Estimating the Technology Coefficients in Linear Programming Models

Bruce L. Dixon and Robert H. Hornbaker

Bruce L. Dixon is a Professor of Agricultural Economics at the University of Arkansas, Fayetteville and Robert H. Hornbaker is an Assistant Professor of Agricultural Economics at the University of Illinois, Urbana-Champaign.

#6757

Estimating the Technology Coefficients in Linear Programming Models

ABSTRACT

Linear constraints for mathematical programming models are demonstrated to be random coefficient regression (RCR) models when estimating constraint coefficients from samples. Monte Carlo experiments show an RCR estimator preferable to least squares although least squares is also acceptable. Dependence between output levels and technical coefficients can lead to biased estimates.

Estimating the Technology Coefficients in Linear Programming Models

Estimation of the coefficients of mathematical programming models has received some attention in the past. Sengupta considered the problems of estimating the objective function coefficients and, in passing, noted that it might be worthwhile estimating the coefficients of the technology matrix in a random coefficient regression (RCR) framework. More recently, Ray also addressed the problem of estimating the coefficients of a linear technology matrix by RCR but rejected it in favor of alternative methods of estimation. A major estimation problem noted by Ray is that for most applications it is required that the technology coefficients be nonnegative.

In this paper the problem of estimating the technical coefficients is reexamined. We first argue conceptually that random coefficients is the only defensible way to specify the coefficients of deterministic mathematical programming problems. We then consider in a Monte Carlo context if it is reasonable to expect RCR estimation to yield satisfactory estimates of the means of the technology coefficients from a sample of N firms. The major problems from an estimation point for RCR are to obtain estimates of the covariance matrix associated with the random coefficients and to assure that the estimated coefficients are nonnegative. We suspect negative estimates are rare for most empirical problems and test this hypothesis. The RCR estimates are also compared with ordinary least squares (OLS). Finally, as observed by Ray, the technical coefficients in mathematical programming models could likely be distributed jointly with output levels. Ramsey's RESET test is proposed as a means for checking for such dependence and its reliability is measured in the Monte Carlo experiments.

A Conceptual View of Technology Coefficients

Consider a linear technology writtten in canonical form as:

$Ax = b$

where, as in Ray, A is a mxn matrix of known technology coefficients, x is a nx1 vector of output levels and b is the vector of available resources. If it is assumed that each component of x is an output level and that the technology is concave in the input space, then each element of A must be positive.

The estimation problem envisaged in this paper, and also for some of the estimators in Ray, is that there exists a sample of data from N firms, with observations on x and b and the object is to estimate the unknown elements of A. Making the usual deterministic assumptions for the linear programming model, it follows that any row of $Ax = b$ for any of the N firms holds identically as an equation. The implication of this fact is that it is inappropriate to estimate a row of A in the usual regression context. For example, Ray suggests three estimators that assume an additive error term can be appended to the left hand side of any of these equations to account for the fact that any given set of estimates for A will most likely not satisfy the linear constraints exactly for all the N firms.

However, the resource constraints are, by their very definition, identities. That is, each firm uses a given amount of the ith component of b in its production activities. If one firm uses more of a resource to produce the same vector x of outputs then it must be that its associated technological coefficient for that output is higher than for other firms. Thus, the only explanation for firms using different levels of b_i for the same level x_j is that at least some of the coefficents of the rows of A vary from firm to firm. Thus to add a stochastic error term to each row and estimate the elements in A under the assumption that the elements of A are population constants as recommended in the latter part of Ray is to estimate a structure that is not the same as the structure generating the observed data.

This point can be further illustrated by looking at the estimation problem by the columns of A instead of the rows of A. If the individual farmers know how much of each resource is devoted to each output and every farmer produces some of the output then the estimation problem is trivial. If y units of resource b_1 are used to produce x units of activity one then clearly $a_{11} = y/x_1$. In a sample of N firms, N observations on a_{11} could be similarly computed. Most likely there will be some variation in the observations for a given a_{ij} across firms so that it is clear that the elements of A are not population constants. Hence the proper perspective from which to view estimation of the elements of A is one of viewing the elements of A as random variables and that the objective of estimation is to estimate the means of these elements.

It is now posssible to formalize the estimation problem. For each firm it is assumed that the elements of A take on some value but for a given element there is variation across firms. In addition, every firm satisfies the set of canonical constraint equations with A varying from one firm to another. Therefore, from the point of view of the population, the constraints can be viewed as stochastic identities. That is, each constraint for each firm is the sum of random variables. In such a situation the focus of estimation are the means and variances of these coefficients although the focus here is primarily on the means.

In the estimation method proposed below it is assumed that the individual firms being sampled do not possess sufficiently detailed records of production to be able to identify the amount of each resource devoted to each activity. If this were the case then estimation would be handled as discussed in Ray for this situation. Our conjecture is that such problems are very rare, since few farmers record input usage by each type of output. Hence a random coefficient approach is considered for the situation where only observations on the x_j and b_i are available. These are the same assumptions made by Ray in his development of the L_p estimators which are constrained to make all estimates of the a_{ij} nonnegative.

Random Coefficient Estimation

Assume that it is desired to estimate the ith row of A, i.e.

(1) $b_{ik} = a_{i1k}x_{1k} + a_{i2k}x_{2k} + \ldots + a_{ink}x_{nk} \qquad k = 1,\ldots,N$

where it is assumed, letting a_i be the ith row of A, that

(2) $E(a_{ik}) = \bar{a}_{ik}$

(3) $E(a_{ik}a_{ik}') = D$

(4) $E(a_{ik}a_{ik'}') = 0; \quad k \neq k', \text{ all } k,k'.$

These are exactly the assumptions of the Hildreth and Houck random coefficients model. It can be shown, as in Judge et al. that estimation of $\bar{a}_i$ is essentially a problem of applying generalized least squares (GLS) to a model with heteroscedastic error terms. This latter fact gives support to the Ray approach that assumes an additive error term. However, it is a much different justification than suggested by Ray.

Two problems potentially arise in trying to obtain estimates of $\bar{a}_i$ using a random coefficient approach. The first is that the matrix D must be estimated since it is not known in any real world application. In research to date, the estimation of D has not been totally satisfactory and numerous approaches have been suggested in the literature, see Swamy and Methta, Swamy and Tinsley and Dixon, Batte and Sonka.

The second problem is keeping the estimates of the means positive. Ray suggest various methods that directly constrain the estimates to be positive by using a mathematical programming approach. If the matrix D is known then the model in (1) can be transformed to be homoscedastic and the methods of Ray could be applied to the transformed data.

However, it is our contention that negativity problems are likely to be rare in practice. Since all observations on the b_i and x_j must be nonnegative by definition and the a_{ij} are assumed nonnegative, it seems unlikely that negative estimates would often arise. This hypothesis is explored in the empirical section.

An assumption implicitly made for the estimation of $\bar{a}_i$ in (1) is that the distribution of the a_i is identical for any level of x. However, in practice it seems plausible that if activity x_j requires more b_i for a given firm than most of the other firms in the industry, then that firm is more likely to have a lower x_j. Such a functional dependence would invalidate the properties of the estimators. To explore the implications of such behavior Monte Carlo experiments are conducted to estimate the sampling distribution of the estimators and to determine if Ramsey's RESET test is robust in identifying this type of misspecification. It should be pointed out that if the means of the a_{ij} are dependent on observable exogenous variables, then this information can be used to dispense with the dependence problem as discussed in Dixon, Batte and Sonka.

The Random Coefficients Estimator

Assuming that the matrix D is known, the GLS estimator of $\bar{a}_i$ is:

(5) $\tilde{\bar{a}}_i = (X'GX)^{-1}(X'GB)$

where G is a diagonal matrix whose ith diagonal element is the reciprocal of $x_t'Dx_t$ where t denotes row t of X, the matrix of observations on x_t. B is a vector of the N observations on b_i.

There are a variety of estimators for D. Some of these are given in Swamy and Mehta. The procedure used in this study is quite detailed. One of the estimators given in Swamy and Tinsley is used with the modifications suggested in Havenner and Swamy. Basically an estimator of the elements of D is obtained using the residuals of repeated iterations of (1). This estimator is a least squares estimator so that the variances estimated are not necessarily positive. In such cases the modification in Havenner and Swamy is used with the values appropriately scaled for our particular estimation problem. Readers wishing complete detail should contact the senior author.

The Monte Carlo Experiments

Three hypotheses tested in the Monte Carlo experiments are (1) RCR is a more efficient estimator of the a_i than OLS, (2) negativity of the estimates is a rare occurence if the process generating the data conforms with the assumptions of RCR models, and (3) dependence of the coefficients with the activity levels will result in biased estimates.

To test the above hypotheses the experiments are structured in the following way. First, all of the observations are generated according to a random coefficients model specified as: $b_{1k} = a_{11k}x_{1k} + a_{12k}x_{2k}$ where the a_{ij} are random with means $\bar{a}_{11}$ and $\bar{a}_{12}$. This is in accordance with the argument in the first part of this paper. The degree of correlation among the regressors is either zero or .75 in the experiments and the correlation of the random coefficients is set at zero or .75 in the various experiments. The x_i are drawn from a multivariate normal distribution and are constant across experiments and samples except for the experiments with dependence between a_{12} and x_2. The values of the means of the coefficients (a_{ij}) are three standard deviations from zero for those generated by a multivariate normal distribution. These are experiments E1 through E8 where E denotes experiment. In these experiments a_{11} has a mean of 9 and a_{12} a mean of 3. This allows for some very nearly zero observations on the a_{ijk} which should aid in the testing of hypothesis (2). Any negative a_{11k} or a_{12k} are truncated to zero. In E9-E12 hypothesis (2) is even more definitively tested by assuming that the two random coefficients have a uniform distribution between zero and ten for a_{11} and zero and one for a_{12}.

A total of 24 experiments are conducted. For the first 12 experiments 100 observations (N) are drawn for each sample and 80 samples are drawn. In the second 12 experiments N = 25 and 80 samples are drawn. RCR and OLS estimates are obtained for each sample and their means and standard deviations are computed. Also, the number of coefficients for each method of estimation that is negative is reported. Each sample is tested using Ramsey's RESET test with the set of regressors (the second, third, and fourth powers of included variables) suggested by Thursby and Schmidt. This test was performed on the RCR estimates of the model after adjusting for heteroscedasticity. The number of F statictics exceeding the critical value for the test at a 95 percent level is also reported.

The number of F's in excess of the critical value is important in the experiments where the level of one of the two regressors for an observation is dependent on the value of the a_{ij} drawn. In E5-E8 and E11 and E12 the second regressor is set to zero if the realization on a_{12} is one standard deviation or more above its mean, thus establishing a dependence between a_{12k} and x_{2k}. This is in accordance with the earlier conjecture that firms with large coefficients would tend to produce less of the output.

Results and Implications

The results in Table 1 for N = 100 and Table 2 for N = 25 suggest several conclusions regarding the hypotheses. For hypothesis (1) dealing with the superiority of RCR estimators over OLS, E1-E4 and E9-E10 provide the most illuminating evidence. Both estimators are unbiased for both sample sizes in these experiments. For N = 100 RCR appears slightly more efficient for E1-E4. The greatest superiority is in E4 with both high regressor and coefficient correlation. With the uniform distribution there is no substantial difference in efficiency. These results also hold for N = 25. Hence our conclusion is that if the process generating the data conforms to the RCR model, then using an RCR estimator is to be preferred over OLS but not by a great margin. RCR is definitely more efficient in the experiments with regressor and coefficient dependence for normally distributed coefficients. The estimates of the elements of D vary widely. However, even if they were known with certainty so the true variances of the RCR estimates could be computed, these variances would be reduced by no more than 15 percent.

The figures in the columns under NEG COEFF indicate the number of estimates that were negative. For E1-E8 for both sample sizes which satisfy the RCR assumptions, a negative estimate shows up only 46 times out of 5120 possibilities or 1 percent of the time. For the experiments with a uniform distribution 10 percent of the estimates are negative and with the small sample estimates having more than twice as many negative estimates as the large sample estimates. Thus, on a relative basis, negativity of estimates does not appear to be a serious problem. It is not surprising that the uniform distribution experiments show some negative estimates because both coefficients have zero as a lower bound.

For those experiments having a dependence between regressors and coefficients, both the RCR and OLS estimators are biased in every such experiment for the second coefficient except E12 for N = 25. The estimators are only biased for the first coefficient when the coefficients are correlated, i.e. E7 and E8. Not surpisingly, the bias in the first coefficient is positive since it has numerous observations in each such sample where $x_{2t} = 0$ and this leads to biasing upwards because such observations are at extreme values of one of the regressors and therefore convey more information to the estimator. This tends to occur when a_{11} is greater than its average. It is somewhat encouraging to note that when there is no coefficient correlation the degree of bias seems modest but this, of course, is a function of the degree of dependence and would undoubtedly become more severe as the strength of dependence increased. The RESET test is not a very promising test, at least for the degree of dependence between coefficients and output levels hypothesized here. For both sample sizes rejection is more likely in the misspecified models where the coefficients have a normal distribution but not by a great margin and not nearly as frequently as it should. The test seems to indicate better for larger samples but not with any great assurance.

Conclusions

It has been demonstrated for mathematical programming models that the OLS assumptions do not conform with those of the programming model. However, RCR assumptions do conform. RCR estimation has some superiority over OLS but the latter estimator is an acceptable substitute for RCR and is not likely to do much worse. Monte Carlo experiments suggest that negativity of coefficients is not a frequent occurence. Dependence between coefficients and output levels can lead to large biases, particularly when the coefficients are correlated with each other. The RESET test is not robust in identifying regressor and coefficient dependence in these experiments.

Table 1. Random Coefficient Regression and Ordinary Least Squares Estimates Resulting from Monte Carlo Simulation, Large Samples (N = 100).

Experiment	r_a	r_x	RCR COEFFS a_{11}	RCR COEFFS a_{12}	NEG COEFFS	OLS COEFFS a_{11}	OLS COEFFS a_{12}	NEG COEFFS	F
E1	0	0	9.037	2.997	0	8.933	3.036	0	8
			1.6412	0.5461		1.6160	0.5338		
E2	0	.75	9.176	2.924	0	8.977	2.995	1	8
			3.0751	1.0532		3.4179	1.1659		
E3	.75	0	8.996	3.049	0	9.093	3.025	0	9
			2.1591	0.7686		1.9504	0.7016		
E4	.75	.75	9.032	2.994	1	9.186	2.955	5	5
			4.0838	1.3460		4.4607	1.4625		
E5	0	0	8.872	2.766*	0	8.832*	2.780*	0	7
			0.6415	0.2599		0.7507	0.2785		
E6	0	.75	9.056	2.692*	0	8.976	2.725*	0	12
			0.7848	0.3011		0.8564	0.3272		
E7	.75	0	11.697*	1.621*	0	11.198*	1.773*	0	21
			0.6816	0.3435		0.6647	0.2976		
E8	.75	.75	12.342*	1.386*	0	12.019*	1.500*	0	22
			0.6166	0.2792		0.7587	0.3204		
E9	0	0	5.251	0.442	7	5.258*	0.443	7	5
			1.2216	0.3890		1.1902	0.3814		
E10	0	.75	5.108	0.454	26	4.969	0.506	25	9
			2.2699	0.7410		2.2743	0.7419		
E11	0	0	4.961	0.389*	2	4.972	0.380*	3	7
			0.6270	0.2035		0.6245	0.2041		
E12	0	.75	5.021	0.399*	3	5.029	0.391*	3	7
			0.6501	0.2204		0.6738	0.2361		

a. In the RCR and OLS COEFS columns the means of the 80 estimates are given and immediately below them are their estimated standard deviations.

b. An asterisk denotes the estimator is biased at the 95 percent level. The column under r_a indicates the degree of correlation of the slope coefficients and r_x indicates the degree of regressor correlation. The columns NEG COEFFS indicate the number of coefficient estimates that were negative out of 160 estimates. The column F is the number of RESET specification tests that were significant out of a possible 80.

Table 2. Random Coefficient Regression and Ordinary Least Squares Estimates Resulting from Monte Carlo Simulation, Small Samples (N = 25).

Experiment			RCR COEFFS		NEG COEFFS	OLS COEFFS		NEG COEFFS	F
	r_a	r_x	a_{11}	a_{12}		a_{11}	a_{12}		
E1	0	0	9.123 2.5735	2.999 .8998	0	8.947 2.4633	3.051 .8758	0	6
E2	0	.75	9.929 4.9187	2.714 1.7015	6	9.764 4.9386	2.765 1.6973	3	10
E3	.75	0	9.266 3.2929	2.901 1.1109	1	9.216 3.3735	2.927 1.1318	1	7
E4	.75	.75	10.098 6.3753	2.641 2.1678	12	9.878 6.7122	2.709 2.3322	15	14
E5	0	0	8.732 1.5757	2.841* .6090	0	8.809 1.7252	2.794* .6372	0	11
E6	0	.75	8.914 1.8044	2.806* .6872	0	8.830 1.9506	2.826* .7663	0	7
E7	.75	0	11.581* 1.5407	1.600* .6657	1	10.968* 1.6503	1.815* .6323	0	13
E8	.75	.75	12.065* 1.1772	1.414* .5724	0	11.677* 1.5074	1.582* .6534	0	18
E9	0	0	5.414 2.1331	.361 .7112	28	5.298 2.1212	.399 .7000	24	8
E10	0	.75	5.067 3.3512	.473 1.1300	32	4.984 3.8247	.486 1.2970	35	11
E11	0	0	5.056 1.2105	.378* .4660	20	5.122 1.1692	.346* .4246	18	9
E12	0	.75	4.952 1.3421	.411 .4983	16	4.827 1.3095	.478 .4887	10	11

a. In the RCR and OLS COEFFS columns the means of the 80 estimates are given and immediately below them are their estimated standard deviations.
b. An asterisk denotes the estimator is biased at the 95 percent level. The column under r_a indicates the degree of correlation of the slope coefficients and r_x indicates the degree of regressor correlation. The columns NEG COEFFS indicate the number of coefficient estimates that were negative out of 160 estimates. The column F is the number of RESET specification tests that were significant out of a possible 80.

REFERENCES

Dixon, Bruce L., Marvin T. Batte and Steven T. Sonka. "Random Coefficients Estimation of Average Total Product Costs for multiproduct Firms." J. Bus. Econ. Stat. 2(1984):360-66.

Havenner, A. and Swamy, P. A. V. B. "A Random Coefficient Approach to Seasonal Adjustment of Economic Time Series." J. Econometrics 15(1981):177-209.

Hildreth, C. and J. P. Houck. "Some Estimators for a Linear Model with Random Coefficients." J. Amer. Stat. Assoc. 63(1968):583-595.

Judge, G. G., W. E. Griffiths, R. C., Hill, H. Lutkepohl, T. C. Lee. The Theory and Pactice of Econometrics. New York: Wiley and Sons, 1985.

Ramsey, J. B. "Tests for Specification Errors in Classical Linear Least Squares REgression Analysis." J. Roy. Stat. Soc. Series B, 31(1969):350-371.

Ray, S. "Methods of Estimating the Input Coefficients for Linear Programming Models." Amer. J. Agr. Econ. 67(1985):660-665.

Sengupta, J. K. "Estimating Parameters of a Linear Programming Model." J. Cybernetics. 6(1976):309-328.

Swamy, P. A. V. B. and J. S. Mehta. "Bayesian and Non-Bayesian Analysis of Switching Regressions and of Random Coefficient Regression Models." J. Amer. Stat. Assoc. 70(1975):593-602.

Swamy, P. A. V. B and P. Tinsley. "Linear Prediction and Estimation Methods for Regression Models with Stationary Stochastic Coefficients." J. Econometrics 12(1980):103-42.

Thursby, J. G. and P. Schmidt. "Some Properties of Tests for Specification Error in a Linear Regression Model." J. Amer. Stat. Assoc. 72(1977):635-641.