REGRESSION ANALYSIS WITH COMPLEX SURVEY DATA:

A COMPARISON OF ESTIMATION TECHNIQUES

BY

STAN LDABERKOW

Submitted for delivery at the American Agricultural Economics Association Meetings, August, 1984. Cornell University, Ithaca, New York.

ABSTRACT

"Regression Analysis With Complex Survey Data: A Comparison of Estimation Techniques." Stan Daberkow (U.S. Department of Agriculture)

A complex survey design often necessitates the use of some approximation technique when calculating regression standard errors. Two approximation procedures, the Taylor expansion method (TEM) and balanced repeated replication (BRR) technique, were compared to ordinary least squares (OLS) and weighted least squares (WLS). While the regression coefficients were similar across all procedures, TEM and BRR produced larger standard error estimates than did OLS and WLS.

REGRESSION ANALYSIS WITH COMPLEX SURVEY DATA:

A COMPARISON OF ESTIMATION TECHNIQUES

Contemporary large sample data collection efforts rarely utilize a simple random sample survey design. Cost considerations, the availability of auxiliary information about the population, oversampling, or the desire to obtain highly precise point estimates encourage the use of complex survey design. However, using such techniques as chi-squared tests, analysis of variance or regression with data collected through complex survey designs has raised a number of methodological problems. Foremost among these issues is whether the survey design must be included in the analysis and if so, how? Some researchers argue that survey design is irrelevent for most analytical work (Cramer), while others suggest some weighting scheme is necessary (Porter). Several writers propose an explicit recognition of the survey design through the use of dummy variables (Smith). Other authors suggest that most complex sample designs inherently introduce violations of the basic assumptions underlying such techniques as regression analysis (Kish and Frankel). What <u>is</u> clear is that most econometric textbooks do not address this issue.

I gratefully acknowledge the programming assistance of John Fritsvold, Data Services Center, ERS.

The purpose of this paper is to 1) briefly discuss various problems, controversies, and assumptions associated with using complex survey data in analytical studies; 2) examine selected regression techniques which incorporate the survey design into the analysis; and 3) test these techniques with data derived through a complex survey design which includes stratification, clustering, multiple stages and multiples sampling frames. Ordinary least square (OLS), weighted least squares (WLS), the balanced repeated replication (BRR) approach and the Taylor expansion method (TEM) are the regression techniques compared. The results emphasize how the regression coefficients and standard error estimates differ across these analytical methods.

BACKGROUND

The seminal 1974 article by Kish and Frankel is widely cited as a comprehensive empirical investigation of the use of complex survey data in analytical studies. Their schematic of survey designs cross-referenced by type of statistic desired is a helpful guide (figure 1). Row A, a simple random sample survey design, poses few problems as the complexity of statistics increases from left to right. Similarly, column (1) quantities are easily calculated for most survey designs. The remaining boxes are less straight forward with complex analytical statistics from complex survey designs generating the most difficulties.[1]

---

[1] The following discussion assumes the use of the regression model is not in doubt and that the data has been scrutinized using, for example, the techniques suggested by Belsley, et. al.

STATISTICS

Selection Methods	1. Means and totals of entire samples	2. Subclass means and differences	3. Complex analytical statistics (e.g. coefficients in regression)
A. Random sample selection of elements	No Problem	No Problem	No Problem
B. Stratified selection of elements	No Problem	Available	Conjectured
C. Complex cluster sampling	No Problem	Available	Difficult

Source: Kish and Frankel

Figure 1. The present status of sampling errors.

If one makes the assumption that the coefficients of a particular model are homogeneous across the entire population, the regression results are not affected by the survey design (Porter).[2] In such cases OLS becomes defensible. If this assumption is not valid, then some way of incorporating the survey design into the regression analysis must be found. In certain situations WLS is one alternative. Advocates of WLS argue, by analogy, that the estimates for population totals, means and ratios requires weighting, hence regression techniques also require weighting. In the case of a stratified sample design different sampling and response rates lead to a weighting scheme "...which attempts to give each stratum the same relative importance in the sample that

---

[2] Holt, et. al. also present conditions under which selection probabilities can be ignored and OLS is appropriate. Basically they require a simple random sample survey design which is comparable to the assumption underlying OLS.

it has in the population" (DuMouchel and Duncan, p. 535).[3]

Regression coefficients and their standard errors are based on the assumption of independent selection of elements. However, complex survey designs often introduce correlations between the element values particularly when clusters are involved. Therefore, Kish and Frankel argue that (p.2) "Standard errors should be computed in accord with the complexity of the sample design; neglect of that complexity is a source of serious mistakes." They recommend either the BRR and TEM approach to account for these positive correlations between element values. They suggest that OLS coefficient estimates show little bias but that the estimate of variance underestimates the true sampling variance.[4]

SURVEY DATA

In late 1979 and early 1980 the Economic Development Division of the Economic Research Service, in cooperation with the Statistical Reporting Service, conducted a survey of employers and households in a 9-county area of south central Kentucky. The survey design was a multiple-frame, multi-stage, stratified, clustered design (Kleweno, 1980). A list frame of employers, through which employee households were identified for interview, was supple-

---

[3] DuMouchel and Duncan explore the different underlying assumptions behind OLS and WLS when stratified survey data is involved and they suggest a criterion by which a choice can be made between the two sets of regression estimates.

[4] A number of other procedures for incorporating the survey design in the analysis are available but not tested in this paper. For example, Smith suggests a dummy variable for each strata (except one). However in surveys with a large number of strata, the number of dummy variables can become prohibitive. Porter suggests a random coefficients model because errors arise from two sources: random error plus the random selection of individuals. Holt, et. al. and Nathan and Holt have shown that standard errors of OLS regression coefficients are biased when data originate with a stratified or multi-stage survey design. They show that if a design variable is correlated with either the dependent or independent variables, an alternative estimator to OLS is needed. They suggest a maximum likelihood estimator.

mented by an area frame of households. The area frame was necessary because of the incompleteness of the employer list.

Employers were regarded as primary sample units in the list frame while land segments served the same purpose in the area frame. The first stage of the list frame randomly selected employers (i.e. clusters of employees) which had been stratified by employee size and industry. The second stage involved subsampling households from a list of the chosen employer's current employees. Various socio-economic data were collected from members, age 16 and over, of the employee's household (Daberkow, et. al.).

The first stage of the area frame involved randomly selecting area segments (i.e. clusters of individuals age 16 and over) which had been stratified by household density. Within a chosen segment, a subsample of households was identified and an enumerator screened the households for possible overlap with the list frame households. The probabilities at each stage were originally chosen to generate a self-weighting sample but because of refusals and incompleted questionnaires at one or more stages this goal was not attainable. As a result, the probabilities of selection varied by strata. Collapsing strata with zero or only one primary sample unit resulted in 23 strata.

MULTIPLE REGRESSION PROGRAMS

OLS results are from the SAS (Statistical Analysis System) software package. WLS results are given for the SAS and SPSSX (Statistics Package for Social Sciences) weighted least squares procedures.[5] The weights used in this

---

[5] The finite population correction factor (FPCF) is rarely considered in most analyses either theoretically or operationally. The assumption of an infinite population is convenient but in some strata we may be dealing with a small

(Footnote continued)

paper are based on the proportion of the population in each stratum (Pj) and the size of the sample in each stratum (Nj) where j = 1,..., k and k is the number of strata. The weight for the ith observation is proportional to Pji/Nji, where i=1,..., n, and n is the total sample size.

The TEM algorithm used here is contained in a program called SUPERCARP (Cluster Analysis and Regression Program) developed at Iowa State University (Hiridroglou, et. al.). TEM "... produces an approximate estimate for the variance of a first-order statistic based on variances of the linear terms of the Taylor expansion of the statistic" (Kish and Frankel, p.14). The primary assumption of TEM is that the linear terms of the expansion are appropriate approximations of the sampling variance.

The BRR approach is a SAS oriented package which was written for the National Center for Education Statistics (Wise). With the (BRR) approach, the variance of a statistic is estimated using the variability among replicates of the full study. "Each replicate is created by excluding a sub-sample of primary selections (PS) in the dataset. The idea is for each replication to reproduce, except for size, the design of the entire study. The statistic of interest is then estimated for the whole sample and for each replication. The variability among these estimates is used to estimate the variance of the statistic." For large designs with many strata and PS's," (t)he replicates are selected in such a way that they are mutually orthogonal." (Brandt, p. 9-10). Brandt compares TEM and BRR with respect to documentation, ease of set-up,

---

[5](continued)
population of which we sample all or nearly all of the units. Smith points out (p.2) "...omitting the (FPCF) for calculations for samples from small populations can often result in appreciable overestimate of the variances." The popular statistical packages including the two used here, assume an infinite population or sampling with replacement. TEM and BRR specifically allow for the FPCF.

options and the cost of acquiring and running the programs.

RESULTS

A labor supply model, using annual hours worked as the dependent variable, is used to illustrate the differences among the various estimation techniques (Table 1). Each independent variable is defined in the appendix.

The OLS estimation method ignores the different sampling rates and/or different responses rates among strata which produces different selection probabilities for each observation. Hence, no attempt is made to give the same importance to each observation as it has in the population from which it was drawn. In a severe case where the bulk of the sample is drawn from one or a small number of strata, the regression coefficients will apply primarily to individuals in those strata not to the population as a whole.

The weighting process in SPSSX merely replicates each case by the size of the expansion factor (i.e., inverse of the probability of being selected) and then uses an OLS algorithm to solve for the regression coefficients. Although the estimated coefficients are equal to those found in SAS's WLS program, the greatly reduced standard errors on the coefficients are based on a expanded number of observations which has no theoretical justification. A re-scaling of the expansion factors is necessary to overcome this nonsensical result in SPSSX.

WLS regression coefficients differed from the OLS estimates in nearly all cases although the signs remained consistent between the two techniques.[6] If one uses the t-statistic as a criterion for ascertaining the statistical

---

[6] Note that the use of WLS to correct for heteroscedasity is not the problem in this case. "In the usual homoscedastic regression model, [the OLS estimator] is minimum variance unbiased whether or not the strata are sampled proportional to size" (p.535, DuMouchel and Duncan.)

Table 1-- Results from selected regression packages using complex sample data 1/

	OLS	SPSSX/WLS	SAS/WLS	TEM	BRR
Intercept	713.0 (4.21)*	628.0 (25.83)*	628.0 (3.58)*	627.4 (3.09)*	627.4 2/
AGE	28.3 (3.26)*	29.8 (24.44)*	29.8 (3.39)*	29.8 (2.98)*	29.8 (2.65)*
AGESQ	-.3 (3.02)*	-.3 (23.26)*	-.3 (3.22)*	-.3 (2.87)*	-.3 (2.93)*
DED2	69.3 (1.76)***	47.2 (8.27)*	47.2 (1.15)	47.2 (.91)	47.2 (.70)
DED3	-15.6 (.34)	-17.3 (2.52)**	-17.3 (.35)	-17.3 (.31)	-17.3 (.24)
DJTR	93.1 (2.04)**	168.6 (24.95)*	168.6 (3.46)*	168.6 (2.48)**	168.6 (1.75)***
DHLT	-67.3 (1.31)	-50.3 (6.71)*	-50.3 (.93)	-50.3 (.64)	-50.3 (.40)
DMAR	41.3 (.93)	61.7 (10.05)*	61.7 (1.39)	61.8 (1.08)	61.7 (1.09)
DEMG	-68.0 (1.38)	-112.1 (15.13)*	-112.1 (2.10)**	-112.1 (1.76)***	-112.2 (1.83)***
DRMG	-55.8 (1.25)	-84.4 (13.41)*	-84.4 (1.89)***	-84.5 (1.43)	-84.5 (1.55)
DSEX	103.9 (2.93)*	148.3 (27.86)*	148.3 (3.86)*	148.3 (2.88)*	148.3 (2.90)*
DSCH	-736.1 (8.03)*	-665.3 (54.14)*	-665.3 (7.50)*	-665.3 (5.46)*	-665.3 (5.01)*
NEWEMP	-126.3 (2.88)*	-134.9 (21.03)*	-134.9 (2.91)*	-135.0 (2.32)**	-134.9 (2.54)**
ACHLD10	-6.6 (.34)	-26.7 (8.77)*	-26.7 (1.21)	-26.7 (.88)	-26.7 (1.08)

(cont'd)

Table 1-- Results from selected regression packages using complex sample data 1/ continued

	OLS	SPSSX/WLS	SAS/WLS	TEM	BRR
MILE	-1.9	-2.6	-2.6	-2.6	-2.6
	(1.40)	(13.35)*	(1.85)***	(1.89)***	(1.55)
WKWG	3.5	3.7	3.7	3.7	3.7
	(12.66)*	(95.15)*	(13.18)*	(9.61)*	(2.29)**
WKWGSQ	-.0018	-.0019	-.0019	-.0019	-.0019
	(7.50)*	(62.40)*	(8.65)*	(7.69)*	(6.33)*
OTH	.0018	.0058	.0058	.0058	.0058
	(.13)	(3.07)*	(.42)	(.31)	(.32)
OTHSQ	-.00000020	-.00000014	-.00000014	-.00000014	-.00000014
	(.23)	(1.34)	(.19)	(.20)	(.18)
D.F.	1116	58,143	1116	1116	1116
R-SQUARE	.40	.41	.40	.41	.41
MSE	264,827	297,600	297,599	302,584	2/

Dependent Variable: Total hours worked in 1979; mean = 1973.1 hrs/yr; standard deviation = 710.1.

1/ t-statistics are in parentheses.

* significant at .01 level.
** significant at .05 level.
*** significant at .10 level.

2/ Not reported

reliability of a particular coefficient, then some differences between OLS and WLS appear. DED2 becomes "less significant" while DEMG, DRMG, DJTR and MILE are "more significant" using the WLS estimate rather than OLS. The remaining coefficients which were significant in the OLS model were also significant in WLS. In most cases, the standard errors of the regression coefficient were underestimated by OLS compared to WLS.

TEM regression coefficient estimates are nearly equal to those found in WLS (table 1). What does change with TEM are the standard errors of the regression estimates. By accounting for the "positive correlations between the errors of the model" (p.10, Kish and Frankel) introduced by clusters (i.e. employers, segments and households in this case), TEM produces larger estimates of the standard errors of regression coefficients than either OLS or WLS. This is consistent with the conclusion of Kish and Frankel.

As with TEM, BRR produced regression coefficients equal to those found in WLS. MILE was no longer significant at even the .10 level of significance. DJTR and WKWG became "less significant" than in the TEM algorithm. In general, BRR calculated even larger standard errors than did TEM.[7]

---

[7] If the survey design does not naturally produce exactly two primary sample units per stratum then BRR requires the user to artifically create such a design. This study did require some manipulation to create such a BRR design. TEM has no such constraint. Some of the difference between BRR and TEM may be attributable to the artifically created BRR design.

CONCLUSIONS

The regression program chosen for analysis with complex survey data does make a difference. Both the size of the regression coefficients and their standard errors change depending on whether one compares OLS with WLS or WLS with TEM. TEM and BRR were shown to be approximately equivalent for the data used in this analysis although BRR tended to produce the largest standard errors. Either TEM or BRR account for the design effect when clustering is present in the survey design. In doing so, both techniques produce more precise standard error estimates. Compared to BRR and TEM, WLS produced similar regression coefficients and only slightly underestimated the coefficient's standard errors. WLS appears preferable to OLS. With WLS, each sample observation reflected the importance it had in the overall population, while OLS consistently underestimated regression standard errors.

These results suggest that researchers faced with analyzing complex survey data should critically evaluate their choice of analytical techniques.

REFERENCES

1. Belsley, D., Kuh and R. Welsch, Regression Diagnostics: Identifying Influential Data and Sources of Collinearity, New York: John Wiley, 1980.

2. Brandt, D., Comparison of Computer Programs Which Compute Sampling Errors for Complex Samples, Technical Report No. 26, Submitted to the National Center for Education Statistics by American Institutes for Research, Palo Alto, California, June 1982.

3. Daberkow, S., D. Larson, R. Coltrane and T. Carlin, Impacts of Employment Growth in Nonmetropolitan Areas: A Case Study of Nine Kentucky Counties, RDRR No._____, Economic Research Service, U.S. Department of Agriculture, Washington, D.C., 1984.

4. DuMouchal, W. and G. Duncan, "Using Sample Survey Weights in Multiple Regression Analyses of Stratified Samples," Journal of the American Statistical Association, Vol. 78, No. 383, September, 1983.

5. Hidiroglou, M., W. Fuller and R. Hickman, SUPERCARP, Sixth Edition, Survey Section, Statistical Laboratory, Iowa State University, Ames, Iowa, 1980.

6. Holt, D., and T. Smith and P. Winter, "Regression Analysis of Data from Complex Surveys," Journal of the Royal Statistical Society, Series A, No. 143, 1980.

7. Kish, L. and M. Frankel, "Inference from Complex Samples", Journal of the Royal Statistical Society, Series B, Vol. 36, 1974.

8. Kleweno, D., Application of the Multiple Frame Design In An Economic Distribution Effect Study. ESS No. AGESS 801222.3, U.S. Dept. of Agriculture, January, 1981.

9. Nathen, G. and D. Holt, "The Effect of Survey Design on Regression Analysis," Journal of the Royal Statistical Society, Series B, Vol. 42, 1980.

10. Porter, R. "On the use of Survey Sample weights in the Linear Model," Annnals of Economic and Social Measurement, No. 212, 1973.

11. Smith, K., "Analysing Disportionately Stratified Samples with Computerized Statistical Packages," Sociological Methods and Research, Vol. 5, 1976.

12. Wise, L. The BRRVAR Procedure: Documentation, Technical Report 28, Submitted to the National Center for Education Statistics by American Institutes for Research, Palo Alto, California, March 1983.

Appendix-Definition of Variables

AGE	Age in years
AGESQ	AGE squared
DED2	Equals one if high school was completed
DED3	Equals one if education exceeded high school
DJTR	Equals one if enrolled in a job training program
DHLT	Equals one if a health problem exists
DMAR	Equals one if married
DEMG	Equals one if migrated to the area within the last 5 years.
DRMG	Equals one if migrated to the area within the last 5 to 10 years.
DSEX	Equals one if male
DSCH	Equals one if currently enrolled in school
NEWEMP	Equals one if not a member of the labor force in 1974
ACHLD10	Number of children less than 10 years old in the household
MILE	Number of miles to work one-way
WKWG	Weekly wage in dollars
WKWG	WKWG squared
OTH	Non-employment income
OTHSQ	OTH squared

SURVEY DESIGN