Reliability of Relative Standard Errors Computed from NHDS Public Use Data Files

Bill Cai[1], Iris Shimizu[1]
[1]National Center for Health Statistic, 3311 Toledo Road, Hyattsville, MD 20782

Abstract

The National Hospital Discharge Survey (NHDS) is conducted to produce nationally representative estimates of hospital discharges. Micro data files containing data collected in the NHDS are released to the public, but sampling design variables required for computing the NHDS variances are omitted from these files to maintain confidentiality of respondents' identities. To enable data users to compute variances for estimates derived from the public use data files, generalized variance functions (GVF) are supplied in the public use data file documentation. This paper discusses the quality of the GVF results for aggregate estimates. It compares relative standard errors (RSEs) (aka coefficients of variation or CVs) derived by using GVF from the public data use files with RSEs computed by applying SUDAAN software to in-house data files.

Key Words: RSE, NHDS, SUDAAN, GVF, public use data file

1. Introduction

The National Hospital Discharge Survey (NHDS) conducted by the National Center for Health Statistics (NCHS) annually, collects medical and demographic information from a sample of inpatient discharge records from a national probability sample of non-Federal, general and short-stay hospitals. NHDS is a principal source of information on inpatient hospital utilization in the United States.

With advice and support from the American Hospital Association, the American Medical Association, individual experts, other professional groups, and officials of the U.S. Public Health Service, the NCHS initiated the National Hospital Discharge Survey in 1964. [1]

Prior to 1988, the NHDS used a two stage sample of discharges. In 1988 and thereafter, a stratified and modified three stage sample has been used. The sampling strata are defined, by region, hospital specialty and bed size groups. The first stage units consist of either hospitals or geographic areas that are counties or groups of counties (or townships in New England). Within sampled areas, hospitals are selected at the second stage. Discharges are selected within sampled hospitals.

NHDS public use data files are released to the public every year with sampled records and analysis weight used to obtain weighted estimates. Confidentiality concerns prevent the NHDS public use data files from including the design variable (for strata and clusters) needed to correctly compute estimates of variance for NHDS estimates. Generalized variance functions (GVFs) which can be used to predict estimated variances, are routinely supplied in the public use data file documentation.

Section 2 introduces the relative standard error (RSE) of the estimates, GVFs, and GVF derived variance estimates. Section 3 describes the procedures to compare NHDS in-house RSEs with GVF derived RSEs. Section 4 describes the observations from data set and plots. Section 5 summarizes the findings from 2006 NHDS data and discusses directions for further research about GVF derived variance estimates for NHDS.

2. Relative Standard Error and Generalized Variance Functions

The standard error (SE) or variance of the estimate is a measure of the accuracy of predictions. The RSE of the estimate (X) is obtained by dividing the standard error by the estimate itself [1]

$$\text{RSE (X)} = \text{SE (X)} / \text{X} \qquad (1)$$

so, that

$$\text{SE (X)} = \text{X} * \text{RSE (X)} \qquad (2)$$

Estimates with large RSEs are considered less reliable than estimates with small RSEs. NCHS recommends that estimates with RSEs above 30 percent should be considered unreliable. [2]

Confidentiality concerns prevents the NHDS public use data files from including the complex sample design information needed to correctly compute estimates of variance from NHDS. However GVFs supplied with public use data files can be used to approximate RSEs. Standard errors can be approximated from formula (2).

Using the GVFs for aggregate estimates:

$$\text{RSE (X)} = \text{SQRT (a + b/X)} \qquad (3)$$

with values for a and b provided in NHDS public use data documentation. [1]

To derive error estimates that would be applicable to a wide variety of statistics, numerous estimates and their variance were produced. A regression model then used these data to produce best-fit curves, based on an empirically determined relationship between the size of an estimate X and its relative variance (standard error).

The standard error estimates in National Health Statistics Reports are computed using the first-order Taylor series approximation of the deviation of estimates from their expected values as applied in the SUDAAN software package. However, the public can only use GVFs supplied in public use data documentation for standard error calculation when using NHDS public use data.

How reliable are the standard errors derived from GVFs when compared with those computed from in-house data using SUDAAN? It is a known fact that variances estimated by GVFs, in general, tend to overstate variances for large estimates and understate variances for small estimates. We give quantitative analysis results of in-house variances (in the form of RSEs) versus GVF derived variances for 2006 NHDS aggregate estimates.

3. Analysis Procedures

NHDS in-house data files and NHDS public data use files for 2006 were used for our analysis. In-house RSEs for 731 aggregate estimates of number of discharges typically published in NCHS National Health Statistics Reports based on NHDS data were computed using SUDAAN. [3] The GVF derived RSEs were calculated for the same 731 estimates by using coefficients a and b provided in Table 1 in the 2006 NHDS public use file documentation. [1] Two examples of aggregate estimates are: 1.) total number of discharges for black females from the South; and 2.) total number of discharges for children under age 15 with a first-listed diagnosis of asthma (ICD-9-CM code 493).

In Figure 1, in-house RSEs and GVF derived RSEs are ploted against size of estimates (natural logarithm) for those 731 visit aggregate estimates.

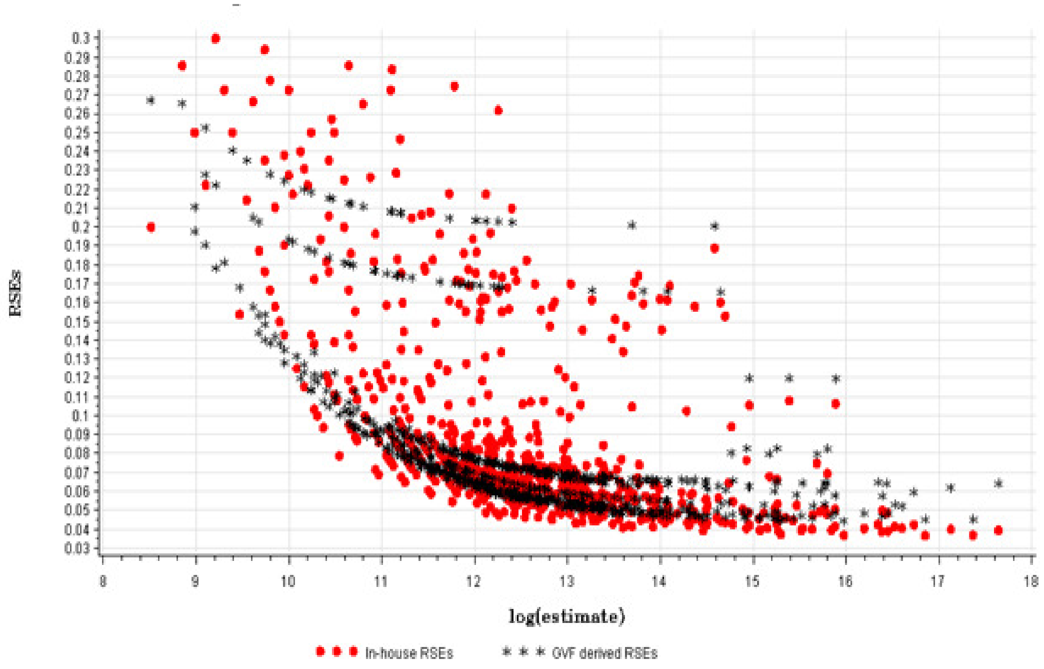


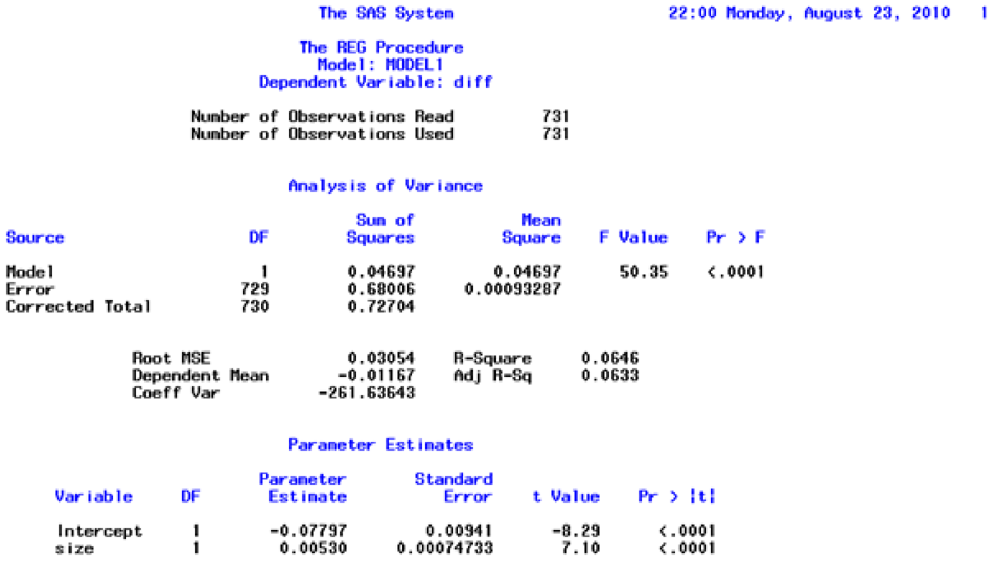
Figure 1: In-house RSEs and GVF derived RSEs, 2006 NHDS

The difference (referred to hereafter as Diff) between GVF derived RSEs and in-house RSEs tells whether the GVF derived RSEs overstate (Diff > 0) or understate (Diff < 0) the in-house RSEs. A linear regression model of the form

$$(\text{Diff})^{\wedge} = A + B{*}\log(X) \qquad (4)$$

was fitted to the data points [Diff, log(X)].

Figure 2 shows the SAS output for the linear regression model in equation (4) while Figure 3 shows a plot of the points [Diff, log(X)], the fitted regression line, and the intercept point. It can be seen that $\widehat{(\text{Diff})}=0$ at log(X) =14.71 which corresponds to aggregate estimate X=2,446,087 (referred to hereafter as the *intercept point*).



```
                              The SAS System                    22:00 Monday, August 23, 2010   1

                             The REG Procedure
                               Model: MODEL1
                         Dependent Variable: diff

                    Number of Observations Read          731
                    Number of Observations Used          731

                            Analysis of Variance

                                    Sum of           Mean
Source                   DF        Squares         Square    F Value    Pr > F

Model                     1        0.04697        0.04697      50.35    <.0001
Error                   729        0.68006     0.00093287
Corrected Total         730        0.72704

              Root MSE              0.03054    R-Square     0.0646
              Dependent Mean       -0.01167    Adj R-Sq     0.0633
              Coeff Var          -261.63643

                            Parameter Estimates

                         Parameter       Standard
     Variable     DF      Estimate          Error    t Value    Pr > |t|

     Intercept     1      -0.07797        0.00941      -8.29      <.0001
     size          1       0.00530     0.00074733       7.10      <.0001
```

Figure 2: SAS Output when Applying Regression Model

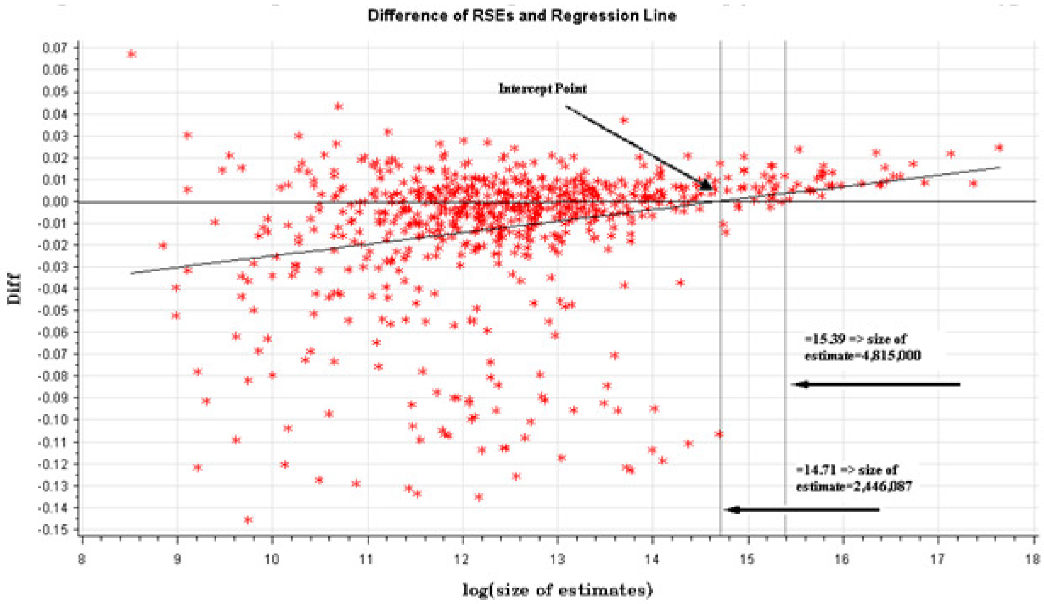


Figure 3: Linear Regression Model [Diff = A + B*log(size of estimates)]

Two sign tests were applied in SAS to the Diff values; one for points with X > *intercept point* and one for points with X < *intercept point*. According to those sign tests, the GVF derived variance estimates over (under) estimated in-house variance estimates more often than not when the estimates were greater (less) than the *intercept point*.

4. Some Observations

When aggregate estimates are more than 4,815,000 in size, GVF derived RSEs overstated in-house derived RSEs. All 33 points have Diff more than 0.

When aggregate estimates are more than the *intercept point*, GVF derived RSEs overstated about 8 times more often than they understated in-house RSEs (55 versus 7 times).

When aggregate estimates are less than the *intercept point*, GVF derived RSEs understated in-house RSEs a little less than 2 times more often than they overstated in-house RSEs (412 versus 257 times).

When aggregate estimates are less than 131,000 in size, GVF derived RSEs understated in-house RSEs slightly more than 2 times more often than they overstated in-house RSEs (155 versus 71). (The point 131,000 was investigated to show patterns when aggregate estimates are substantially below the *intercept point*.)

The minimum value of Diff observed in this study was -0.145534139; the maximum value of Diff was 0.0673338736. That is, when RSEs are expressed as percents of the corresponding estimates, the GVF derived RSE understated the in-house RSE for some published estimate by up to 14.5 percentage points (almost half of the 30 percent maximum RSE allowed to satisfy NCHS standards for reliability of the estimate). In addition to the need for conservative methods when testing for significance, the magnitude of understatement implies a need for caution in determining which estimates are reliable.

5. Conclusion and Summary

The general conclusion when using GVF derived RSEs is that variances estimated by GVFs tend to overstate variances for estimates with large size and understate variances for estimates with small size.

When aggregate estimates are greater than 2,446,087, RSEs derived from GVFs are highly likely (8:1) to overstate in-house RSEs.

When the estimates are less than 2,446,087, GVF derived RSEs are more likely (2:1) to understate in-house RSEs.

In the current study, which included aggregate estimates typically published from the National Hospital Discharge Survey, the magnitude of differences between their GVF derived RSEs and their in-house RSEs ranged from -0.14 to 0.067. While larger magnitude differences may exist, it is believed the differences observed in this study are probably typical for most aggregate estimates based on NHDS data.

The analysis presented here is only for 2006 NHDS and only includes aggregate estimates for a single year. Though the results are meaningful and might be put into good use, future research may consider the quality of GVF derived variances for multi-year NHDS estimates. Because ratio estimates based on NHDS data are also important, this research could also be extended to examine the quality of GVF derived variance for NHDS ratio statistics.

Acknowledgements

We thank Esther Hing for her paper [4], from which we got some inspiration; we thank Aleksandr Golosinskiy for his work in [3], from which we got in-house standard errors for estimates.

References

[1] National Center for Health Statistics (NCHS). 2008. Public Use Data File Documentation: 2006 National Hospital Discharge Survey. Hyattsville, Md.:NCHS.

[2] National Center for Health Statistics (NCHS). 2008. Public Use Data File Documentation: 2006 National Hospital Ambulatory Medical Care Survey. Hyattsville, Md.:NCHS.

[3] National Center for Health Statistics (NCHS). 2008. National Health Statistics Reports (NHSR): 2006 National Hospital Discharge Survey. Hyattsville, Md.:NCHS.

[4] Hing E, Gousen S, Shimizu I, Burt C. 2003/2004 Guide to Using Masked Design Variables to Estimate Standard Errors in Public Use Files of the National Ambulatory Medical Care Survey and the National Hospital Ambulatory Medical Care Survey. Inquiry. 40(4):416-415.