



The World's Largest Open Access Agricultural & Applied Economics Digital Library

This document is discoverable and free to researchers across the globe due to the work of AgEcon Search.

Help ensure our sustainability.

Give to AgEcon Search

AgEcon Search

<http://ageconsearch.umn.edu>

aesearch@umn.edu

*Papers downloaded from **AgEcon Search** may be used for non-commercial purposes and personal study only. No other use, including posting to another Internet site, is permitted without permission from the copyright owner (not AgEcon Search), or as allowed under the provisions of Fair Use, U.S. Copyright Act, Title 17 U.S.C.*

No endorsement of AgEcon Search or its fundraising activities by the author(s) of the following work or their employer(s) is intended or implied.

THE AUSTRALIAN JOURNAL OF AGRICULTURAL ECONOMICS

VOL. 19

APRIL 1975

NO. 1

USING REGRESSION ANALYSIS TO REDUCE AGGREGATION BIAS IN LINEAR PROGRAMMING SUPPLY MODELS*

JOHN O. S. KENNEDY

La Trobe University

Because methods of eliminating aggregation bias are impracticable, an alternative method is suggested for reducing aggregation bias which uses regression estimates of farm resource availabilities as functions of farm size. The estimates are incorporated in a parametric run of the LP problem in which size is parameterized. The method is applied to a case study problem, and the results compared with other methods of demarcating representative farms.

Introduction

Normative estimates of aggregate supplies of outputs from N resource situations may be obtained by running a linear programming (LP) model for each resource situation and summing the resulting N output vectors. However, using this method, N may be too large to make the aggregate estimate obtained worth the cost of N LP runs. It may be possible to obtain the same result, or an approximation to the result, by summing the results of q ($< N$) LP models which are in some sense representative of the N original LP models. In general, the extent of the approximation, or aggregation bias as it is known, is inversely proportional to q . The trade-off between the costs saved in decreasing q at the expense of increasing the inaccuracy of estimates of aggregate supply remains a study in its own right. The costs involved in using biased estimates of normative supply are difficult to quantify, depending upon the purpose for which they are to be used. They are likely to be a complex function of the size of the bias and the aggregate supply level. In practice, the funds available for building models generally determine q , and it is hoped that the method of selecting the q representative resource situations results in low aggregation bias.

Various attempts have been made to propose theoretical methods for eliminating aggregation bias. Some of these are discussed in the following section. They all imply an impracticably high q , but they do suggest criteria for classifying the N farms in such a way as to minimize or at least reduce aggregation bias. Later in the paper an approach is suggested, based on a simplification of one of these methods, which enables aggregation bias to be reduced by running just one parametric LP model for each farm type.

* Work on this study was financed by a grant from the Wool Research Trust Fund of the Australian Wool Corporation. The author is grateful to John Guise and Ross Drynan for useful discussions on an earlier draft of this article.

Sufficient Conditions as Guidelines

Sufficient conditions for eliminating aggregation bias in aggregative LP studies were first put forward by Day [3]. These were that all farms be classified into groups with identical technical matrices and with net revenue expectations and resource vectors proportional to the respective aggregate vectors. The conditions were seen as highly restrictive, and attempts were made by Miller [7] and Lee [6] to relax them.

Paris and Rausser [8] have recently added to the literature on sufficient conditions, and also place their discussion in historical perspective. They make the point that Day's specifications are not only impossible to meet in practice but are also the most restrictive set of conditions conceivable. Paris and Rausser show that it is not mathematically necessary to require that farms have either equality or proportionality amongst the elements of the LP matrices of the N farms, nor even that they have technical matrices of the same dimensions. However, whilst they succeeded in finding less restrictive sufficient conditions, it is not clear to what extent this development promotes a more operational approach to the problem of aggregation bias. As they comment, one of the crucial factors in the argument is the data available.

Paris and Rausser state that usually the only type of information available to the empirical researcher is aggregate information on the technical matrix A_o , the vector of resources available r_o , and the vector of expected net revenues c_o . They argue that it is only because of the requirements of Day's sufficient conditions that it is usually *assumed* that the i -th farm allocated to the one group has $A_i = A_o$, and r_i and c_i proportional to r_o and c_o respectively. By showing that Day's sufficient conditions may be relaxed they suggest that these assumptions are not necessary. Less restrictive assumptions may be made.

However, just because there is usually a lack of data on individual farms, it is often reasonable to assume that farms of a particular type do have identical A matrices and proportional c vectors. From this point of view, these two conditions of Day do not seem unduly restrictive, nor the relaxation permitted by the work of Paris and Rausser immediately relevant. It is the way in which to deal with the variability between the r vectors of the farms that is most problematic. Often there are data on resources used on individual farms, which may be taken as an approximation to resources available on individual farms. For example, such data may be available either in sample form from farm consultants, or in census form from government statistical bureaus. The assumption that a number of farms may be discovered with proportional r vectors is the most tenuous of Day's conditions.

Buckwell and Hazell [2], in an attempt to devise an operational method of reducing aggregation bias, proposed the use of cluster analysis. They decided to base their method on Day's requirements, and used cluster analysis to achieve the requirement of resource proportionality as closely as possible. For all q from $q = N$ to $q = 1$, a system of allocating resources to q groups may be found which by some criterion minimizes the differences in resource proportions within groups. The method is attractive because an attempt is made to formally minimize aggregation bias. However, there is little guidance on what should be the optimal number of groupings, q .

The operational method put forward in this article is based on conditions related more closely to Miller's [7] as extended by Lee [6] than to Day's. Lee's method and its criticism by Day are discussed in the Appendix. It is concluded that whilst Lee's method does not provide a practical system of eliminating aggregation bias, it does suggest a method of attempting to reduce aggregation bias provided one is prepared to make some simplifying assumptions. The principle is the allocation of farms to groups based on the ratios of the resources available on each farm. Each representative group contains farms which have the same LP solution basis.

An Alternative Method Based on Regression Analysis

The method proposed in this article has scope where the farms to be aggregated may be allocated to type groups on the basis of common A and proportional c , but where the ratios of farm resources available within groups vary on account of size of farm. Ahn [1] has pointed out that it is often important to classify farms by size to take account of different resource mixes on farms of different sizes when using representative farms to study adjustment processes in a developing country.

Regression results from data on resources available on a sample of farms are incorporated in a parametric LP run. The procedure permits both the determination of the number of representative farm categories within the type-group, and of the cut-off points for demarcating each category.

The starting assumption is that the availability of the m farm resources specified increases approximately linearly with farm size δ , as measured by the amount used of one of the m resources on the farm, say the first. That is, that the availability of the i -th resource r_i for $i = 2, \dots, m$, may be expressed in the form $r_i = a_i + b_i\delta$, where a_i and b_i are constants. If the availability of any resource r_j does not vary with size significantly then obviously $b_j = 0$. As δ is varied through the range of farm sizes, resource availability ratios r_i/r_h will change if a_i or $a_h \neq 0$. Thus if such expressions for resource availabilities are built into an LP matrix, and δ is varied parametrically, basis changes will occur in the LP run when resource ratios reach critical values. An aggregate supply estimate may be obtained by scaling the solution vector obtained at each basis appropriately, using information on the number of farms within the ranges of sizes for which no basis change occurs.

Suppose that the derivation of aggregate supply requires the solution of the following LP problem:

$$\begin{aligned} &\text{Maximize} && c' x_k \\ &\text{subject to} && Ax_k \leq \begin{bmatrix} \delta \\ r \end{bmatrix} && k = 1, \dots, N \end{aligned}$$

where

- x is an $(n \times 1)$ vector of activities;
- c is the corresponding $(n \times 1)$ vector of net revenues;
- r is an $((m - 1) \times 1)$ vector of resource levels; and
- A is an $(m \times n)$ matrix of technical coefficients.

The aggregate supply vector, S , is defined as $\sum_{k=1}^N x_k$.

If r can be written as a linear function of δ ,

$$r = a + b\delta + e,$$

then, ignoring the error term vector e , the equivalent formulation for obtaining the estimate of aggregate supply by running one LP parameterizing δ is:

$$\begin{array}{ll} \text{Maximize} & c'x \\ \text{subject to} & Ax - \begin{bmatrix} 1 \\ b \end{bmatrix} \delta \leq \begin{bmatrix} 0 \\ a \end{bmatrix} \end{array}$$

In order to calculate the estimate of aggregate supply, list the N farms in ascending order of δ available and identify each farm by subscript $k = 1, \dots, N$. Let the resource availability of δ on the k -th farm be δ_k . The scale resource δ is parameterized from δ_1 to δ_N . Suppose that solution vectors x_j , $j = 1, \dots, Q$ are obtained by this procedure for basis changes at which $\delta = h_j$, $j = 1, \dots, Q$. Let the number of farms for which $\delta \leq h_{j+1}$ be n_j . Then, by linear interpolation, the estimate of aggregate supply is obtained as

$$\hat{S} = \sum_{j=1}^{Q-1} (n_j - n_{j-1})x_j + \sum_{k=f(j)}^{g(j)} \{(\delta_k - h_j)/(h_{j+1} - h_j)\} (x_{j+1} - x_j)$$

where $n_0 = 0$;
 $f(j) = n_{j-1} + 1$; and
 $g(j) = n_j$.

If the error term vector e in the regression equations for the resources is zero, then $\hat{S} = S$ and there is no aggregation bias. However, this is not generally the case. Aggregation bias is therefore some function of the residual error e_i for the regression of resource availability r_i . The problem of reducing aggregation bias becomes one of deciding how best to minimize the residual error.

Ordinary least squares regression entails the minimizing of $\sum e_{ik}^2$. This is an appropriate procedure if the loss function is quadratic in e_i . However, if the loss function is linear in e_i , then minimization of $\sum |e_{ik}|$ is the more appropriate criterion for fitting the regression line [5]. Minimization of $\sum |e_{ik}|$ in obtaining a linear regression fit may be effected using LP [9].

For all those e_{ik} which are small enough so that they imply no change in the LP solution for the k -th farm, aggregation bias is a linear function of e_{ik} . For larger e_{ik} it is impossible to generalize. A large negative e_{ik} may lead to greater aggregation bias than a large positive e_{ik} . In other words, aggregation bias may not be a symmetric function of e_{ik} about zero, and neither the least squares nor the least absolute deviation regression methods may be strictly appropriate. The question is essen-

tially an empirical one for each case. All that can be said is that the larger the proportion of e_{ik} that are small enough not to cause a basis change, the lower will be the aggregation bias resulting from using the least absolute deviation regression method.

Limitations of the Method

It has been assumed that the size of the representative farm for a type-group can be varied without altering A , the technical matrix. If economies of size exist in the industry, this is an untenable assumption. If the returns to size are increasing, then resort would have to be made to some technique such as separable programming, approximating the returns to size effect by linear segments. If the returns to size are decreasing, then the approximating method of linear segments may be handled in the usual way in the LP. An example of dealing with decreasing returns to size is given in the following case study, in which livestock stocking rates decrease with increasing grazing acreage available.

A disadvantage of the proposed method of reducing aggregation bias is that it is static. Different categories are required for different c vectors. However, as a new set of categories can be demarcated by one parametric LP run with a revised c vector, this is not a great problem. A more serious drawback is the difficulty of allowing for resource carryover if the LP model is to be run recursively over several periods. For example, if it were assumed that operating capital available at the beginning of each year was dependent upon income generated in the previous year on the N farms, then adjustments would have to be made in the regression equation for operating capital for each successive year that estimates were required. This would probably entail fitting successive regression lines for the initial observations of operating capital available, plus or minus some adjustment for assumed cash carryover between years.

A Case Study

In order to obtain some information on the relative degrees of aggregation bias that may be introduced by different systems of classifying farms into representative farm groups, a case study approach was adopted. Net revenues and a technical matrix for a typical multi-purpose farm in the Southern Tablelands Statistical Division of New South Wales were derived. Limiting resources were assumed to be acreage, seasonal labour, and operating capital. The records of a sample of 52 farms which could be reasonably represented by such data were obtained on an anonymous basis from the Agricultural Business Research Institute (ABRI) of the University of New England. Data on resources available on each of these farms were obtained from the records.

Five methods of grouping the farm data were used. Estimates of the aggregate supply of products and aggregate resource use were derived by running LP's and scaling the results appropriately for each of the five methods. Aggregate estimates were compared with the results obtained by summing the solution vectors for all of the 52 individual farm LP's. The five methods are described below:

(1) *All resources pooled*

Land, labour and operating capital were aggregated for all the 52 farms. One LP was run using the aggregate resource vector obtained.

(2) *Three acreage classes*

Representative farms have been based on size amongst other characteristics in some studies. In order to investigate this method resources were aggregated within three acreage size categories. Each size category represented roughly equal total acreages. One LP was run for each category, and the output vectors were scaled by the appropriate number of farms in each category.

(3) *Three operating capital to labour ratio classes*

Classifying farms on the basis of the most limiting resource as suggested by Miller [7] is impracticable unless preliminary sensitivity analysis has been carried out to determine in which situations what resource is most limiting. The approach, however, does suggest that ratios of resource availability may be a pragmatic criterion for grouping farms. Resources were aggregated within three resource-ratio categories, each category again representing roughly equal acreages. One LP was run for each category, and the output vectors appropriately scaled.

(4) *Least squares regression and parametric LP*

Acreage was chosen as the size resource, although there was no reason why operating capital or labour was not chosen. Scatter diagrams of labour versus acreage, and operating capital versus acreage showed that positively sloped, straight-line relationships did exist approximately, but that deviations from any fitted straight line increased markedly with acreage, implying the presence of heteroscedasticity. Operating capital divided by acreage, and labour divided by acreage, were therefore regressed on the reciprocal of acreage in order to obtain efficient estimates of the linear relationships between operating capital and acreage, and labour and acreage.¹ A parametric LP problem was run and estimated aggregates obtained in the manner described above. Ten LP solutions were obtained for the range of acreage which was parameterized.

(5) *LP regression and parametric LP*

Relationships between operating capital, labour and acreage were obtained as in (4), except that LP was used to minimize the sum of the absolute errors of the regression fit. Table 1 shows the regression results obtained from using the two different regression procedures. Eight LP solutions were obtained when the LP regressions were used in the parametric LP run.

It was apparent from the ABRI records that stocking rates for live-stock decreased markedly with increasing acreage. This effect is largely accounted for by the less intensive pasture management of larger holdings. To allow for this, four pasture activities were included with

¹ The problem of heteroscedasticity is only eliminated in this way if the variance of the error is proportional to acreage squared. Although this assumption was not tested statistically it seemed to be a reasonable approximation in this case.

TABLE 1

Resource Regressions using Least Squares (LS) Regression and Linear Programming (LP) Regression

Regression	Method	Constant	Acreage coefficient	\bar{R}^2
Operating Capital (\$) on acreage	LS	955	4.34	0.76
	LP	144	4.33	—
Labour (man-hours) on acreage	LS	519	0.233	0.85
	LP	347	0.305	—

Note: Acreage was parameterized in the farm LP matrix from 500 to 7,500.

different stocking rates. The pasture activity which permitted the highest stocking rate was allowed into the solution at levels up to 1000 acres. Further pasture demands carried progressively less stock per acre, as determined by similar acreage bands. Thus parameterizing total acreage, which varied grazing acreage available proportionately, automatically allowed for changing stocking densities.

The technical matrix allowed for the augmenting of some of the farm's resources. Cash could be borrowed, and seasonal labour hired. Thus methods 4 and 5 are still feasible when addition to resources is permitted, whilst Lee's method would not be. (See comments on Lee's method in the Appendix.)

The results obtained from the five methods of aggregation are displayed in Table 2, and may be compared with the results obtained from aggregating the solution vectors of all the 52 LP's.

Two statistics were used for measuring the aggregation bias involved in each method, so that comparisons could be made between the aggregation bias in all cases for multi-product LP results. The magnitude of errors in estimates of production or resource use may be more important for certain products or resources than others, depending upon how the estimates are to be used. If the interest is in obtaining an aggregate supply function for wool, then errors in sheep numbers may be the most significant errors. Alternatively the interest may be in estimating total farm income, in which case errors in total farm income may be the most significant ones. However, in order to enable a general comparison, errors for each activity in Table 2 are weighted by the net revenue contribution of the item to total net revenue for the aggregate of the 52 LP results.

The two aggregation bias statistics used are

$$B_1 = \sum_i W_i |Y_i - Z_i| / Z_i$$

and

$$B_2 = \sum_i U_i (Y_i - Z_i)^2 / Z_i^2$$

where Y_i is the aggregated level of the i -th activity using one of the five methods;

Z_i is the aggregated level of the i -th activity using the 52 LP results;

TABLE 2
Results from using different methods of Aggregation

Activity	1	2	3	4	5	Aggregation of 25 LP results
	All resources pooled	3 acreage classes	3 crop/lab classes	LS regression	LP regression	
Sheep (numbers)	131,300	125,300	129,100	124,800	125,100	116,300
Steers (numbers)	0	174	388	79	0	1,121
Wheat (acres)	27,920	26,590	25,780	24,780	24,180	22,630
Barley (acres)	20,230	21,560	22,370	23,370	22,960	22,790
Borrowing (\$)	36,180	43,840	79,560	71,225	96,252	103,698
HHL1* (man-hours)	296	3,741	5,827	7,246	5,707	6,335
HHL2* (man-hours)	43,180	40,790	42,570	43,850	37,560	36,890
HHL4* (man-hours)	1,683	3,006	5,934	7,717	5,285	6,790
Total gross margin (\$)	1,312,000	1,267,000	1,269,000	1,221,000	1,221,000	1,167,000
Error criterion						
$B_1 = \Sigma W Y - Z /Z$	0.1926	0.1306	0.1179	0.0937	0.0786	0.0000
$B_2 = \Sigma U(Y - Z)^2/Z^2$	0.0539	0.0161	0.0141	0.0072	0.0058	0.0000

* HL1, 2 and 4 are activities for hiring labour in Summer, Autumn and Spring respectively.

W_i is the proportion that the net revenue of the i -th activity contributes to total net revenue using the 52 LP results; and U_i is the proportion that the net revenue squared of the i -th activity contributes to total net revenue squared using the 52 LP results.

The criteria B_1 and B_2 would be relevant if the loss functions were linear and quadratic respectively. Table 2 shows that the rank ordering of the bias of the methods is the same whichever criterion is used. The order in which the methods are displayed is the order of decreasing aggregation bias.

All methods result in substantial overestimation of total gross margin. This is 12 per cent when all resources are pooled, 9 per cent when resources are aggregated into three classes, and 5 per cent when regression is used.

Conclusion

Despite the additional flexibility in specifying sufficient conditions for eliminating aggregation bias indicated by Paris and Rausser, there still does not appear to be a practical method of eliminating aggregation bias. It therefore seems that an attempt has to be made to minimize rather than eliminate aggregation bias. One such approach involves the use of cluster analysis. An alternative, put forward in this article, is the use of regression analysis for determining availabilities of resources as functions of size of farm, and using this information in a parametric run of the LP problem in which the size of farm is parameterized.

Results from the case study suggest that with this latter method it is possible to reduce aggregation bias compared with more *ad hoc* methods in at least some cases. They further suggest that obtaining regression fits for resource availabilities by minimization of absolute deviations is worthy of consideration. The question which remains is whether the additional computation required compared with that for the other methods is justified by the higher accuracy of the results. This depends upon the costs of making wrong decisions as a result of relying on estimates which include various levels of aggregation bias.

The regression-parametric LP method could be easily programmed for computer operation. For that matter, so could the method of aggregating the LP solutions for all farms. Facilities exist in most LP packages for running the same technical matrix with a large number of successive righthand sides. All that is required in addition is a subroutine for aggregating the solution vectors. For about 50 farms there might be little difference in the total amount of computation involved. However, for hundreds of farms, the regression-parametric LP method would have the computational advantage.

The amount of aggregation bias resulting from this method depends upon the error in the regression fits. No generalization can be made on the size of this error because it is an empirical matter. Resource indivisibilities and different resource endowments on farms even with the same technical opportunities will ensure that the availabilities of resources are not a simple linear function of size. However, it is to be expected that where such functions can be approximated by linear functions, the method will produce superior results to methods which simply average resources within farm categories.

APPENDIX

Day [4] gives the maximum number of groups to which farms must be allocated for Lee's exact aggregation method for the multi-product, multi-resource case as

$$v = C(m + n, n)$$

where $C(p, q)$ = number of combinations of p things taken q at time;

m = number of rows of the representative LP matrix;

and,

n = number of activities of the representative LP matrix.

This formula defines the number of basic feasible solutions to the LP problem. However, Lee proposes a method of allocating firms before running the LP, based solely on the comparison of the ratio of resource availability on the individual farm to the ratio of resource use in the technical matrix of the representative farm.

Lee's method may be extended to the $m > 2$ case as follows: The number of resource availability ratios possible are $M = C(m, 2)$. For each pair of resources there are a maximum of n resource use ratios from a technical matrix filled with positive elements, and consequently $n + 1$ categories to which to allocate farms on the basis of resource availability ratios on the continuum of ratios between zero and infinity. This implies a maximum total of resource use ratios $w = (n + 1)^M$. If $m > 2$, then $w > v$.

Typically, however, the technical matrix will not consist solely of positive elements. Some will be zero, and some negative. Allowing for the occurrence of zeros first, let $p_i \leq n$ be the number of ratios of resources used in the technical matrix which are non-zero but finite for the i -th ratio. The maximum number of categories required for the i -th ratio is then $p_i + 1$. The maximum number when account is taken of all M resource ratios is the combinatorial function

$$\prod_{i=1}^M (p_i + 1)$$

As an example, consider the following 4×5 technical matrix:

$$\begin{array}{ccccc} a_{11} & 0 & a_{13} & 0 & a_{15} \\ 0 & a_{22} & a_{23} & 0 & a_{25} \\ a_{31} & a_{32} & a_{33} & a_{34} & 0 \\ 0 & 0 & a_{43} & a_{44} & 0. \end{array}$$

There are $M = 6$ possible resource ratios. Considering the ratio of the first two resources, the technical matrix provides two ratios which are non-zero but finite. Three categories must therefore be allowed for the range of ratios of the first two resources available on farms. For each of these categories, three further categories must be allowed for the range of ratios of the first and third resources available on farms. In all, the maximum provision is $3 \times 3 \times 2 \times 3 \times 2 \times 3 = 324$ categories.

If, however, there are negative coefficients in the matrix, the system breaks down, even for Lee's original two resource case. In effect, negative coefficients allow the original resource situation for the individual

farm, as specified by the resource vector for the LP matrix of the farm, to change depending on the relative profitability of introducing the 'supply' activities into the solution. In this case, more information is required before farms can be allocated to categories in which the marginal value product of each resource is equal for all farms in the category.

To conclude, Lee's method does not directly provide a practical method for eliminating aggregation bias. Even for the two resource case it may fail. For the multi-resource case, for cases where it does not fail, the maximum number of categories is very large, and in general larger than the number of feasible bases of the LP.

References

- [1] Ahn, C. Y., 'A Recursive Programming Model of Regional Agricultural Development in Southern Brazil (1960-1970): An Application of Farm Size Decomposition', Unpublished Ph.D. thesis, Ohio State University, 1972.
- [2] Buckwell, A. E. and P. B. R. Hazell, 'Implications of Aggregation Bias for the Construction of Static and Dynamic Linear Programming Supply Models', *J. Agr. Econ.*, 23: 119-134, May 1972.
- [3] Day, R. H., 'On Aggregating Linear Programming Models of Production', *J. Farm Econ.*, 45: 797-813, Nov. 1963.
- [4] ———, 'Exact Aggregation With Linear Programming Models—A Note on the Sufficient Conditions Proposed by R. H. Day—Reply', *Am. J. Agr. Econ.*, 51: 686-688, Aug. 1969.
- [5] Jones, H. B. and J. C. Thompson, 'Squared Versus Unsquared Deviations for Lines of Best Fit', *Agr. Econ. Res.*, 20: 64-69, April 1968.
- [6] Lee, J. E., 'Exact Aggregation—A Discussion of Miller's Theorem', *Agr. Econ. Res.*, 18: 58-61, April 1966.
- [7] Miller, T. A., 'Sufficient Conditions for Exact Aggregation in Linear Programming Models', *Agr. Econ. Res.* 18: 52-57, April 1966.
- [8] Paris, G. and Rausser, G. C., 'Sufficient Conditions for Aggregation of Linear Programming Models', *Am. J. Agr. Econ.*, 55: 659-666, Nov. 1973.
- [9] Wagner, H. M., 'Linear Programming Techniques for Regression Analysis', *Am. Stat. Assoc. J.*, 54: 206-212, March 1959.