APPROXIMATION OF RELATIVE STANDARD ERRORS OF MULTI-YEAR ESTIMATORS IN THE NATIONAL HEALTH INTERVIEW SURVEY

Joe Fred Gonzalez, Jr., Carrie Jones, Chris Moriarity, and Van Parsons, NCHS
Joe Fred Gonzalez, Jr., NCHS, 6525 Belcrest Road, Room 915, Hyattsville, MD 20782

KEY WORDS: Minority Subpopulations, Regression, Year-to-Year Correlation

1. Introduction

The National Health Interview Survey (NHIS) is one of the major national population-based surveys conducted by the National Center for Health Statistics (NCHS) to monitor the health of the U.S. civilian noninstitutionalized population. The NHIS collects information throughout the year via household interviews. Each year approximately 40,000 households, containing about 100,000 persons, are sampled for the NHIS. In the current (1995-2004) NHIS sample design, Non-Hispanic Blacks and Hispanics are oversampled to increase the precision of health-related estimates for these subpopulations. The NHIS has been a continuous data collection program since 1957, and the sample is redesigned after each decennial census.

Research is underway for the next NHIS sample design cycle scheduled for 2005-2014. A major focus of the research is on improving the precision, as measured by the relative standard error (RSE), of prevalence estimates for several minority subpopulations. One new subpopulation being considered for more precise annual estimates is Non-Hispanic Asians. Additionally, more precise estimates are desired for selected Hispanic and Non-Hispanic Asian subgroups (e.g., Mexican-American, Chinese-American) after pooling two or three years of data. The starting point for this research was an assessment of the precision of annual as well as multi-year estimates for various subpopulations from the current NHIS design. At the time that the research was carried out, only two years of NHIS data were available. Prevalence estimates and their RSEs were computed directly using one and two years of data. An approximation formula, which included an estimate of the year-to-year correlation in the NHIS, was used to predict the RSEs for two-year and three-year prevalence estimates. The correlation results from revisiting the same primary sampling units (PSUs) every year within a 10 year cycle. To examine the validity of the approximation formula, direct RSEs of two-year estimates were compared with predicted RSEs of two-year estimates.

The main purpose of this paper is to discuss the performance of the model for producing estimates of RSEs of multi-year estimates for Hispanic and Non-Hispanic Asian subpopulations when actual data are unavailable. In addition, results are presented for several other subpopulation groups. Section 2 presents the derivation of the model. Section 3 compares the predicted RSEs of two-year estimates using the assumed multi-year correlation with direct RSE estimates based on two years of NHIS data for a selected group of health-related characteristics.

The approximation formula for RSEs of multi-year estimates can be interpreted as a simple linear model, where the RSE of a multi-year estimate is the dependent variable, and the RSE of a one year estimate is the independent variable. Section 4 discusses the results of simple linear regressions for several subpopulations using a full linear model. Section 5 discusses the results of simple linear regressions for several subpopulations using a no intercept linear model, followed by a summary in Section 6.

2. Derivation of the Model for the RSE of Prevalence Estimators For Multiple Years of Data

Let RSE(n) equal the RSE of a prevalence estimator based on n years of data and RSE(1) represent the RSE of a prevalence estimator based on one year of data. The following formula was used to approximate the RSE of a prevalence estimator based on n years of data:

$$RSE(n) = RSE(1)\bullet\sqrt{\frac{1 + (n - 1)D}{n}}$$

where ρ represents the correlation between two years of data. Previous analysis of NHIS data has suggested that ρ = 0.2 is a reasonable (conservative) estimate of the correlation for all race/ethnic groups between two consecutive years of data in the NHIS. Of course, this assumed value of ρ is likely to decrease within the 10 year NHIS design cycle as time increases. For example, we would expect ρ to decrease over time because of changes in the distribution of race/ethnic groups. The 1997 NHIS data were used to produce the 1 year estimates and their RSEs.

The derivation of the formula to approximate the RSE of prevalence estimators for multiple years of NHIS data follows:

Let $\hat{X}_i$ represent the estimator based on data from year *i* of the NHIS, and let $\hat{X}_1, \ldots, \hat{X}_n$ represent estimators based on n years of data. Assume that for each year *i*, $\hat{X}_i$ is distributed with mean μ and variance σ^2. For a 10 year cycle of the NHIS sample design, the PSUs are fixed; thus, there is a correlation between annual

estimators which will be denoted as $D_{ij} = corr(\hat{X}_i , \hat{X}_j)$. For simplicity, assume $D_{ij} = D$ (constant) for all *i, j*. (As mentioned above, this assumption is overly simplistic if i and j are far apart.) Further assume that for a specific health characteristic, the annual estimate values and sample sizes are approximately the same for all *i, j*.

A multi-year estimator can be viewed as a linear combination of the $\hat{X}_i's$ with variance as follows:

$$Var(\sum_{i=1}^{n} a_i \hat{X}_i) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j cov(\hat{X}_i, \hat{X}_j)$$

(Equation 1)

Since a multi-year estimator usually would be computed as the simple average of annual estimators, this implies that

$$a_1 = a_2 = \ldots = a_n = \frac{1}{n} .$$

Also reexpressing the right-side of Equation 1 for *i = j* and *i ≠ j*,

$$Var\left[\frac{\sum_{i=1}^{n} \hat{X}_i}{n}\right] =$$

$$\frac{1}{n^2}\left[\sum_{i=1}^{n} Var(\hat{X}_i) + \sum\sum_{i \neq j} Cov(\hat{X}_i, \hat{X}_j)\right]$$

(Equation 2)

Recall that, given the assumptions above,

$$D = \frac{cov(\hat{X}_i , \hat{X}_j)}{F_{\hat{X}_i} F_{\hat{X}_j}} .$$

(Equation 3)

Since the variances for annual estimators are assumed to be the same (σ^2) from year to year, Equation 3 becomes

$$D = \frac{cov(\hat{X}_i , \hat{X}_j)}{F^2} ,$$

$$or\ DF^2 = cov(\hat{X}_i , \hat{X}_j) .$$

(Equation 4)

Since the number of covariance terms in Equation 2 is the number of permutations of n years taken two at a time, the number of terms is *P(n,2) = n (n - 1)*. Also, using the assumption that the annual variances are the same, Equation 2 becomes:

$$Var\left(\frac{\sum_{i=1}^{n} \hat{X}_i}{n}\right) =$$

$$\frac{1}{n^2}\left[nF^2 + n(n-1)F^2D\right]$$

(Equation 5)

Equation 5 may be reexpressed as

$$Var\left(\frac{\sum_{i=1}^{n} \hat{X}_i}{n}\right) =$$

$$\frac{F^2}{n}\left[1 + (n-1)D\right] =$$

$$F^2\left(\frac{1 + (n-1)D}{n}\right) .$$

(Equation 6)

Thus,

$$RSE\left(\frac{\sum_{i=1}^{n}\hat{X}_i}{n}\right) = \sqrt{\frac{F^2}{\mu^2}\ \frac{1 + (n-1)D}{n}}\ .$$

(Equation 7)

In other words, the RSE of a multi-year NHIS estimator can be expressed as:

$$RSE(n) = RSE(1)\bullet\sqrt{\frac{1 + (n-1)D}{n}}$$

(Equation 8)

where n represents the number of data years under consideration and ρ represents the correlation between two years of data.

3. Comparison of Direct and Predicted Two-Year RSE Estimates

Direct estimates of RSEs were computed for fourteen health-related characteristics using 1997 and 1998 NHIS data, and SUDAAN [1] (Taylor series linearization approach) was used to approximate variances for the prevalence estimators ($\hat{p}$). Predicted two-year RSE estimates were produced using Equation 8 with ρ = 0.2, n = 2, and the 1997 NHIS data were used to produce the one year RSE estimates. Table 1 shows direct two-year RSEs and predicted two-year RSEs of the prevalence estimates for a subset of three of the fourteen health characteristics for Hispanic and Non-Hispanic Asian subpopulations as well as certain other subpopulations. The three characteristics were chosen to display a range of results for high, medium, and low prevalences. The characteristics chosen were: "Have you ever smoked (variable name: ever smoke)?" (medium prevalence); "Are you overweight, i.e., body mass index ≥ 25 (variable name: overweight)?" (high prevalence); and "Have you ever been told you had asthma (variable name: asthma)?" (low prevalence). Table 1 shows direct two-year RSEs and predicted two-year RSE estimates. Predicted two-year RSEs are smaller than direct two-year RSEs for Chinese-Americans reporting they have ever smoked or had been told they had asthma. However, predicted two-year RSEs are larger than two-year direct RSEs for Filipino-Americans reporting that they had ever smoked and for Cuban-Americans who had been told they had asthma. Not shown in Table 1 are certain variables, such as questions about "ever been told had diabetes," "trouble hearing," "ever been told had heart disease," where the predicted two-year RSEs are consistently larger than the direct two-year RSE estimates. Therefore, at least for predicting two-year RSEs, while assuming ρ = 0.2, the approximation formula usually predicted slightly larger two-year RSEs except for Chinese-Americans, where the predicted RSE estimates are less than the direct two-year RSEs. While extreme caution must be used in interpreting direct estimates of RSE based on small sample sizes, it appears that, in general, the use of ρ = 0.2 in the approximation formula produces conservative predictions of two-year RSE estimates.

4. Multi-Year RSE Approximation Formula Viewed as a Simple Linear Full Model

The multi-year RSE approximation Equation 8 can be interpreted as a linear model as follows:

$$y = \beta_0 + \beta_1 x + \varepsilon,$$

where y=dependent variable=RSE(2-year estimate); x=independent variable=RSE(1-year estimate); β_0=y-intercept; ε=error term; and, the slope,

$$\beta_1 = \sqrt{\frac{1+D}{2}}.$$

SAS PROC REG [2] was used to perform the regression analyses. After the linear regression is performed, the regression estimate of β_1 can be set equal to

$$\sqrt{\frac{1+D}{2}}$$

and an estimate of the year-to-year correlation, ρ, can be solved for. The estimate of the standard error of β_1 can be used to get an assessment of the variability in the estimate of ρ by forming a confidence interval about β_1, and then transforming this to a confidence interval about the estimate of ρ.

The same fourteen health-related characteristics referred to in Section 3 were included in the analysis as well as the characteristics by age (<18 years, 18-44 years, 45-64 years, and 65+ years), sex, and sex-age. Therefore, the analysis included a maximum number of 210 (14 characteristics + 2 sexes x 14 characteristics + 4 ages x 14 characteristics + 2 sexes x 4 ages x 14 characteristics) data points for each race/ethnic group. Since a research goal for the 2005-2014 redesign of the NHIS is to produce reliable prevalence estimates that are not too rare (e.g., p ≥ 5%), and with an

RSE ≤ 30%, data points with prevalence levels, p < 5%, or RSEs > 30% were excluded from the regression analyses. Also, data points with missing prevalence estimates were deleted from the analyses by default. A second regression analysis was conducted but prevalence levels of p < 10%, instead of p < 5%, were excluded from the analyses. For each prevalence level criterion and race/ethnicity shown, Table 2 displays the number of estimates (n) that met the prevalence level and RSE criteria and had non-missing prevalence estimates; the R-squared (R^2); and the estimated year-to-year correlation (ρ). Although a goal of the NHIS redesign research is to produce multi-year prevalence estimates with acceptable precision levels for Hispanic subgroups and Non-Hispanic Asian subgroups, the results of linear regressions for several other major race/ethnicity groups are shown here for comparison purposes only.

As shown in Table 2, for prevalence levels p ≥ 5% for: Total, Non-Hispanic Whites, Non-Hispanic Blacks, Hispanics, Mexican-Americans, Puerto Ricans, Other Hispanics, and Non-Hispanic Asians, the assumed linear model fits the data very well (R^2 ≥ .90). A similar pattern is shown in Table 2 for the same race/ethnicity groups when p ≥ 10%. However, for p ≥ 5%, the linear model did not fit as well for Chinese-Americans and Filipino-Americans where R^2 was equal to .7756 and .6346, respectively. A similar pattern is shown in Table 2 for Chinese-Americans and Filipino-Americans when p ≥ 10%.

As shown in Table 2, all of the estimated intercepts (b_0) are between -0.02 and 1.11, for both prevalence levels, except for Filipino-Americans, where the intercept (=3.64) for p ≥ 5% and the intercept (=3.40) for p ≥ 10%. As for Chinese-Americans, the sample size for Filipino-Americans was small (n=35) for p ≥ 5%.

Table 2 shows that the estimated slopes (b_1) for p ≥ 5% and p ≥ 10% ranged from 0.56 to 0.84 and 0.58 to 0.82, respectively.

Also, it should be noted that the residual plots (residuals versus 1997 RSEs), which are not shown in this paper, displayed a nonconstant variance to some degree for the x-values (1997 RSEs) for all groups.

Table 2 also gives the estimates of ρ for each race/ethnicity shown for p ≥ 5%. For instance, ρ is -0.10, 0.17, -0.06, -0.03 for Hispanics, Mexican-Americans, Puerto Ricans, and Cuban-Americans, respectively. The year-to-year correlation ρ is -0.08, 0.41, and -0.36 for Non-Hispanic Asians, Chinese-Americans, and Filipino-Americans, respectively. Except for Chinese-Americans (ρ = 0.41), Equation 8 produces conservative estimates of two-year RSE estimates with ρ = 0.2. Of course, the sample size for Chinese-Americans is rather small (n = 39) for p ≥ 5%; a 95% confidence interval of the $\$_1$ estimate for Chinese-Americans transforms to a confidence interval for ρ of [-.04, .96], demonstrating that point estimates of ρ based on small sample sizes have sizable variability associated with them. Results from the regression analysis for prevalence levels, p ≥ 10%, provide similar results.

5. Multi-Year RSE Approximation Formula Viewed as a Simple Linear Model With No Intercept

Strictly speaking, Equation 8, the multi-year RSE approximation formula discussed in section 2, should be fitted using a no intercept model. This was done by using the "NOINT" option in the Model statement in SAS PROC REG. The estimated slopes from corresponding no intercept models usually were similar to those estimated using the full model discussed in section 4. As expected, the no intercept model, containing one less model parameter, did not fit the data as well as the model containing an intercept term, as shown by increases in the error sum of squares for corresponding no intercept models. Thus, as expected, synthetic "R^2" values, computed using the corrected sums of squares from the corresponding full models, were lower for no intercept models than the corresponding full models.

6. Conclusion

The approximation formula presented in this paper is a method to consider when one needs to approximate multi-year RSEs for prevalence estimates when multiple years of data are not available. Based on the R-squared criterion, the assumed linear model fit the data well for certain population groups, but not equally well for all groups. Residual plots indicated nonconstant variance, a violation of an underlying assumption of the linear model that is important if hypothesis testing procedures were to be applied. On the other hand, the results show that, in general, the approximation formula produces conservative estimates of two-year RSE estimates with ρ = 0.2.

Table 1. Comparison of Direct and Predicted Two-Year RSE Estimates Using Approximation Formula, 1997 and 1998 NHIS.

Health Characteristic	Race/Ethnicity	Direct 2-year p(%)	Direct 2-year RSE (%)	Direct 1-year p(%)	Predicted 2-year RSE(%)
ever smoke	Hispanic	35.2	1.7	35.3	1.7
	Mexican-American	33.6	2.4	34.4	2.3
	Puerto Rican	43.9	4.0	42.2	4.3
	Cuban-American	37.3	6.6	33.2	6.9
	Non-Hispanic Asian	28.0	5.0	30.0	4.8
	Chinese-American	24.6	12.1	24.2	9.9
	Filipino-American	30.4	11.8	28.4	13.4
overweight*	Hispanic	60.5	1.0	59.3	1.0
	Mexican-American	63.9	1.3	62.4	1.3
	Puerto Rican	60.8	2.9	61.6	3.3
	Cuban-American	52.0	4.8	48.6	5.3
	Non-Hispanic Asian	29.8	4.4	29.1	5.0
	Chinese-American	23.2	11.1	21.9	11.3
	Filipino-American	43.0	8.1	45.1	8.9
asthma	Hispanic	8.1	3.2	7.8	3.4
	Mexican-American	6.4	4.9	5.9	5.2
	Puerto Rican	16.3	6.2	14.8	6.3
	Cuban-American	9.3	13.5	7.1	17.7
	Non-Hispanic Asian	6.2	8.7	7.0	8.7
	Chinese-American	6.2	17.7	7.9	15.9
	Filipino-American	9.8	15.4	11.0	15.0

* ”overweight” is defined as having a body mass index ≥ 25.

Note: Differences in direct and predicted RSEs could be due, at least in part, to differences in the one-year and two-year prevalence rates.

Table 2. Linear Regression Results For Prevalence Estimates With RSE ≤ 30% Assuming Full Simple Linear Model, 1997 and 1998 NHIS.

	$p \geq 5\%$					$p \geq 10\%$				
Race/Ethnicity	**n**	**b_0**	**b_1**	**R^2**	**ρ**	**n**	**b_0**	**b_1**	**R^2**	**ρ**
Total	154	.09	.70	.9927	-.02	110	.08	.70	.9899	-.01
Non-Hispanic White	153	.07	.71	.9911	.02	105	.07	.71	.9904	.00
Non-Hispanic Black	151	.15	.72	.9895	.02	119	.14	.71	.9866	.02
Hispanic	149	.32	.67	.9842	-.10	100	.36	.65	.9860	-.16
Mexican-American	139	-.02	.76	.9771	.17	100	.06	.75	.9764	.12
Puerto Rican	124	.45	.68	.9236	-.06	105	.50	.68	.9229	-.09
Cuban-American	86	.75	.70	.8627	-.03	70	1.11	.67	.8661	-.10
Other Hispanic	133	.41	.66	.9603	-.13	88	.33	.65	.9592	-.15
Non-Hispanic Asian	92	.80	.68	.8988	-.08	66	.93	.66	.9057	-.13
Chinese-American	39	.64	.84	.7756	.41	32	.93	.82	.7378	.34
Filipino-American	35	3.64	.56	.6346	-.36	29	3.40	.58	.6270	-.33
Other Asian	79	.11	.73	.8475	.06	62	-.08	.75	.8411	.12
Non-Hispanic Other	112	.63	.74	.8473	.10	98	.67	.75	.8475	.11

References

1. Shah, B., Barnwell, B., Bieler, G., *SUDAAN Users's Manual, Release 7.0*, Research Triangle Park, NC: Research Triangle Institute (1996).

2. SAS Institute Inc., SAS/STAT User's Guide, Version 8, Cary, NC: SAS Institute Inc (1999).