Estimation of Production Functions using Average Data

Matthew J. Salois
Food and Resource Economics Department, University of Florida
PO Box 110240, Gainesville, FL 32611-0240
msalois@ufl.edu

Grigorios Livanis
Center for International Business and Education Research, University of Florida
315 Matherly Hall, Gainesville, FL 32611-7140
livanis@ufl.edu

Charles B. Moss
Food and Resource Economics Department, University of Florida
PO Box 110240, Gainesville, FL 32611-0240
cbmoss@ufl.edu

ABSTRACT

Agricultural economists rely on aggregated data at various levels depending on data availability and the econometric techniques employed. However, the implication of aggregation on economic relationships remains an open question. To examine the impact of aggregation on estimation, Monte Carlo techniques and data are employed on production practices.

ADDITIONAL KEY WORDS: Monte Carlo, measurement error, aggregation, production.

JEL CODES: C150, C430, D200, Q100

Selected Paper prepared for presentation at the
Southern Agricultural Economics Association Annual Meetings,
Orlando, Florida, February 5-8, 2006

I. INTRODUCTION

Applied work by agricultural economists have increasingly utilized aggregated or averaged data sets, however the statistical properties and the biases introduced by such data has yet to be adequately explained. More importantly, implications on the economic behavior and relationships predicted by estimates obtained from aggregated samples remain ambiguous. This paper investigates the potential impact of aggregation using Monte Carlo methods and real data on estimation of production functions.

The specific level of aggregation often depends on the research question being addressed. For example, developments in agricultural technology have refocused attention on estimation of the primal production function, which often rely on data aggregated by field levels. Often though, the level of aggregation faced by the agricultural economist depends on the data available which represents numerous levels of aggregation. For example, the USDA's *Agricultural Chemicals and Production Technology: Recommended Data Products* (USDA 2005) dataset is typically averaged by field level observations, while the *Agricultural Chemicals* dataset is aggregated by either crop reporting district or state. The *Agricultural Resource Management Survey* (ARMS) has been aggregated either by farm typology or by production regions, and data is aggregated from a variety of survey instruments to produce state and national level observations.

The demands of the econometric techniques employed may also decide the necessary level of aggregation in a dataset. For instance, while the data in the *Agricultural Chemicals* dataset are appropriate for simple ordinary least squares estimation, these data cannot be pooled over time because observations in one sample period cannot be matched with observations in subsequent or preceding periods. Thus, panel techniques that estimate firm and time series techniques are not supported by the data. To rectify this difficulty, these field or firm level datasets are typically aggregated to a larger unit that maintains confidentiality of sources.

Clearly, aggregation is often necessary and cannot be avoided in agricultural economics. The problems associated with aggregation are not new to economics and date back to Theil (1954), however, nor have they been completely resolved. Stoker (1993) provides an excellent general discussion of the empirical problems related to aggregation with respect to individuals. Felipe and Fisher (2003) focus on the issue of aggregation in production functions and provide both an historical and methodological discussion of the problem.

This paper addresses whether aggregation affects the statistical results of estimation using Monte Carlo analysis and data from the *Agricultural Chemicals* survey. Results from both Monte Carlo methods and data analysis provide convincing support that aggregation leads to biased estimates. The next section reviews the problem of aggregation and discusses some of the relevant literature. Section III reviews the theoretical aspects of aggregation and specifically it's relation to the classic error-in-variables problem. Next the estimation procedure and empirical results are discussed in section IV. Concluding remarks are offered in section V as well as consideration of possible extensions.

II. AGGREGATION IN ECONOMIC ANALYSIS

The phrase "aggregation bias," originally coined by Theil (1954), refers to the estimation problem of an aggregate variable, namely when an aggregate parameter estimate differs from its true value. His work using regressions based on simple average aggregated data revealed slope estimates that were in fact equal to the individual estimates with an additional covariance term. Hence, the aggregation bias was equivalent to the covariance between the aggregate variables and the individual level variables.

While many studies have attempted to deal with aggregation, including the development of several econometric innovations, many have chosen to ignore the issue. This is particularly true in studies that estimate aggregate production functions. Only under highly restrictive conditions on individual or firm level behavior will aggregate parameter estimates be consistent with the individual parameters (Theil 1954). Many criticisms have ensued of aggregate studies that fail to acknowledge that a problem may be present in their results due to aggregation problems or a failure to associate the aggregation bias as a potentially serious problem.

Despite the fact that aggregate production functions have little theoretical foundation compared to their microeconomic counterparts, they remain prevalent in the literature. At the individual or firm level, micro production functions are well behaved. However, aggregates of micro-level production functions into a single aggregate function involve many difficulties, not the least of which is a difficulty in interpreting the properties of an aggregate production function. Recent work has focused on better describing the microeconomic properties of the aggregate production function.

Koebel (2002) described the microeconomic implications of aggregated production functions questioning whether the same optimization framework used for disaggregated production function can be used for their aggregated counterparts. The theoretical model outlined provided support for the notion that the use of aggregated goods and prices will not conflict with orthodox microeconomic theory, though a loss of information does occur in the aggregation process. A possible consequence of this includes biased estimates. The empirical results presented in Koebel (2002) are less optimistic than the theoretical model. Using panel data from 1978-1990 of 27 German industries, the author estimates the input demand system and profit function. He finds that not all microeconomic properties apply to the estimated aggregated function, such as convexity and homogeneity of degree one.

The need for empirical analysis continues, however, as several difficulties remain to be solved. A particular problem in empirical analysis of aggregation is that micro and macro parameters remain largely unknown. Typically, least squares estimates are assumed to coincide with the micro relations true value. However, macro relations are typically approached as a sum of a "true value" composed of both aggregation bias and sampling error. Rather than resorting to mere ad hoc explanation, empirical analysis of the statistical implications from using aggregated data is likely the best method of answering these unresolved issues.

Many recent empirical studies that examine the problem are focused in the investment demand or consumer demand literature. Gordon (1992) used industry aggregated and disaggregated data from the Canadian manufacturing sector to estimate equations on the costs of adjusting inputs in production. His results suggest using aggregated data will result in estimated adjustment cost functions that are greater than industry level estimates (hence an upward bias). Park and Garcia (1994) investigate the effects of aggregating micro-level data on acreage response equations. Their data was obtained from Illinois crop reporting districts from 1960 to 1988. They find that the problem of aggregation is less severe than the problem of specification error in the micro-level data. Although their econometric response equations imply that aggregation bias is present in the state level data, as opposed to the crop reporting district level data, they find that the aggregation bias largely depends on the degree of homogeneity of farms at the CRD level. However, the affect on statistical properties from aggregation is mostly ignored. Gilbert (1986) examines how the use of averaged data effects the testing of the efficient market hypothesis. He finds that not only does averaging data complicate estimation, but also leads to inefficiency. Chung and Kaiser (2002) investigate the presence of aggregation bias using cross-sectional data on U.S. liquid milk advertising and household consumption. Their parameter estimates on the price, income and advertising variables indicated that the aggregated macro model were not only biased, but performed poorly when compared to an alternative disaggregated micro model.

The above studies are just a sampling of the research attempting to deal with issues of aggregation. While all have contributed to our comprehension of the problem, albeit in different ways, the need for better understanding specifically in the context of statistical properties of aggregate estimators is still warranted. This paper further examines the issue of going from micro-level data to macro-level data in production analysis.

III. THEORY AND METHODS

We start by formulating the regression model within a measurement error framework (Fuller 1987). Specifically, we are interested in estimating a regression model

$$y_t = x_t\beta + \varepsilon_t \qquad (1)$$

where y_t is an endogenous variable hypothesized to be a linear function of a set of predetermined exogenous variables, x_t, and ε_t is the resulting residual from the estimated relationship. Given this formulation we hypothesize a set of averages based on some grouping of the original data

$$\begin{aligned} \tilde{y}_g &= \tilde{\beta}\tilde{x}_g + \varepsilon_g \\ \tilde{y}_g &= \frac{1}{N_g}\sum_{t \in g} y_t \\ \tilde{x}_g &= \frac{1}{N_g}\sum_{t \in g} x_g \end{aligned} \qquad (2)$$

where N_g denotes the count or number of observations in group g. The relevant question is then whether $E\left[\beta_g\right] = E\left[\beta\right]$ or $E\left[\beta_g\right] \to E\left[\beta\right]$.

Following the measurement error literature, we return to the original sample

$$\begin{aligned} \tilde{y}_g + \nu_t &= \beta\left(\tilde{x}_g + \varsigma_t\right) + \varepsilon_t \\ y_t &= \tilde{y}_g + \nu_t \\ x_t &= \tilde{x}_g + \varsigma_t \end{aligned} \qquad (3)$$

where we are simply replacing the original values with a measure (some average value of both the dependent and independent variable) plus a measurement error. Under typical assumptions when the errors of each measure are uncorrelated replacement of the actual data with proxies attenuates the regression coefficients. However, in this case, the assumption that the errors are uncorrelated may be erroneous.

Consider the bivariate case of a constant term and a single regressor; if the explanatory variable has been badly measured then the least squares coefficient will be biased towards zero. Extension to multivariate regressions with only a single badly measured variable reveals that the coefficient on that variable is still attenuated, while the others are biased but in unknown directions. Fuller (1987) defines the size of the bias as the reliability ratio which is given by

$$\frac{\mathrm{var}(\tilde{x}_g)}{\mathrm{var}(x_t)} \qquad (4)$$

where the numerator is the true variance and the denominator is the total variance.

IV. ESTIMATION PROCEDURE

To examine the possible effects of aggregation on the estimates, we first turn to Monte Carlo techniques. We start by generating a sample of 200 observations based on random draws from a uniform distribution of three variables. Initially we assume that the true β vector is a vector of ones including a term for the intercept. Based on the random draw and this β vector, we generate a sample of dependent variables by adding a vector of 200 random normal deviations. This combination (the random vector of exogenous variables and the resulting endogenous value) represented our true sample.

Given this sample, we used ordinary least squares to estimate a sample observation for β. Next, we aggregated the sample by taking the average of every group of n observations. We then apply ordinary least squares to the aggregated sample, resulting in an estimated vector $\tilde{\beta}$. Given these two estimates we form two error vectors, one for the difference between the full sample estimates of β and the true unity vector and the other between the aggregated estimate $\tilde{\beta}$ and the unity vector, denoted by μ and $\tilde{\mu}$, respectively. Under the measurement error problem, we expected these errors to be unbiased (which we take to be symmetric around zero).

Table 1 presents the results for the Monte Carlo estimation using values of 5, 10, and 20 for g, hence aggregating every five, ten, and twenty observations. In order to test whether the vectors μ and $\widetilde{\mu}$ were unbiased or not, we used Hotelling's T^2 statistic, an extension of the univariate *t*-statistic, to test $H_0 : \mu = \mu_0, \widetilde{\mu} = \mu_0$ *versus* $\mu \neq \mu_0, \widetilde{\mu} \neq \mu_0$, where μ_0 is a 3×1 vector of zeros.[1] This was completed for three different sample sizes: 100, 500 and 1,000.

From Table 1, several results are clear. First, it appears that the full sample is not biased, an expected result. However, the various aggregated samples produce mixed results. When the sample size is 100, aggregating every five observations does not appear to bias the results, however aggregating every 10 and 20 results in rejection of the null hypothesis. Additionally, the magnitude of the test statistic increases with the level of aggregation. This may suggest that higher levels of aggregation result in increasing bias. Results are similar when the sample size is 500. However, when the sample size is increased to 1,000 rejection of the null hypothesis occurs for all levels of aggregation. Hence, a potential cause for this result is the increasing measurement error that occurs over a larger sample size which has been aggregated.

To examine if aggregation bias persists in real data, we turn to the *Agricultural Chemical* dataset for corn production in 1991. This data set provided 1082 observations across 10 states after observations containing zero yields were dropped from the analysis.[2] Overall, 253 observations for the inputs contained zeros. To circumvent the issue of estimating a logarithmic production function containing zero-level observations

[1] See Rencher (2002) for a good discussion on Hotelling's T^2 statistic.

[2] The 10 states included were: Illinois, Indiana, Iowa, Michigan, Minnesota, Missouri, Nebraska, Ohio, South Dakota, and Wisconsin.

on the inputs, we follow the technique described in Moss (2000) where the zeros are replaced by 0.1, a small positive number.[3]

Using the traditional Cobb-Douglas specification, the production function provides estimates of corn yields as a function of nitrogen, phosphorous, and potassium

$$y = Ax_1^{\alpha} x_2^{\beta} x_3^{\gamma} \qquad (5)$$

where y is the level of corn, x_1 is the level of nitrogen, x_2 is the level of phosphorous, $x_3$and is the level of potassium, and A is the constant term. A linear production function was also estimated.

To examine the implications of aggregation bias on the data set, first the full sample production function was estimated. Next, observations were aggregated according to state and the aggregated production function was then estimated. According to the null hypothesis of no aggregation bias, the estimated parameter vectors for both sets of observations should not be significantly different from one another. That is, we wish to test $H_o : \beta = \tilde{\beta}\ versus\ H_1 : \beta \neq \tilde{\beta}$, where β is the estimated vector of parameters from the full sample and $\tilde{\beta}$ is the estimated vector of parameters from the state-aggregated sample. A Hotelling's T^2 statistic was used to test the null hypothesis of equality across the estimated parameter vectors.

Table 2 presents the results for this estimation procedure for both the linear production function and the Cobb-Douglas production function. Based on the test statistic, we can strongly reject the null hypothesis that the estimated parameters are

[3] Moss (2000) reports that as an observed input level approaches zero a preferred treatment of such observations is to substitute a small positive number as opposed to a bootstrapping technique.

statistically equivalent. This implies that aggregating across states imparts a general bias in the estimated coefficients. However the bias is not systematic, that is, initial results do not indicate the direction of the bias, upwards or downwards.

V. CONCLUSION

The use of averaged or aggregated data cannot be avoided by the agricultural economist, and cannot likely be avoided by any applied economist. Driven by privacy constraints, econometric techniques, or mere data availability, the use of aggregated data is commonplace. In spite of this, the statistical properties of parameter estimates from the use of aggregated data in production analysis remain unresolved. This paper provides evidence that the use of averaged data results in biased parameter estimates.

Monte Carlo experiments indicate that the error terms from an aggregated sample were significantly different from zero. This result conflicts with one of the basic Gauss-Markov assumptions, namely that the error terms have an expected value of zero, and hence indicates biased results. Turning to the *Agricultural Chemical* dataset also provide evidence that aggregated data results in biased parameter estimates. Namely, parameter estimates from a disaggregated dataset are not statistically similar to an aggregated version.

These results have profound implications for agricultural policies and farm decisions based on results from an aggregated dataset. For example, precision agriculture has allowed producers to manage much smaller tracts of land by permitting fertilizer application rates down to the yard. However if yield and input ratios are decided by a production function estimated from averaged data then the potential for mis-specifying the optimal level of fertilizer becomes a serious issue.

This paper addresses the implications of using aggregated data on production practices. Much further consideration is warranted, however. For instance, robustness of the results can be reached through increasing sample size asymptotically. Further investigation on aggregating by crop reporting district or field level observations should be conducted. Additional estimators should also be considered other than OLS. For example, Richter and Brorsen (2006) show the FGLS estimator to be successful in reducing the aggregation bias that occurs in the estimation of school quality measures. Finally, determining the direction of the bias will help in developing methods and possibly new estimators to correct for the problem.

Table 1. Monte Carlo error estimates for full sample and aggregated samples

	Sample Size											
	N=100[1]				N=500[2]				N=1,000[3]			
	Level of Aggregation				Level of Aggregation				Level of Aggregation			
Error Vector	Full	5	10	20	Full	5	10	20	Full	5	10	20
β_0-β_t	.0061	.0495	.1083	.0190	.0146	.0511	.0400	-.1010	.0088	-.0750	-.0717	-.1501
β_1-β_t	-.0062	-.0741	-.1502	-.1195	-.0132	-.0083	-.0070	.1255	-.0072	.0658	-.0114	-.0143
β_2-β_t	.0134	-.0171	-.0823	.0672	-.0161	-.0954	-.0860	.0811	-.0051	.0925	.1582	.3155
T^2 statistic	2.600	2.414	9.462	11.1846	4.020	3.991	9.002	14.3823	2.627	11.239	10.594	50.408

[1] The critical value of the T^2 statistic is 8.257 for 3 variables and 200 observations.

[2] The critical value of the T^2 statistic is 7.922 for 3 variables and 500 observations.

[3] The critical value of the T^2 statistic is 7.857 for 3 variables and 1,000 observations.

Table 2. Linear and Cobb-Douglas parameter estimates

	Functional Form of Production Function			
	Linear Function[1]		Cobb-Douglas Function[2]	
Input	Full Sample	State Aggregation	Full Sample	State Aggregation
Constant	94.266	88.415	4.318	4.159
	(2.697)	(19.126)	(.0488)	(.4483)
Nitrogen	0.1501	0.4854	.0737	.1985
	(.0202)	(.2236)	(.0111)	(.0917)
Phosphorous	-.0739	-1.312	-.0103	-.2440
	(.0301)	(.8089)	(.0081)	(.1456)
Potassium	.0287	0.4488	.0042	.1104
	(.0228)	(.3750)	(.0064)	(.0710)
T^2 statistic	2709.37		2747.63	

[1] The critical value of the T^2 statistic is 9.488 for 4 variables and 1082 observations.

[2] The critical value of the T^2 statistic is 23.545 for 4 variables and 10 observations.

REFERENCES

Ashenfelter, O. and A. Krueger (1994) Estimates of the Economic Return to Schooling from a New Sample of Twins, *American Economic Review*, **84**, 1157-1173.

Boot, J.C.G., and G.M. de Wit (1960) Investment Demand: An Empirical Contribution to the Aggregation Problem, *International Economic Review*, **1**: 3-30.

Chung, C., and H.M. Kaiser (2002) Advertising Evaluation and Cross-Sectional Data Aggregation, *American Journal of Agricultural Economics*, **84**: 800-806.

Felipe, Jesus, and Franklin M. Fisher (2003) Aggregation in Production Functions: What Applied Economists Should Know, *Metroeconomica*, **54**: 208-262.

Fuller, W.A. (1987) *Measurement Error Models*. New York: John Wiley and Sons.

Gilbert, C.L. (1986) Testing the Efficient Markets Hypothesis on Averaged Data, *Applied Economics*, **18**: 1149-1166.

Gordon, Stephen (1992) Costs of Adjustment, the Aggregation Problem and Investment, *Review of Economics and Statistics*, **74**: 422-429.

Greene, W.H. (2003) *Econometric Analysis, 5th Edition.* New York: Prentice Hall.

Imbens, G. and D. Hyslop. (2001) Bias from Classical and Other Forms of Measurement Error, *Journal of Business and Economic Statistics*, **19**, 475-481.

Koebel, B.M. (2002) Can Aggregation Across Goods be Achieved by Neglecting the Problem? Property Inheritance and Aggregation Bias, *International Economic Review*, **43**: 223-255.

Klepper, S. and E. Leamer (1983) Consistent Sets of Estimates for Regressions with Errors in All Variables, *Econometrica*, **52**, 163-183.

Levi, M. (1973) Errors in the Variables in the Presence of Correctly Measured Variables, *Econometrica*, **41**, 985-986.

Moss, C.B. (2000) Estimation of the Cobb-Douglas with Zero Input Levels: Bootstrapping and Substitution." *Applied Economics Letters*, **7**, 677-679.

Rencher, A.C. *Methods of Multivariate Analysis, 2nd Edition*. New York: John Wiley and Sons, 2002.

Richter, F.G.-C. and B.W. Brorsen (Forthcoming). Journal of Productivity Analysis.

Stroker, Thomas M. (1993) Empirical Approaches to the Problem of Aggregation Over Individuals, *Journal of Economic Literature*, **31**: 1827-1874.

Theil, H. (1954) *Linear Aggregation of Economic Relations*. North-Holland Publishing: Amsterdam.

United States Department of Agriculture, Economic Research Service "Agricultural Chemicals and Production Technology: Recommended Data Products." Website http://www.ers.usda.gov/Briefing/AgChemicals/data.html Accessed September 9, 2005.