Developing Generalized Variance Functions for Estimates of Recidivism Rates

Marcus E. Berzofsky[1], David Heller[1], Rick Williams[1]
[1]RTI International, 3040 Cornwallis Rd., Research Triangle Park, NC 27709

Abstract
This paper discusses the methods used to develop a generalized variance function (GVF) to properly compute standard errors (SEs) for recidivism rates for a cohort of inmates released from prison in 2005. GVFs use model-based parameters in a relatively simple function to calculate the SEs for an estimate while properly taking into account the complex survey design without requiring the use of statistical software. Thus, GVFs provide users a computationally easy method to calculate SEs for estimates from a survey with a complex design. The Bureau of Justice Statistics (BJS) has developed an online tool for estimating recidivism rates over time for a cohort of inmates released in 2005. The cohort of inmates was selected using a stratified design. Due to computational constraints the online tool could not use a statistical software package to dynamically compute requested SEs. Therefore, BJS wanted to incorporate a GVF into the tool to produce SEs along with the recidivism estimates. We discuss the models considered and the challenges associated with developing three parameter GVFs for each estimate type for this population.

Key Words: Generalized Variance Function (GVF), complex survey design, design effect, recidivism rates, online data tools

1. Introduction

Estimating sampling variances from complex survey data can be done through one of two broad methods: (1) direct estimation, or (2) generalized variance functions (GVFs). Direct estimation use point variance estimators or conduct replication procedures to obtain variance estimates individually for survey statistics (Salvucci, Weng, & Kaufman, 1995). GVFs attempt to model the variance of a survey estimator as a function of the estimate and possibly other variables (Wolter, 1985). As such the GVF model will depend on the type of estimate – total, mean, or proportion.

Survey statisticians develop GVFs for three main reasons: (1) statistical software that could account for a complex survey design is not readily available, (2) computational time required to process complex survey data is too onerous in some circumstances, and (3) allow analysts that may not be statistically savvy to account for the complex survey design through a simple formula. Given these reasons, unlike direct variance estimates, a single GVF will apply to an entire class of estimates for a given survey (e.g., all proportions) (Salvucci, Weng, Kaufman, 1995). Furthermore, if the modeling is successful then direct variance techniques for each individual estimate is unnecessary potentially resulting in considerable cost savings (Johnson & King, 1987). Moreover, for published reports, GVFs reduce the volume of what needs to be included because standard errors can be calculated directly from the published point estimates (Bieler & Williams, 1990).

2. Understanding the problem

2.1 Background

The Bureau of Justice Statistics (BJS) periodically does a retrospective study of a cohort of released inmates from prison to determine the rate of recidivism in the United States. The latest iteration of the study reviewed a cohort released from prison in 2005. The study determined the recidivism rates over a 60 month period. The results from this study can be found in *Recidivism of Prisoners Released in 30 States in 2005: Patterns from 2005 to 2010* (Durose, Cooper & Snyder, 2014).

The 2005 cohort came from 30 states. The states included in the study can be seen in ***Figure 1***. Rather than tracking all inmates released in 2005, BJS drew a stratified simple random sample of inmates. Inmates were stratified by state, gender, and controlling offense. Inmates released for more violent crimes (e.g., inmates released for homicide) were selected with a higher probability of selection (i.e., the analysis weights across inmates are not equal).

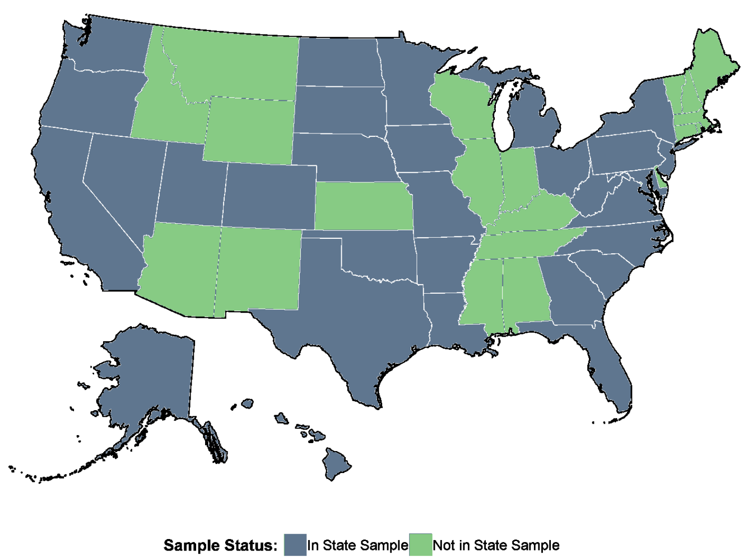
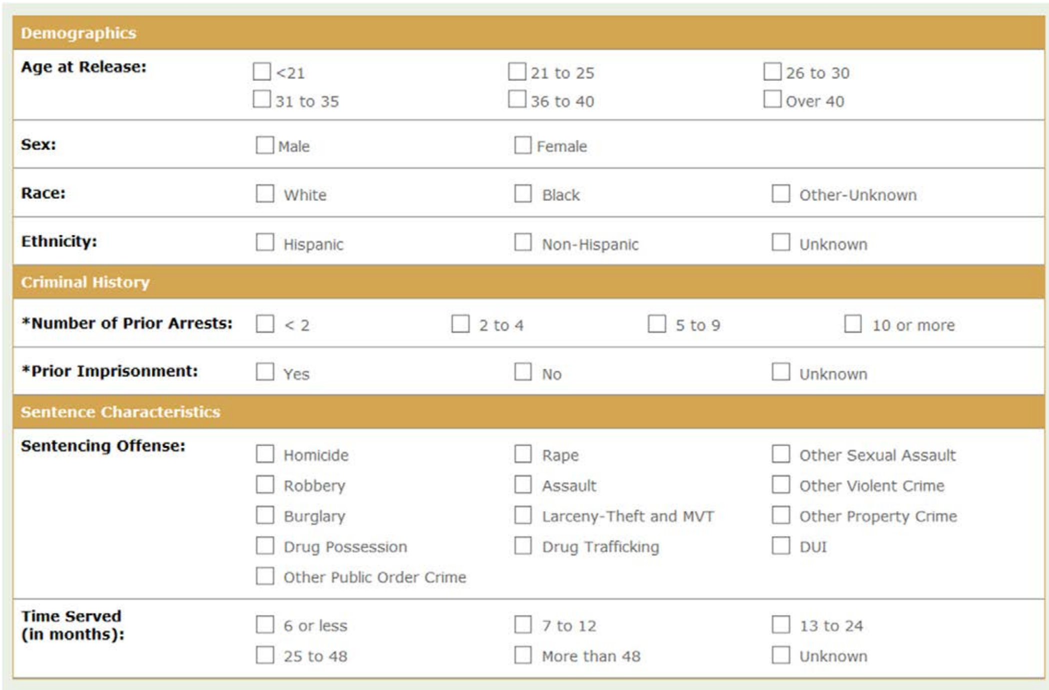


Figure 1: States included in the 2005 Recidivism Cohort Study

The outcome of interest in the study was the time until the released inmate recidivated. During the 60 month period, BJS recorded five different points within the criminal justice process a person in the cohort would experience if they committed another crime – (1) arrested, (2) adjudicated (3) convicted (4) incarcerated, and (5) imprisoned,.

2.2 Need for a GVF

In addition to the published report, BJS wanted to develop an online data analysis tool to allow users the ability to obtain recidivism rates and corresponding standard errors for all subsets of the data possible. Given the number of estimates possible (over one million) it was not possible to preload the estimates and corresponding standard errors into tables the tool would utilize as needed (see ***Figure 2*** for listing of all characteristic types from which a user can select). Therefore, the online tool houses the survey data and dynamically calculates the requested recidivism rate.



Demographics				
Age at Release:	<21	21 to 25	26 to 30	
	31 to 35	36 to 40	Over 40	
Sex:	Male	Female		
Race:	White	Black	Other-Unknown	
Ethnicity:	Hispanic	Non-Hispanic	Unknown	
Criminal History				
***Number of Prior Arrests:**	< 2	2 to 4	5 to 9	10 or more
***Prior Imprisonment:**	Yes	No	Unknown	
Sentence Characteristics				
Sentencing Offense:	Homicide	Rape	Other Sexual Assault	
	Robbery	Assault	Other Violent Crime	
	Burglary	Larceny-Theft and MVT	Other Property Crime	
	Drug Possession	Drug Trafficking	DUI	
	Other Public Order Crime			
Time Served (in months):	6 or less	7 to 12	13 to 24	
	25 to 48	More than 48	Unknown	

Figure 2: Demographic, Criminal History, and Sentence Characteristics Available on 2005 Recidivism Cohort Online Tool

Ideally, the tool would dynamically calculate the corresponding standard errors to mimic the methods used in Durose, Cooper & Snyder (2014). While computing rates is relatively simple and does not require special software, calculating the standard error directly requires running a software package that can account for the complex survey design (e.g., SUDAAN, Stata) in the background of the tool. After investigating this possibility, it was determined that the online tool did not have sufficient power to run one of these software packages.

Therefore, BJS determined that a GVF was the best option to provide standard errors that properly account for the complex survey design. The GVF and accompanying parameters can be written into the online tool without causing any computational problems.

2.3 Recidivism Outcomes

The 2005 Recidivism Cohort online tool provides two general classes of outcomes: (1) cumulative recidivism rates, and (2) failure rates. The cumulative recidivism rates are the proportion of released prisoners who recidivated by the end of each month during the 5-year follow-up period. In other words, for month m, the cumulative recidivism rate (CR) is defined as

$$CR_{jm} = \frac{R_{jm}}{N_j}$$

where R_{jm} is the number of released prisoners in characteristic set j who recidivated by month m and N_j is the total number of released prisoners in characteristic set j. The failure rate is the number of released prisoners who recidivated in year Y divided by the number of released prisoners who had not recidivated by year Y-1. In other words, for year Y, the failure rate (FR) is defined as

$$FR_{jY} = \frac{R_{jY}}{N_j - \sum_{t=1}^{Y-1} R_{jt}}$$

where R_{jY} is the number of released prisoners in characteristic set j who recidivated in year Y and N_j is the number of released prisoners in characteristic set j.

2.3 Study goals

Given the need for a GVF, BJS had the following goals

1. Determine the most appropriate GVF model for each of the outcome types
2. Determine the simplest set of covariates for the model that provide accurate standard errors and compared to the corresponding direct estimate.

3. Methods

The methods to achieve our study goals consisted of three components: (1) determining a set of potential models for the GVF, (2) estimating the variances directly and determining the model parameters for each GVF model, (3) assessing the fit of the resulting GVF standard errors for each potential GVF model.

3.1 Potential GVF Models

In order to determine potential models to use for the GVF we first determined that both outcomes of interest (i.e., the cumulative recidivism rate and failure rate) are proportions. Given this, we reviewed GVF models used for similar types of outcomes; namely, we looked at the models used in the National Crime Victimization Survey (NCVS) (see, Ash et. al.,, 2008 for review of the NCVS GVF models). In addition we considered more traditional models used for proportions. Based on this review we identified three potential models.

Two-parameter Census Model (Tomlin, 1974): $v(\hat{p}) = b \times \frac{\hat{p}(1-\hat{p})}{\hat{N}}$

Three-parameter Census Model (Chand, 1993): $v(\hat{p}) = b \times \frac{\hat{p}(1-\hat{p})}{\hat{N}} + c \times \frac{\hat{p}(\sqrt{\hat{p}}-\hat{p})}{\sqrt{\hat{N}}}$

Log-transform of proportion (Bieler & Williams, 1990):
$ln[v(\hat{p})] = a + b \times ln[\hat{p}(1-\hat{p})] + c \times ln(\hat{N})$

3.2 Estimation Variances Directly

To estimate the standard errors, for each outcome type, all possible estimates and their resulting standard errors were computed (i.e., all characteristic sets by time periods) using SUDAAN. Standard errors for cumulative recidivism rates were estimated using PROC DESCRIPT. Standard errors for failure rates were estimated using PROC KAPMEIER (while there was no censoring in the recidivism data, the outcome was treated as a time to event outcome for ease of programming).

Once the directly estimated standard errors were computed, the direct standard errors were regressed against each GVF formula to obtain model parameters.

3.3 Assessing Model Fit

In order to assess each of the models, fit statistics were produced based on all possible estimate domains. This included

- 734 non-gender specific domains;
- 1,458 male specific domains; and
- 1,458 female specific domains.

Plots comparing the log variances computed directly and through the GVF models were produced. For these plots, estimates that fall along the 45-degree line have the same variance under both approaches (direct and GVF). Estimates above the 45-degree line indicate that the direct variance is larger. Estimates below the 45-degree line indicate the GVF model produced larger variances.

4. Results

4.1 Initial Results

The initial results found that the log-transform of proportions model fit the data the best. However, as ***Figure 3*** shows, the fit of even this model shows an upward curve as the log variance gets larger.

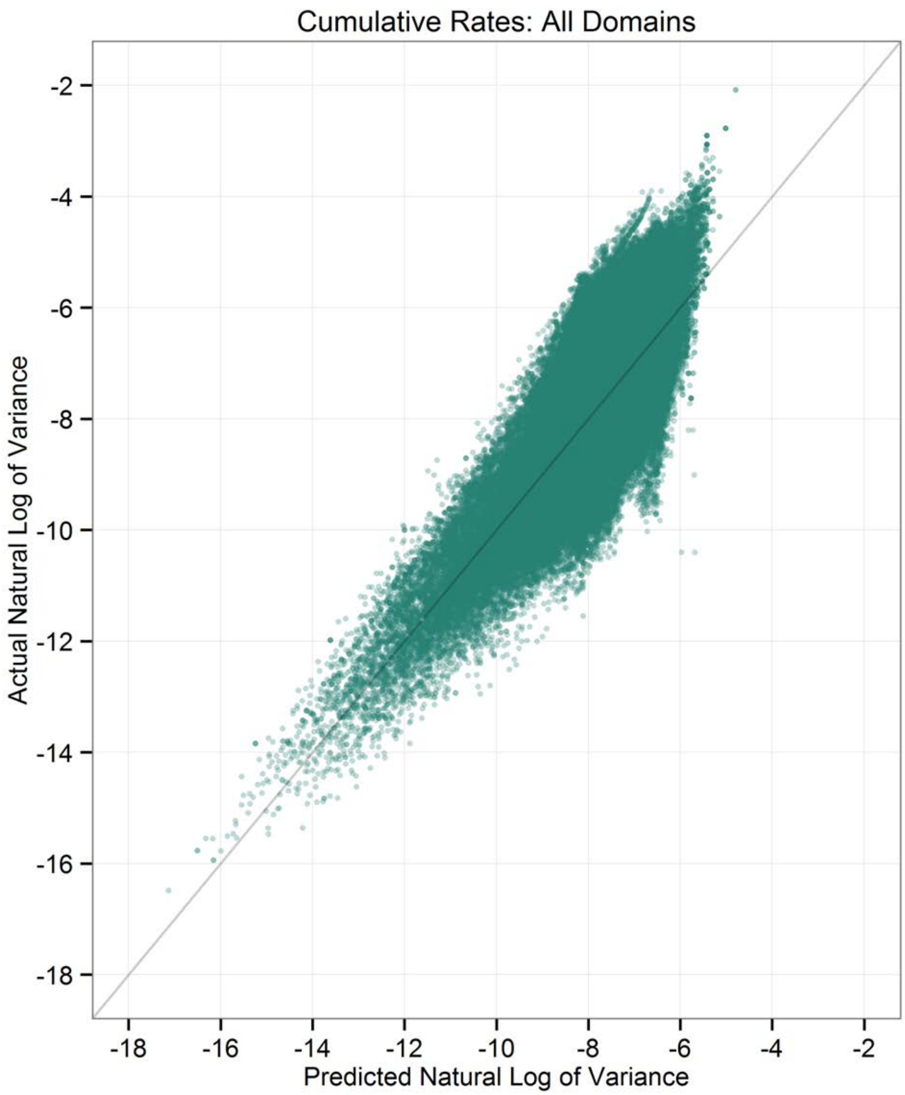


Figure 3: Initial Results comparing Direct Variance Estimates to Log-Transformation GVF Estimates for Cumulative Recidivism Rate Estimates.

After investigating which estimates were causing the upward curve, it was determined that prisoners released for homicide had variances that did not fit as well as the variances for estimates from non-homicide offense prisoners (see ***Figure 4***). The variance estimates for estimates involving homicide were consistently overestimated by the GVF. Furthermore, when isolated, the estimates involving non-homicide prisoners fit very well. In reviewing the differences between the homicide cohort and the non-homicide cohort, two differences appear to be the cause for the difference in model fit: (1) the homicide cohort was selected with certainty from all states giving these inmates very different weights than non-homicide cohort members, and (2) the recidivism rate for persons in the homicide cohort appear to be consistently lower than other persons. Since more persons are in the non-homicide cohort the GVF model parameters produced may not be appropriate for the homicide cohort.

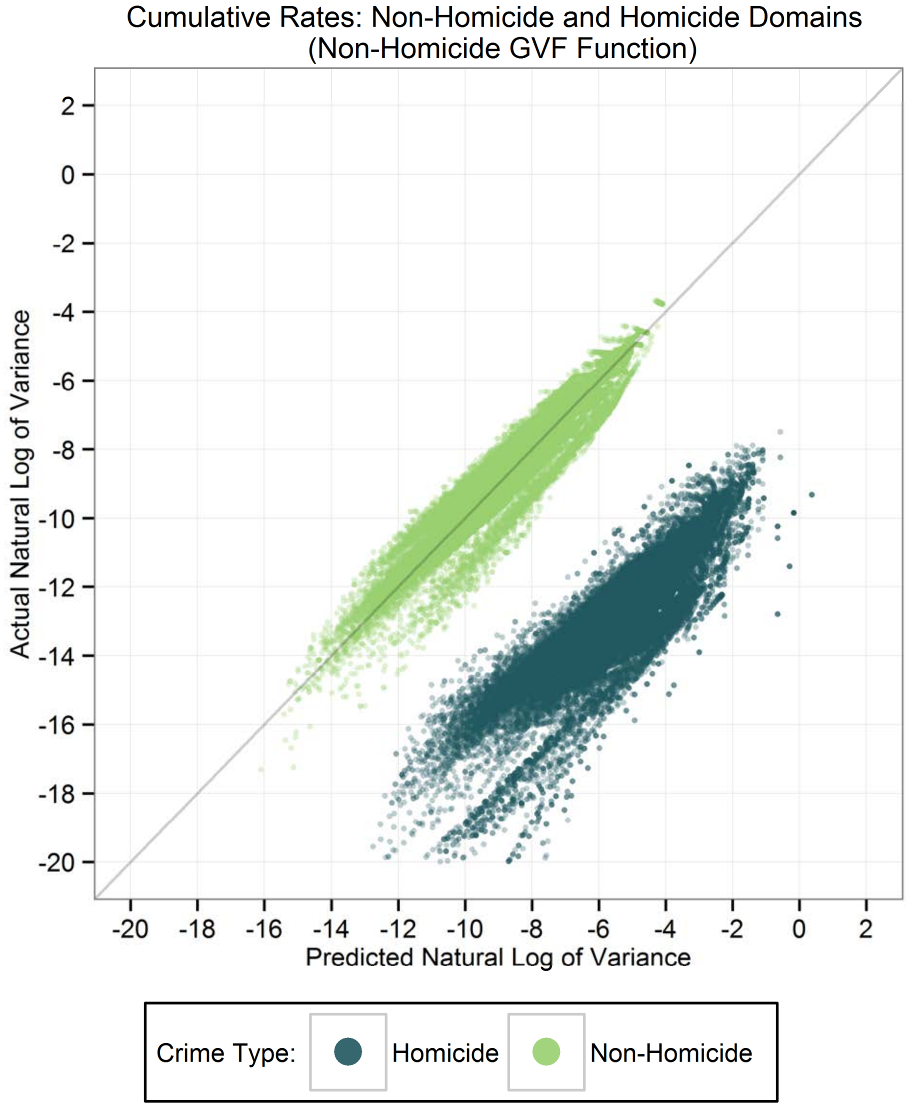


Figure 4: Comparison of Log Variances between Direct Variance and GVF Variance by Homicide and non-Homicide Characteristics for Cumulative Recidivism Rate Estimates

Based on this finding, an alternative model for estimates involving homicide was developed. Namely, a four-parameter log-transformation model was assessed. This model had the form:

$$ln[v(\hat{p})] = a + b \times \hat{p}(1-\hat{p}) + c \times ln(\hat{N}) + d \times ln(\hat{N})^2$$

4.2 Final Results

Based on the dual models – three parameter log-transform for estimates involving non-homicide prisoners and four parameter log transform for estimates involving homicide prisoners – the models fit the data well. ***Figure 5*** and ***Figure 6*** display the final fit for non-homicide and homicide cumulative recidivism rate estimates, respectively. The estimates for failure rates had a similar fit.

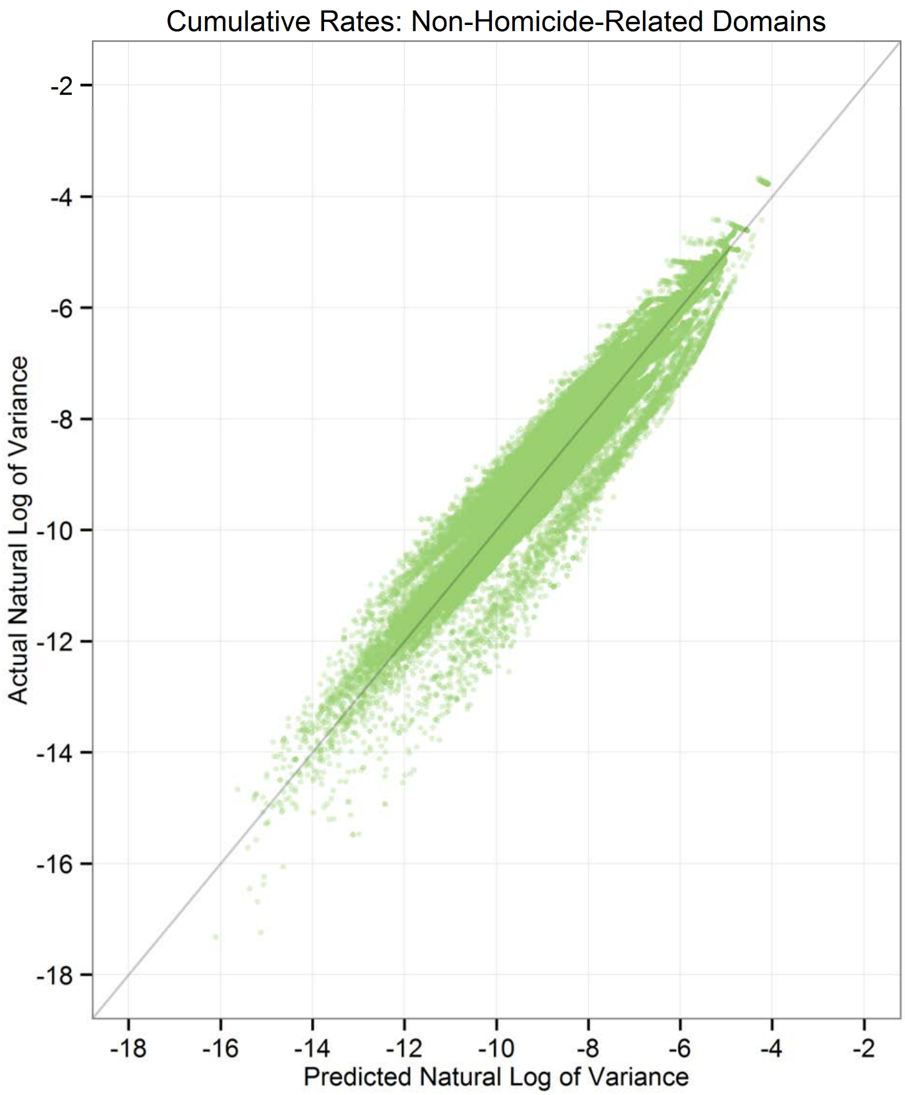


Figure 5: Comparison of Log Variances for Direct Estimates and GVF Estimates (Based on Three-Parameter Log-Transform Model) for Non-Homicide Cumulative Recidivism Rate Estimates

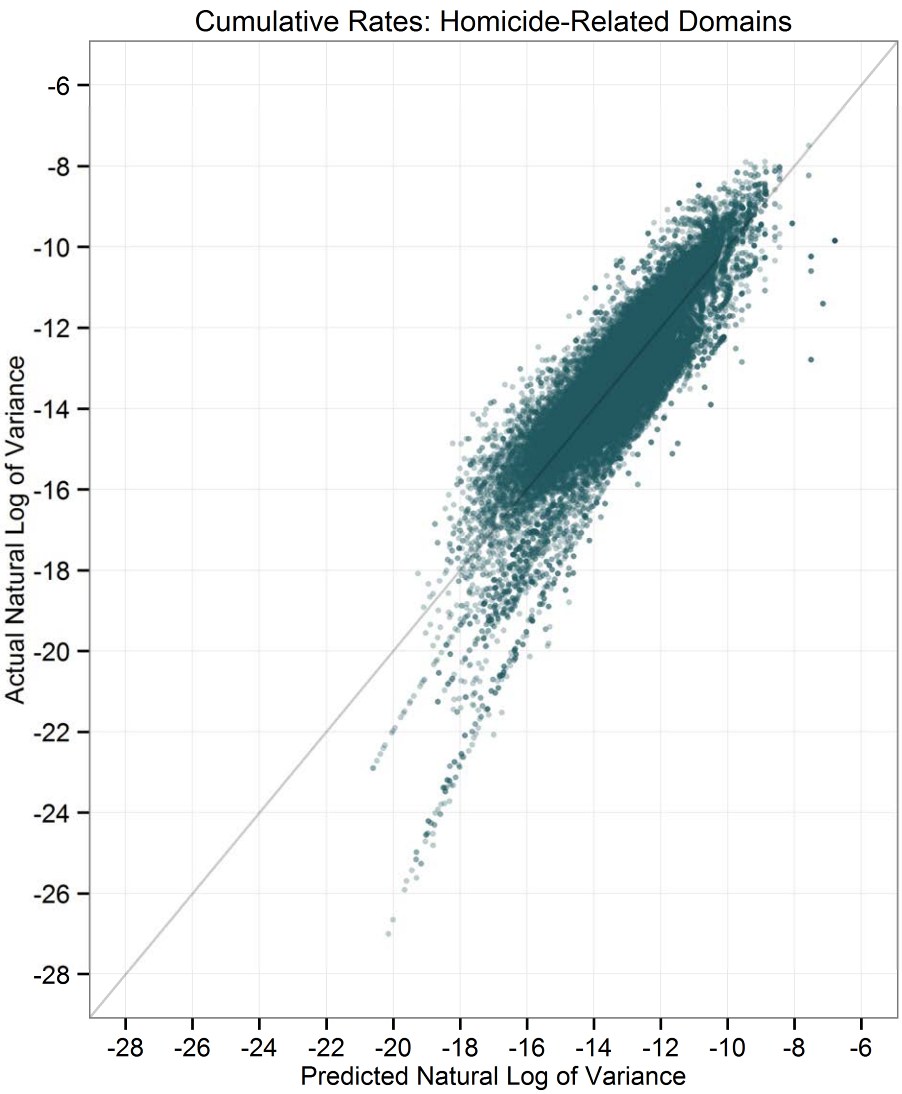


Figure 6: Comparison of Log Variances for Direct Estimates and GVF Estimates (Based on Four-Parameter Log-Transform Model) for Non-Homicide Cumulative Recidivism Rate Estimates

Based on these findings parameters for four models were produced:

- Cumulative recidivism rate – non-homicide
- Cumulative recidivism rate – homicide
- Failure rate – non-homicide
- Failure rate – homicide

5. Conclusions

GVFs can be a useful tool for producing variance estimates that account for the complex survey design when direct methods are not feasible. For the 2005 recidivism cohort online analysis tool, producing direct variance estimates for all possible estimates was not possible without compromising the website's performance. This paper illustrates how a GVF can be developed for such a web tool.

Our results found that a single model may not be possible for all estimates. Therefore, it is important to assess the models fit and see if any subdomains cause outlier results. In the case of the recidivism data, GVF variance estimates involving homicide prisoners were being over estimated by the initial model.

Acknowledgements

The authors would like to thank BJS for sponsoring this research and comments provided by Matt Durose and Howard Snyder. However, we would like to note that the views expressed in this paper are those of the authors only and do not reflect the views or position of BJS or the Department of Justice.

References

Ash, S., Hornick, D., McCune, J., Rottach, R., Sanduckchi, M. , & Warner, T. (2008). *Variance Methods of the Demographic Statistical Methods Division*. U.S. Census Bureau.

Bieler, G.S., Williams, R.L., (1990). *Generalized Standard Error Models for Proportions in Complex Design Surveys*. Proceedings for the 1990 Joint Statistical Meetings, Section on Survey Research Methods, 272 – 277.

Chand, N. (1993). U.S. Census Bureau memorandum "SIGMA Programs Based on the New Generalized Variance Function (GVF)," from Nanak Chand.

Durose, M.R, Cooper, A.D., & Snyder, H.N. (2014). *Recidivism of Prisoners Released in 30 States in 2005: Patterns from 2005 to 2010*. Special Report. U.S. Department of Justice, Bureau of Justice Statistics, NCJ 244205

Johnson, E.G., and King, B.F. (1987). Generalized variance functions for a complex sample survey. Journal of Official Statistics, 3, 235-250.

Tomlin, P. (1974). U.S. Census Bureau memorandum "Justification of the Functional Form of the GATT Curve and Uniqueness of Parameters for the Numerator and Denominator of Proportions," from Paul Tomlin to the Record, dated December 11, 1974.

Salvucci, S., Weng, S., & Kaufman, S. (1995). Design Effects and Generalized Variance Functions for the 1990-91 School Staffing Survey (SASS). U.S. Department of Education, National Center for Education Statistics, NCES 95-341-1.

Wolter, K.M. (1985). Introduction to Variance Estimation. New York: Springer-Verlag.