Eastern Economic Journal, Volume XVII, No. 1, January–March 1991

On Using a "Patched" Data Base: An Illustration of the Conundrum

Samuel Schwarz*

INTRODUCTION

It is well-known that "(t)ime series data sometimes contain missing observations. . . . Observations can be missed for a variety of reasons, such as clerical error, malfunction of recording equipment, deletion of apparently 'bad' observations, and the inability to observe the process at certain times, for example at night-time or on weekends." (Robinson, 1984) Consequently, empirical economists sometimes make use of data bases whose gaps and omissions have been filled, based upon those data points that do exist. For example, in a recent study on the growth of nonprofit arts organizations in the U.S. during the decade of the 1970's, Schwarz and Peters (1983), attempted to measure the growth of organizations in the different arts disciplines and their subgroups, stratified by budget size. One of these subgroups was the class of orchestras labeled "Metropolitan" by the American Symphony Orchestra League (ASOL), the requirement being a budget of $100,000–$499,999 in FY79. ASOL collects data from its member orchestras annually. However, unlike the larger major orchestras which had a nearly perfect response rate over the decade, the smaller metropolitans had a greatly varying response rate. Hence, when we attempted to create a uniform data base, of necessity, the gaps had to be patched. Our experience with the patched data base provides a novel approach for detecting self-created bias and offers a clear lesson about the pitfalls and dangers of using such a data base.

CREATING A DATA BASE

Table 1 lists the number of metropolitan orchestras that reported in each year of the decade. The number of orchestras reporting not only varied from year to year, but also more than doubled over the decade. In all, 83 metropolitans reported for at least one year.

Faced with so many gaps in the data, we followed two tracks: one using a subsample of 12 metropolitans which reported every year during the decade and the other by blowing up the samples of each year to the full complement of 83 orchestras, implicitly assuming that the average of those reporting during any given year was representative of all the 83 orchestras. Table 2 presents the growth rates of five key variables—earned income, support income, total income, expenditures, and earnings gap (the difference between earned income and expenditures)—over the decade for these two groups.

It is clear that the "12" had a much larger growth rate than the "83." Moreover, the growth of the former was in line with the growth of the rest of the orchestra universe.[1] What caused the apparently smaller growth of the remaining 71 metropolitans? Or was it, perhaps, a product of our own data

*The College of Staten Island, CUNY.

The research for this paper was done as part of a larger study, Schwarz and Peters (1983), performed under a contract from the National Endowment for the Arts, NEA PC 80-29. I am especially grateful to Harold Horowitz, NEA director of research, for encouraging and appreciating sound research. The sound advice of my able research associate, Mary G. Peters, is gratefully appreciated. Thanks are also due to the American Symphony Orchestra League (ASOL) for granting us use of the most complete and comprehensive time-series data available in the performing arts. The very helpful comments of the editor of this journal on enhancing the presentation and the comments of two anonymous referees are gratefully appreciated.

TABLE 1
Number of Metropolitan Orchestras Reporting To the American Symphony Orchestra League

FY70	FY71	FY72	FY73	FY74	FY75	FY76	FY77	FY78	FY79
26	36	39	32	43	50	48	59	60	63

Source of Data: American Symphony Orchestra League

creation? A deeper probe into the data will reveal an interesting phenomenon that will not only clarify this puzzle but will also yield a lesson for future research.

DISSECTING THE PATCHED DATA BASE

The data base for the "83" was created by blowing up the figures from an ever-growing subsample, its size going from 26 to 63 orchestras. Were the "26" as representative of the "83" as were the "63?"

Table 3 presents the fraction (percentage) of orchestras reporting that have *total expenditures less than the average (mean)* for all orchestras reporting that particular year. The numerals along the diagonal are for the total number of orchestras reporting that year. The off-diagonal numerals along each row (horizontal) are for the orchestras reporting in each year listed on the top that have *also* reported in the year listed on the left for that given row. For example, proceeding along the second row (FY71), the first numeral (21/36) indicates that 36 orchestras reported that year and 21 (or 58 percent) of them had expenditures less than the average; the second figure (16/34) indicates that in FY72 only 34 of the original 36 reported, and 16 (or 47 percent) of the 34 had expenditures less than the average figure for *all 39* orchestras reporting in FY72, as is indicated by the bottom figure in that column; and

TABLE 2
Growth Rates of Five Key Variables

	12 Metropolitans	83 Metropolitans
Earned Income	12.1(.27)[1]	8.5(.35)
	5.0(.26)[2]	1.6(.30)
	4.8(.27)[3]	1.4(.30)
Support Income	11.2(.22)	7.6(.21)
	4.1(.23)	.7(.24)
	3.9(.24)	.5(.27)
Total Income	11.5(.18)	8.0(.20)
	4.4(.19)	1.1(.19)
	4.2(.21)	.9(.21)
Expenditures	11.2(.18)	7.6(.20)
	4.1(.12)	.7(.15)
	3.9(.13)	.5(.17)
Earnings Gap	10.7(.28)	6.9(.23)
	3.6(.21)	.1(.20)
	3.4(.21)	−.1(.22)

[1]Growth rate of actual dollars and corresponding standard deviation in parantheses.
[2]Growth rate as deflated by the Implicit GNP Deflator with standard deviation.
[3]Growth rate as deflated by the Consumer Price Index with standard deviation.

Source: Figure 5-15, Schwarz and Peters (1983), p. 5–32.

TABLE 3
Percentage of Reporting Metropolitan Orchestras with Total Expenditures Below Average

	Reporting Years									
	FY70	FY71	FY72	FY73	FY74	FY75	FY76	FY77	FY78	FY79
Numbers (Percentage) of Organizations That Began Reporting In:										
FY70	15/26 (58%)	13/26 (50%)	9/24 (38%)	5/19 (26%)	8/20 (40%)	11/25 (16%)	9/23 (39%)	11/26 (42%)	5/22 (23%)	7/23 (30%)
FY71		21/36 (58%)	16/34 (47%)	9/25 (36%)	13/27 (48%)	17/33 (52%)	14/30 (47%)	16/34 (47%)	8/29 (28%)	10/31 (32%)
FY72			21/39 (54%)	12/29 (41%)	16/32 (50%)	19/36 (53%)	14/33 (42%)	17/37 (46%)	8/31 (26%)	10/32 (31%)
FY73				15/32 (47%)	14/29 (48%)	14/31 (45%)	11/28 (39%)	12/30 (40%)	6/28 (21%)	6/27 (22%)
FY74					27/43 (63%)	24/39 (61%)	18/36 (50%)	21/40 (53%)	8/33 (24%)	10/34 (29%)
FY75						31/50 (62%)	21/41 (51%)	24/45 (53%)	10/37 (27%)	13/38 (34%)
FY76							26/48 (54%)	24/47 (51%)	12/39 (31%)	10/36 (28%)
FY77								33/59 (56%)	20/50 (40%)	20/48 (42%)
FY78									30/60 (50%)	25/50 (50%)
FY79										34/63 (54%)

Source of Data: American Symphony Orchestra League.

TABLE 4
Percentage of 12 Metropolitan Orchestras With Total Expenditures Below Average

FY70	FY71	FY72	FY73	FY74	FY75	FY76	FY77	FY78	FY79
7/12 (58%)	7/12 (58%)	5/12 (42%)	5/12 (42%)	5/12 (42%)	5/12 (42%)	4/12 (33%)	3/12 (25%)	1/12 (8%)	2/12 (17%)

Source of Data: American Symphony Orchestra League

the third figure (9/25) indicates that in FY73, only 25 of the original 36 reported, while 9 (or 36 percent) of the 25 had expenditures less than the average for *all 32* reporting in FY73.

A look at Table 3 reveals that the orchestras reporting in the early years of the decade were larger-expenditure orchestras on the average than those reporting in the late years of the decade. This pattern is most clearly seen in the columns for FY78 and FY79. Of all the orchestras reporting in each of these two years, 60 in FY78 and 63 in FY79, 50 and 54 percent, respectfully, had expenditures less than the average. Of those that had already reported in FY75 or before, only 27 percent (10/37) and 34 percent (13/38), respectively, had expenditures less than the average amount for all the organizations reporting in FY78 or FY79, respectively.

In this case we have information on the underlying cause for this pattern.[2] Although, because of inflation, $100,000 in FY70 was equivalent to $181,000 in FY79, the minimum requirement for entry into the class of metropolitans remained fixed at $100,000 over the entire decade. Thus, many smaller orchestras entered the ranks of "Metropolitan" toward the end of the decade because of inflation creep; i.e., inflation drove their budgets above $100,000. If so, it was only natural that the newly reporting orchestras were lower-budget orchestras than those reporting at the beginning of the decade. Consequently, the *average* total real expenditures for all orchestras reporting in a particular year were constantly being lowered.

Because the orchestras reporting at the beginning of the decade were higher-budget orchestras than those reporting at the end, the growth rates that have been established are underestimates of the growth rates that would have been obtained if all 83 orchestras had reported for each year of the decade. The total expenditures for the early years was blown up from an average for higher-budget orchestras; the total figure for the later years was blown up from an average for lower-budget orchestras.

If the data are taken at face value—and that is what the calculated growth rates have done, they would indicate that total expenditures did not grow much over the decade. However, if the figures for all the years, for either the group of orchestras that reported at the start of the decade or the group that reported at the end, were obtainable, they would indicate a larger growth for each of the two groups. In fact, the subsample of 12 metropolitans is a group that is continuous over the decade. Table 4 compares the 12 in the same format as Table 3 showing the fraction (percentage) of the 12 that had expenditures less than the average of all metropolitans reporting in each year. It shows that, although during the first two years, this fraction was *identical* to that of all metropolitans reporting in those years (see Table 3), it decreased over the decade, apparently as more and more smaller-budget orchestras reported data. The same fractions indicate that the 12 are a representative sample of all the metropolitans reporting in the first two years, at least in terms of expenditure levels. The latter, in turn, represents 43 percent of all the 83 metropolitans.

In conclusion, all the data presented seem to indicate that the actual growth rates of five key variables for all of the 83 metropolitans are larger than the ones calculated and are probably close to those for the subsample of 12.

CONCLUDING REMARKS: A FLASHING SIGNAL

In this paper, we provided one example of a patched data base, where missing elements in the original data matrix were filled by some method. For this particular case, our mode of analysis allowed us to detect a self-created bias and, therefore, better interpret the estimated growth rates. This type of analysis can be replicated for similar cases, where one has use of the basic data elements from which the data base was derived.

While our data matrix had many elements missing, it is not uncommon for the basic data matrix to have at least some elements missing. The preparer of the data base then has two choices: either delete those observations that have some points missing or attempt to fill in the missing elements, i.e., create a patched data base. In both cases, there exists the risk of biasing the data base.

The problem is even more complex when we are not given the basic data elements, but rather an already completed data base. For example, Baumol and Baumol (1980) used a data base which they then thought contained "figures for a fixed sample of some 160 orchestras for the six years 1973–74 through 1978–79." (p. 2) Only later did this author (Schwarz, 1981) make them aware that: "The sample size varies from year to year as the number of orchestras reporting changes. . . . The data are adjusted at ASOL to correct these omissions." (Baumol and Baumol, 1981, p. 90) Although the Baumols stated "that our description of the orchestral data was misleading, *nostrae culpa,*" (p. 89), it is clear to this author, who was also in possession of the same table which they used, that were it not for the fact that he also had the raw data from which the table was derived, it would also never have dawned on him to question the table.

The main point of this paper is, therefore, to place a "flashing signal" before every data base: slow down and handle with extreme care; ask questions about the origin of the data and imagine everything that could be wrong with the data base. Move on only after such a thorough probe.

NOTES

1. See Figure 5-15, Schwarz and Peters (1983), p. 5–32.
2. See Schwarz and Peters (1983), p. 3–26.

REFERENCES

Baumol, Hilda and Baumol, William J. 1980. "On Finances of the Performing Arts During Stagflation: Some Recent Data," *Journal of Cultural Economics,* 4, no. 2, 1–14.

_____, 1981. "A Rejoinder to Schwarz," *Journal of Cultural Economics,* 5, no. 2, 89–90.

Robinson, P.M. 1984. "Kernel Estimation and Interpolation for Time Series Containing Missing Observations," *Ann. Inst. Statist. Math.,* 36, Part A, 403–417.

Schwarz, Samuel. 1981. "The Facts First: A Reply to Baumol and Baumol," *Journal of Cultural Economics,* 5, no. 2, 85–87.

Schwarz, Samuel and Peters, Mary G. 1983. *Growth of Arts and Cultural Organizations in the Decade of the 1970's,* Rockville, Md.: Informatics General Corporation. ERIC No. ED 241 398.