates are already known. Therefore, a non-trival unbiased linear transformation of our estimates cannot but increase the variance. The existence of a biased linear estimate or an unbiased non-linear estimate that might have attractive properties is not known at this time.

Accuracy of Price Levels



FIGURE

XI. Data Collection

One of the advantages in having a model like that which has been constructed is that, given a data set of a certain size, it allows us to compute how fine an index is allowed in order to achieve a specified reliability in the index numbers. For example, with 10,000 buy-sell observations over a 15-year interval, the model would tell us how fine an index could be constructed (e.g., quarterly) that would achieve a specified accuracy.

Alternately, given a certain fineness in the index (e.g., quarterly) and a requirement for reliability, we would be able to compute the size of the data set necessary to achieve these joint objectives. For example, the model would tell us how many buy-sell pairs would be needed to compute a quarterly index with standard errors of a certain size.

Unfortunately, the absolute size of the standard errors is not an independent variable because the variance of the y variable is a scalar factor in the variance-covariance matrix. This is unavoidable however because the movement of the index is ultimately controlled by the movements of the prices of each of the underlying properties. And the y variable is a measure of these underlying individual prices.

In Section VII this was made precise where, for example, the variance of the ith coefficient, $1 < i < n$, was shown to be, $n(1 + a^2)/N(1 - a^2)$ multiplied by the variance of the y variable. (a has the value $(n-3)/(n-1)$.) If we let K be the mean of all of these factors over all the coefficients, then

$$(10)\quad K = \frac{n^2 - 3n + 5 + 1/(n-2)}{2N} .$$

K turned out to be 0.0110 for Cochise County (unedited) and 0.0021 for Pima County.

The important thing about (10) is that it relates sample size (N), the fineness of the index (n), and the reliability of the index numbers (K). Thus if two of these variables are known, the third can be computed.

The Figure is a graph of the level curves of this relationship corresponding to K = 0.0025, 0.0050, 0.0100, 0.0200, and 0.0400.

The following Table 4 is a representative set of information from the Figure. The data sets analyzed in this paper and in [W1] are points in the upper lefthand corner of this Table (i.e., the lower lefthand corner of the Figure). Data sets analyzed in [CS1] and [AS] are points off to the lower right of Table 4 (i.e., off to the upper right of the Figure).

ACCURACY OF PRICE LEVELS

n	N				
	K=.0025	K=.005	K=.01	K=.02	K=.04
4	1900	950	475	238	119
5	3067	1533	767	383	192
6	4650	2325	1163	581	291
7	6640	3320	1660	830	415
8	9033	4517	2258	1129	565
9	11829	5914	2957	1479	739
10	15025	7513	3756	1878	939
11	18622	9311	4656	2328	1164
12	22620	11310	5655	2828	1414
13	27017	13508	6754	3377	1689
14	31815	15908	7954	3977	1988
15	37014	18507	9254	4627	2313
16	42613	21307	10653	5327	2663
17	48613	24306	12153	6077	3038
18	55012	27506	13753	6876	3438
19	61811	30906	15453	7726	3863
20	69011	34505	17253	8626	4313

TABLE

BIBLIOGRAPHY

[W1] Webb, C. 1988. "A Probabilistic Model for Price Levels in Discontinuous Markets," in Measurement in Economics, W. Eichhorn, editor, Physica-Verlag, Heidelberg, 1988.

[BMN] Bailey, M.J., Muth, R.F., Nourse, H.O. 1963. "A Regression Method for Real Estate Price Construction," J. of Amer. Stat. Assoc., 58:933-42.

[C] Court, A.T. 1939. "Hedonic Price Indexes with Automobile Examples," in The Dynamics of Automobile Demand, based on a joint meeting of the American Statistical Association and the Econometric Society in Detroit, Dec. 27, 1938. New York, General Motors Corp.

[CS1] Case, K.E. and Shiller, R.J. 1987. "Prices of Single-Family Homes since 1970: New Indexes for Four Cities," New England Economic Rev., Sept./ Oct. 1987, 45-56.

[K] Kendall, M. and Stuart, A. The Advanced Theory of Statistics, vol. 3, 3rd edition, Macmillan, New York, 1976.

[T] Theil, H. Principles of Econometrics, John Wiley, 1971.

[HJ] Horn, R.A. and Johnson, C.A. Matrix Analysis, Cambridge, 1985.

[AS] Abraham, J.M. and Schauman, W.S. 1990. "New Evidence on Home Prices from Freddie Mac Repeat Sales," presented at the AREUA Midyear Meetings, May 30, 1990.