LEAST SQUARES ROBUSTNESS WHEN THE ERROR TERM IS MISSPECIFIED IN COBB-DOUGLAS TYPE FUNCTIONS

By Terry N. Barr and James F. Horrell*

The Cobb-Douglas type function has a multiplicative deterministic form, $Y = \beta_0 X_1^{\beta_1} X_2^{\beta_2}, \ldots, X_k^{\beta_k}$, which can be used to analyze certain economic structures. The usual error specification for this function is that the error term enters multiplicatively with lognormal distribution. This specification permits use of a natural logarithmic transformation and the usual linear least squares estimation techniques with slight modifications. The study reported on investigates the effect of an error misspecification on the estimates of the parameters. In particular, is there a significant bias in the estimates of the parameters $\beta_1, \beta_2, \ldots \beta_k$ when the logarithmic transformation is applied to data generated using additive errors? Monte Carlo techniques were used to generate empirical distributions of the estimates to be analyzed with analysis of variance (ANOVA) techniques. Evidence suggests bias would exist but not of the magnitude, in most cases, to greatly challenge previous economic evaluations based on regressions using Cobb-Douglas type functions. Though this may be a source of consolation for many analysts, it should be noted that any conclusions that loglinear estimation yields satisfactory estimates in the presence of additive errors must still depend on the peculiarities of the sampling experiment.
Keywords: Production functions, econometrics, error misspecification.

INTRODUCTION

Economists have traditionally used the following Cobb-Douglas type functions in studies of demand and production (*4*, pp. 464-72).[1]

$$Y = \beta_0 X_1^{\beta_1} X_2^{\beta_2} \ldots X_n^{\beta_n} \tag{1}$$

The error specification typically facilitates the use of the logarithmic transformation and the usual least squares estimation techniques with slight modifications (*6*, pp. 1034-38). The multiplicative error assumption can be rationalized through the multiplicative central limit theroem. But its pragmatic justification, namely allowing the Cobb-Douglas form to be intrinsically linear, has been eroded with the advance of nonlinear estimation techniques (*2*, pp. 101-102).

The question for the researcher is how serious is the bias in estimation if the assumption of the multiplicative error is incorrect and additive errors are present? Goldfeld and Quandt (*5*, pp. 251-57) have proposed a method of estimation devised to account simultaneously for additive and multiplicative errors. Although the technique appeared manageable and the results reasonable, a number of interesting statistical questions remained open. In particular is the question of the need to know more about the consequences of misspecification of the error terms. The problem is also discussed in many econometrics texts (*3*, pp. 213-18) (*11*, pp. 591-94). This article is designed to provide some insights into this question.

Two error specifications which are the most prevalent in the literature are:

$$Y_i = \beta_0 X_i^{\beta_1} e^{u_i} \qquad U \text{ is } N(0, \sigma^2) \tag{2}$$

$$Y_i = \beta_0 X_i^{\beta_1} W_i \qquad W \text{ is lognormal with mean } e^{\frac{1}{2}\sigma_v 2} \tag{3}$$

Suppose the *true model* which generated the Y_i's is

$$Y_i = \beta_0 X_i^{\beta_1} + V_i \qquad V_i \text{ is } N(0, \sigma^2 I) \tag{4}$$

But the specification used for estimation is the traditional model of (2). The conclusions will apply equally well to (3) but notational convenience dictates the alternative form. Employing the natural logarithmic transformation, equation (2) becomes

$$\ln Y_i = \ln\beta_0 + \beta_1 \ln X_i + u_i \tag{5}$$

Upon examination it is clear that (5) is a linear model and the OLS estimates of $\ln\beta_0$ and β_1 are given by

$$\ln b_0 = \frac{1}{n}\sum_{i=1}^{n} \ln Y_i - b_1 \frac{1}{n}\sum_{i=1}^{n} \ln X_i = \overline{\ln Y} - b_1 \overline{\ln X} \tag{6}$$

*Terry N. Barr is an economist with the National Economic Analysis Division, ERS, and James F. Horrell is an associate professor of business administration at the University of Oklahoma.

[1] Italicized numbers in parentheses refer to items in References at the end of this article.

$$b_1 = \sum_{i=1}^{n} (\ell n X_i - \overline{\ell n X})\ (\ell n Y_i - \overline{\ell n Y}) \Big/ \sum_{i=1}^{n} (\ell n X_i - \overline{\ell n X})^2 \tag{7}$$

By noting that

$$\sum_{i=1}^{n} (\ell n X_i - \overline{\ell n X})k = k \sum_{i=1}^{n} (\ell n X_i - \overline{\ell n X}) = k \cdot 0 = 0 \tag{8}$$

we have, upon simplifying (7)

$$b_1 = \sum_{i=1}^{n} (\ell n X_i - \overline{\ell n X})\ell n Y_i \Big/ \sum_{i=1}^{n} (\ell n X_i - \overline{\ell n X})^2 \tag{9}$$

Note from (4) that

$$\ell n Y_i = \ell n \beta_0 X_i^{\beta_1} (1 + V_i / \beta_0 X_i^{\beta_1}) = \ell n \beta_0 + \beta_1 l_n X_i + \ell n (1 + V_i / \beta_0 X_i^{\beta_1}). \tag{10}$$

Substituting (10) into (9) and using (8) yields

$$b_1 = \left[\beta_1 \sum_{i=1}^{n} (\ell n X_i - \overline{\ell n X})\ \ell n X_i\right] / D + \left[\sum_{i=1}^{n} (\ell n X_i - \overline{\ell n X})\ \ell n (1 + V_i / \beta_0 X_i^{\beta_1})\right] / D$$

where $D = \sum_{i=1}^{n} (\ell n X_i - \ell n X)^2$. The coefficient on β_1 reduces to 1, thus

$$E(b_1) = \beta_1 + \sum_{i=1}^{n} (\ell n X_i - \overline{\ell n X}) E\left[\ell n (1 + V_i / \beta_0 X_i^{\beta_1})\right] \Big/ \sum_{i=1}^{n} (\ell n X_i - \overline{\ell n X})^2 \tag{11}$$

Looking at the term on the right-hand side of (11) that has the expected value operator, E, applied to it, we note the following. *Analytically* it is impossible to assume a multiplicative error term and make a logarithmic transformation when the error term is actually additive and normally distributed unless the error term, when negative, does not dominate the deterministic part of the model. For analytical purposes, it is necessary to truncate the error term distribution so that taking the logarithm of negative numbers can be avoided. Thus to obtain the expected value of b_1, it is necessary to evaluate

$$E\left[\ell n (1 + V_i / \beta_0 X_i^{\beta_1})\right] = K \frac{\int \infty l_n}{-\beta_0 X_i \beta_1} (1 + V_i / \beta_0 X_i^{\beta_1}) \exp\left[-1/2 \frac{(V_i - \mu)^2}{\sigma}\right] dV_i$$

This evaluation will be carried out empirically with the use of Monte Carlo and ANOVA techniques. The Monte Carlo technique will be employed to generate empirical distributions of the estimates in accord with two specific stochastic functional forms.

Functional Form I: $Y_i = \beta_0 X_i^{\beta_1} + V_i$

Functional Form II: $Y_i = \beta_0 X_{1i}^{\beta_1} X_{2i}^{\beta_2} + V_i$

Once construction of the distributions is complete, analysis of variance (ANOVA) techniques employing factorial models can be utilized to analyze the effects on parameter bias of varying the sample characteristics.

DATA GENERATION AND MONTE CARLO APPLICATION

To illustrate the generation of the sample data for regression analysis, consider the single-variable model

$$Y_i = \beta_0 X_i^{\beta_1} + V_i$$

The value of the deterministic part of this model, namely $\beta_0 X_i^{\beta_1}$, is generated for each possible combination of β value, X range, and sample size identified in table 1. For all combinations the value of β_0 remains at 4.0.

For example, utilizing the first variant of each characteristic, a set of Y's would be constructed for $\beta_0 = 4$, $\beta_1 = .5$ in which X_i takes on values uniformly in the interval [5, 15] with a sample size of twenty, ($i = 1, \ldots, 20$). All of the generated (X, Y) points would lie on the graph of the function $Y = 4X^{.5}$. For example, if X = 5, the nonstochastic Y = 8.94. The Monte Carlo technique is used to obtain the probability distribution of estimators of any parameter for comparison with the true value of the parameter. By repeatedly adding random error values V_i from a known distribution to the deterministic portion of each combination and performing regression analysis, we derive an empirical distribution of each estimator.

Table 1.—Sample characteristics for functional form I

β values	X range	Sample size	Error variance
1. $\beta_1 = .5$	1. $5 \leqslant x \leqslant 15$	1. $n = 20$	1. $\sigma^2 = 4$
2. $\beta_2 = 1.0$	2. $5 \leqslant x \leqslant 85$	2. $n = 40$	2. $\sigma^2 = 9$
3. $\beta_3 = 1.5$			3. $\sigma^2 = 16$
			4. $\sigma^2 = 25$
			5. $\sigma^2 = 36$
			6. $\sigma^2 = 49$
			7. $\sigma^2 = 64$
			8. $\sigma^2 = 81$
	(96 possible combinations)		

Although the generation of the independent, normally distributed random observations V_i with desired variance would appear conceptually easy, in actuality it involves a considerable amount of justification. The reader interested in this problem and its subtleties is referred to (*13*, pp. 39-61).

For each combination of sample characteristics and the corresponding unique deterministic data set, a total of 100 different sets of independent error values based on the error variance are created and added. The result is 100 different data sets for each of the 96 combinations. Applying least squares analysis to each set yields 100 estimates of β_1 for each of the 96 possible sample combinations enumerated in table 1. (The total is 9,600 regressions, not including one replication for $\sigma^2 = 16$). Each of these sets of 100 estimates constitutes, under slightly different conditions, a small empirical distribution of the estimator $\widehat{\beta}_1$, and each provides a sample for estimating $E(\widehat{\beta}_1) = \beta_1 +$ bias. Since we know the actual value of β_1, we can subtract it from our estimate of β_1, namely, $\bar{\beta} = 1/100 \sum_{i=1}^{100} \widehat{\beta}_{1i}$. The difference, b_i, provides a good estimate of the bias created in estimating β_1 under conditions of error misspecification. The $\bar{\beta}$ are presented in table 2 for each of the respective combinations.

A similar process was followed in generating Monte Carlo results for functional form II, $Y_i + \beta_0 X_{1i}^{\beta_1} X_{2i}^{\beta_2} + V_i$. However, the 1,444 total combinations were based on fewer error variance levels and dual β-values and X-range values (table 3). The $\bar{\beta}$ corresponding to each combination are shown in table 4 (14,400 regressions).

HYPOTHESES OF INTEREST

In examining the results of the Monte Carlo experiment a number of hypotheses are of interest. Those listed below that do not have an alternative hypothesis explicitly stated should be assumed to have, for an alternative hypothesis, H_a: There exists at least one inequality. Bias refers to the estimates of the respective β's.

1. H_0: All b_{ijkl}/s_{ijkl} are equal[2]
2. H_0: The biases due to the levels of "return to scale" averaged over all other factors are equal
3. H_0: The biases due to the different sample sizes averaged over all the other factors are equal
4. H_0: The biases due to the different "regions of the independent variables" averaged over all other factors are equal
5. H_0: The biases due to the different "error variances," averaged over all other factors, are equal
6. H_0: There are interactions in the factors of the experiment

vs H_a: There are no interactions in the factors of the experiment

The hypothesis stated in 6 is ambiguous because a number of different interaction hypotheses can be stated. At this point, it is unnecessary to state them specifically; it is only essential to point out that such hypotheses need special consideration.

If equality is accepted in these hypotheses, it is pertinent to ask if the common value is different from zero and, if so, in what direction is it different. If equality is rejected, then obviously some of the biases are different from zero and only the differences need to be investigated.

It can be shown that rejection of any of the hypotheses, 2, 3, 4, 5, and 6, will not indicate the acceptance or rejection of hypothesis 1. If there is interest in other hypotheses, the model must be reclassified and an additional analysis of variance performed.

A completely randomized design involving a factorial treatment arrangement was taken from statistical theory to assess the general impacts and to test hypotheses 2 through 6. Cost factors in generation of the data have prevented obtaining a complete replicate of the experiment that would be necessary to test hypothesis 1.

[2] This is a standardizing transformation of the $\widehat{\beta}_{ijkl}$ to satisfy assumptions of ANOVA factorial models. A more detailed justification appears later in the article.

Table 2.—Monte Carlo results for functional form I

		$\beta = .5$		$\beta = 1$		$\beta = 1.5$	
	Variance	Mean of βs	Standard deviation	Mean of βs	Standard deviation	Mean of βs	Standard deviation
	4	.5160	.1373	.9959	.0508	1.4995	.0191
	9	.5214	.2156	1.0137	.0715	1.5010	.0267
	16	.5393	.2709	.9995	.0914	1.5025	.0376
	16	.5560	.2952	1.0113	.0950	1.5013	.0383
n = 20	25	.5429	.3855	1.0337	.1308	1.5060	.0524
	36	.5332	.5973	1.0713	.1689	1.5135	.0575
	49	.5667	.6756	1.0295	.2114	1.5201	.0761
	64	.4736	.5968	1.0195	.2273	1.4980	.0796
	81	.4268	.5701	1.1304	.3394	1.5039	.0869
$5 \leq x \leq 15$							
	4	.5134	.0945	1.0075	.0352	1.5001	.0144
	9	.5324	.1554	1.0163	.0590	1.5013	.0207
	16	.5215	.2344	1.0010	.0684	1.5022	.0266
	16	.5854	.2319	1.0196	.0714	1.5061	.0278
n = 40	25	.5993	.3348	1.0170	.0904	1.5039	.0319
	36	.5377	.3163	1.0377	.1124	1.5091	.0456
	49	.4646	.3410	1.0582	.1358	1.5162	.0422
	64	.4825	.4285	1.0681	.1744	1.5060	.0561
	81	.3916	.4183	1.0596	.1859	1.5158	.0682
	4	.5100	.0512	1.0015	.0130	1.4996	.0042
	9	.5191	.0795	1.0035	.0182	1.5011	.0058
	16	.5201	.0953	1.0030	.0285	1.5010	.0075
	16	.5340	.1108	1.0020	.0234	1.4993	.0083
n = 20	25	.5334	.1278	1.0013	.0279	1.5007	.0096
	36	.5636	.1810	1.0075	.0451	1.5004	.0123
	49	.5876	.2075	1.0039	.0500	1.5002	.0133
	64	.5411	.1762	1.0126	.0529	1.5040	.0172
	81	.5555	.2275	1.0041	.0610	1.5016	.0184
$5 \leq x \leq 85$							
	4	.5029	.0274	1.0010	.0093	1.4999	.0031
	9	.5098	.0526	1.0006	.0147	1.5000	.0054
	16	.5324	.0800	1.0011	.0237	1.4991	.0067
	16	.5151	.0625	1.0039	.0205	1.4995	.0073
n = 40	25	.5537	.1111	1.0077	.0250	1.5009	.0088
	36	.5548	.1149	1.0064	.0299	1.5029	.0107
	49	.5556	.1330	1.0029	.0364	1.5032	.0118
	64	.5526	.1415	1.0134	.0412	1.5009	.0137
	81	.5610	.1806	1.0058	.0477	1.4996	.0149

Table 3.—Sample characteristics for functional form II

β values	X range	Sample size	Error variance
1. $\beta_1 = .25$ $\beta_2 = .25$	(5x15) x (5x15)	$n = 20$	$\sigma^2 = 9$
2. $\beta_1 = .4$ $\beta_2 = .1$	(5x15) x (5x85)	$n = 40$	$\sigma^2 = 25$
3. $\beta_1 = .5$ $\beta_2 = .5$	(5x85) x (5x85)		$\sigma^2 = 49$
4. $\beta_1 = .9$ $\beta_2 = .1$			$\sigma^2 = 81$
5. $\beta_1 = .75$ $\beta_2 = .75$			
6. $\beta_1 = 1.4$ $\beta_2 = .1$			
	(1,444 possible combinations)		

EMPIRICAL RESULTS

Empirical Distributions and ANOVA Models

Consider the data layout in tables 2 and 4, in which the entries are $\bar{\beta}$ values. Each is in a cell characterized by four subscripts, i, j, k, and l, where i denotes the ith "return to scale," j denotes the jth "region of the independent variable," k denotes the kth "sample size," and l denotes the lth "error variance."

From an examination of the data layout, it is apparent that a completely randomized design involving a factorial treatment arrangement should be used. For functional form I, ANOVA will be applied, and, recognizing the presence of two responses under form II, MANOVA techniques will be used. For more information on both techniques, see (*7, 8, 10*). The distribution of $\bar{\beta}_{ijkl}$ indicates that certain transformations must be made to these data to realistically satisfy the normality of the factorial model. Since $\bar{\beta}_{ijkl}$ is a mean of 100 estimates, $\bar{\beta}_{ijkl}$ is normally distributed with variance $\frac{\sigma^2_{ijkl}}{100}$, where σ^2_{ijkl} is the variance of $\hat{\beta}_{ijkl}$. The variance of $\hat{\beta}_{ijkl}$ is not known, but a good estimate of σ^2_{ijkl} based on 100 observations is s^2_{ijkl}. So $\bar{\beta}_{ijkl}$ is distributed approximately $N(\beta_i + b_{ijkl}, \sigma^2_{ijkl}/100)$ and, consequently, $(\beta_{ijkl} - \beta_i)/s_{ijkl}$ is approximately distributed $N(b_{ijkl}/s_{ijkl}, 1/100)$.

Since there are no *a priori* reasons for making strong assumptions concerning interaction and there is generally only one observation per cell, the analysis will proceed in stages. The first stage will be to use a three-way factorial analysis of variance model on a transformation of the data in table 2 that corresponds to:

- The three levels of "return to scale" $\beta = .5$, 1.0, and 1.5;
- The two "regions of the independent variable" $5 \leqslant X \leqslant 15$, $5 \leqslant X \leqslant 85$;
- The two "sample sizes" $n = 20$, $n = 40$.

The "error variance" is $\sigma^2 = 16$ with two replications.

Based on the conclusions of the first stage, we will use a modified four-way factorial model on the transformed data excluding only that part of the data that was generated to use as a replicate in the factorial model of stage one.

On the basis of the conclusions of the first two stages, the selection of a multivariate factorial model for analyzing the second functional form $Y_i = \beta_0 X_{1i}^{\beta_1} X_{2i}^{\beta_2} + V_i$ can be made. The results of the Monte Carlo application to this functional form appear in table 4.

Functional Form I $Y = \beta_0 X^{\beta_1}$

In the first stage the concern is the degree of interaction of effects among the sample characteristics considered:

1. Return to scale $= \alpha_i$ $(i = 1, \ldots, 3)$
2. Region of independent variable $= \nu_j$ $(j = 1,2)$
3. Sample size $= \delta_k$ $(k = 1,2)$
4. Error variance $= \delta_l$ $(l = 1, \ldots, 9)$

To find out if there are interactions among the characteristics that will require making allowances for them, the analysis begins by holding the error variance fixed and varying factor 1, factor 2, and factor 3 where the assumed model is

$$Y_{ijkm} = \mu + \alpha_i + \upsilon_j + \gamma_k + (\alpha\upsilon)_{ij} + (\alpha\gamma)_{ik} + (\upsilon\gamma)_{jk} + (\alpha\upsilon\gamma)_{ijk} + \epsilon_{ijkm} \qquad (12)$$

$i = 1, 2, 3 \qquad j = 1, 2 \qquad k = 1, 2 \qquad m = 1, 2$

Since this model provides no estimate of error if there is no replication, two sets of $\bar{\beta}$ values were obtained for the error variance $\sigma^2 = 16$. Consequently, we have a three-way factorial model with two observations per cell.

It is not realistic to assume that the ϵ's are distributed independently $N(0, \sigma^2)$. More realistically, we should assume that $\underline{\epsilon} \sim N(\underline{0}, \sigma^2\Sigma)$, where Σ is a diagnonal matrix. Consequently, the $\bar{\beta}$ values which appear in the tables have been transformed for ANOVA by subtracting out the actual value of the parameter and dividing by the estimated standard deviation.

Table 4.—Monte Carlo results for functional form II

		σ^2	$\beta = .25$		$\beta = .25$		$\beta = .4$		$\beta = .1$		$\beta = .5$		$c = .5$	
		ν	Mean of βs	Standard deviation	Mean of βs	Standard deviation	Mean of βs	Standard deviation	Mean of βs	Standard deviation	Mean of βs	Standard deviation	Mean of βs	Standard deviation
5x15 x 5x15	n = 20	9	.3107	.2050	.2889	.2176	.4151	.1674	.1061	.1947	.4950	.0669	.4984	.0568
		25	.3281	.4320	.2391	.4595	.4504	.3980	.1874	.3705	.5131	.1213	.5019	.1002
		49	.2296	.5820	.3609	.5020	.4314	.5992	.1893	.5359	.5056	.1686	.5337	.1387
		81	.2367	.6684	.2203	.5896	.2713	.5980	.0635	.6270	.5364	.2364	.5389	.2217
	n = 40	9	.2811	.1500	.2770	.1473	.4289	.1621	.1297	.1436	.5050	.0468	.5054	.0479
		25	.2544	.2738	.2920	.2972	.3973	.2811	.2031	.2488	.5161	.0811	.5069	.0714
		49	.2783	.3626	.2225	.3834	.2918	.3817	.1512	.4389	.5383	.1299	.5142	.1263
		81	.2815	.4328	.2100	.4513	.3086	.4532	.1609	.4001	.5429	.1563	.5446	.1472
5x15 x 5x15	n = 20	9	.2483	.0708	.2534	.1292	.4168	.0813	.0937	.1109	.5027	.0204	.4989	.0428
		25	.2708	.1588	.2591	.2411	.4275	.1318	.1153	.2119	.5035	.0325	.5033	.0638
		49	.2846	.2163	.2894	.3713	.4138	.1871	.1034	.3632	.5052	.0491	.4948	.0910
		81	.3085	.2960	.3020	.4954	.4508	.2279	.1133	.4389	.5136	.0623	.5146	.1067
	n = 40	9	.2589	.0458	.2764	.1138	.4102	.0514	.1045	.0888	.5020	.0129	.5014	.0249
		25	.2702	.1061	.2702	.2025	.4387	.1175	.1025	.1642	.5012	.0233	.5032	.0454
		49	.2748	.1356	.3041	.2750	.4604	.1557	.0740	.2388	.5077	.0378	.5043	.0570
		81	.2618	.1853	.2786	.3963	.4372	.1629	.1022	.2997	.5163	.0450	.5247	.0839
5x15 x 5x15	n = 20	9	.2556	.0560	.2594	.0545	.4099	.0693	.0988	.0603	.4999	.0125	.5014	.0135
		25	.2734	.1017	.2523	.1013	.4256	.1217	.1077	.1134	.5052	.0220	.5007	.0198
		49	.2859	.1721	.2948	.1559	.4658	.1883	.1100	.1427	.5096	.0366	.5052	.0328
		81	.2943	.2146	.2755	.1951	.4207	.1790	.0897	.1965	.5067	.0425	.4985	.0402
	n = 40	9	.2538	.0430	.2549	.0437	.4071	.0449	.1019	.0371	.4997	.0060	.4943	.0096
		25	.2604	.0707	.2591	.0709	.4254	.1157	.1135	.0791	.5003	.0119	.5009	.0158
		49	.2803	.1047	.2955	.1305	.4438	.1361	.0858	.1177	.5031	.0196	.5043	.0238
		81	.2679	.1557	.2734	.1547	.4600	.1469	.1061	.1364	.5030	.0254	.5120	.0332

		σ^2	$\beta = .9$		$\beta = .1$		$\beta = .75$		$\beta = .75$		$\beta = 1.4$		$\beta = .1$	
		ν	Mean of βs	Standard deviation	Mean of βs	Standard deviation	Mean of βs	Standard deviation	Mean of βs	Standard deviation	Mean of βs	Standard deviation	Mean of βs	Standard deviation
5x15 x 5x15	n = 20	9	.9129	.0765	.1090	.0533	.7509	.0230	.7529	.0216	1.4064	.0266	.1022	.0204
		25	.9113	.1176	.1016	.1053	.7574	.0383	.7535	.0324	1.4082	.0522	.0952	.0383
		49	.9552	.2524	.1182	.1621	.7494	.0533	.7506	.0463	1.4100	.0640	.0926	.0494
		81	.9859	.2462	.1249	.1836	.7643	.0755	.7589	.0607	1.4050	.0794	.0938	.0605
	n = 40	9	.9122	.0569	.0963	.0410	.7477	.0151	.7511	.0159	1.3981	.0220	.0983	.0165
		25	.9214	.0844	.1026	.0654	.7516	.0290	.7538	.0328	1.4066	.0344	.0963	.0246
		49	.9171	.1343	.1110	.0971	.7560	.0329	.7593	.0387	1.4041	.0482	.1022	.0362
		81	.9606	.1964	.1492	.1941	.7518	.0569	.7566	.0539	1.4066	.0676	.1094	.0418
5x15 x 5x85	n = 20	9	.9033	.0219	.0953	.0294	.7500	.0063	.7511	.0103	1.4000	.0060	.0997	.0073
		25	.9009	.0312	.0963	.0419	.7488	.0112	.7494	.0174	1.4002	.0111	.1002	.0138
		49	.9095	.0462	.0897	.0690	.7509	.0159	.7518	.0312	1.4013	.0127	.0997	.0158
		81	.9142	.0733	.0849	.0955	.7525	.0190	.7501	.0305	1.4029	.0208	1.0161	.0269
	n = 40	9	.9014	.0143	.0990	.0158	.7496	.0041	.7505	.0076	1.4004	.0049	.0991	.0038
		25	.9005	.0229	.1053	.0262	.7503	.0079	.7504	.0143	1.4001	.0095	.0990	.0064
		49	.9095	.0340	.0974	.0372	.7497	.0087	.7508	.0160	1.4030	.0107	.1003	.0082
		81	.9144	.0585	.1018	.0507	.7524	.0129	.7505	.0251	1.4015	.0157	.0996	.0138
5x85 x 5x85	n = 20	9	.9006	.0239	.1007	.0112	.7504	.0032	.7500	.0031	1.3984	.0096	.0995	.0031
		25	.9061	.0477	.0985	.0218	.7501	.0052	.7503	.0053	1.3998	.0173	.0993	.0049
		49	.9099	.0618	.1014	.0306	.7504	.0091	.7502	.0065	1.4059	.0223	.0996	.0070
		81	.9233	.0796	.1062	.0354	.7508	.0096	.7495	.0087	1.4017	.0306	.1001	.0081
	n = 40	9	.9008	.0146	.1000	.0087	.7496	.0026	.7497	.0018	1.3995	.0049	.1000	.0028
		25	.8999	.0281	.0995	.0171	.7501	.0043	.7500	.0031	1.4002	.0077	.0997	.0044
		49	.9035	.0313	.1036	.0212	.7500	.0066	.7502	.0049	1.4010	.0102	.0998	.0054
		81	.9167	.0466	.0972	.0300	.7510	.0084	.7506	.0051	1.4001	.0152	.0994	.0078

The analysis of variance results in table 5 indicate that no significant interaction effects of the second or third order occur in the three variables "return to scale," "sample size," and "region of independent variable." There exists the possibility that a "return to scale - region of independent variable" interaction exists, since the F-ratio of 2.28 would be significant at a probability level of approximately .175. The F-ratio 7.88 indicates a probability greater than .99 that "return to scale" affects the bias.

Table 5.—ANOVA for a three-factor factorial with replication: functional form I

Factor	Degrees of freedom	Sum of squares	Mean squares	F-ratio
1	2	.18863	.09431	[a]7.88
2	1	.00124	.00124	.10
3	1	.00716	.00916	.60
12	2	.05454	.02727	2.28
13	2	.00676	.00338	.28
23	1	.01348	.01348	1.13
123	2	.01400	.00700	.58
Error	12	.14366	.01197	
Total	23	.42947		

[a]Significant at .01 level.

The means for the levels of "return to scale" are .2449, .1051, and .0310 for $\beta = .5$, $\beta = 1$, and $\beta = 1.5$, respectively. Hence, the bias is greatest when the return to scale is lowest and least when the return to scale is greatest. The means for the levels of "return to scale" are independent, unbiased estimates of the biases due to their respective levels. In theory, the error variance in the ANOVA is independent of the means; consequently, Tukey's T method can be used as a followup analysis to determine the nature of the differences in levels of bias corresponding to the levels of "return to scale." The T method indicates that the mean response for $\beta = .5$ and for $\beta = 1$ differ at approximately the .06 significance level and the mean response for $\beta = .5$ and for $\beta = 1.5$ differ at the .01 significance level.

Since the "return to scale" factor is quantitative, we can consider the functional form of the response curve for this factor. When there are just three treatments, the sum of squares can only be split into two quantities, linear and quadratic responses. The total sum of squares for "return to scale" is .18863, and the sum of squares for linearity and for a quadratic response are .18286 and .00577, respectively. This would suggest a linear response curve.

Since the results of the ANOVA suggest that a significant interaction exists between "return to scale" and "region of independent variable" we now incorporate it and a fourth variable "error variance" into the model. The new model is

$$Y_{ijkl} = \mu + \alpha_i + \upsilon_j + \gamma_k + \delta_l + (\alpha\upsilon)_{ij} + (\alpha\delta)_{il} + (\upsilon\delta)_{jl} + (\gamma\delta)_{kl} + \epsilon_{ijkl} \qquad (13)$$

$i = 1, 2, 3 \qquad j = 1, 2 \qquad k = 1, 2 \qquad l = 1, 2, \ldots, 8$

Note that third and fourth order interactions do not appear in the model. There are two reasons: (1) in the first stage, higher order interactions seemingly did not exist, and (2) each higher order interaction that can be safely assumed out of the model causes an increase in the stability of the estimate of the error variance. It should also be noted that the significant interaction effect found in the first stage, $(\alpha v)_{ij}$, remains in this model.

The distribution of $\underline{\epsilon}$ is again most realistically assumed to be $N(\underline{0}, \sigma^2\Sigma)$; consequently, the $\bar{\beta}$ must be transformed for the analysis.

The results of the analysis of variance for the second stage are presented in table 6.

Table 6.—ANOVA for a four-factor factorial without replication: functional form I

Factor	Degrees of freedom	Sum of squares	Mean squares	F-ratio
1	2	.15087	.07544	[a]5.825
2	1	.04364	.04364	3.370
3	1	.04205	.04205	3.247
4	7	.29436	.04205	[a]3.247
12	2	.53364	.26682	[a]20.604
14	14	.24608	.01758	1.357
24	7	.06162	.00880	.680
34	7	.05405	.00772	.596
Error	54	.69913	.01295	
Total	95	2.12544		

[a]Significant at .01 level.

Table 6 F-ratios indicate significant main effects for "return to scale" and "error variance" and a very significant "return to scale - region of independent variable" interaction.

The presence of the interaction complicates the making of inferences concerning main effects. Mean values for a "return to scale" across "region of independent variable" classification appear below:

Data range	$\beta = .5$	$\beta = 1.0$	$\beta = 1.5$
$5 < x < 15$	.0568	.2233	.1192
$5 < x < 85$	.3090	.1408	.0775

The value that conspicuously contributes to the interaction effect is .0568. This mean is the result of averaging 16 $\bar{\beta}$ values generated when $\beta = .5$, $5 < x < 15$, $n = 20$, $n = 40$ and $\sigma = 2, 3, \ldots, 9$. When σ took on the values 7, 8, and 9, the distribution of the error term at the lower end of the region $5 < x < 15$ was a truncated normal distribution. The truncation was in the neighborhood of 1 standard deviation away from the mean. To illustrate, consider the error distribution when $x = 5.0625$ and $\sigma = 9$. Since $Y = 4X^{.5} = 9.00$, it is clear that if negative values were allowed, the distribution of $Y = 4X^{.5} + \epsilon$ would be normal with mean 9 and variance 81. However, the logarithm transformation demands positive Y values, so the generation process would not allow any error values smaller than -9. Thus, the error density at $\sigma = 9$ is

$$f(\epsilon) = \left\{ [1-F(-9)]3\sqrt{2\pi} \right\}^{-1} \exp\left\{ -\epsilon^2/18 \right\}.$$

Such an error distribution will tend to increase the level of the Y values in the lower part of the region of the independent value. The problem is not as severe at the upper end of the region since the deterministic portion of Y is larger. Consequently, the estimated slope in the linear equation is lower than it would be without the truncation problem.

The truncation difficulties can be partially removed by limiting the error standard deviation to values less than 6. Under this subset the mean values become:

Data range	$\beta = .5$	$\beta = 1$	$\beta = 1.5$
$5 < x < 15$	.1501	.1332	.0584
$5 < x < 85$	.2610	.1203	.0296

Even with this restriction of the data, the interaction effect is still present. Thus, the general inferences remain valid, namely that analysis of the main effects due to "return to scale" must be tempered by the effects of the region of the independent variable.

The overall means for $\beta = .5$, $\beta = 1$, $\beta = 1.5$ are respectively .1829, .1820, and .0984. The followup comparison, again using Tukey's T method, indicates the biases for $\beta = .5$ and $\beta = 1$ are not distinguishable, but both are different from the bias for $\beta = 1.5$ at the .01 level. The decrease in error for larger "return to scale" values, particularly when the X region is larger, results from two factors. First, as the error variance decreases, the model becomes deterministic and the logarithmic transformation becomes more appropriate. Second, when the deterministic portion of Y dominates the standard error, there is a negative bias (as pointed out earlier). This bias should tend to counteract the positive bias reflected in the earlier stages of this analysis.

The means for the error variances 4, 9, . . . , 81 are, respectively, .0640, .1438, .0977, .1971, .2399, .2080, .1617, and .1232. An orthogonal breakdown for the "error variance" sum of squares appears below:

linear	.05500
quadratic	.15779
cubic	.01507
remainder	.06650
total	.29436

This indication of a quadratic response for "error variance" reflects the truncation difficulty mentioned earlier. As the error variance increases, the misspecification of the error term becomes increasingly serious. When the error standard deviation effectively overshadows the deterministic portion of Y that corresponds to small values of X, the general level of the Y observation increases and pushes the estimates of β in the opposite direction. Hence, the bias goes down.

The T method followup indicates that adjacent values are not distinguishable at the .1 level.

Functional Form II $Y = \beta_0 X_1^{\beta_1} X_2^{\beta_2}$

Analyzing the empirical distributions in table 4 for functional form II presents a new difficulty. Functional form II, $Y_i = \beta_0 X_{1i}^{\beta_1} X_{2i}^{\beta_2} + V_i$, has two "return to scale" parameters, and as a consequence, the data are in two dimensional vector form. It cannot be assumed that the estimates of β_1 and β_2 will be independent; therefore, it will be necessary to use multivariate analysis of variance (MANOVA) techniques to analyze the consequence of misspecifying the error term.[3]

The results, using a four-way multivariate factorial analysis of variance model with two responses, are presented in table 7.

From the MANOVA table the following observations can be made:

- "Return to scale" main effect significant at the .01 level
- "Range of the X variables" main effect significant at the .01 level
- "Error variance" main effect at the .10 level
- "Return to scale-range of the X variable" interaction significant at the .01 level
- "Return to scale-error variance" interaction significant at the .10 level
- "Range of the X variable-error variance" interaction significant at the .10 level

The presence of the interactions, of course, complicates a straightforward assessment of the main effects. It does seem reasonable, however, that the "return to scale"

[3] There are several possible tests for analyzing data under MANOVA assumptions. The likelihood ratio tests are used in this study. For a complete treatment see (*1*, chapter 8).

Table 7.—MANOVA for a four-factor factorial: functional form II

Factor	U-statistic	Approximate F-statistic	Degrees of freedom	
1	0.250339	[a]5.7921	10	58
2	.629761	[a]3.7717	4	58
3	.912625	1.3882	2	29
4	.687525	[b]1.9916	6	58
12	.226585	[a]3.1923	20	58
13	.911772	.2741	10	58
23	.906933	.7258	4	58
14	.305862	[b]1.5624	30	58
24	.539751	[b]1.7455	12	58
34	.888048	.5912	6	58
123	.499383	1.2038	20	58
124	.338052	.6959	60	58
134	.524836	.7353	30	58
234	.872698	.3405	12	58

[a]Significant at the .01 level. [b]Significant at the .10 level.

factor will affect the bias and, further, that the "region of the X variables" would similarly affect the bias in the estimates of return to scale. It is not clear, however, that changing the error variance will affect the bias as long as the ratio of the deterministic part of Y and the error standard deviation remains at a reasonable level.

The error standard deviation may be so large that a truncated error distribution is effectively at work on the lower end of the deterministic curve. If so, certainly the "error variance" will affect the bias of the estimates of return to scale. The mean vectors for error variances 9, 25, 49, and 81, respectively, are [.0788, .0277], [.1068, .0397], [.1437, .0673], and [.1507, .0745]. Realizing the effect of the truncated error distribution and consequently allowing for its effect, we can note that the means for 49 and 81 error variances should be somewhat greater than they are. Hence, discounting the truncation factor, the "error variance" main effect will be more significant, and increasing the error variance does increase the bias resulting from error misspecification.

Table 8 gives the mean vectors associated with the "return to scale" and "region of the X variables" classification. To see the interaction effect in table 8, note the behavior of the vectors in the two "wide range" rows as the "return to scale" increases. Contrast this behavior with the behavior of the vectors in the "narrow range" row as the "return to scale" increases. This interaction effect can be partially explained by the truncation problem and partially by the domination of the standard deviation of the error by the deterministic part of Y for the higher "return to scale" values. Hence, this interaction effect is at least somewhat inherent in the misspecification of the error.

The cell means for the "range of the X variable" levels 5x15 x 5x15, 5x15 x 5x85, and 5x85 x 5x85, are respectively, [.1062, .0931], [.1202, .0436] and [.1337, .0315]. The vector [.1062, .0931] is distinguishable from the other two vectors. And although some interaction effects are immersed in these means, it seems that areas of limited size for the "range of the X variables" will increase the bias in estimating the return to scale when the error is misspecified. Note that to speak of the size of the region is somewhat misleading, since the principal determinant of the effect is the level of the deterministic portion of the curve.

The cell means for the subsets of "return to scale" levels .5, 1.0, and 1.5 are, respectively, [.1365, .1175], [.1467, .0668], [.1388, .1009], [.1701, .0233], and [.0527, .0529], [.0751, -.0475]. An examination of

Table 8.—Mean vectors associated with return to scale and region interaction

Region of independent variables	Returns to scale					
	$\beta_1 = .25$ $\beta_2 = .25$	$\beta_1 = .4$ $\beta_2 = .1$	$\beta_1 = .5$ $\beta_2 = .5$	$\beta_1 = .9$ $\beta_2 = .1$	$\beta_1 = .75$ $\beta_2 = .75$	$\beta_1 = 1.4$ $\beta_2 = .1$
5x15 x 5x15	.0995	-.0328	.1369	.2164	.1040	.1131
	.0738	.1728	.1293	.0935	.1187	†.0297
5x15 x 5x85	.1370	.2322	.1671	.1530	.0216	.0910
	.1095	.0024	.0753	-.0564	.0371	-.0465
5x85 x 5x85	.1731	.2408	.1125	.1409	.0325	.0213
	.1692	.0252	.0983	.0327	.0027	-.0664

these means reveals the overall decrease in bias resulting for higher levels of "return to scale." Thus, seriousness of the bias created by error misspecification for an increasing return to scale is less than for a constant or a decreasing return to scale.

CONCLUSIONS

The factors considered in the foregoing analysis were "return to scale," "region of the independent variables," "sample size," and "error variance." The sample sizes examined gave no indication of having distinguishable effects on estimate bias. The "error variance" does affect the bias and, generally speaking, it causes an increase in the bias as the error variance increases. The "region of the independent variable" affects the bias in the estimates in a couple of ways. First, if the region is small and the level of the deterministic portion of Y does not totally dominate the error standard deviation, then the bias is positive and can be substantial. Second, if the region is large and, consequently, the deterministic portion of Y dominates the error standard deviation for a good portion of the region, the bias effect will decrease. In cases where the deterministic portion dominates almost entirely, the bias becomes quite small and it can be negative. The "return to scale" affects the bias in a manner similar to the "region of the independent variable" in that return to scale affects the deterministic portion of Y. Consequently, as the "return to scale" level increases, it causes the deterministic portion to dominate the error standard deviation, and the bias is not serious. As the "return to scale" level decreases, the bias can be substantial.

These conclusions are by no means unexpected, but one is impressed by the robustness of the least squares technique under error misspecification of this type. A cursory review of tables 2 and 4 indicates that, except under the most difficult conditions with regard to variance, the bias generally is not significant enough to affect the empirical application of the corresponding elasticities. The evidence suggests bias would be present but not of the magnitude, in most cases, to greatly challenge previous economic evaluations based on regressions that use Cobb-Douglas type functions. While this is a source of consolation for many analysts, it should be noted that any conclusion that loglinear estimation yields satisfactory estimates in the presence of additive errors must still depend on the peculiarities of the sampling experiment (*9*, p. 445).

REFERENCES

(1) Anderson, T. W. *An Introduction to Multivariate Statistical Analysis.* John Wiley & Sons, Inc., 1958.

(2) Edwards, Clark. "Non-Linear Programming and Non-Linear Regression Procedures." *Jour. Farm. Econ.*, XLIV, No. 1, February 1962.

(3) Goldberger, A. S. *Econometric Theory*. John Wiley & Sons, Inc., 1963.

(4) Goldberger, A. S. "The Interpretation and Estimation of Cobb-Douglas Functions." *Econometrica*, XXXV, No. 3-4, July-October, 1968.

(5) Goldfeld, Stephen M. and Richard E. Quandt. "The Estimation of Cobb-Douglas Type Functions With Multiplicative and Additive Errors." *International Econ. Rev.*, XI, No. 2, 1970.

(6) Heien, Dale S. "A Note on Log-linear Regression." *Jour. Amer. Statis. Assoc.* Vol. 63, No. 323, September 1968.

(7) Morrison, Donald F. *Multivariate Statistical Methods*. McGraw Hill Book Co., 1967.

(8) Ostle, Bernard. *Statistics in Research*. Iowa State Univ. Press, 1963.

(9) Quandt, Richard E. "The Estimation of Constant Elasticities: Comment." *Southern Econ. Jour.* Vol. XXXIX, No. 3, January 1973.

(10) Scheffe, Henry. *The Analysis of Variance*. John Wiley & Sons, Inc., 1959.

(11) Theil, H. *Principles of Econometrics*. John Wiley & Sons, Inc., 1971.

(12) Theil, H. "Specification Errors and the Estimation of Economic Relationships." *Rev. International Statis. Inst.*, XXV, 1957.

(13) Tocher, K. D. "The Application of Automatic Computers to Sampling Experiments." *Jour. Royal Statis. Soc.*, Series B, Vol. 16, 1954, pp.39-61.

(14) Weber, J. E. and Hawkins, C. A. "The Estimation of Constant Elasticities." *Southern Econ. Jour.*, October 1971, pp. 185-192.