Variance Estimation in US Census Data from 1960-2010

Kathryn M. Coursolle

Lara L. Cleveland

Steven Ruggles

Minnesota Population Center

University of Minnesota-Twin Cities

September, 2012

This paper was prepared for submission to the annual meeting of the Population Association of America. Research for this project was supported by the National Institutes of Health (grant number R01HD043392). Opinions expressed herein are those of the authors. Please do not cite or quote without permission of the authors. Address correspondence to Kathryn M. Coursolle, Minnesota Population Center (MPC), University of Minnesota, 50 Willey Hall, 225 19th Avenue South, Minneapolis, MN 55455, e-mail: kathrync@umn.edu.

Abstract

Modern census microdata feature complex sample designs that clustered within households and incorporate stratification. Yet, researchers often calculate standard errors utilizing methods designed for simple random samples. Variance estimates can differ dramatically adjusting for complex survey design clustering and stratification relative to estimates assuming simple random sampling. Examining potential differences in variance estimation in recent IPUMS-USA samples is essential because US census microdata are among the most heavily used data sources for social, historical, demographic, and policy research. This project uses decennial census data from 1960-2000 and American Community Survey data from 2000-2010 to compare standard errors under the assumptions of simple random sampling to estimates which adjust for clustering and stratification, and subsample replicate weights for recent ACS data. We conclude by discussing potential implications of these techniques on statistical inference.

Background

The Integrated Public Use Microdata Series (IPUMS-USA) consists of more than fifty high-precision samples of the US population drawn from decennial censuses from 1850-2000 and the American Community Surveys from 2000-2010. These samples represent the richest source of US microdata and have been heavily used in demographic scholarly research. For example, census microdata was used in more articles of *Demography* than any other data source in recent decades. Census microdata are gathered using complex sampling designs that are clustered by households, incorporate stratification, and sometimes have differential probability of selection. However, most researchers apply methods of variance estimation designed for simple random samples. Failure to adjust for clustering and stratification in the sample design may lead to incorrect standard errors and invalid statistical inferences (Davern & Strief; Kish, 1995; Lohr, 2000).

The impact of sample design on standard errors has been documented on historical census data from 1850-1950 (Davern, Ruggles, Swenson, Alexander, & Oakes, 2009). However, differences in standard errors after adjusting for clustering and strata has not been tested in modern census data from 1960-2010 and sampling techniques in modern census data differ substantially from historical census samples. Using decennial census data from 1960-2000 and American Community Survey (ACS) data from 2000-2010 we evaluate the impact of sample design on standard error estimates. We compare standard error estimates under the assumption of simple random sampling to variance estimates accounting for clustering and strata using Taylor series linearization. In the ACS 2005-2010 samples we also compare standard error estimates to the Census Bureau's subsample replicate weights. We conclude by discussing strategies for

estimating standard errors in modern census microdata and potential directions for future revisions of this research.

Background

The sample designs of modern census microdata are individual-level data clustered by households that incorporate stratification. For variables which tend to be similar within households, like race and birthplace, adjusting for clustering may produce standard errors that are larger than variance estimates assuming a random sample of the same size (Cleveland, Davern, & Ruggles, 2011; Graubard & Korn, 2002). In the worst case scenario, standard errors would be inversely proportional to the square root of the number of households rather than individuals if the characteristics of the people in the household are identical. However, variance estimates of variables that tend to be heterogeneous within households such as age and sex may actually be smaller than estimates under simple random sampling. Stratification, on the other hand, tends to have the opposite effect of clustering. Standard errors can be smaller than simple random sampling adjusting for stratification when the characteristics of individuals or clusters are homogenous within strata.

Stratification in IPUMS-USA samples from 1960-2010

For the 1960-2000 decennial IPUMS-USA samples, strata were based on the criteria the Census Bureau used to select PUMS samples. For 1960, 38 strata were defined on the basis of various characteristics of household size, home ownership, race, and group quarters residence. The procedures used to select cases for inclusion in the 1970 public use samples were similar to those used in 1960, but were slightly more elaborate. Seventy-five strata were created based on

home ownership, race, sex of head, household size, presence of own children, inmate status, and other residence in group quarters. For the 1980 samples, strata were created based on race, Spanish origin, home ownership, and presence of own children, producing 51 strata. For the 1990 and 2000 samples, strata were created based on presence of own children, race, Spanish origin, and home ownership. In addition, to avoid singletons Asian race categories were collapsed into one category and this criteria was also used for the 2000 samples (White/Other Race/Two or More Races Hispanic; White/Other Race/Two or More Races Non-Hispanic; Asian and Pacific Islander; Black and American Indian). For the 1990 samples age was not used for non-institutional group quarters to avoid singletons. Any remaining singletons were collapsed into the White Non-Hispanic Origin strata. These methods produced 119 strata for the 1990 samples and 131 strata for the 2000 samples. For the American Community Survey samples, strata are based on the lowest level of geography available in the sample. For the 2000-2004 samples, each state forms a stratum. In the 2005 onward ACS samples, strata are defined as unique Public Use Micro-data Areas (PUMA). For more detailed information see: http://usa.ipums.org/usa/complex_survey_vars/strata_historical.shtml.

Subsample Replicate Weights in the ACS

Replicate weights were added to the ACS starting in 2005. These weights are produced by the Census Bureau and allow the sample to mimic multiple samples, which can produce more informed standard error estimates and reflect relevant sample design information. Standard errors produced using replicate weighting techniques are usually larger, and produce more conservative statistical inferences, than those under the assumption of simple random sampling (Davern & Strief). The Census Bureau recommends using replicate weights to obtain unbiased standard

error estimates (US Census Bureau, 2005). However, using these procedures is often cumbersome and takes substantially more computing time relative to Taylor series estimates. It is worthwhile to know whether standard errors produced adjusting for clustering and strata are similar to those obtained utilizing the ACS replicate weights.

Results

Table 1 presents the comparison of standard errors using several methods of selected variables in census data from 1960-2000 and ACS data from 2000, 2004, 2005, and 2010.[1] The first column shows the population parameter estimate from the IPUMS sample and the second column presents the standard error estimates based on the assumption of that the survey design was based on a simple random sample.[2] This estimate uses the person weight only. The third and fourth columns display the ratio of the standard error using Taylor series and replicate weight methods relative to the standard errors assuming simple random sampling methods. A ratio above one indicates that the standard error is larger than variance from a simple random sample of the same size, and a ratio below one indicates that the standard error would be smaller than a simple random sample.

Turning first to the results for the decennial census 1960-2000 samples, we can see that for aspects of individuals that tend to be homogeneous within households such as foreign-born, socioeconomic index, and race often produce larger standard errors than techniques which assume simple random sampling survey techniques for several of the sample years. This suggests

[1] Results were very similar for other ACS samples. To present simplified results, only these samples are included.

[2] In some census years the person weight also adjusts for aspects of probability sampling, such as 1990 and 2000. See: http://usa.ipums.org/usa/chapter2/chapter2.shtml for more information.

that for research examining those characteristics using standard errors calculated under the assumption of simple random sampling may produce less conservative criteria for statistical significance. However, the opposite is the case for other characteristics such as age, gender, marital status, school enrollment, and labor force participation, which are characteristics more likely to be heterogeneous within households. Generally the variance of these parameters tends to be smaller after adjusting for clustering and stratification. Indeed prior research suggests that standard errors produced that adjust for clustering and stratification may be smaller than the simple random sample standard error estimates when the effects of stratification are more pronounced (Davern & Strief, but see also Kish, 1995).

We next present the results of the comparison of variance estimates for the American Community Survey Samples. Although these samples also have clustering by households similar to the decennial sample design, pseudo-strata are calculated by the lowest level of geography available in each survey year. For these samples all of the standard error using Taylor series methods are larger than standard errors than would be obtained from a simple random sample of the same size, with the exception of gender. In the 2005 and 2010 ACS the table presents ratios of the standard error calculated using the subsample replicate weights. For marital status, foreign-born, and socioeconomic index the variance estimates were larger utilizing the subsample replicate weights than under the assumption of simple random sample, and the opposite was true for the other measures. Differences between the ratios of the Taylor series and replicate weight methods were fairly modest, with the exception of age, but computing burden was substantially less with the Taylor series techniques. In future revisions, we plan to analyze differences between the Taylor series and replicate weight methods in greater detail.

Discussion

This paper documents the comparison of standard error calculations under the assumption of simple random sampling, clustering and stratification, and utilizing ACS replicate weights in the IPUMS 1960-2010 samples. For the decennial samples, Taylor series standard error estimates were often smaller than standard errors obtained from a simple random sample of the same size, except for variables that tend to be highly corrected within households which are not included in the design of strata, such as foreign-born. On the other hand estimates obtained from pseudo geography-based strata in the ACS samples led to generally larger standard errors than under the assumption of simple random sampling. For these samples, utilizing Taylor series methods would lead to more conservative criterion for statistical inference. However, it is important to remember that for most IPUMS data, the samples are quite large, and there is little risk of drawing incorrect conclusions due to underestimated standard errors. However, for analysis that examines only small subpopulations, the risk could be higher. Providing examples of when this may be the case seems like a logical next step for this research. Future revisions of this project will also compare in more depth differences in standard errors computed using the ACS replicate weights to the Taylor series estimates. In addition, it may be useful to create subsample replicates for the decennial census samples to compare Taylor series variance results to a "gold standard."

References

Cleveland, L. L., Davern, M., & Ruggles, S. 2011. "Drawing Statistical Inferences from International Census Data." IPUMS-International Working Paper: https://international.ipums.org/international/resources/misc_docs/cleveland_davern_ruggles_variance.pdf

Davern, M., Ruggles, S., Swenson, R., Alexander, J. T., & Oakes, J. M. 2009. "Drawing Statistical Inferences from Historical Census Data, 1850-1950." *Demography*. 46: 429-449.

Davern, M. & Strief, J. "IPUMS User Note: Issues Concerning the Calculation of Standard Errors (i.e., variance estimation) Using IPUMS Data Products" Ipums.org: http://usa.ipums.org/usa/resources/complex_survey_vars/UserNote_Variance.pdf

Graubard, B., & Korn, E. 2002. "Inferences for superpopulation parameters using sample surveys." *Statistical Science* 17: 73-96.

Kish, L. 1995. *Survey Sampling*. Wiley Classics Library Edition. New York, NY: Wiley and Sons.

Lohr, S. 2000. *Sampling: Design and Analysis*. Pacific Grove, CA: Duxbury Press.

US Census Bureau. 2005. "PUMS Accuracy of the Data (2005)" Washington, DC: US Census Bureau.

Table 1. Standard Errors Assuming Simple Random Samples Compared with Taylor Series and Subsample-Replicate Estimates: Selected Person Characteristics

Selected Person Characteristics	Sample Mean or %	Standard Error Assuming Simple Random Sampling	Ratio of Standard Error Estimate to Simple Random Sample	
			Taylor Series Adjusting for Clustering and Strata	Subsample Replicate Method (2005-2010 ACS)
1960				
Age (mean)	31.03	0.0165	0.93	
Male (%)	49.12	0.0004	0.75	
Married (%)	46.38	0.0004	0.71	
Nonwhite (%)	11.47	0.0002	0.33	
Foreign-born (%)	6.02	0.0002	1.38	
Socioeconomic Index (mean)	16.88	0.0174	0.95	
Enrolled in School (%)	24.43	0.0003	0.94	
Labor force participant (%)	38.87	0.0004	0.86	
1970				
Age (mean)	31.83	0.0111	0.84	
Male (%)	48.57	0.0002	0.71	
Married (%)	45.04	0.0002	0.49	
Nonwhite (%)	12.33	0.0002	0.79	
Foreign-born (%)	5.72	0.0001	1.44	
Socioeconomic Index (mean)	19.93	0.0124	0.95	
Enrolled in School (%)	29.43	0.0003	0.89	
Labor force participant (%)	40.64	0.0002	0.89	
1980				
Age (mean)	33.44	0.0066	0.88	
Male (%)	48.47	0.0001	0.75	
Married (%)	44.71	0.0001	0.83	
Nonwhite (%)	14.46	0.0001	1.06	
Foreign-born (%)	7.15	0.0001	1.44	
Socioeconomic Index (mean)	22.04	0.0078	1.04	
Enrolled in School (%)	27.70	0.0001	0.97	
Labor force participant (%)	46.61	0.0001	0.99	

(Continued on next page)

Table 1 (Continued)

1990			
Age (mean)	34.83	0.0069	0.83
Male (%)	48.73	0.0002	0.77
Married (%)	43.60	0.0002	0.91
Nonwhite (%)	19.63	0.0001	1.28
Foreign-born (%)	9.23	0.0001	1.45
Socioeconomic Index (mean)	24.86	0.0086	1.00
Enrolled in School (%)	26.12	0.0001	1.02
Labor force participant (%)	50.35	0.0002	0.92
2000			
Age (mean)	35.81	0.0066	0.87
Male (%)	48.99	0.0002	0.78
Married (%)	42.72	0.0001	0.93
Nonwhite (%)	24.90	0.0001	1.32
Foreign-born (%)	12.31	0.0001	1.39
Socioeconomic Index (mean)	25.48	0.0083	0.99
Enrolled in School (%)	27.23	0.0001	0.98
Labor force participant (%)	49.33	0.0002	0.94
2000 American Community Survey			
Age (mean)	35.62	0.0422	1.38
Male (%)	48.82	0.0010	0.80
Married (%)	42.19	0.0010	1.14
Nonwhite (%)	25.52	0.0009	1.87
Foreign-born (%)	12.33	0.0007	1.60
Socioeconomic Index (mean)	26.39	0.0538	1.15
Enrolled in School (%)	26.63	0.0009	1.13
Labor force participant (%)	50.54	0.0010	1.06

(Continued on next page)

Table 1 (Continued)

2004 American Community Survey				
Age (mean)	36.22	0.0243	1.38	
Male (%)	48.94	0.0006	0.80	
Married (%)	42.36	0.0006	1.15	
Nonwhite (%)	24.38	0.0005	1.81	
Foreign-born (%)	13.23	0.0004	1.50	
Socioeconomic Index (mean)	26.30	0.0312	1.14	
Enrolled in School (%)	26.42	0.0005	1.12	
Labor force participant (%)	50.92	0.0006	1.06	
2005 American Community Survey				
Age (mean)	36.38	0.0164	1.37	0.85
Male (%)	49.02	0.0004	0.81	0.50
Married (%)	42.16	0.0004	1.14	1.59
Nonwhite (%)	25.36	0.0004	1.74	1.06
Foreign-born (%)	13.64	0.0003	1.48	1.15
Socioeconomic Index (mean)	26.36	0.0211	1.11	1.37
Enrolled in School (%)	26.04	0.0003	1.12	0.88
Labor force participant (%)	51.07	0.0004	1.05	0.92
2010 American Community Survey				
Age (mean)	37.35	0.0152	1.36	0.48
Male (%)	49.15	0.0003	0.83	0.23
Married (%)	39.09	0.0003	1.14	1.58
Nonwhite (%)	25.83	0.0003	1.72	0.97
Foreign-born (%)	14.27	0.0003	1.44	1.10
Socioeconomic Index (mean)	25.32	0.0191	1.10	1.53
Enrolled in School (%)	26.74	0.0003	1.13	0.61
Labor force participant (%)	50.76	0.0003	1.05	0.86

Source: 1960, 1970, 1980, 1990, 2000, 2005, 2010 IPUMS samples.