Evaluating Prospective Integration of Survey and Administrative Record Data: The Impact of Uncertainty in Measurement of Data Quality and Cost Factors October 2010

Randall Powers and John L. Eltinge
Office of Survey Methods Research, U.S. Bureau of Labor Statistics

Abstract: In evaluating the prospective integration of survey data with administrative record data, one generally needs to balance several factors related to cost and data quality. In practical applications, however, these factors are not known with perfect precision, and measurement of them often requires a substantial effort by the statistical organization. This paper evaluates some ways in which the balance between data quality and cost can be affected by errors in estimation of these quality and cost factors. Principal emphasis centers on a simulation study involving sample allocation in the presence of uncertainty regarding stratum-level variances and per-unit costs.

Key Words: Dual Frame Estimator; Fixed and Marginal Costs; Measurement Error; Sample Allocation; Simulation Study; Stratum-Level Unit Variances

1. Introduction

The previous paper in this session provided a broad overview of potential approaches to the comparison and use of multiple sources of survey and administrative record data. This paper extends those ideas by exploring the impact of uncertainty in design parameters related to data quality and cost factors. We will cover these ideas through four main steps. First, we will present a general approach to the optimal allocation of resources for collection of survey and administrative data. Second, we highlight some ways in which the optimal allocation work becomes more complex when some important parameters of the cost-quality balance are estimated with error. Third, we will describe the general design of a simulation study to evaluate the ways in which the performance of optimal-allocation methods may be sensitive to errors in estimated cost or quality parameters. Fourth, we will review the simulation results and suggest some areas for future study.

1.1. Multivariate Cost and Quality Functions

Optimal allocation of resources for collection of sample survey and administrative record data generally requires us to balance several cost and quality factors. For example, we can consider functions $C(D, X, Z, \gamma_C)$ and $Q(D, X, Z, \gamma_Q)$ where D is a vector of design factors that we can control – for example, sample size; X is a vector of population characteristics that we observe but cannot control – for example, the proportion of households covered by a given set of administrative records; and Z is a second vector of population characteristics that we cannot control and also cannot observe before data collection – for example, recent changes in purchasing behaviour. In addition, we can

think of these cost and quality functions as depending on parameters and , respectively. For example, may include per-unit costs of observation within each stratum, and may include stratum-level variances and design effects.

1.2. Example: Optimal Allocation of Sample Sizes Across Strata

In a classical sampling example of this optimal-allocation work, we seek to minimize our cost for a given fixed quality measure – in this case, the variance A standard derivation (e.g., Cochran, 1977, Chapter 5) leads us to compute stratum-level sample sizes as a function

of the overall sample size , the prespecified , the population size , and stratum-level weights , standard deviations , and unit costs . Similar approaches can be used in allocating resources in collection of data from both sample surveys and administrative record sources.

2. Uncertainty in Measurement of Data Quality and Cost Factors

With some exceptions (for example, Isaki, 1983), formal optimization methods for sample designs generally use the assumption that the cost and quality parameters – the and terms in our development – are known and fixed. However, in practical applications with both surveys and administrative record systems, our cost and quality parameters often are unknown, and may vary over time.

For example, in the sample-allocation case considered earlier, the stratum-level variances and per-unit sampling costs often are estimated with a substantial amount of error. In addition, these quantities may be subject to change over time due to changes in population conditions.

Similarly, in work with administrative records, there are often nontrivial errors, and we often have limited information regarding the mean and covariance structure of those errors. In addition, statistical work with administrative records generally will involve substantial costs, and we often have limited information regarding variable-cost components related to specific steps in our statistical work. Examples include the costs of data management, record linkage, or edit and imputation.

3. Design of the Simulation Study

To study the ways in which errors in cost and quality parameters can affect the performance of nominally optimal procedures, we carried out a simulation study.This study focused on methods to combine sample survey and administrative records to produce an efficient estimator of an overall population mean

$$\bar{Y} = \sum_{h=1}^{H} W_h \bar{Y}_h$$

In estimation of each subpopulation-level mean we combined data from surveys and administrative records, respectively.

For the survey component, our estimators contain sampling error. We used data from a standard without-replacement stratified random sample. The allocation of sample size was based on the standard approach described earlier, with the goal of minimizing the cost, subject to the constraint that the variance for the *sample survey* estimator would be equal to one. For the administrative-record component, we assumed that we obtained data for each unit in each stratum, but that the administrative data had substantial measurement errors.

The final estimator of each stratum mean was a weighted least squares average of the survey- and administrative-based elementary estimators, where the weights were proportional to the estimated variances for these respective elementary estimators. We produced different results for two design approaches. Under design "C" sample allocation and estimation were *constrained* to use only the nominal stratum-level cost and variance estimation. On the other hand, design "D" adjusted its allocation to use the true cost and variance information.

In addition, we used a total of seven strata, with population sizes and nominal variances based on an example from Cochran (1977, p. 111). Table D below presents these from the Cochran example, as well as the per-unit costs that we have added for each stratum. Note especially that the stratum-level variances displayed moderate variability, while the unit costs (added for purposes of this analysis) display more substantial variability across strata.

Table 0: Stratum-level properties used in the simulation study. (Adapted from Cochran, 1977, p. 111)

Stratum	Population	Mean	Std Dev	Cost
1	394	5.4	8.3	1
2	461	16.3	13.3	2
3	391	24.3	15.1	3
4	334	34.5	19.8	4
5	169	42.1	24.5	5
6	113	50.1	26.0	6
7	148	63.8	35.2	7

To study the effects of error in our design parameters, we considered true stratum-level variances $S_h^2 = d_\varsigma S_{h,\text{nominal}}^2$ and costs $c_h = (1+d_c)c_{h,\text{nominal}}$.

The variance-disturbance term $D\ d_S$ followed a chi-square distribution on D degrees of freedom, and in addition, the cost-disturbance term $d_C \sim N(0, s^2)$ Also, the measurement errors in the administrative records were treated as following a $N(0, ARxS_h^2)$ distribution, where AR is a scale parameter equal to the ratio of the variance of administrative record measurement error, divided by the stratum-level variance S_h^2. that describes the variability in the administrative-error distribution, relative to the variability of the true values within each stratum.

We developed simulation results separately for each combination of the values s = 0.1, 0.2; D = 2, 4, 8; and AR = 0.25, 1.0. For each of these combinations, we ran the simulation with 1000 replications; results for selected combinations of s, D and AR are presented in the tables. In addition, to reduce visual clutter, results from only the first 100 replications are displayed in each of the graphs.

4. Simulation Results

Figures 1 through 5 present plots of variance against cost for various combinations of s, D and AR. In each of these plots, the blue squares represent cases in which the sample allocation was not adjusted for the new (correct) unit-level cost and variance estimation. Thus, the blue squares represent the performance of a design and estimation procedure that one would have to use if one did not obtain improved information on the true values of the unit-level costs and variances for each stratum. Also, the green triangles represent the cases in which the sample allocation *was* adjusted for the updated information on unit-level costs and variances.

Recall that for the current study the sample allocation was based on the goal of minimizing total cost, conditional on the constraint that the sample-based estimator would have a variance less than or equal to one. Under ideal conditions, incorporating the administrative record data should lead to further reduction of the variance to a value below one, at some additional cost. This is the approach one would use when the top priority is the precision of the estimator, with cost as an important but secondary consideration.

Figure 1 presents results for the case in which our design information contained errors, but was relatively good, with 8 degrees of freedom for the stratum-level variances, a standard deviation on costs equal to 0.1, and an administrative-record variance ratio equal to 0.25. In this case of relatively mild degradation of design information, the unadjusted design (blue squares) led to some inflation in variance for the overall estimator, including some realizations that had an overall variance exceeding the nominal constraint of one. On the other hand, the adjusted design (green triangles) had consistently low and stable variances across the 100 replications displayed here, but had much higher dispersion of the true costs. In essence, the constrained-optimization approach used the improved design information to maintain tight control over the variance, but at the price of increased uncertainty on costs.

Figure 2 covers a similar scenario, but with the administrative-record variance ratio equal to one. Figure 3 displays results for the same conditions, except that the variability in cost information increases from s=0.1 to s=0.2.

Figures 4 and 5 display results obtained when D=2, i.e., the stratum-level variance information had a high degree of uncertainty. For these cases, note especially the substantial proportion of unadjusted-design cases that had an estimator variance greater than one.

To explore further the case with the most uncertainty in design information (D=2, s=0.2, AR = 1.0) represented by Figure 5, Figure 6 presents a plot of cost for the unadjusted design (on the vertical axis) against cost for the adjusted design (on the horizontal axis). In keeping with comments above, the improved design information, in conjunction with the sample-design variance constraint, has led to a substantial increase in the dispersion of cost for the adjusted procedure. This figure indicates relatively little association between the costs for the adjusted and unadjusted cases.

Finally, for the same extreme case of design conditions, Figure 7 plots the variance for the unadjusted design (on the vertical axis) against variance for the adjusted design (on the horizontal axis). Since the allocation procedure was based on a variance

constraint, this final plot is of special interest. Note especially that substantial proportion of unconstrained-design realizations that produce a variance greater than one, and the relatively strong association between variances for the unconstrained and constrained cases, respectively.

Tables 1 through 6 explore in additional detail the ways in which variance and cost results may be sensitive to changes in values of D and s. Table 1 displays the mean and selected quantiles of cost for the unadjusted design with s=0.2 and D=8, 4 and 2. Note that these values are quite stable over changing values of D. Table 2 displays corresponding results for variances for the unadjusted design. Note that as D decreases from 8 to 2, the average variance stays about he same, but the tail quantiles become more dispersed. Tables 3 and 4 present parallel results for the adjusted design. Note especially in Table 3 for D=2 the mean cost is 1472, or about 10 percent less than the mean cost for the unadjusted design in Table 1. Also, in Table 4 for D=2, the mean variance is about 5 percent less than for the unadjusted design in Table 2. These results give an indication of the value provided by adjusting the design to account for uncertainty in cost and variance information. Finally, for the unadjusted-design case, Table 5 displays the sensitivity of cost to changes in s when D=2, and Table 6 displays the corresponding variance results. Thus, adjustment for the true cost and variance information produces some improvements in both cost and variance overall.

5. Closing Remarks

In summary, this paper considered the ways in which uncertainty in the data-quality and cost information may affect the performance of optimization methods for surveys and administrative record systems. The numerical work centered on the variance and cost of weighted least squares estimators that combined data from, respectively, a stratified random sample and an administrative record source subject to measurement error. For the cases studied, the results indicated that standard optimization methods are relatively sensitive to uncertainty in unit-level cost and variance information.

One could consider several potential extensions of this work. First, the current work focused on allocation intended to minimize cost, subject to a constraint on the variance of the sample-based estimator. In future work, we will also evaluate allocation methods that minimize variance, subject to a cost constraint. In addition, the current numerical work restricted attention to cases in which the estimators of cost and variance parameters are unbiased. In practice, some cost and variance information is biased, and one could extend our simulation work to include the bias cases. In addition, administrative record sources often incorporate data only from subpopulations, and not from the full population of interest. For these partial-population cases, methods for integration of survey and administrative records has similarities to methods developed previously in the multiple frame literature, for example Lohr and Rao (2000, 2006). Extension of our simulation work to the partial-population case would also be of interest.

Acknowledgements

The authors thank Mark Denbaly for helpful comments on an earlier version of this paper. The views expressed in this paper are those of the authors and do not necessarily reflect the policies of the U.S. Bureau of Labor Statistics.

References

Cochran, W.G. (1977). *Sampling Techniques, Third Edition*. New York: Wiley.

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

Isaki, C.T. (1983). Variance estimation using auxiliary information. *Journal of the American Statistical Association*, 78, 117-123.

Kish, Leslie (1988). Multipurpose sample designs. *Survey Methodology*, 14, 19-32.

Lohr, Sharon L. and Rao, J. N. K. (2000). Inference from dual frame surveys. *Journal of the American Statistical Association*, 95, 271-280

Lohr, Sharon and Rao, J.N.K. (2006). Estimation in multiple-frame surveys. *Journal of the American Statistical Association*, 101, 1019-1030

Figure 1:

Plot of Variance vs. Cost
D=8, std=0.1 "AR" mult=0.25
Green Triangle=Adjusted
Blue Square=Not Adjusted

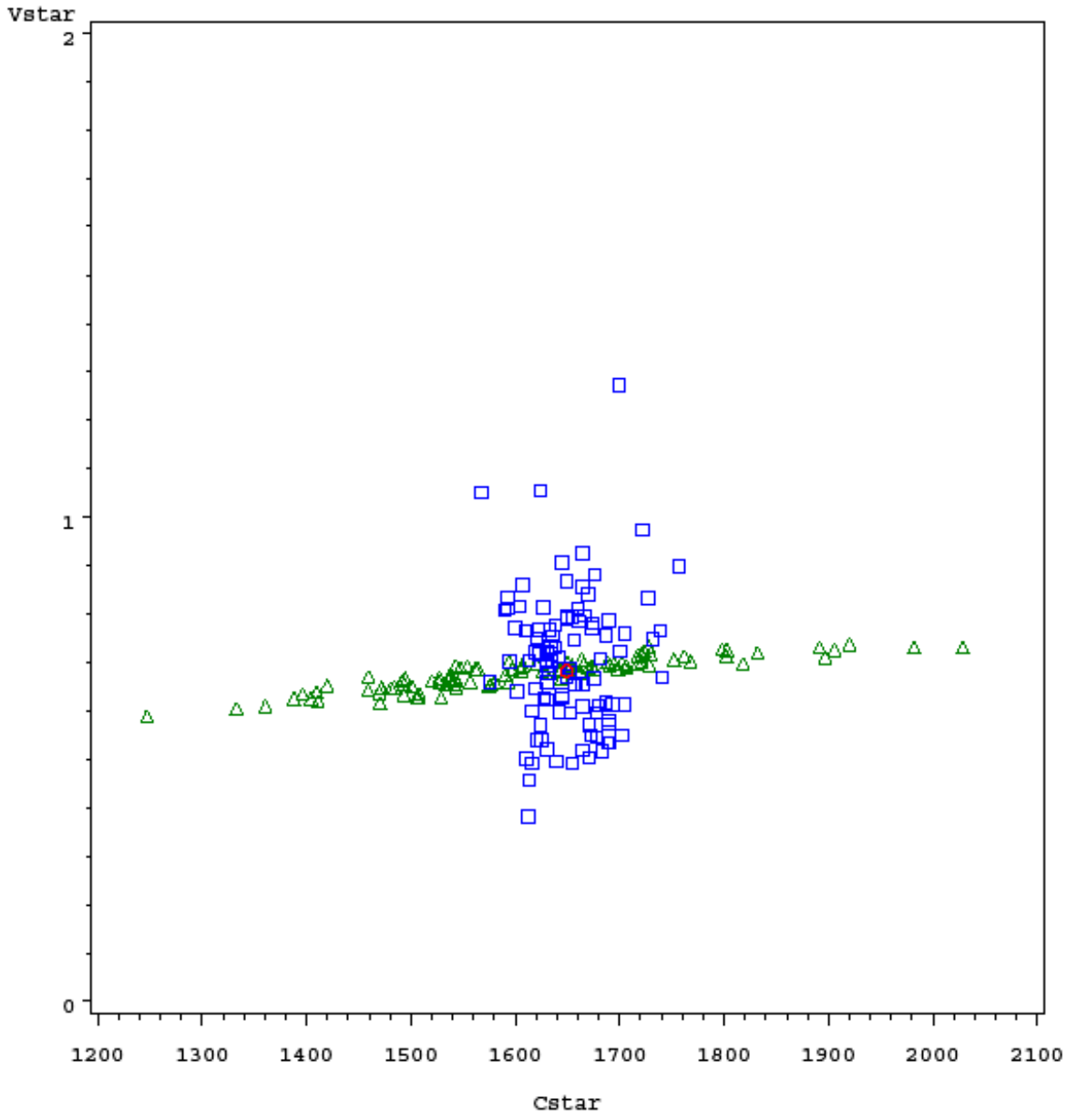

Figure 2:

Plot of Variance vs. Cost
D=8, std=0.1 "AR" mult=1.00
Green Triangle=Adjusted
Blue Square=Not Adjusted

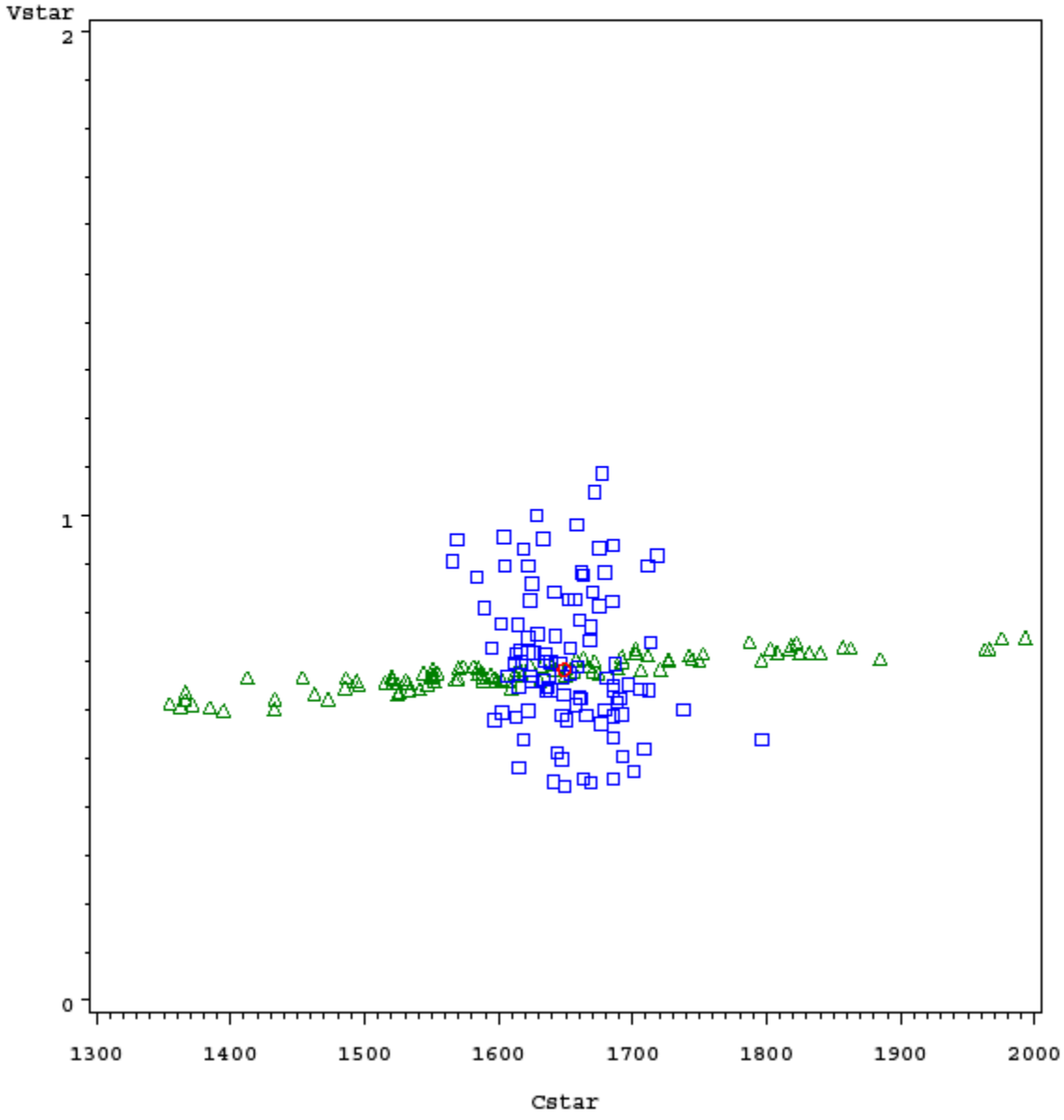

Figure 3:

Plot of Variance vs. Cost
D=8, std=0.2 "AR" mult=1.00
Green Triangle=Adjusted
Blue Square=Not Adjusted

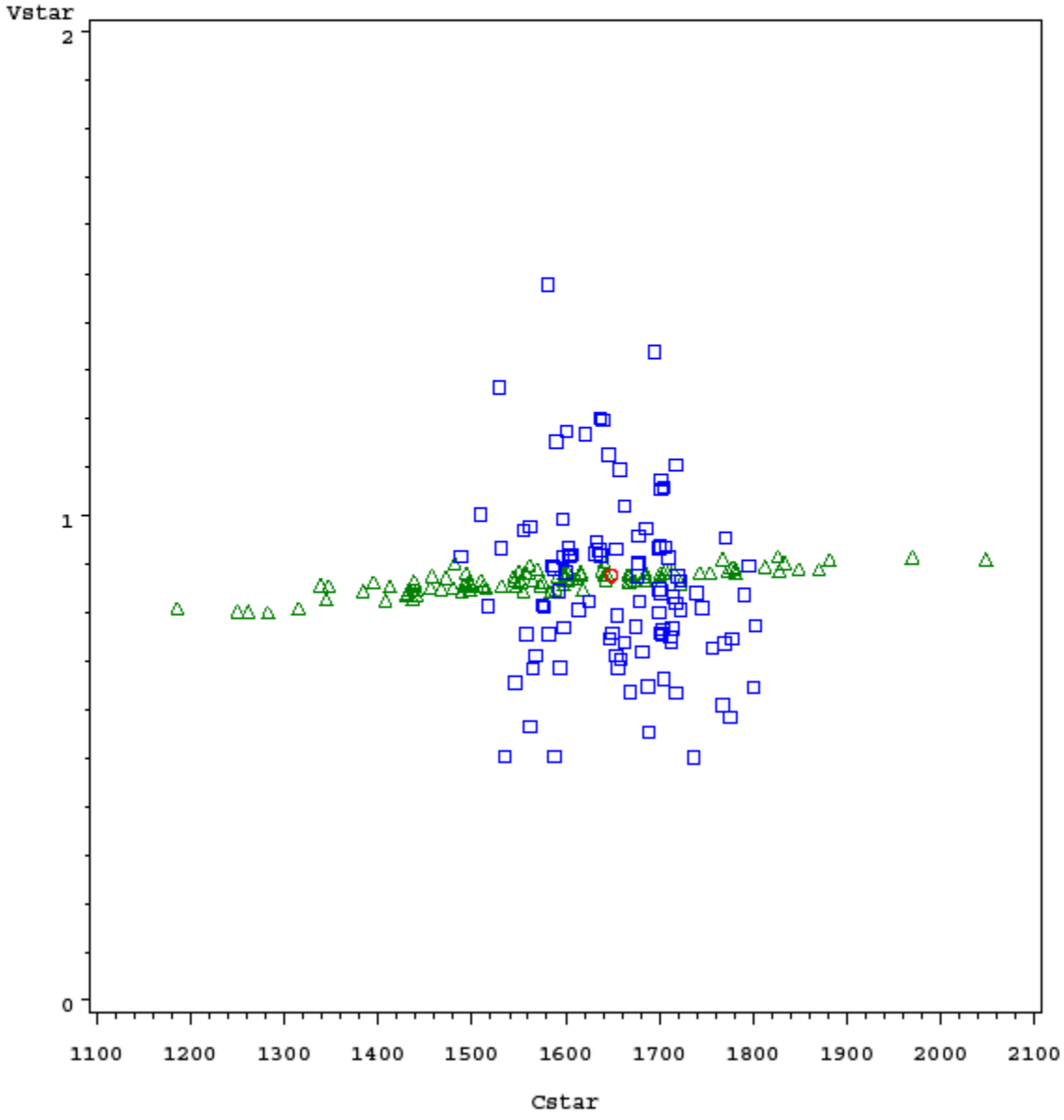

Figure 4:

Plot of Variance vs. Cost
D=2, std=0.1 "AR" mult=1.00
Green Triangle=Adjusted
Blue Square=Not Adjusted

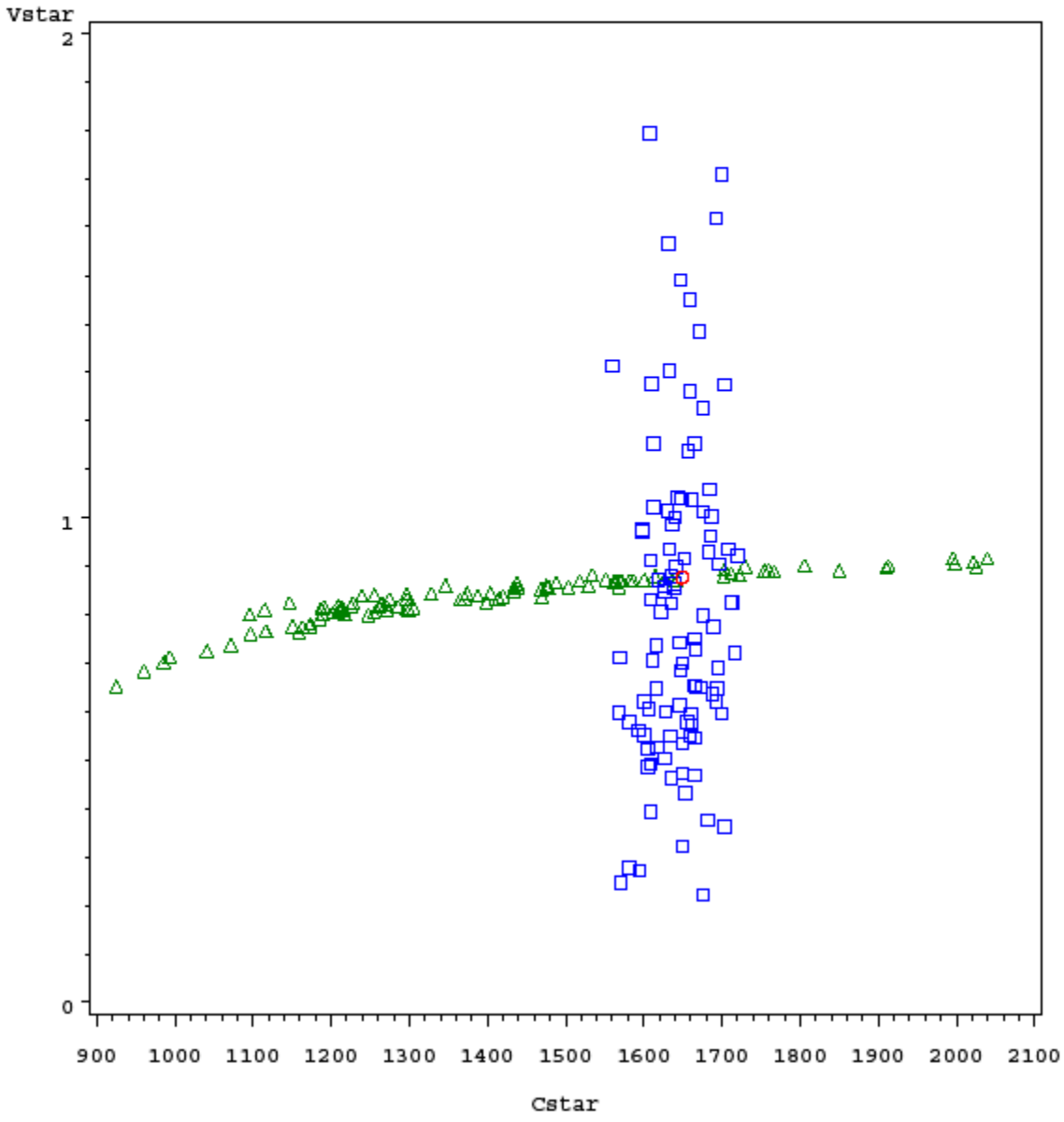

Figure 5:

Plot of Variance vs. Cost
D=2, std=0.2 "AR" mult=1.00
Green Triangle=Adjusted
Blue Square=Not Adjusted

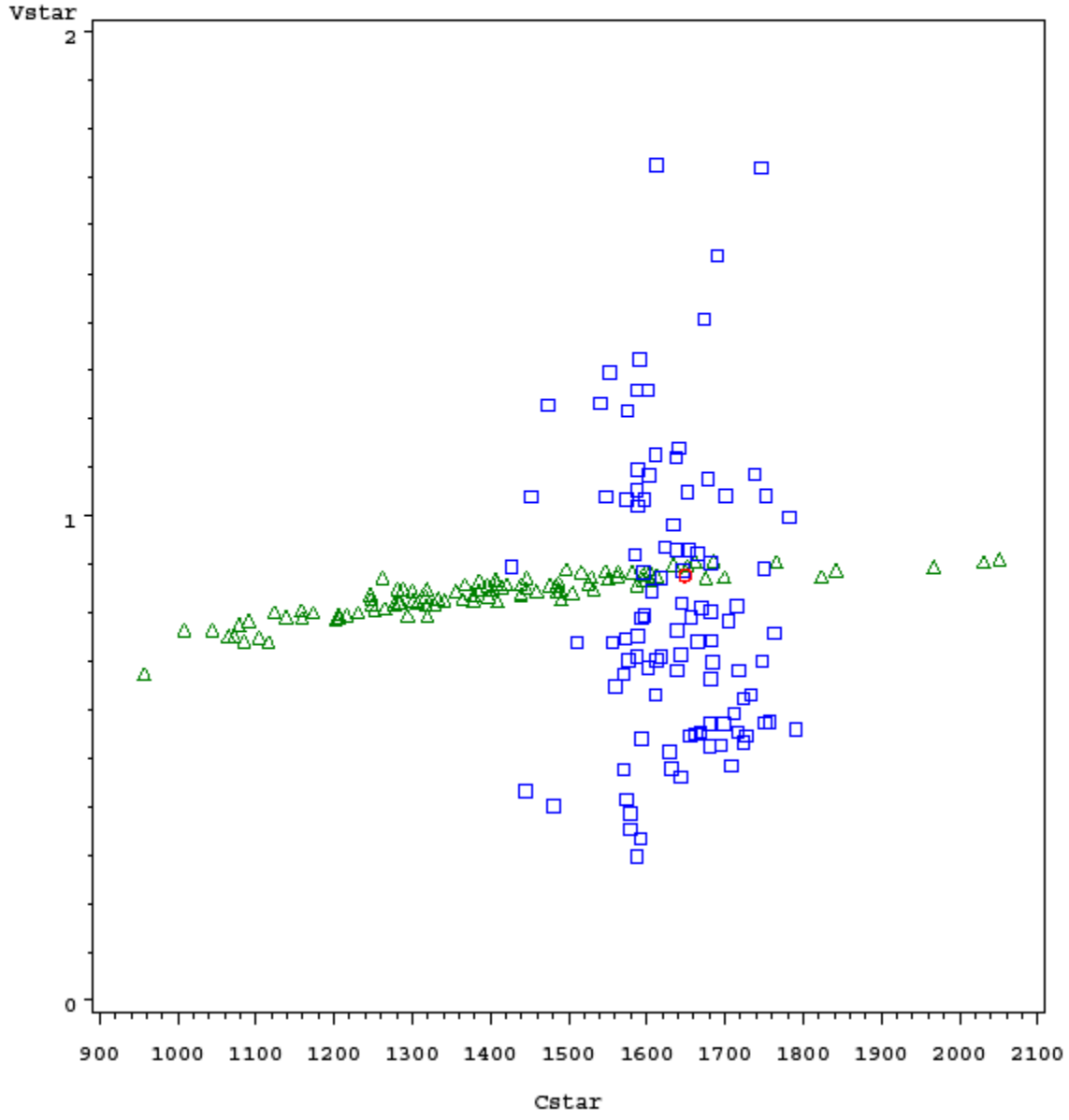

Figure 6:

Plot of Cost for Unadjusted vs. Adjusted Allocation
D=2, std=0.2 "AR" mult=1.00

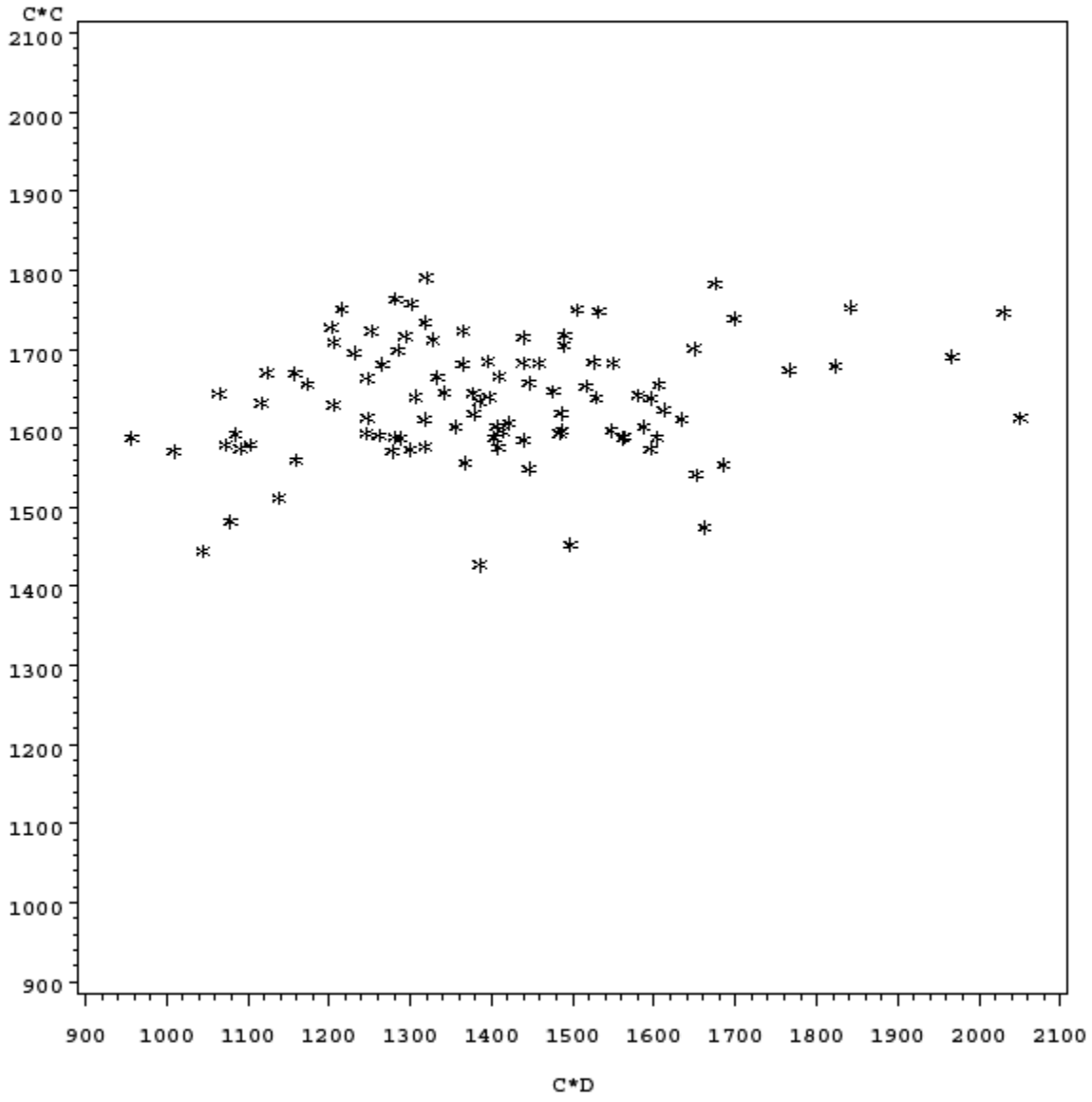

Figure 7:

Plot of Variance for Unadjusted vs. Adjusted Allocation
D=2, std=0.2 "AR" mult=1.00

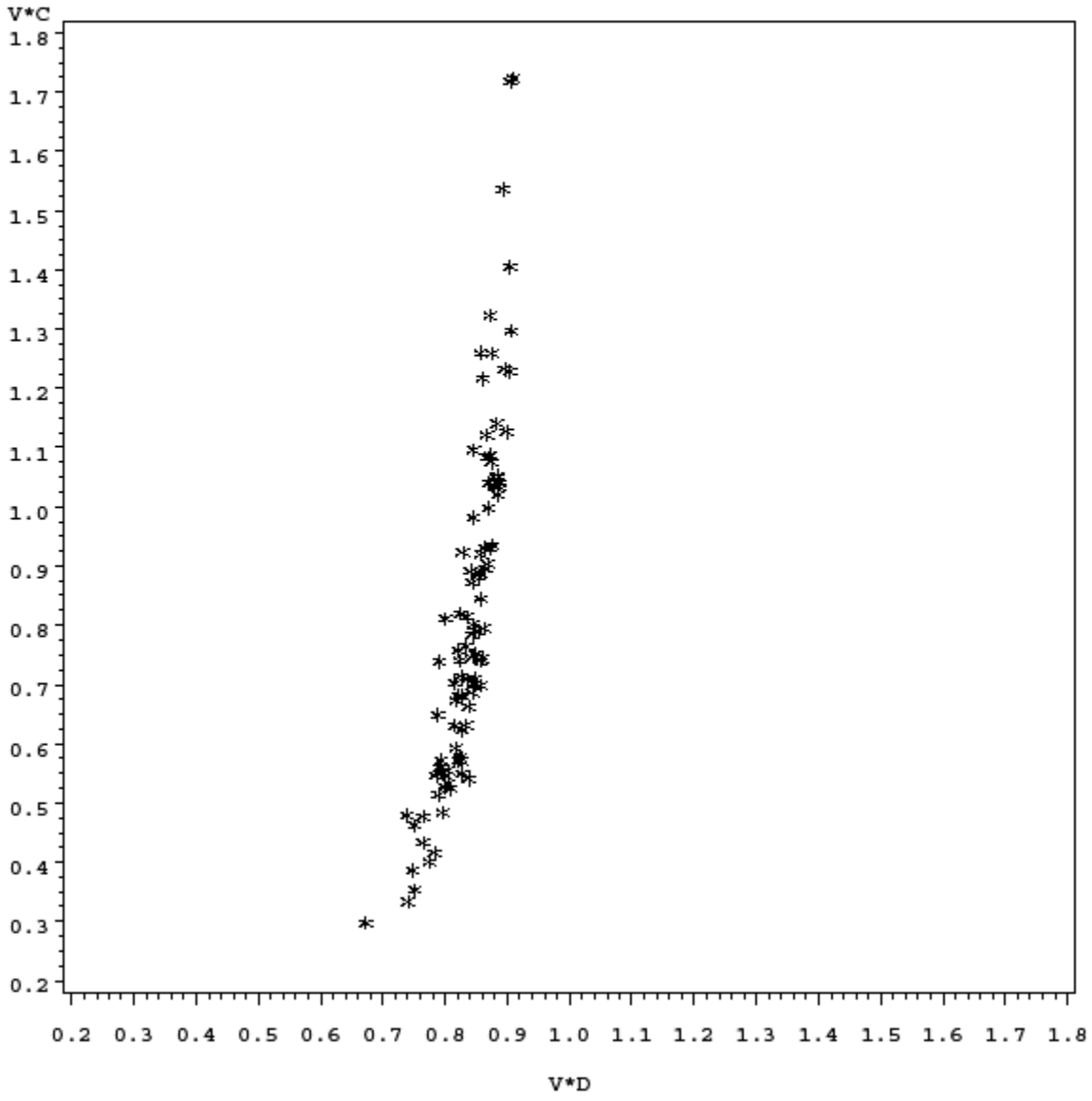

Table 1: Sensitivity of Cost C_C^* to Changes in D, ("degrees of freedom" for variance estimator) when s=0.2 and AR=1.00

	$Q_{0.05}$	$Q_{0.25}$	Mean	$Q_{0.75}$	$Q_{0.95}$
D=8	1527.46	1601.54	1648.73	1696.69	1763.47
D=4	1528.75	1600.41	1648.56	1694.76	1762.30
D=2	1528.14	1601.87	1648.61	1695.15	1764.35

Table 2: Sensitivity of Estimator Variance V_C^* to Changes in D, ("degrees of freedom" for variance estimator) when s=0.2 and AR=1.00

	$Q_{0.05}$	$Q_{0.25}$	Mean	$Q_{0.75}$	$Q_{0.95}$
D=8	0.6151	0.7492	0.8824	0.9970	1.2113
D=4	0.4941	0.6783	0.8581	1.0048	1.2949
D=2	0.4127	0.6261	0.8874	1.0857	1.5024

Table 3: Sensitivity of Cost C_D^* to Changes in D, ("degrees of freedom" for variance estimator) when s=0.2 and AR=1.00

	$Q_{0.05}$	$Q_{0.25}$	Mean	$Q_{0.75}$	$Q_{0.95}$
D=8	1347.92	1473.08	1595.65	1715.09	1872.24
D=4	1214.11	1377.32	1529.43	1661.63	1888.28
D=2	1092.51	1280.72	1472.45	1630.21	1935.30

Table 4: Sensitivity of Estimator Variance V_D^* to Changes in D, ("degrees of freedom" for variance estimator) when s=0.2 and AR=1.00

	$Q_{0.05}$	$Q_{0.25}$	Mean	$Q_{0.75}$	$Q_{0.95}$
D=8	0.8267	0.8544	0.8679	0.8840	0.9003
D=4	0.7982	0.8371	0.8565	0.8801	0.9011
D=2	0.7533	0.8205	0.8456	0.8787	0.9062

Table 5: Sensitivity of Cost C_C^* to Changes in s, (standard deviation of error in per-unit cost) when D=2.0 and AR=1.00

	$Q_{0.05}$	$Q_{0.25}$	Mean	$Q_{0.75}$	$Q_{0.95}$
S=0.1	1589.29	1622.64	1648.41	1672.83	1711.53
S=0.2	1528.14	1601.87	1648.61	1695.15	1764.35

Table 6: Sensitivity of Estimator Variance V_C^* to Changes in s, (standard deviation of error in per-unit cost) when D=2.0 and AR=1.00

	$Q_{0.05}$	$Q_{0.25}$	Mean	$Q_{0.75}$	$Q_{0.95}$
S=0.1	0.3966	0.6222	0.8787	1.0711	1.5554
S=0.2	0.4127	0.6261	0.8874	1.0857	1.5024