THE FINITE-POPULATION LINEAR REGRESSION ESTIMATOR AND ESTIMATORS OF ITS VARIANCE-AN EMPIRICAL STUDY

Richard M. Royall, Johns Hopkins University
William G. Cumberland, U.C.L.A.

1. INTRODUCTION

A previous paper [1] reported on an empirical comparison of two general finite population inference theories as they apply to the ratio estimator. That study showed how prediction theory, by conditioning on the chosen sample, can reveal relationships which are important for inference, but which are concealed in the analyses of random sampling theory. It confirmed prediction theory's warnings about biases in both the weighted least-squares and the conventional variance estimators. Performance of these two statistics was clearly inferior to that of two bias-robust variance estimators generated by prediction models.

The theory for these robust variance estimators has been extended to general linear regression models with independent errors [2]. Application of this extended theory gives bias-robust variance estimators for the standard linear regression estimator. Here we compare these variance estimators with the conventional and least-squares alternatives. Section 2 contains the theory, the populations are described in Section 3 and the empirical results appear in Section 4.

2. TWO THEORIES FOR THE LINEAR REGRESSION ESTIMATOR

With each of the N population units we have two numbers. One is the variable of interest, y, and the other is an auxilliary variable, x, whose value is known. The values associated with unit i are (x_i,y_i); s is the set of n units in the sample; and $\bar{x}_s$ and $\bar{y}_s$ denote the sample means. Similarly, $\bar{x}_r$ and $\bar{y}_r$ denote the averages over the set r of non-sample units, while $\bar{x}$ and $\bar{y}$ are the population averages. The linear regression estimator for the population total, $T = N\bar{y}$, is $\hat{T}=N[\bar{y}_s+b(\bar{x}-\bar{x}_s)]$, where $b= \Sigma_s(x_i-\bar{x}_s)y_i/\Sigma_s(x_i-\bar{x}_s)^2$.

2.1 Theory Using Prediction Models. If the numbers $(y_1,y_2,\ldots,y_N)$ are realized values of random variables $(Y_1,Y_2,\ldots,Y_N)$ then after sample s has been observed, estimating $T = \Sigma_s y_i + \Sigma_r y_i$ is equivalent to predicting the value, $\Sigma_r y_i$, of the unobserved random variable $\Sigma_r Y_i$. Under the model

$$E(Y_i) = \beta_0 + \beta_1 x_i,\ \mathrm{var}(Y_i) = \sigma^2,\ \mathrm{cov}(Y_i,Y_j) = 0,\ i \neq j \qquad (2.1)$$

the least-squares predictor of $\Sigma_r y_i$ is $\Sigma_r[\bar{y}_s + b(x_i-\bar{x}_s)]$ and the sum of this predictor and the known sum $\Sigma_s y_i$ gives the regression estimator $\hat{T}$. Under this model the estimator is unbiased, $E(\hat{T}-T) = 0$, with error variance

$$\mathrm{var}(\hat{T}-T) = (N/f)(1-f)\sigma^2[1 + (\bar{x}_s-\bar{x})^2/(1-f)g(s)] \qquad (2.2)$$

where $f = n/N$ and $g(s) = \Sigma_s(x_i-\bar{x}_s)^2/n$. The least-squares variance estimator, denoted by v_L, is obtained when σ^2 in (2.2) is replaced by $\hat{\sigma}^2 = \Sigma_s d_i^2/(n-2)$, where $d_i = y_i-\bar{y}_s-b(x_i-\bar{x}_s)$. A simple variance estimator found in standard sampling textbooks is

$$v_C = (N/f)(1-f)\hat{\sigma}^2.$$

The least-squares estimator v_L is simply the product of v_C and the factor $1 + h(s)$, where $h(s) = (\bar{x}_s-\bar{x})^2/(1-f)g(s)$. Since h(s) is non-negative, v_L is never less than v_C. Note that if the sample is balanced on x ($\bar{x}_s=\bar{x}$), then h(s) equals zero, the variance (2.2) is minimized, and v_L equals v_C.

The least-squares variance estimator is unbiased under model (2.1); that is $E(v_L) = \mathrm{var}(\hat{T}-T)$. But under models in which the variance is not constant, this is no longer true. For instance, if the true model is

$$E(Y_i)=\beta_0+\beta_1 x_i,\ \mathrm{var}(Y_i)=\sigma^2 x_i,\ \mathrm{cov}(Y_i,Y_j)=0,\ i\neq j \qquad (2.3)$$

then the actual error-variance is

$$\mathrm{var}(\hat{T}-T) = (N/f)\sigma^2\{(2-f)\bar{x} - \bar{x}_s + (\bar{x}-\bar{x}_s)^2[(\Sigma_s x_i^3 - 2\bar{x}_s\Sigma_s x_i^2 + n\bar{x}_s^3)/ng(s)^2]\} \qquad (2.4)$$

while

$$E(v_L) = (N/f)(1-f)\sigma^2[1 + h(s)](\bar{x}_s + [\bar{x}_s-[(\Sigma_s x_i^3 - 2\bar{x}_s\Sigma_s x_i^2 + n\bar{x}_s^3)/ng(s)]]/(n-2)). \qquad (2.5)$$

Of course $E(v_C)$ is given by (2.5) with the term h(s) replaced by zero.

The biases in v_C and v_L when the constant variance condition in model (2.1) fails can be serious. The following rough approximations show the directions and magnitudes of these biases under model (2.3). When n is large and f is small $E(v_C)$ is approximately $(N/f)\sigma^2\bar{x}_s$, while if g(s) is replaced by $g = \Sigma_1^N(x_i - \bar{x})^2/N$, $E(v_L)$ is approximately $(N/f)\sigma^2\bar{x}_s[1 + (\bar{x}_s - \bar{x})^2/g]$. Making similar approximations in the actual variance (2.4) yields $(N/f)\sigma^2[2\bar{x} - \bar{x}_s + (\bar{x} - \bar{x}_s)^2c]$, where c is the positive constant $c = (\Sigma_1^N x_i^3 - 2\bar{x}\Sigma_1^N x_i^2 + N\bar{x}^3)/Ng$. The approximate variance has its minimum at $\bar{x}_s = \bar{x} + (1/2c)$, while the approximations for both $E(v_C)$ and $E(v_L)$ are increasing functions of $\bar{x}_s$ in an interval about $\bar{x}$. Since all

three, $\text{var}(\hat{T}-T)$, $E(v_C)$, and $E(v_L)$ are approximately the same when $\bar{x}_s = \bar{x}$, we see that when the sample is one where $\bar{x}_s < \bar{x}$, both v_C and v_L can have serious negative biases. Although we have explicitly considered only the case of variance proportional to x here, similar results can easily be established when the variance is a more general increasing function of x.

Royall and Cumberland [2] studied some bias-robust variance estimators for linear regression models. These are approximately unbiased estimators of the true error-variance even when the variances, $\text{var}(Y_i)$, are not those specified by the model. For the linear regression estimator $\hat{T}$, one of these variance estimators is

$$v_D = [(1-f)/f]^2[n/(n-1)]\Sigma_s d_i^2\{[1+(x_i-\bar{x}_s)(\bar{x}_r-\bar{x}_s)/g(s)]^2+(1-f)/f\}/\{1-[(x_i-\bar{x}_s)^2/(n-1)g(s)]\}.$$

Another variance estimator, which is asymptotically equivalent to v_D [2], is the jackknife statistic

$$v_J = (1-f)(n-1)\Sigma_s(\hat{T}_{(j)}-\hat{T}_{(\cdot)})^2/n ,$$

where $\hat{T}_{(j)}$ is the regression estimate based on the sample of size n-1 obtained by deleting unit j from the sample, and $\hat{T}_{(\cdot)}$ is the average of these n estimates, $\Sigma_s\hat{T}_{(j)}/n$.

Under model (2.1), $E(v_D) = \text{var}(\hat{T}-T)$. When the variances are non-constant, as in (2.3), this relation continues to hold as an approximation whose accuracy improves as n and N/n increase.

Although failure of the constant-variance condition in model (2.1) can introduce a serious bias in the variance estimator v_L, it does not bias $\hat{T}$. On the other hand failure of the linear regression condition can introduce a bias in this statistic [3]. When the true regression function is a polynomial of degree $k > 1$, the linear regression estimator is unbiased only if the sample is balanced on k moments of x: $\Sigma_s x_i^j/n = \Sigma_1^N x_i^j/N$ for $j = 1,2,\ldots,k$. For example, if in fact $E(Y_i) = \beta_0+\beta_1 x_i + \beta_2 x_i^2$, then the bias vanishes in samples which are balanced on both x and x^2.

Failure of the working model's condition that $E(Y_i) = \beta_0+\beta_1 x_i$ increases the expected value of each of the variance estimators as well as the mean-square error. If the sample is balanced on x and sufficiently well-balanced on other variables (or other powers of x) that the regression estimator is essentially unbiased, then the variance estimators v_C, v_L and v_D are all conservative in that their expected values exceed the actual mean-square error.

2.2 Theory Using the Random Sampling Distribution. A fundamental tenet of standard probability sampling theory is the Randomization Principle, which asserts that inferences should be based, not on prediction models, but on the probability distribution created by the sampler when he uses a random device to choose which units will be observed. The use of a simple random sampling plan creates a distribution under which the regression estimator is biased. But in many populations this bias is negligible in large samples, and it is generally ignored. The variance is approximately $V = (N/f)(1-f)\Sigma_1^N[y_i - \bar{y} - B(x_i - \bar{x})]^2/(N-2)$ where $B = \Sigma_1^N(x_i-\bar{x})y_i/\Sigma_1^N(x_i - \bar{x})^2$. The variance estimator v_C is also biased, as an estimator of the sampling variance of $\hat{T}$, but its bias too is ignored in larges samples. It is sometimes suggested that v_C be adjusted for bias in small samples, and the adjusted estimator is very nearly v_L. Analysis under the random sampling distribution entails averaging over all possible samples s. Thus while the prediction approach studies properties of estimators for specific samples, the randomization theory looks at properties averaged over all possible samples. That is why the negative bias of v_C in samples where $\bar{x}_s < \bar{x}$, which is revealed by prediction theory, has been concealed by randomization analysis. This bias is displayed clearly in the empirical results which follow.

3. THE STUDY POPULATIONS

The preceding theoretical results were studied empirically in the six real populations described in Table 1. These are the same populations which we used in a previous study of the ratio estimator and these are described more fully in [1]. We do not claim that these populations are ones where the regression estimator and a simple random sampling plan would be anyone's strategy of choice. They are populations where a straight line through the origin regression model, with variance proportional to x, would be a reasonable first approximation. By studying them we can see how the variance estimators perform under conditions different from those described by the constant-variance model. This gives us a chance to study the robustness of the estimator appropriate under the constant-variance model, v_L, the new estimator v_D, and the estimator v_C whose "validity" is affirmed by randomization theory without reference to prediction models.

4. EMPIRICAL RESULTS

From each of the six populations we selected two sets of samples for study. These were chosen using (i) simple random sampling, and (ii) two purposive (non-random) samples, of which prediction theory identifies one as particularly bad for the linear regression estimator, and one as potentially useful.

TABLE 1: STUDY POPULATIONS

Symbol	Description	x	y	T	$\rho(X,Y)$
Cancer	301 Counties in N.C., S.C. and Ga.	adult female population 1960	breast cancer mortality 1950-69	11994	0.967
Cities	125 U.S. cities	population 1960	population 1970	35.691×10^6	0.947
Counties 60	304 Counties in N.C., S.C. and Ga.	households 1960	population 1960	10.007×10^6	0.998
Counties 70	304 Counties in N.C., S.C. and Ga.	households 1960	population 1970	11.243×10^6	0.982
Hospitals	National sample of 393 hospitals	number of beds	number of patients discharged	320159	0.910
Sales	331 U.S. Corporations	gross sales 1974	gross sales 1975	796.986×10^9	0.997

TABLE 2: RESULTS FOR 1000 SIMPLE RANDOM SAMPLES OF n=32 (LINEAR REGRESSION ESTIMATOR)

		(Average in 1000 samples)$^{\frac{1}{2}}$				
Population	Average Error	$(\hat{T}_s - T)^2$	v_C	v_D	v_J	v_L
Cancer	-41	696	600	656	757	612
Cities (millions)	-0.02	1.35	1.27	1.33	1.40	1.30
Counties 60 (thousands)	27	153	129	150	177	132
Counties 70 (thousands)	-6	487	417	453	554	426
Hospitals	1.9	16.5	15.8	16.6	17.5	16.0
Sales (billions)	-3.1	21.6	16.4	19.1	24.0	17.1

4.1 Simple Random Sampling. From each of the six study populations we drew 1000 simple random samples of n = 32. These are in fact the same 1000 random samples used in the previous study [1]. For each sample we calculated the regression estimate $\hat{T}$, and the actual error, $\hat{T}$-T, as well as the four variance estimates v_C, v_D, v_J and v_L. The average values are shown in Table 2. Note that v_C and v_L underestimate the mean-square error in all six populations. The estimate v_D performs slightly better while still underestimating the actual mean-square error. The jackknife statistic is the only one showing a tendency to overestimate.

The prediction theory sketched in Section 2.1 suggests that performance of the variance estimates will depend strongly on $\bar{x}_s$. To examine performance as a function of $\bar{x}_s$, we arranged the 1000 samples from each population in order of increasing values of $\bar{x}_s$. We then grouped the samples in 20 sets of 50, so that the first group contains the 50 samples whose values of $\bar{x}_s$ are the smallest, the next group contains the samples with the next 50 smallest $\bar{x}_s$ values, etc. For each of these 20 groups we calculated the average value of $\bar{x}_s$, $\Sigma_1^{50}\bar{x}_s/50$, and the average error $\Sigma_1^{50}(\hat{T}-T)/50$, as well as the mean-square error (mse) $\Sigma_1^{50}(\hat{T}-T)^2/50$, and the averages of each of the four variance estimates, $\bar{v}_C = \Sigma_1^{50} v_C/50$, etc. We then plotted the average errors, the values of $(\text{mse})^{\frac{1}{2}}$, $(\bar{v}_C)^{\frac{1}{2}}$, $(\bar{v}_D)^{\frac{1}{2}}$, etc., against the average values of $\bar{x}_s$.

Figures A-F show the results. Each plotted point is obtained by averaging 50 samples, and each figure summarizes the results for the 1000 random samples of n = 32 from one population.

For each population there are six trajectories. The trajectory showing average error plotted against average value of $\bar{x}_s$ is labelled error. The one showing the root mean-square error is labelled $\sqrt{\text{mse}}$, and those showing $(\bar{v}_C)^{\frac{1}{2}}$, $(\bar{v}_D)^{\frac{1}{2}}$, etc. are labelled C, D, etc. The population mean $\bar{x}$ is shown on the abscissa.

The degree to which these empirical results agree with those of prediction theory is remarkable. The variance estimator v_C increases rapidly with increasing $\bar{x}_s$, while the actual mse

decreases until $\bar{x}_s$ exceeds $\bar{x}$, then begins to increase. In all of these populations $\bar{v}_C$ tends to be smallest when the actual squared error is largest. It is clear that $v_C^{\frac{1}{2}}$ represents a gross under-estimate of the standard error in the 10-20% of randomly-selected samples whose $\bar{x}_s$ values are smallest. The least-squares variance estimate v_L is not much better. By contrast, the bias-robust estimates v_D and v_J show remarkable tenacity in "tracking" the actual mse as $\bar{x}_s$ varies.

4.2 Non-Random Sampling. From each of the six study populations we drew two non-random samples. One of these came from the lower extreme of the x-distribution and one was obtained by fitting the sample's x-distribution to the population's as closely as possible.

The extreme sample consisted of the thirty-two units whose x-values are smallest (the low sample). This sample is badly balanced on x and x^2 and is, according to prediction theory, one of the worst possible samples (large variance, no bias protection) for $\hat{T}$ under most reasonable models for these populations.

The other purposive sample consists of those thirty two units for which the sample cumulative distribution function (c.d.f.) of x best approximates the population c.d.f. That is, if $F_s(x_o)$ is the proportion of units in sample s whose x-values are no greater than x_o, and $F(x_o)$ is the corresponding proportion in the whole population, the sample for which $\max|F_s(x) - F(x)|$ is minimized is chosen. This we call the best-fit sample. It is "like the population" with respect to the distribution of the known variate x. It does not necessarily provide the best possible balance on x or on x^2, but in many populations it does provide a sample which is reasonably well-balanced on these and other important parameters of the size variate's distribution. Results for the two samples, low and best-fit, are shown in Table 3.

While the entries in Table 2 represent averages over 1000 samples and the points plotted in Figures A-F represent averages over 50 samples, the entries in Table 3 are subject to the variability of individual samples. Although they must be interpreted cautiously, these results have some interesting features.

As expected from the analysis in Section 2.1, the variance estimator v_C gives a gross underestimate in the low samples. In fact, the standardized errors $|\hat{T}-T|/v_C^{\frac{1}{2}}$ range from the minimum of 8.91 in Counties 70 to a maximum of 126.58 in the Sales population. In every population the value of $v_C^{\frac{1}{2}}$ in the low sample is much smaller than the root mean square of the estimate's values in the thousand simple random samples.

The errors produced in the low samples were, as predicted, enormous, compared with the errors in better-balanced samples. In these worst of all possible samples the variance estimators v_D, v_J, and v_L provide surprisingly accurate measures of the uncertainty in $\hat{T}$.

In every population the best-fit sample produced an estimation error smaller than the root mean square error under simple random sampling. These best-fit samples are sufficiently well-balanced on x that the four variance estimates are quite comparable, and all four give standard error estimates, $v^{\frac{1}{2}}$, whose magnitudes appear appropriate. The largest of the standardized errors in these samples was 1.08.

5. DISCUSSION

Simple random sampling gives all samples an equal chance of being chosen. It does not make them equally informative. Yet conventional theory associates with $\hat{T}$ a bias and variance which are defined by averaging over all samples, and which say nothing about the estimate's reliability in in any one particular sample. On the other hand, prediction theory identifies important differences between samples and shows how these differences should influence our inferences. Although average performance over all possible samples might be of interest before we select a sample at random, after sampling we must make inferences from the unique sample we have observed. If we have the bad fortune to draw a sample with $\bar{x}_s < \bar{x}$, our inferences must be more cautious than if the sample is well-balanced. We believe these empirical results illustrate once again the fallacy of the Randomization Principle.

ACKNOWLEDGEMENTS

This work was supported by grant MCS78-02471 from the National Science Foundation.

REFERENCES

[1] Royall, R. M., and Cumberland, W. G. (1978a), "An Empirical Study of Prediction Theory in Finite Population Sampling: Simple Random Sampling and the Ratio Estimator," Ch. 18 in Survey Sampling and Measurement, N.K. Namboodiri, ed. New York: Academic Press.

[2] _____, and Cumberland, W. G. (1978b), "Variance Estimation in Finite Sampling," Journal of the American Statistical Association, 73, 351-8.

[3] _____, and Herson, J. H. (1973), "Robust Estimation in Finite Populations I," Journal of the American Statistical Association, 68, 880-9.

TABLE 3: RESULTS FOR TWO SPECIAL SAMPLES OF n = 32

Population	Sample	$\bar{x}_s$	Error	$(v_C)^{\frac{1}{2}}$	$(v_D)^{\frac{1}{2}}$	$(v_J)^{\frac{1}{2}}$	$(v_L)^{\frac{1}{2}}$
Cancer	low	1314	1334	116	3006	2927	2961
	best-fit	11078	501	462	524	619	462
Cities	low	0.110	-11.20	0.42	13.38	11.88	12.70
(millions)	best-fit	0.268	- 0.38	1.00	1.09	1.25	1.00
Counties 60	low	1.252	502	13	280	274	342
(thousands)	best-fit	8.808	116	128	130	134	128
Counties 70	low	1.252	401	45	673	663	1146
(thousands)	best-fit	8.808	40	261	262	271	261
Hospitals	low	0.0245	-75.0	2.2	56.1	55.3	84.8
(thousands)	best-fit	0.2741	10.5	12.9	13.0	13.3	12.9
Sales	low	0.539	-746.8	5.9	515.6	502.9	491.1
(billions)	best-fit	2.188	1.7	15.4	16.7	22.1	15.4

FIGURES A-F: RESULTS FROM 1000 SIMPLE RANDOM SAMPLES
OF 32 FROM EACH OF SIX POPULATIONS
LINEAR REGRESSION ESTIMATOR

FIGURE A: CANCER

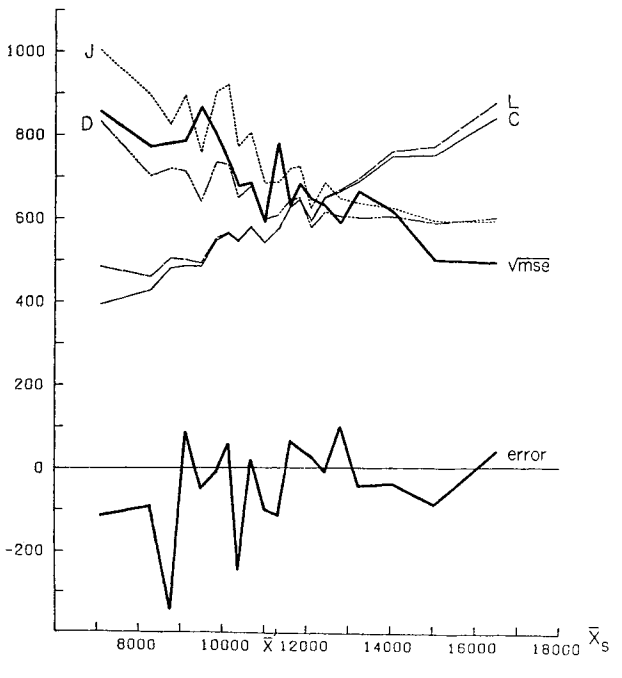

FIGURE B: CITIES

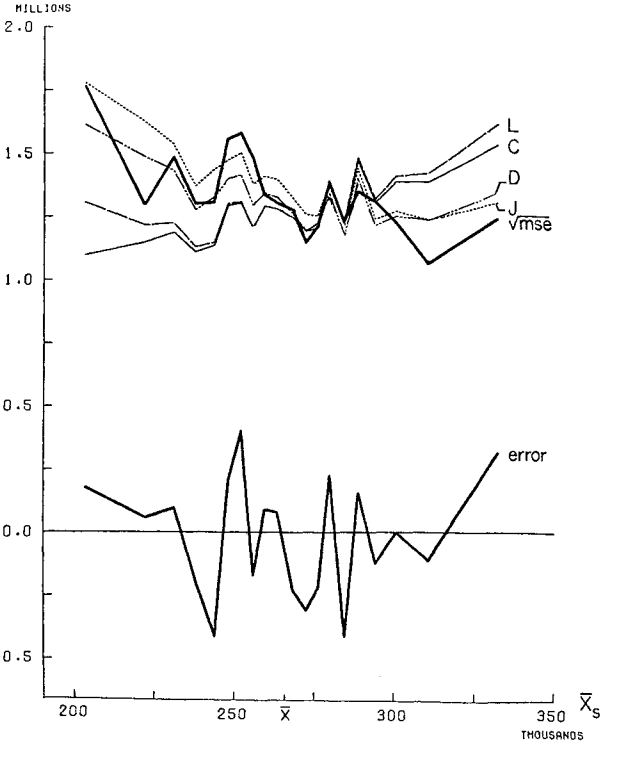

FIGURE C: COUNTIES 60

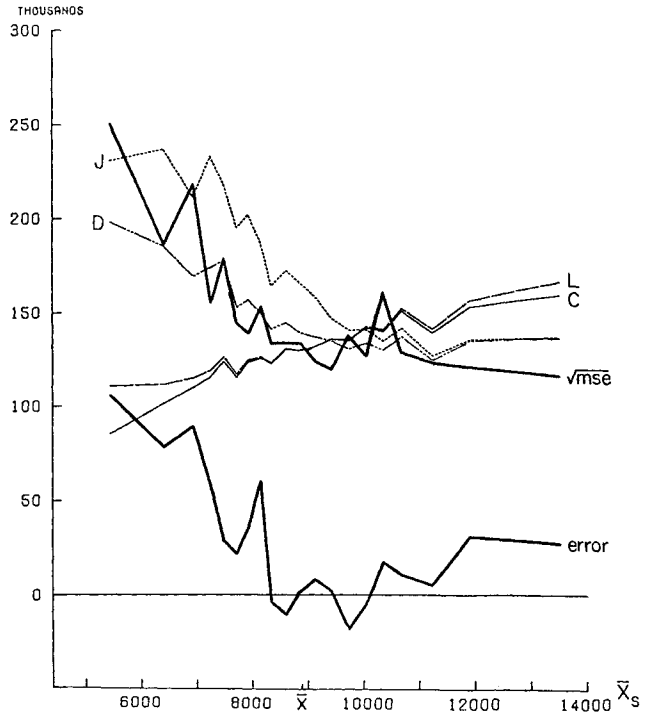

FIGURE D: COUNTIES 70

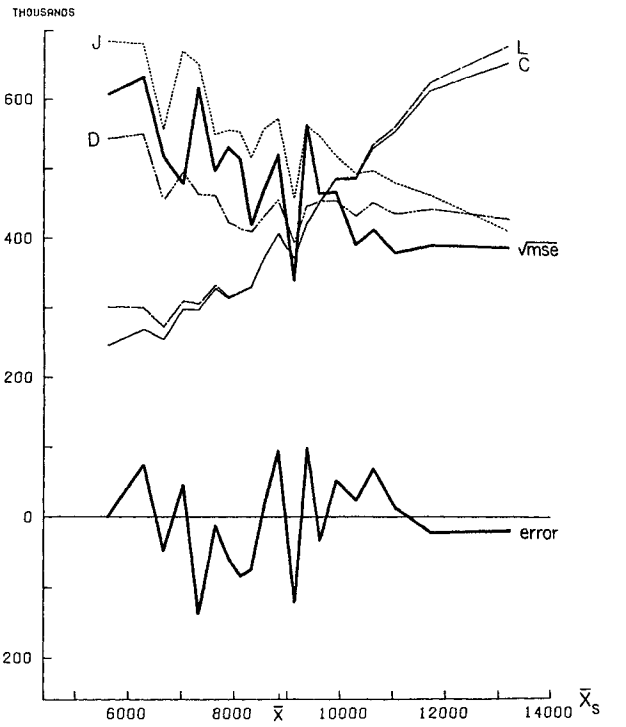

FIGURE E: HOSPITALS

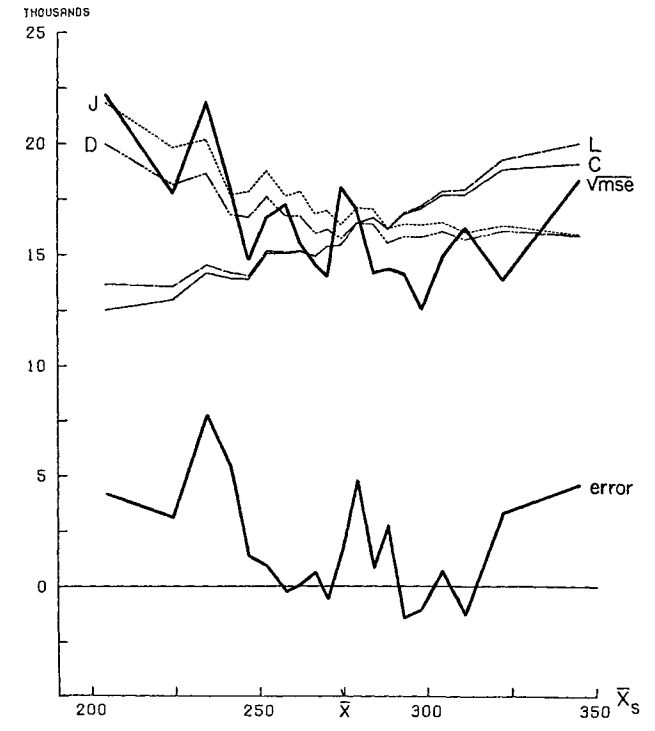

FIGURE F: SALES