Robust Inferences from Random Clustered Samples: Applications Using Data from the Panel Survey of Income Dynamics

John Pepper
Assistant Professor
Department of Economics
University of Virginia
114 Rouss Hall
Charlottesville, VA 22903

e-mail: jvp3m@virginia.edu

June 6, 2000

Keywords: Clustered Samples; Design Effects; PSID.

Abstract:
Many large data sets are created using clustered, rather than random sampling schemes. Clustered data arise when multiple observations exist on the same respondent, as in panel data, and when respondents share a common factor, such as a neighborhood or family. In the presence of clustered data, methods that rely on random sampling to measure the precision of an estimator may be incorrect. Many researchers, however, continue to treat respondents from the same sampling cluster as independent observations and thus implicitly ignore the potential intracluster correlation. In this paper, I use a robust method for drawing inferences and data from the Panel Survey of Income Dynamics, to examine the implications of clustered samples on inference. Consistent with the previous survey sampling literature, important differences are revealed in comparisons between the estimated asymptotic variances derived assuming random and clustered sampling, even when there are only a few observations per cluster. The estimates derived under random sampling are generally biased downward.

*I wish to thank Robert Haveman, John Karl Scholz, Steven Stern, James Walker, and the participants at the University of Wisconsin and the University of Virginia Econometrics Workshops for their helpful comments. I am especially grateful to Charles Manski for his many insightful comments, and Terry Adams for helping me understand the sampling scheme used to create the PSID. Also, I wish to thank the Wisconsin Alumni Research Foundation for financial support.

1.) Introduction

Empirical analyses often use data consisting of independent clusters of dependent random variables. In fact, most large surveys in the social sciences, including the National Longitudinal Survey of Youth, the Panel Survey of Income Dynamics, and the Current Population Survey, use some type of clustered sampling scheme. These data arise when multiple observations exist on the same respondent, as in panel data, and when respondents share a common factor, such as a neighborhood or family. In the presence of clustered data, methods that rely on random sampling to measure the precision of an estimator may be incorrect (Kish and Frankel, 1974; Scott and Holt, 1982; and Moulton, 1990).

Consider, for instance, the Panel Study of Income Dynamics (PSID), one of the most important and widely cited surveys in the social sciences. In 1968, The University of Michigan's Survey Research Center selected approximately 4,800 families to interview for the PSID. This sample of families is composed of two sub samples: The Survey Research Center's (SRC) sample of 2,930 families is representative of the households in the United States in 1968 and the Survey of Economic Opportunity (SEO) sample of 1,872 families over-represents the low-income minority population (Survey Research Center, 1984).[1] Each year since 1968, the members and offspring of these families have been surveyed. Thus, the 1968 wave of the PSID includes socioeconomic data on 18,224 individuals and the 1992 panel includes information on 41,420 individuals and 7,561 households.

Since each wave of the PSID includes multiple individuals and households that can all be connected to an original 1968 family, these data are clustered. That is, for each 1968 household, the SRC collects potentially dependent information on the associated individuals and derivative households. Furthermore, to reduce surveying costs the SRC utilized a complex geographic clustering scheme so that groups of respondents share the same 1968 block, city, or county.

Arguably, the intracluster correlation in surveys such as the PSID is not zero and traditional methods of inference made under the random sampling assumption are inappropriate. Most researchers, however, continue to treat respondents from the same sampling cluster as independent observations and thus implicitly ignore potential intracluster correlation.[2]

---

[1] To avoid complications that arise from unrepresentative stratification, this analysis only uses the SRC subsample.

[2] A random survey of articles published from 1986 through 1995 in the American Economic Review, the Quarterly Journal of Economics, the Journal of Labor Economics, and the Journal of Human Resources, reveals that nearly 80% of the analyses using the PSID treat respondents who share the same 1968 household as independent observations..

Researchers who do account for the clustering of individuals sharing the same 1968 household or geographic region often either arbitrarily exclude observations from the analysis or impose strong prior information about the intracluster dependencies. One approach has been to exclude all but a single observation from each 1968 household (see, Solon (1992)). Conventional panel data methods have also used to account for the intracluster dependencies (see, for example, Moulton, 1990). While these models can be used to formalize the intracluster relationships, they only apply to particular parametric specifications and typically require prior information about the form of the within cluster dependence. Certainly, a more general approach seems useful.

This paper examines the practical implications of clustered sampling processes for statistical inference. Focusing on the limiting distribution of method of moments estimators, Section 2 formalizes the notion of clustered sampling and describes a robust method of statistical inference. This method, which generalizes the White (1980) variance estimator, allows for both arbitrary intracluster dependence and applies to the large class of method of moments problems. To evaluate the finite sample properties of this variance estimator, a series of simple Monte Carlo simulations are examined in Section 3. Using a bivariate linear mean regression model, these simulations demonstrate that inferences drawn under the assumption of random sampling can be highly misleading if the true sampling process is clustered. In contrast, the test statistic derived using the generalized variance estimator performs relatively well in finite samples.

In Section 4, I apply this method to various parametric regressions using clustered data from the PSID. Consistent with the previous survey sampling literature, important differences are revealed in comparisons between the estimated variances derived assuming random and clustered sampling. In general, the estimates derived under random sampling are biased downward. Finally, Section 5 summarizes the main findings.

2.) Statistical Inference in Clustered Sampling

Let the population be divided into mutually exclusive and exhaustive clusters (e.g, 1968 households) with a finite number N_c of observations in each cluster c (e.g., individuals residing in the 1968 household). Assume that a random sample of C independent realizations of the random vector Z is observed, where Z_c characterizes the N_c observations within cluster c. In particular, $Z_c \equiv [\, Z_{ci},\ i = 1,..., N_c \,]$ is the N_cx 1 vector of variables for observations i within cluster c. Thus, the sample includes C

independent clusters, N_c respondents per cluster, and N observations, where N equals $\sum_{c=1}^{C} N_c$.

This clustered sampling scheme leads to a random sample of clusters with information on each individual or respondent within the cluster. Since the random variable Z_{ci} is observed for the N_c respondents within a cluster, the sample is self-weighting. That is, the probability of observing any individual or cluster in the sample is the same.

Furthermore, notice that while the clusters are statistically independent, no assumptions are imposed on the dependence between observations within a cluster. For $i \neq j$, the random variables Z_{ci} and Z_{cj} may be independent or they may be identical. Random sampling of clusters, however, implies that Z_c and Z_d are independent random variables for all $c \neq d$. In this framework, random sampling is the special case where each cluster contains a single respondent.

With unknown dependence between respondents within a cluster, methods of inference that rely on random sampling may be inappropriate. However, by treating the cluster rather than the individual as the unit of observation, the standard random sampling results are easily generalized. As the number of clusters C goes to infinity, the limiting distribution of the method of moments estimator will be normal with mean zero and finite variance. In addition, the analog estimator of the variance of this distribution, which is a generalization of White's (1980) variance estimator, will be consistent. Full proofs are available from the author.

3.) Monte Carlo Simulations

A series of Monte Carlo simulations are used to evaluate the finite sample properties of this robust estimator. After drawing 10,000 random clustered samples of size N from a bivariate linear mean regression model, I repeatedly test the hypothesis that the estimated slope coefficient equals the population parameter, β, at the 5% significance level. For each pseudo-sample, a test statistic is computed under the assumption of both random and random clustered sampling. Then, the size of the test statistic is estimated as the fraction of samples where the null hypothesis is rejected. Of course, in the limiting case where the number of clusters C approaches infinity, this hypothesis should be rejected in exactly 5% of the pseudo-samples.

Formally, assume a linear mean regression model

$$Y_{ci} = \alpha + \beta X_{ci} + \varepsilon_{ci} \qquad (1)$$

where the intercept α equals 1, the slope β equals zero, and $E(\varepsilon_{ci}|X) = 0$. Furthermore, let

$$\varepsilon_{ci} = A*V_c + B*V_{ci} \text{ and} \qquad (2a)$$
$$X_{ci} = A*W_c + B*W_{ci} \qquad (2b)$$

where V_c, V_{ci}, W_c, and W_{ci} are independent standard normal random variables. Notice that V_c and W_c vary across clusters while V_{ci} and W_{ci} are individual specific random variables. The parameters A and B characterize the intracluster correlation. In particular, for $i \neq j$,

$$\text{Corr}(\varepsilon_{ci},\varepsilon_{cj}) = \text{Corr}(X_{ci},X_{cj}) = A^2 / (A^2 + B^2). \qquad (3)$$

As A goes to infinity or B goes to zero, the intracluster correlation approaches one. In contrast, as A goes to zero or B goes to infinity, the individual random variables become increasingly important and the intracluster correlation approaches zero. If A equals zero, the sampling process is random.

Using this model, simulations were run on a total of 36 random clustered samples which varied by the number of observations per cluster, the number of clusters, and the intracluster correlation. Table 1 displays the results for clusters of size 5, 20, and 50, for samples of size 500, 1000, and 5,000, and for intracluster correlation between 0.10 and 1.00. In each sample, the number of observations per cluster is constant. Under the assumptions of random and random clustered sampling, the table records the faction of type I errors in the 10,000 Monte Carlo experiments. The program used to perform the simulations was written in Gauss for Windows, Version 3.2.22.

These simulation results clearly demonstrate that in finite random clustered samples, inferences drawn under the traditional random sampling assumption can be highly misleading. Under random sampling, the empirical size of the t-test statistic substantially exceeds the nominal 5% level, especially as the cluster size N_c and the intracluster correlation increase. In fact, when the number of observations per cluster and the intracluster correlation are small, the probability of a type I error nearly matches the 0.05 benchmark. In 26 of the 36 experiments, however, the size of the test under random sampling exceeds 0.10, and in 15 cases the size exceeds 0.35. In the worst cases, when the intracluster correlation equals one and the cluster size equals 50, the true null is rejected in over 80% of the simulations.

Under the assumption of random clustered sampling, the simulation results are markedly different.

Again, the probability of a type I error tends to deviate from the 5% nominal level as the sample size decreases and as the number of observations per cluster and the intracluster correlation increase. However, in general the probability of a type I error lies between 0.05 and 0.10, exceeding 0.10 in only 6 of the 36 cases. In the worst case, the size of the test equals 0.142, substantially less than analogous 0.817 found under the random sampling assumption.

4.) Applications

To empirically investigate the effects that clustered sampling can have on inferences, I use clustered data from the PSID to estimate two econometric models. The first is a mean wage regression model similar to the one estimated by Hill (1981); the second is a probit model to examine the relationship between birth weight and various background factors including whether the mother smoked cigarettes. While the effects of clustering on inferences for both the unconditional mean and the linear mean regression have been previously examined (see, for example, Scott and Holt, 1992 and Hill 1981) little is known about the practical effects of clustering in nonlinear regressions. These models are estimated using data from the SRC representative sample of the PSID. Individuals and households in the SRC subsample are self-weighting.

Using these data I apply the methods of inference described above. In particular, the asymptotic standard errors of MOM estimators that apply in both random and clustered sampling are computed. The standard errors derived under the assumption of random sampling are robust to arbitrary heteroskedasticity (White, 1980), while those computed under the assumption of clustered sampling are also robust to arbitrary intracluster dependence. Finally, the ratio of the two estimated variances are reported for each parameter estimate. This ratio of the clustered sampling asymptotic variance to the random sampling asymptotic variance, termed the design effect in the survey sampling literature, measures the relative change in the variance caused by clustered sampling.

Wage Regression Model

Every year since 1968, the SRC records detailed information regarding the job market experiences of heads and spouses. Using these data, numerous studies investigate the relationships between wages and socioeconomic background characteristics. To examine the effects of clustered sampling on a simple wage regression model, I use a sample of 4,059 respondents from the 1992 wave of PSID (Morgan, et al., 1992). Included are heads and spouses who were at least 25 years old in 1992 and who worked over 500

hours in 1991. These data are clustered. In particular, the 4,059 respondents are associated with 1,253 original PSID households. Thus, on average, this sample contains 3.2 observations per 1968 household.

A standard linear wage regression model assumes that $E[Y_c|X_c] = X_c\beta$ for all c =1,..., C, where Y_c is the N_cx1 random vector of log wages for individuals in cluster c, X_c is the N_cxK vector of observed covariates, and β is a Kx1 vector of unknown coefficients. This moment condition implies that in random clustered sampling, the least squares estimator $b_c = (X'X)^{-1}X'Y$ is a method of moments estimator of β, where Y is the Nx1 observed vector of log wages and X is the NxK vector of observed covariates. Letting the number of clusters C go to infinity, the limiting distribution of this estimator is

$$C^{1/2}(b_c - \beta) \text{ -->}_d \text{ } N[0 , \text{ } E^{-1}[X_c'X_c] * E[X_c'\varepsilon_c \varepsilon_c'X_c] * E^{-1}[X_c'X_c] \text{ }], \qquad (4)$$

where ε_c is the N_c x 1 vector of prediction errors$(Y_c - X_c\beta)$. Thus,

$$V_c = \sum_{c=1}^{C}(X_c'X_c)^{-1} \sum_{c=1}^{C}(X_c'e_c)(e_c'X_c) \sum_{c=1}^{C}(X_c'X_c)^{-1}$$

$$= (XX)^{-1} \sum_{c=1}^{C}\{ (X_c'e_c)(e_c'X_c) \} (X'X)^{-1} \qquad (5)$$

where e_c is the N_cx1 matrix of residuals $(Y_c - X_cb_c)$ is a consistent estimators of the asymptotic variance of b_c in clustered sampling,

The estimated coefficients for this linear mean regression model are presented in Table 1. Like much of the past literature, these estimates suggest that expected log wage is highest for union men living in large cities and lowest for black women living in the south. Table 1 also displays the estimated asymptotic standard errors, which in certain instances are substantively altered by the clustered sampling design. By failing to account for the intracluster correlation, the estimates derived under random sampling are generally understated. In certain cases, the differences are negligible. The sampling scheme has little effect on the estimated asymptotic errors associated with the age, race and gender coefficients. In contrast, the estimated design effects for the coefficients on union status, years of schooling, and whether the respondent lived in the south are all greater than 1.15, implying that the estimated standard errors in clustered sampling exceed those in random sampling by at least 7% (i.e., the square root of 1.15). At 1.30, the design effect for the "south" coefficient implies that the random sampling standard errors are understated almost 15%.

Birth Weight Regression Model

In 1985, the SRC began recording information on the birth weight of individuals born to a PSID head or spouse. Low birth weight is an important health concern in the United States, accounting for nearly 10% of the medical expenses of all children and costing more than $5.4 billion per year (Future of Children, 1995). Here, using a sample of 3,416 observations from the 1992 PSID, I examine the association between low birth weight and various characteristics of the respondent's mother. Included are respondents in the 1992 PSID whose parents were PSID heads or spouse for at least one year from 1985 to 1992, and for whom there exists information on birth weight and the mother's cigarette consumption. These data are clustered, with the 3,416 respondents linked to 1,038 original PSID households. Thus, on average this sample contains 3.3 observations per 1968 household.

To predict the probability of being a low birth weight baby conditional on the various observed characteristics of the mother, I use a standard probit model. In particular, assume that

$$Y_c = \begin{cases} 1 & \text{if } X_c\beta + \varepsilon_c > 0 \\ 0 & \text{otherwise} \end{cases} \qquad (6)$$

for all c = 1,...,C. Here Y_c is a N_cx 1 vector indicating whether individuals in cluster c were low birth weight babies, X_c is a N_cxK vector of the observed background characteristics, and ε_c is the N_c x 1 vector reflecting the unobserved determinants of birth weight.

Let the marginal distribution of ε_c given X_c be normal with mean zero and unit variance. This "probit" model implies that the mean regression is $E[Y_c| X_c] = F[X_c\beta]$, where $F(\cdot)$ is the standard normal cumulative distribution function. Thus, the parameter β satisfies the moment condition

$$E\left[\left\{\frac{f[X_c\beta]}{(1-F[X_c\beta])*F[X_c\beta]}\right\}'\eta_c\right] = E[V(X_c)'\eta_c] = 0 \qquad (7)$$

where $\eta_c = (Y_c - F[X_c\beta])$ is the N_cx1 vector of prediction errors and $f(\cdot)$ is the standard normal

probability distribution function.[3] For this model, the standard method of moments estimator of β will be consistent and have an asymptotically normal distribution with a finite variance.

Table 2 displays the coefficient estimates along with the associated standard errors. The estimates suggest that the probability of being a low birth weight baby decreases if the mother is married and if she graduated from high school and increases if she smoked cigarettes prior to birth. The results also confirm previous findings that blacks are much more likely to be low birth weight than whites, ceteris paribus.

The estimated design effects displayed in Table 2 reflect the strong intracluster correlation in birth weights among families. With only an average of three individuals from each 1968 family, the clustered sampling scheme still has important effects on inferences in this nonlinear regression model. Again, all of the design effects are larger than one except for the measures associated with whether the child was the first born, which is negatively correlated within a family. Otherwise, the design effects exceed 1.15 and for the coefficients associated with the race variable, the design effect is 1.7. Thus, the asymptotic standard errors derived under the random sampling assumption are understated by as much as 30% (i.e., the square root of 1.70).

To isolate the effect of clustered sampling, both sets of standard errors are estimated using the analogs to the general limiting distribution for MOM estimators. Alternative estimators of the standard error derived under random sampling imply larger design effects. For instance, the design effect for the coefficient associated with the race indicator variable increases from 1.7 to 2.1 if the random sampling standard errors are computed using the negative of the inverse of the Hessian matrix, and to 2.5 if instead the asymptotic standard errors under random sampling are computed using the inverse of the information matrix.

Geographic Clusters in the PSID

To reduce interviewing costs, the SRC relied on a geographic clustering scheme which first selected counties, then areas within each county, and eventually households along particular blocks (see Kish and Hess (1965) for additional details). By design, the sample is self-weighing at the individual and

[3] With dependence between observations sharing the same cluster, the standard likelihood equation for the probit model will not apply. However, the moment conditional in Equation (7) is equivalent to the moment condition satisfied by maximizing the log-likelihood of the standard probit model under iid sampling (see Avery, Hansen, and Hotz (1983)). Thus, this MOM estimator can be interpreted as a quasi-maximum likelihood estimator, and can be implemented using standard MLE routines. Of course, the variance estimator cannot rely on standard MLE results, but instead must account for the clustered sampling.

household levels and thus the standard MOM estimators apply. However, research that ignores the potential correlation between respondents sharing the same geographic cluster may draw distorted inferences. To evaluate the effects of geographic clustering on inferences from the PSID, I reevaluate the design effects for the two models examined above. Two geographic clusters can be identified in the data. The first are the primary sampling units (PSU) that include the 80 counties selected in the first stage of the sampling process, and the second are the detailed place codes that identify 329 distinct neighborhoods of households selected in an intermediate stage.

Table 3 displays the estimated design effects for the estimated regression coefficients using the household, the detailed place code, and the PSU as the unit of observation. In general, the results show that the estimated asymptotic standard errors increase with the number of observations per cluster and the intracluster correlation. Inferences drawn regarding outcomes that are likely to exhibit strong intra-regional correlation, such as wages, appear to be sensitive to geographic clustering. The design effect for the coefficient on whether the respondent lived in the south, for example, increases from 1.31 when the household is the unit of analysis, to 3.20 when accounting for the clustering of respondents within counties. In contrast, inferences drawn about outcomes that are arguably independent of geographic locale, such as birth weight, appear unaffected. The design effect associated with the coefficient associated with race, for instance, only moves from 1.67 when accounting for household clustering, to 1.77 when accounting for PSU's. Still, the design effects substantially increase for the coefficients associated with the marital status and education of the mother.

5.) Conclusion

In this paper, I provide several empirical illustrations of a method of inference that is robust to the clustering of individuals who share the same sampling cluster (e.g., 1968 household). While the specific results cannot be generalized to other models, they do provide evidence that in clustered samples inferences made under the random sampling assumption can be misleading. Simulations in Section 3 demonstrate that in finite samples, the empirical size of a simple t-test made under the assumption of random sampling are often substantially different than the nominal size. Applications presented in Section 4, reveal design effects that exceed 2.00 in samples that contain only a few observations per cluster and design effects of over 3.00 in samples with nearly 50 observations per PSU. In general, the estimated standard errors derived under the random sampling assumption are biased downward. These findings should not come as a surprise: the underlying sampling process affects statistical inferences. Certainly,

researchers using clustered samples should be wary of drawing inferences under the assumption that the sampling process is random. However, for self-weighting clustered samples like the PSID, the general procedures outlined above can be applied.

References

Avery, R.B., P.L. Hansen, and V.J. Hotz. (1983). "Multiperiod Probit Models and Orthogonality Condition Estimation," *International Economic Review*, 24 (1), 21-35.

Hill, M. (1981). "Some Illustrative Design Effects: Proper Sampling errors Versus Simple Random Sample Assumptions." In Five Thousand American Families -- Patterns of Economic Progress - Vol. IX, edited by Martha S. Hill, Daniel H. Hill and James M. Morgan. Ann Arbor: Institute for Social Research, the University of Michigan, 1981.

Kish, L. and M.R. Frankel. (1974). "Inferences from Complex Samples," *Journal of the Royal Statistical Association*. Ser. B, 36, pp. 1-37.

Kish, L. and I. Hess. (1965). The Survey Research Center's National Sample of Dwellings. Institute for Social Research, No. 2315, The University of Michigan, Ann Arbor, MI.

Morgan, J.N., Duncan, G.J., Hill, M.S., & Lepkowski, J. (1992). Panel Study of Income Dynamics, 1968-1989, [waves I-XXII], [Computer File]. Ann Arbor: University of Michigan, Survey Research Center. Ann Arbor: Inter-University Consortium for Political and Social Research (distributor).

Moulton, B.R. (1990). "An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables on Micro Units", *The Review of Economics and Statistics*, 72, 334-8.

Scott, A.J. and D. Holt. (1982). "The Effect of Two-Stage Sampling on Ordinary Least Squares Methods". *Journal of the American Statistical Association*. 77. pp.848-54.

Solon, G. (1992). "Intergenerational Income Mobility in the United States," *American Economic Review*, 82(3), 393-408.

Survey Research Center. (1984). Panel Study of Income Dynamics: User Guide. Ann Arbor, Insititute for Social Research

White, H. (1980). "A Heteroscedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroscedasticity", *Econometrica*, 48(4), 817-38.

Table 1
Monte Carlo Simulations for a t-test
Statistic from a Bivariate Linear Mean Regression

Fraction of Rejections of the True Null Hypothesis that
$\beta = 0$ at the 5% Significance Level

		Nc = 5		Nc = 20		Nc = 50	
N	Intracluster Correlation	CS	RS	CS	RS	CS	RS
500	0.10	0.058	0.059	0.071	0.067	0.101	0.105
	0.20	0.061	0.075	0.070	0.172	0.111	0.211
	0.50	0.061	0.156	0.094	0.433	0.136	0.537
	1.00	0.064	0.394	0.105	0.690	0.142	0.817
1000	0.10	0.050	0.055	0.064	0.068	0.073	0.119
	0.20	0.052	0.070	0.064	0.140	0.077	0.193
	0.50	0.058	0.163	0.074	0.451	0.094	0.559
	1.00	0.052	0.383	0.071	0.668	0.101	0.800
5000	0.10	0.052	0.058	0.051	0.074	0.050	0.100
	0.20	0.054	0.070	0.055	0.157	0.059	0.272
	0.50	0.055	0.174	0.055	0.428	0.060	0.588
	1.00	0.047	0.374	0.054	0.665	0.063	0.787

Note: CS = Clustered Sampling; RS = Random Sampling; N=Number of Observations;
Nc = Number of Observations per Cluster

Table 2
Linear Mean Wage Regression Coefficients

Estimates and the Associated Asymptotic Standard Errors
Assuming Random and Random Clustered Sampling

Y = Natural Log of Wages

	Coefficient	Estimated Standard Error		Design
Variable	Estimate	in RS	in CS	Effect
Constant	-0.132	0.119	0.127	1.144
Years of School	0.107	0.004	0.004	1.204
Age in Decades	0.543	0.053	0.055	1.080
Age in Decades Squared	-0.054	0.006	0.006	1.077
Union Status (1 = in Union)	0.260	0.020	0.022	1.175
Whether in South	-0.029	0.018	0.021	1.311
Whether in Large City	0.158	0.025	0.026	1.125
Female	-0.303	0.016	0.016	0.979
Black	-0.161	0.031	0.032	1.035

Note: RS = Random Sampling; CS = Clustered Sampling
The sample size N = 4,059 and the cluster size C = 1,253.

Table 3
Probit Model Regression Coefficients

Estimates and the Associated Asymptotic Standard Errors
Assuming Random and Random Clustered Sampling

Y = 1 if Low Birth Weight

	Coefficient	Estimated Standard Error		Design
Variable	Estimate	In RS	in CS	Effect
Constant	-1.66	0.17	0.19	1.22
Marital Status of Mother at Birth (1=Married)	-0.04	0.15	0.16	1.22
First Birth	-0.01	0.08	0.08	0.92
Mother Graduated From High School	-0.10	0.09	0.10	1.16
Mother Smoked Cigarettes Before Birth	0.25	0.08	0.09	1.32
Black	0.39	0.12	0.16	1.67

Note: RS = Random Sampling; CS = Clustered Sampling
The sample size N = 3,416 and the cluster size C = 1,038.

Table 4

Estimated Design Effects Under Various Definitions for Clusters

Linear Mean Regression: Y = ln(wage)				Probit Model: Y = 1 if Low Birth Weight, 0 Otherwise			
	Family	Place	PSU		Family	Place	PSU
Constant	1.14	1.27	1.28	Constant	1.22	1.39	1.60
Years of School	1.20	1.35	1.89	Marital Status of Mother at Birth (1=Married)	1.22	1.41	1.56
Age in Decades	1.08	1.19	1.31	First Birth (1 = first birth, 0 otherwise)	0.92	0.77	0.54
Age in Decades Squared	1.08	1.21	1.30	Mother Graduated From High School	1.16	1.30	1.44
Union Status (1 = in Union)	1.18	1.22	1.34	Mother Smoked Cigarettes Before Birth	1.32	1.22	1.24
Whether in South	1.31	1.82	3.20	Black	1.67	1.58	1.77
Whether in Large City	1.13	1.50	2.16				
Female	0.98	0.92	0.95				
Black	1.04	1.20	1.02				
Average Number of Observations per Cluster	3.24	13.44	50.74		3.29	11.86	42.70
Number of Clusters	1253	302	80		1038	288	80

Note: PSU stands for Primary Sampling Unit, which reflects the original county or SMSA cluster selected by the SRC. Place is the detailed place clusters used to create the SRC subsample. These clusters reflect a block or street, so that two households with the same place code share the same neighborhood.