U.S. Department
of Transportation

National Highway
Traffic Safety
Administration

NHTSA

DOT HS 813 041

December 2020

Crash Report Sampling System: Generalized Variance Functions

DISCLAIMER

This publication is distributed by the U.S. Department of Transportation, National Highway Traffic Safety Administration, in the interest of information exchange. The opinions, findings, and conclusions expressed in this publication are those of the authors and not necessarily those of the Department of Transportation or the National Highway Traffic Safety Administration. The United States Government assumes no liability for its contents or use thereof. If trade or manufacturers' names or products are mentioned, it is because they are considered essential to the object of the publication and should not be construed as an endorsement. The United States Government does not endorse products or manufacturers.

Suggested APA Format Citation:

Zhang, F., Diaz, E. (2020, December). *Crash Report Sampling System: Generalized variance functions* (Report No. DOT HS 813 041). National Highway Traffic Safety Administration.

Technical Report Documentation Page

1. Report No. DOT HS 813 041	2. Government Accession No.	3. Recipient's Catalog No.	
4. Title and Subtitle Crash Report Sampling System: Generalized Variance Functions		5. Report Date December 2020	
		6. Performing Organization Code NSA-210	
7. Author(s) Fan Zhang [1], Eliseo Acevedo-Diaz [2]		8. Performing Organization Report No.	
9. Performing Organization Name 1. Mathematical Analysis Division, National Center for Statistics and Analysis National Highway Traffic Safety Administration U.S. Department of Transportation 1200 New Jersey Avenue SE Washington, DC 20590 2. Bowhead Logistics Solutions, LLC. 6564 Loisdale Court Springfield, VA 22150		10. Work Unit No. (TRAIS)	
		11. Contract or Grant No.	
12. Sponsoring Agency Name and Address Mathematical Analysis Division, National Center for Statistics and Analysis National Highway Traffic Safety Administration 1200 New Jersey Avenue SE Washington, DC 20590		13. Type of Report and Period Covered NHTSA Technical Report	
		14. Sponsoring Agency Code	
15. Supplementary Notes			
Abstract This study establishes the generalized variance functions for NHTSA's Crash Report Sampling System estimates.			
17. Key Words NHTSA, CRSS, GVF	18. Distribution Statement This document is available from the National Technical Information Service, www.ntis.gov.		
19. Security Classif. (of this report) Unclassified	20. Security Classif. (of this page) Unclassified	21. No. of Pages 45	22. Price

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

TABLE OF CONTENTS

Acronyms

CDS	Crashworthiness Data System
CISS	Crash Investigation Sampling System – a replacement of CDS
CRSS	Crash Report Sampling System – a replacement of GES
FARS	Fatality Analysis Reporting System
GES	General Estimates System
GVF	generalized variance function
MOS	measure of size
NASS	National Automotive Sampling System
PCR	police crash report
PJ	police jurisdiction
PPS	probability proportional-to-size, a sampling method
PSU	Primary Sampling Unit
SE	standard error
SSU	Secondary Sampling Unit
TSU	Tertiary Sampling Unit

1 Introduction

NHTSA has collected crash data since the early 1970s to support its mission to reduce motor vehicle crashes, injuries, and deaths on our Nation's highways. In 2016 NHTSA implemented two new annual surveys to improve crash data collection, the Crash Report Sampling System (CRSS) that replaced the General Estimates System (NASS-GES), and the Crash Investigation Sampling System (CISS) that replaced the Crashworthiness Data System (NASS-CDS).

Selecting a nationwide simple random sample of police crash reports requires access to all PCRs in the Nation. Therefore, it is cost prohibitive to directly select a national simple random sample of PCRs. Instead CRSS data was collected under a complex survey design with features such as multistage sampling, stratification, and unequal selection probabilities to ensure it was a nationally representative sample. However, due to its complex survey design features, CRSS was not a simple random sample. Therefore, CRSS case weights were derived corresponding to its complex design features in order to produce unbiased and robust estimates.

As for any probability-based sample, the estimates generated from CRSS data are subject to sampling errors. The sampling error is a measure of the variability of an estimator from its mean under repeated sample selections. The magnitude of sampling error depends on the study variable, the estimator used, and the sample design.

Failing to consider the complex survey design features in CRSS estimation can bias both point estimates and their associated standard error estimates.

Estimation methods and computer software have been developed to make estimates from complex survey data. Specialized procedures for complex survey data analysis, such as SAS PROC SURVEY procedures and SUDAAN procedures, can be used in CRSS data analysis along with proper design statements to take the complex survey design into account. See Zhang et al., (2018) for more details and examples of CRSS data analysis.

For users who do not have access to specialized software and wish to have a quick assessment of the magnitude of the standard errors of CRSS estimates, the generalized variance functions described in this report can be used to generate ballpark standard error estimates for a large quantity of estimates. In this approach, it is assumed the standard error of a point estimate X can be approximated by a known generalized variance function f of X indexed by estimated parameters, say, a, b, and c:

$$SE \approx f(X; a, b, c)$$

The survey statisticians normally provide the estimated parameters and specify the GVF form $f(X; a, b, c)$. Some GVF may have more or fewer estimated parameters. To have a quick assessment of the standard error of X, the data user simply first estimates X and then plugs X into $f(X; a, b, c)$ to calculate SE.

To determine the function f and estimate the parameters in f, first a group of point estimates (X's) and their associated variance estimates $\left(Var(X)\right)$ or standard error estimates (SE's – normally the square root of $Var(X)$ is used) are made from the sample data using specialized software such as SAS PROC SURVEY procedures or SUDAAN procedures. These point estimates and their associated standard error estimates are then used to identify the GVF form f and estimate the associated parameters through regression analysis.

NHTSA provided GVFs for the GES estimates (see for example, Appendix C of NHTSA's Traffic Safety Facts 2015 [NHTSA, 2017]). Over the years, GES GVF's function forms have been stable but the estimated coefficients have been changing, which indicates the need for the regular updating of the coefficients. CRSS is designed independently from GES and has a different PSU formation, sample selection method, and weighting procedure, among other differences. Therefore, the CRSS GVF may have different function forms or coefficients. For this reason, NHTSA conducted this study to determine the CRSS GVFs.

In Chapter 2 of this document, we briefly describe the CRSS sample design and weighting procedures because these features dictate how the CRSS point estimates and standard errors should be estimated.

Chapter 3 is an outline of the study. We propose the GVF's function forms to be considered. We then identify the variables, the data files, the estimators, and the variance estimation method to be used in this study. Finally, we consider the criteria to compare the GVF models.

Chapter 4 describes the GVF fitting process, compares the GVFs, and identifies the final GVF model for CRSS. After the final GVF model is identified, model coefficients for 2016 – 2019 CRSS (2019 CRSS data became available when we were making the final revision of this document) are estimated and GVF standard error estimate tables are provided.

In Chapter 5 we give some examples to show how to use the identified final GVFs to estimate standard errors.

2 The CRSS Sample Design and Weighting Procedure

2.1 The CRSS Sample Design

Unlike the CDS PSU sample was a subsample of GES PSU sample, the CRSS was designed independent of other NHTSA surveys. The target population for CRSS was the same as GES: all police-reported motor vehicle crashes on trafficways. Because a nationwide direct selection of PCRs requires access to all the PCRs in the nation, it was infeasible to select a simple random sample of PCRs. Instead, CRSS PCR sample was selected in multiple stages with unequal selection probabilities to produce a nationally representative probability sample.

At the first stage of selection, 3,117 counties in the United States were grouped into 707 Primary Sampling Units. A PSU in the CRSS was either a county or a group of counties. U.S. territories, some remote counties in Alaska, and small islands of Hawaii were excluded because of the cost and operational inefficiency.

The 707 PSUs in the PSU frame (the collection of all PSUs) were stratified into 50 strata by the four Census regions, urban/rural, vehicle miles traveled, total number of crashes, total truck miles traveled, and road miles. Each of the 707 PSUs in the frame was assigned a measure of size equal to the combination of its estimated nine types of crash counts. There were 101 PSUs selected by a stratified probability proportional-to-size sampling method. Then a sequence of sub-samples was selected from the 101 PSU sample. During this process the strata were collapsed as necessary. This produced a sequence of nested PSU samples with decreasing sample sizes selected from the collapsed strata. These nested PSU samples allow NHTSA to change the PSU sample size without reselecting the sample in the future. Therefore, the final PSU sample was the result of a multiphase sampling, and the PSU sample was selected in such a way that the resulting selection probability was still approximately PPS.

In 2016 there were 60 PSUs selected from 24 PSU strata for CRSS data collection. Because 7 PSUs did not cooperate, the CRSS data were collected from 53 PSUs. A PSU level non-response adjustment was applied to mitigate the potential non-response bias. In 2017, 6 non-responding PSUs were converted to responding PSUs and one replacement PSU was added. Therefore, from 2017 a total of 60 PSUs were used for data collection.

The secondary sampling units were police jurisdictions. Within each selected PSU, PJs were stratified into three PJ strata by their estimated measure of size which is a combination of crash counts in six categories of interest. The Pareto sampling method (Rosén, 1997) was used to select PJ samples from each PJ stratum. The Pareto sampling method produces overlapping samples when a new sample is reselected. This reduces the changes to the existing PJ sample if a new PJ sample would need to be selected because of PJ frame (the collection of all PJs in the selected PSU) changes. The PJ inclusion probability under the Pareto sampling is approximately PPS (Rosén, 1997). In 2016 for example, across the 53 responding PSUs, a total of 350 PJs was selected and 337 PJs cooperated. Weight adjustments were made to mitigate the potential bias caused by the 13 non-responding PJs.

The tertiary sampling units were PCRs. The CRSS samplers periodically received PCRs from the selected PJs. All new PCRs were sequentially stratified into nine PCR strata in the order they became available. These nine PCR strata were formed based on the results of NHTSA's internal data needs and public data needs studies. The PCR stratification was used to over-sample the following important analysis domains to ensure enough cases were selected into the sample:

- Crashes involving killed or injured pedestrians;
- Crashes involving killed or injured motorcycle occupants;
- Crashes involving killed or injured occupants in a late model year passenger vehicle; and
- Crashes involving killed or severely injured occupants in a non-late-model-year passenger vehicle.

From each PCR stratum, a systematic sampling method was used to select the PCR sample. The sampling intervals were determined in such a way that the final weights were approximately equal for all the PCRs in the same PCR stratum to reduce the sampling variance for the domain estimates. The target PCR sample size was around 50,000 every year.

2.2 The CRSS Weighting Procedure

The CRSS sample was the result of probability sampling featuring stratification, clustering, and selection with unequal probabilities. Because of these features, the CRSS sample was not a simple random sample and users need to use proper weights to produce unbiased and robust estimates. The 2016 CRSS weights were created as follows:

- Calculated the base weights (the inverse of selection probabilities) at all three stages (PSU, PJ, and PCR).
- Adjusted the base weights for non-response at all three stages to correct potential non-response bias.[1]
- Calibrated the PJ and the PCR weights using the PSU level total PCR stratum counts to further correct potential non-response bias and coverage bias.
- Adjusted the weights for duplicates.

See Zhang, Noh, Subramanian, and Chen (2019) for more detailed information on CRSS sample design and weighting process.

[1] Non-responding PCRs were incomplete or non-readable PCRs. Non-responding PJs and PSUs were PJs and PSUs refused to cooperate.

3 The Outline of CRSS GVF Study

3.1 GVFs and Fitted Models

According to Wolter (2007): “Most of the GVFs to be considered are based on the premise that the relative variance V^2 is a decreasing function of the magnitude of the expectation X.” Here $V^2 = Var(X)/X^2$. CRSS estimates are mainly domain size estimates, e.g. total number of injury crashes, total number of injured pedestrians, etc. To see how Wolter’s premise holds in domain size estimates, we consider the following simplified scenario: assuming a simple random sample (without replacement) of size n is selected from a population of known size N. Also, let X be the estimated total number of units in a domain with a certain characteristic. Let the estimated proportion of the domain size X to the population size N be $p = X/N$. The estimated sampling variance of p can be written as:

$$var(p) = \left(1 - \frac{n}{N}\right)\frac{p(1-p)}{n-1}$$

Notice $X = Np$, hence we have:

$$\begin{aligned} var(X) &= N^2 var(p) \\ &= \left(1 - \frac{n}{N}\right)\frac{N^2 p(1-p)}{n-1} \\ &= \left(1 - \frac{n}{N}\right)\frac{N^2(p-p^2)}{n-1} \\ &= \left(1 - \frac{n}{N}\right)\frac{1}{n-1}(NX - X^2) \end{aligned}$$

Let $a = -\left(1 - \frac{n}{N}\right)\frac{1}{n-1}$ and $b = \left(1 - \frac{n}{N}\right)\frac{N}{n-1}$, $var(X)$ can be rewritten as:

$$var(X) = bX + aX^2.$$

Therefore,

$$\frac{var(X)}{X^2} = \frac{b}{X} + a$$

This indeed leads to Wolter’s premise,

$$V^2 = a + \frac{b}{X}$$

In the following we list 9 linear models to be considered in CRSS GVF study and their corresponding GVFs. These models are commonly used as GVFs and most of them can be found in Wolter (2007). Although as Wolter pointed out “there is very little theoretical justification for any of the models,” the goal, however, is clear: to find a GVF that fits the estimates well.

In practice, data users are mainly interested in standard error estimates instead of variance estimates. Therefore, in this study, the GVF is referred to the function to be used to calculate standard error estimates $ste(X)$, while the linear model is referred to the linear model to be fitted in order to estimate the coefficients in the GVF. We also examine Wolter’s premise - the related relative variance $V^2 = Var(X)/X^2$ is a decreasing function of the magnitude of the expectation X - when we consider the GVF candidates.

Model-1: $Var(X) = aX^2 + bX$

Dividing both sides of model-1 by X^2 leads directly to the relative variance:

$$V^2 = a + \frac{b}{X}$$

Obviously, the right-hand side is indeed a decreasing function when X increases. Taking square root of both sides of model-1 produces the GVF:

$$ste(X) = \sqrt{aX^2 + bX}$$

Model-2: $Var(X) = aX^2 + bX + c$

Notice model-1 is a special case of model-2. From model-2,

$$V^2 = a + \frac{b}{X} + \frac{c}{X^2}$$

This model is Wolter's (7.2.2). It's obvious the relative variance is a decreasing function when X increases. Taking square roots of both sides of model-2 gives the GVF:

$$ste(X) = \sqrt{aX^2 + bX + c}$$

Model-3: $ln[ste(X)] = a + b * ln(X)$

Alternatively, the GVF is:

$$ste(X) = e^{a+b*ln(X)}$$

Notice:

$$Var(X) = e^{2a+2b*ln(X)}$$

Multiplying both sides by $X^{-2} = e^{-2*ln(X)}$, the relative variance is:

$$V^2 = e^{2a+2b*ln(X)-2*ln(X)} = e^{2a+2(b-1)*ln(X)}$$

Since a and b can be any real numbers, the right-hand side indeed can be a decreasing function of X.

Model-4: $ln[ste(X)] = a + b * ln^2(X)$

From model-4, the GVF is:

$$ste(X) = e^{a+b*ln^2(X)}$$

Notice this is the same as the GES GVF. To obtain the relative variance, notice:

$$Var(X) = e^{2a+2b*ln^2(X)}$$

Multiplying both sides by X^{-2}, the relative variance is:

$$V^2 = e^{2a+2b*ln^2(X)-2*ln(X)}$$

Since a and b can be any real numbers, the right-hand side indeed can be a decreasing function of X.

Model-5: $ln[ste(X)] = a + b * ln(X) + c * ln^2(X)$

Obviously, both model-3 and model-4 are special cases of model-5. To see the corresponding relative variance, notice from model-5 the corresponding GVF is:

$$ste(X) = e^{a+b*ln(X)+c*ln^2(X)}$$

Also,

$$Var(X) = e^{2a+2b*ln(X)+2c*ln^2(X)}$$

Therefore,

$$V^2 = e^{2a+2b*ln(X)+2c*ln^2(X)-2*ln(X)} = e^{2a+(2b-2)*ln(X)+2c*ln^2(X)}$$

Again, the right-hand side can be a decreasing function of X.

Model-6: $V^{-2} = a + bX$

Model-6 can be rewritten as:

$$V^2 = (a + bX)^{-1}$$

which is Wolter's (7.2.3). From this equation:

$$Var(X) = \frac{X^2}{a + bX}$$

Therefore, the GVF is:

$$ste(X) = \frac{X}{\sqrt{a + bX}}$$

Model-7: $V^{-2} = a + bX + cX^2$

It's easy to see model-6 is a special case of model-7. Model-7 can be rewritten as:

$$V^2 = (a + bX + cX^2)^{-1}$$

which is Wolter's (7.2.4). From this equation:

$$Var(X) = \frac{X^2}{a + bX + cX^2}$$

Therefore, the GVF is:

$$ste(X) = \frac{X}{\sqrt{a + bX + cX^2}}$$

Model-8: $ln(V^2) = a + b * ln(X)$

This model is the same as Wolter's (7.2.5). To see the relative variance, from model-8 we have:

$$V^2 = e^{a+b*ln(X)}$$

The right-hand side can be a decreasing function of X. From this relative variance, we can also obtain the GVF as the following:

$$Var(X) = e^{a+b*ln(X)+2*ln(X)} = e^{a+(b+2)*ln(X)}$$

$$ste(X) = e^{(a/2)+[(b+2)/2]*ln(X)}$$

Model-9: $ste(X) = a + b * ln(X)$

From model-9:

$$Var(X) = b^2 * ln^2(X) + 2ab * ln(X) + a^2$$

$$V^2 = b^2 * \frac{ln^2(X)}{X^2} + 2ab * \frac{ln(X)}{X^2} + \frac{a^2}{X^2} \xrightarrow[X\to\infty]{} 0$$

The GVF is model-9 itself.

$$ste(X) = a + b * ln(X)$$

Like GES, the CRSS data have a hierarchical structure: information is collected from crashes, vehicles involved in the crashes, and the persons involved with the vehicles. The GES GVFs were fitted separately for crash, vehicle, and person estimates and showed different coefficient estimates. For this reason, we also fitted model 1 to 9 mentioned above at three different levels: crash, vehicle, and persons.

Sampling variance is the function of the population, the sample design, the point and the variance estimators used, and the variable itself. A GVF fitted to a more specific group of estimates (i.e. under the same design, the same type of estimators defined on the same population) may give better estimated standard errors but it covers a limited range of estimates. On the other hand, a GVF fitted to a more general group of estimates (i.e. under different sample designs such as GES and CRSS, different type of estimators defined on different populations) may cover a wider range of estimates but then there may be more large differences between the GVF estimates and the estimates calculated from specialized software.

3.2 Variables Considered

Table 1 lists all variables used in the GES GVF model fitting. We included all these variables used in GES GVF for CRSS GVF model fitting.

Table 1: Variables Used for GES GVF Model Fitting

Accident	Vehicle	Person
Crash Severity (Property Damage/Injury)* - **INJSEV_IM**	Body Type*- **BDYTYP_IM**	Injury severity* - **INJSEV_IM**
Crash Type (Single/Multi vehicle) – **VE_FORMS**	Initial Contact Point*- **IMPACT1_IM**	Person Type – **PER_TYP**
Month-**MONTH**	Speed Limit* - **CSPD_LIM**	Age* - **AGE_IM**
Time of Day* - **HOUR_IM**	Special Use Vehicle - **SPEC_USE**	Sex* - **SEX_IM**
Day of the Week*- **WKDY_IM**	Most Harmful Event* - **VEVENT_IM**	Passenger Ejected* - **EJECT_IM**
Weather Condition* - **WEATHR_IM**	Rollover - **ROLLOVER**	Correct Restrain Use – **REST_USE**
Light Condition* - **LGTCON_IM**	Fire Occurrence – **FIRE_EXP**	Incorrect Restrain Use – **REST_MIS**
Relation to Roadway – **REL_ROAD**	Pre-Event Movement* - **PCRASH1_IM**	Seating Position* - **SEAT_IM**
Relation with Junction* – **RELJCT2_IM**		Air Bag Deployed – **AIR_BAG**
Manner of Collision*-**MANCOL_IM**		Non-Motorist Location - **LOCATION**

* - Denotes imputed variable

In GES GVF model fitting, all variables were treated as categorical variables. And all estimates are the category sub-population size estimates. Numerical variables such as the number of vehicles involved in the crash or the number of persons in the vehicle were not considered in GES GVF model fitting. For these numerical variables, the related estimates were population or sub-population total estimates. To extend the CRSS GVF application to total estimates, we added the following numerical variables to the CRSS GVF model fitting:

Table 2: Numerical Variables Added to CRSS GVF Model Fitting

Accident	Vehicle
Number of Persons Not in Motor Vehicles – **PEDS**	Number of Occupants - **NUMOCCS**
Number of Persons Not in Motor Vehicles in Transport- **PERNOTMVIT**	Imputed Number Injured in Vehicle – **NUMINJ_IM**
Number of Total Motor Vehicles – **VE_TOTAL**	Number of Lanes in Roadway– **VNUM_LAN**
Number of Total Motor Vehicles in Transport – **VE_FORMS**	
Number of Parked Vehicles – **PVH_INVL**	
Number of Persons in Motor Vehicles in Transport – **PERMVIT**	
Imputed Number of Injured in Crash* – **NO_INJ_IM**	

* - Denotes imputed variable

Therefore, three types of estimates were used in CRSS GVF model fitting: categorical variable estimates, numerical variable estimates, and combined (or mixed) estimates. Person level only had categorical variable estimates.

3.3 Variance Estimation Method

Because of the complex sample design, jackknife variance estimation method was used to estimate the variances in this study.

The jackknife variance estimation method is a replication method. By this method, one PSU is deleted from the original sample so the remaining sample becomes a sub-sample. The analysis weights in the sub-sample are adjusted to compensate the deleted PSU. Then the estimate under consideration is calculated using the sub-sample and the adjusted analysis weights. This process is replicated over all PSUs to obtain a group of estimates calculated from the sub-samples. Then the variation of these sub-sample estimates around the full sample estimate is calculated as the estimated variance of the estimate. See Wolter (2007) for a more detailed description. Zhang et al. (2018) provided examples on CRSS data analysis using the jackknife variance estimation method.

3.4 Point Estimators Considered

As we mentioned above, two types of estimators were considered in the CRSS GVF model fitting. The first type was the sub-population size estimator. If categorical variable X has k possible categories: $1, 2, \dots, k$ (7 days of the week for example), the number of cases in category j (Sunday for example) in the population can be estimated by:

$$\widehat{N}_j = \sum\nolimits_{i \in s_j} w_i$$

Here i refers to a sampled unit, s_j is the set of sampled units in category j (Sunday crashes for example), w_i is the weight of unit i. $\widehat{N}_j$ is the estimate of a sub-population size – the estimated total number of category j units in the population. GES GVF model fitting only considered this type of estimates.

The second type estimator considered in the CRSS GVF study was for the total estimates of numerical variables. Let X be a numeric variable (number of vehicles involved in a crash for example). The population total estimate of a numeric variable X is estimated by:

$$\hat{X} = \sum\nolimits_{i \in s} w_i X_i$$

Here X_i is the numerical value of X for unit i. For a sub-population d (Sunday for example), the corresponding sub-population total estimate is:

$$\hat{X}_d = \sum\nolimits_{i \in s \cap d} w_i X_i$$

Here $\hat{X}_d$ is the total estimate of numerical variable X for sub-population d (total number of vehicles involved in Sunday crashes, for example).

$\widehat{N}_j$ can be viewed as a special case of $\hat{X}_d$ where d is the category j, $X_i = 1$ when $i \in s \cap d$ and $X_i = 0$ otherwise. In this sense, CRSS GVF covered a wider range of estimates than GES GVF.

3.5 Data

We had 3 years of CRSS data at the time the GVF models were fitted: 2016-2018. Because of PSU non-response, 2016 CRSS had 53 PSUs. 2017 and 2018 CRSS had 60 PSUs. We used 2016 and 2017 CRSS files separately to make point estimates (X_i) and variance estimates ($var(X_i)$). Then we combined these estimates of X_i and $var(X_i)$) to fit all proposed models. After we estimated the model coefficients and goodness of fit statistics, we identified a few plausible models. We then used 2018 CRSS data to compare the plausible models and recommended the final model.

3.6 Model Comparison Method

After we fitted model 1-9 using 2016 and 2017 CRSS estimates, we used the goodness of fit statistics to identify a few plausible models for further comparison. We used 2018 CRSS data to compare these plausible models to determine the final model. To this end, we first used 2018 CRSS data to make the corresponding point and standard error estimates $\left(X_i,\ ste(X_i)\right)$. We then used the 2018 point estimates X_i and the GVFs fitted using 2016-2017 CRSS estimates to calculate standard error estimates $ste_{GVF}(X_i)$. These GVF standard error estimates were compared to the actual 2018 CRSS standard error estimates $ste(X_i)$ calculated from SAS SURVEY procedures. We considered the following criteria to compare the plausible models:

$$\frac{1}{k}\sum_{i=1}^{k} \frac{|ste_{GVF}(X_i) - ste(X_i)|}{ste(X_i)}$$

here k is the number of estimates used. This is the average absolute relative errors of the standard error estimates using GVF compared to the actual standard error estimates. We used the actual standard error estimates because the true standard errors were unknown.

4 GVF Model Fitting

We describe the CRSS GVF model fitting process in this chapter. Our goal was to find the best model at each data level following the procedure described in section 3.6. All the estimates in this section were obtained using SAS Studio 3.6 Enterprise Edition.

4.1 Exploratory Analysis

To see if there is any apparent dependence of the standard error estimates on the point estimates, the scatterplots of the standard error estimates and the point estimates for the categorical and numerical variables (person level only has categorical variables) are presented in Figure 1 to 3. These figures show clear dependence between the standard error estimates and the point estimates.

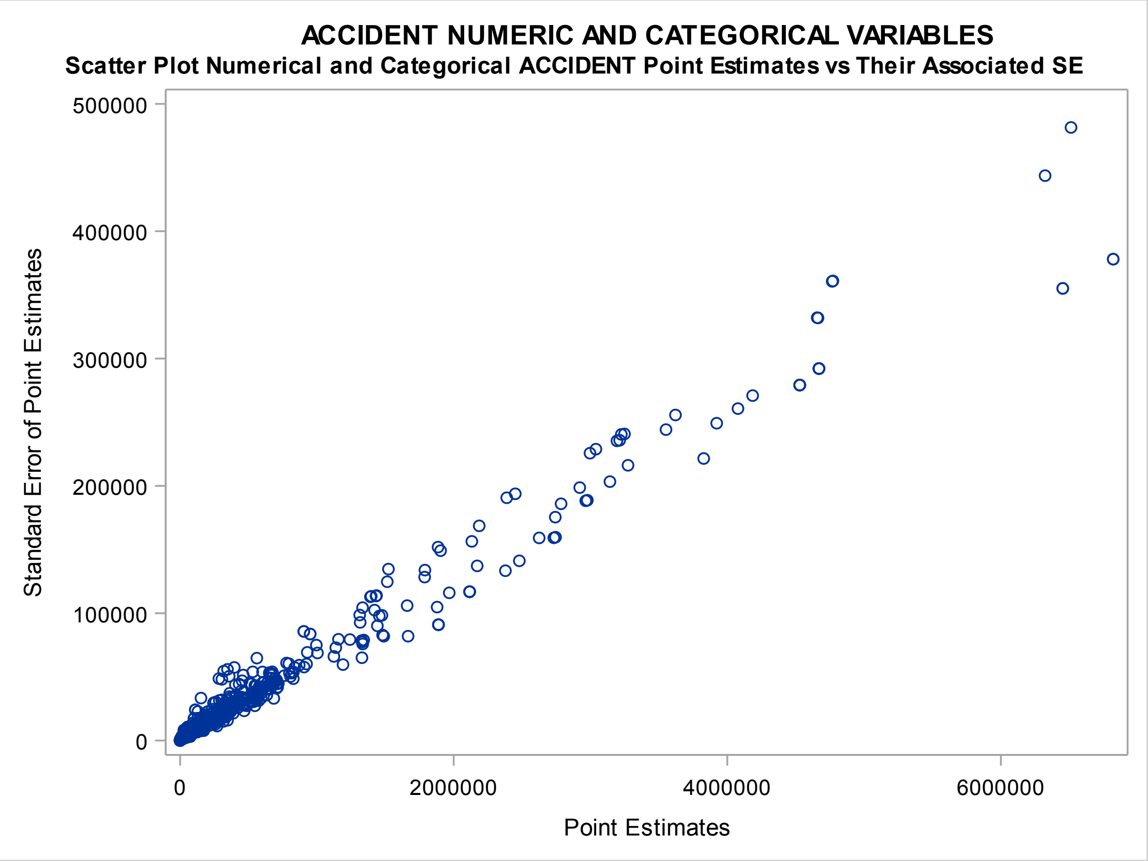


Figure 1: Crash Level Categorical and Numerical Variable Estimate Scatterplot

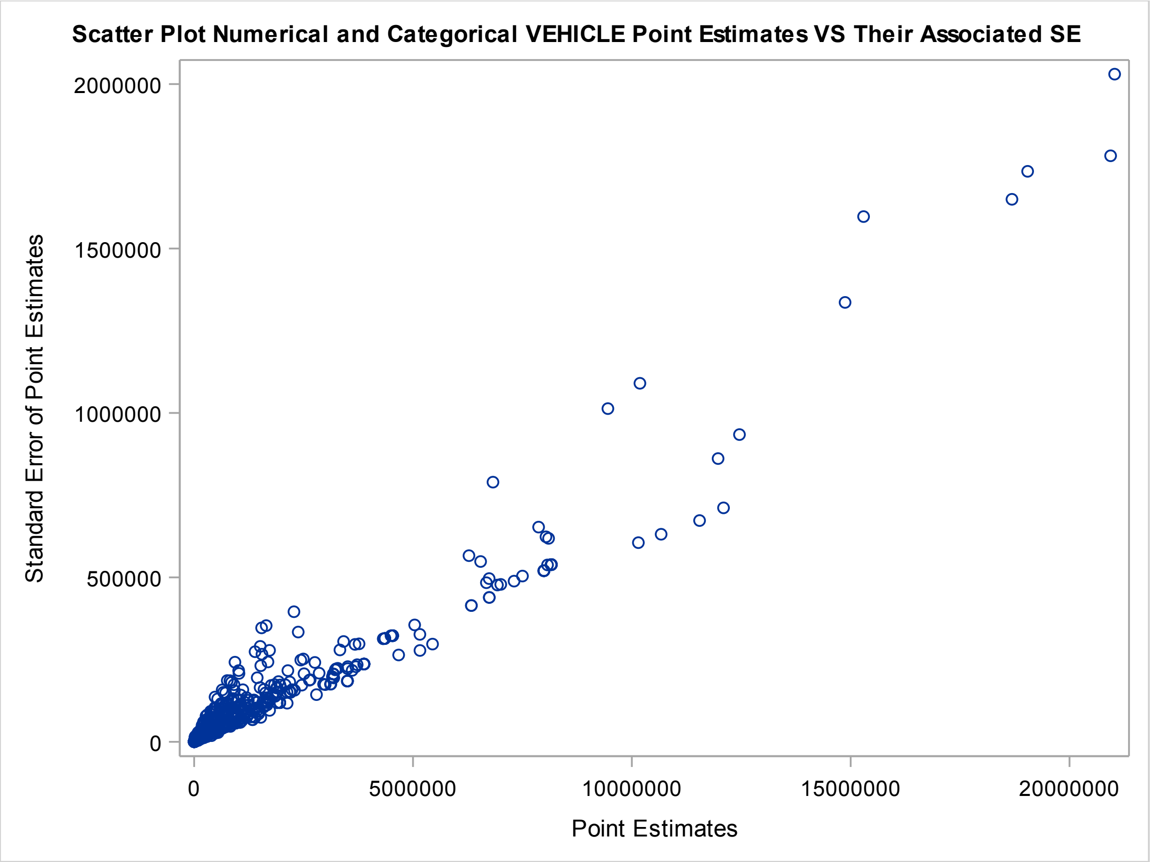


Figure 2: Vehicle Level Categorical and Numerical Variable Estimate Scatterplot

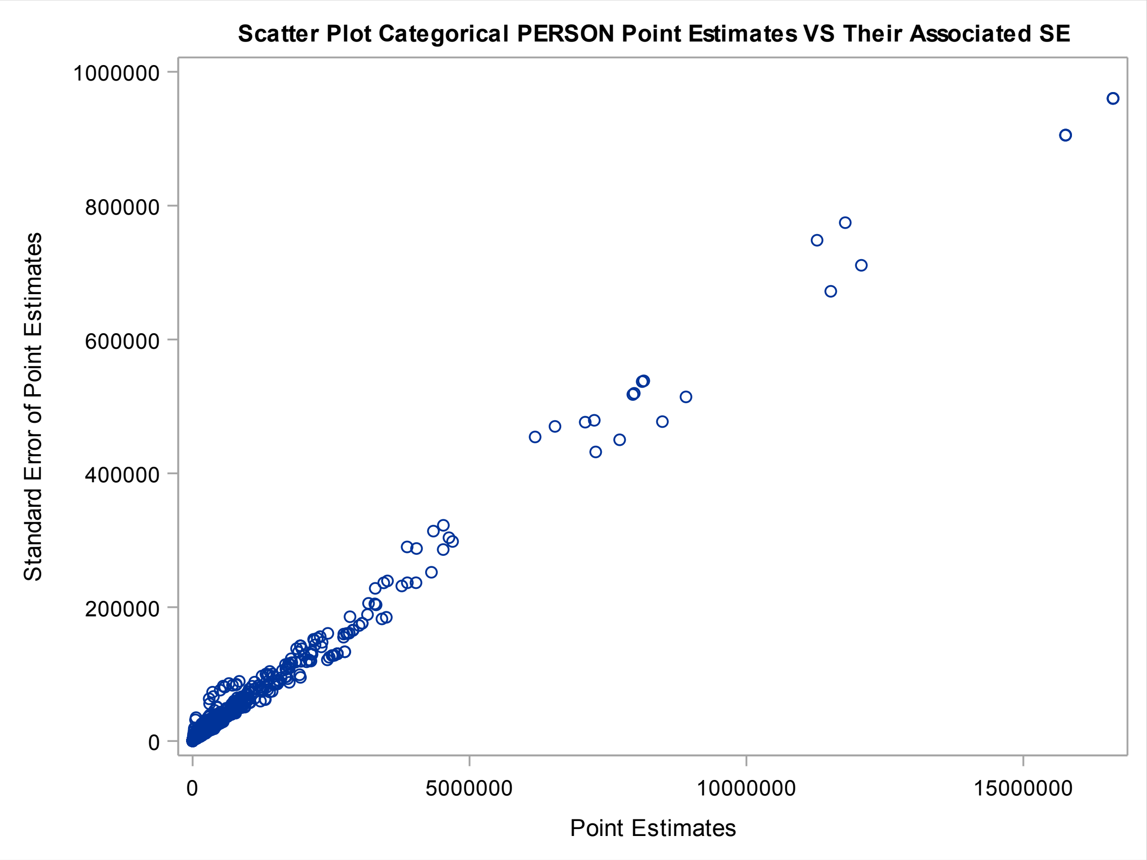


Figure 3: Person Level Categorical Variable Estimate Scatterplot

4.2 Model Fitting

We fitted regression models 1 to 9 to three types of estimates: the categorical variable estimates, numerical variable estimates, and combined (categorical and numerical) estimates at three levels: the accident (crash) level, vehicle level, and person level (categorical variable estimates only). Tables 3 to 5 present the R^2. R^2 measures the proportion of the total variability in the dependent variable that can be accounted for by the fitted model.

Table 3: Accident Level Model Fitting R^2

Model	Regression Model	Categorical	Numerical	Combined
1	$Var(X) = aX^2 + bX$	0.9762	0.9877	0.9474
2	$Var(X) = aX^2 + bX + c$	0.9761	0.9865	0.9449
3	$\ln[ste(X)] = a + b * \ln[X]$	0.9768	0.9873	0.9737
4	$\ln[ste(X)] = a + b * \ln^2[X]$	0.9769	0.9749	0.9760
5	$\ln[ste(X)] = a + b * \ln[X] + c * \ln^2[X]$	0.9805	0.9886	0.9796
6	$V^{-2} = a + bX$	0.1711	0.3986	0.1750
7	$V^{-2} = a + bX + cX^2$	0.2696	0.6511	0.2854
8	$\ln(V^2) = a + b * \ln(X)$	0.6881	0.8891	0.6999
9	$ste(X) = b + a * \ln(X)$	0.3827	0.4260	0.3745

Table 4: Vehicle Level Model Fitting R^2

Model	Regression Model	Categorical	Numerical	Combined
1	$Var(X) = aX^2 + bX$	0.9595	0.9741	0.9477
2	$Var(X) = aX^2 + bX + c$	0.9574	0.9683	0.9477
3	$\ln[ste(X)] = a + b * \ln[X]$	0.9692	0.9633	0.9691
4	$\ln[ste(X)] = a + b * \ln^2[X]$	0.9664	0.9601	0.9668
5	$\ln[ste(X)] = a + b * \ln[X] + c * \ln^2[X]$	0.9710	0.9662	0.9712
6	$V^{-2} = a + bX$	0.2201	0.0775	0.1246
7	$V^{-2} = a + bX + cX^2$	0.2973	0.1214	0.2418
8	$\ln(V^2) = a + b * \ln(X)$	0.5133	0.4827	0.4900
9	$ste(X) = a + b * \ln(X)$	0.4563	0.5220	0.3195

Table 5: Person Level Model Fitting R^2

Model	Regression Model	Categorical
1	$Var(X) = aX^2 + bX$	0.9909
2	$Var(X) = aX^2 + bX + c$	0.9909
3	$\ln[ste(X)] = a + b * \ln[X]$	0.9717
4	$\ln[ste(X)] = a + b * \ln^2[X]$	0.9725
5	$\ln[ste(X)] = a + b * \ln[X] + c * \ln^2[X]$	0.9755
6	$V^{-2} = a + bX$	0.1249
7	$V^{-2} = a + bX + cX^2$	0.2579
8	$\ln(V^2) = a + b * \ln(X)$	0.6786
9	$ste(X) = a + b * \ln(X)$	0.2614

* Person Level Coefficient only includes Categorical variables

The reduction in R^2 by fitting models to the combined estimates is trivial. Therefore, we did not fit separate models to categorical estimates and numerical estimates. This would reduce the number of GVFs to be fitted and published. In addition, the GVF users don't need to differentiate the domain size estimates and the domain total estimates in order to use the GVFs.

The first 5 models are clearly plausible candidates because of their high R^2. Next, we identified the final model from the first 5 models.

4.3 Average Absolute Relative Errors and Model Coefficients

After excluding Models 6 through 9 due to low R^2 values, the average absolute relative errors of the standard error estimates were calculated for the remaining 5 models. These average absolute relative errors are the average absolute relative errors between the standard error ($ste_{GVF}(X_i)$) estimates calculated from the GVF fitted from the combined 2016-2017 estimates and the actual standard error estimates $ste(X_i)$ estimated from the 2018 data using SAS SURVEY procedures. A low average absolute relative error is preferred for a good GVF. Accident level and vehicle level coefficients were fitted from categorical and numerical variable estimates. Person level coefficients were fitted from categorical variable estimates.

Table 6 presents the average absolute relative errors for models 1 to 5. Table 7 presents the estimated coefficients for these models.

Model 1 forces the intercept to be zero. This resulted in bad fit for some small point estimates. For example, one of the accident level point estimates was $X = 2{,}038.30$. By the estimated coefficients of model 1, we have:

$$ste_{GVF}(X) = \sqrt{aX^2 + bX} = \sqrt{5241870} \approx 2290$$

while the actual standard error estimate is:

$$ste(X) = 207$$

This resulted in a relative error of:

$$\frac{2290 - 207}{207} \approx 10$$

From Table 7, model 1 vehicle level GVF and all model 2 GVFs have a negative coefficient. GVFs with negative coefficient may produce negative variance estimates hence imaginary standard error estimates for certain range of point estimates. This is not desirable for practical use. Use accident level model 2 as an example: because coefficient c is negative, small point estimates may have negative GVF variance estimate. For example, when X=57.5521:

$$Var_{GVF}(X) = aX^2 + bX + c = -284{,}120{,}586.9$$

The average absolute relative errors for model 1 and 2 in Table 6 were calculated for those non-negative variances only. Because of these reasons, we do not further consider model 1 and 2.

Models 3 to 5 had similar absolute relative errors. Models 3 and 4 were special cases of model 5. For practical use, it is desirable to use the most parsimonious model. In the next section, we look into the model fitting statistics to determine whether the full model 5 is necessary.

Table 6: Average Absolute Relative Errors of Standard Error Estimates

Model	Crash Avg. Abs. Error	Vehicle Avg. Abs. Error	Person Avg. Abs. Error
1	2.18794	0.23970	4.36351
2	0.40993	72.3294	0.65228
3	0.28096	0.34484	0.26237
4	0.25538	0.33743	0.24666
5	0.24181	0.32592	0.23547

Table 7: Estimated Model Coefficients

Model 1 $ste(X) = \sqrt(ax\hat{}2 + bx)$	**Accident Level Coefficients**	**Vehicle Level Coefficients**	**Person Level Coefficients***

2016/2017 (combined)	a = 0.00350
	b = 2564.554
		a = 0.00922
		b = -24096
			a = 0.00314
			b = 4225.447

2018	a = 0.00271
	b = 2770.673
		a = 0.00661
		b = -7241.786
			a = 0.00272
			b = 3528.647

Model 2 $ste(X) = \sqrt(ax\hat{}2 + bx + c)$			

2016/2017 (combined)	a = 0.00344
	b = 2900.220
	c = -284287512
		a = 0.00935
		b = -26611
		c = 6040496740
			a = 0.0.00312
			b = 4527.446
			c = -470393364

2018	a = 0.00264
	b = 3172.581
	c = -3808957354
		a = 0.00662
		b = -7572.146
		c = 993044170
			a = 0.00270
			b = 3799.326
			c = -409294705

Model 3 $ln(ste(x)) = a + b * ln(x)$			

2016/2017 (combined)	a = -0.06256
	b = 0.79950
		a = -0.46990
		b = 0.85098
			a = -0.15941
			b = 0.80119

2018	a = -0.05183
	b = 0.79146
		a = -0.41156
		b = 0.84049
			a = -0.10564
			b = 0.78968

Model 4 $ln(ste(x)) = a + b * ln^2(x)$			

2016/2017 (combined)	a = 3.90257
	b = 0.03815
		a = 4.05120
		b = 0.03806
			a = 3.81855
			b = 0.03824

2018	a = 3.94026
	b = 0.03728
		a = 4.06559
		b = 0.03749
			a = 3.81822
			b = 0.03770

Model 5 $ln(ste(x)) = a + b * ln(x) +$ $c * ln^2(x)$			

2016/2017 (combined)	a = 2.12560
	b = 0.35394
	c = 0.02144
		a = 1.36664
		b = 0.50254
		c = 0.01571
			a = 1.93305
			b = 0.37707
			c = 0.02038

2018	a = 2.33242
	b = 0.31521
	c = 002258
		a = 1.69299
		b = 0.44262
		c = 0.01787
			a = 2.02774
			b = 0.35777
			c = 0.02075

4.4 The Final Model

We now determine the final model from the following three models:

- Model 3: $\ln[ste(X)] = a + b * \ln(X)$
- Model 4: $\ln[ste(X)] = a + b * \ln^2(X)$
- Model 5: $\ln[ste(X)] = a + b * \ln(X) + c * \ln^2(X)$

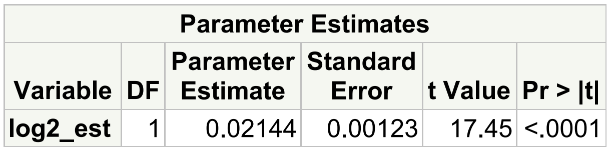
The following SAS output was from PROC REG fitting model 5 to 2016-2017 accident level combined estimates. All estimated coefficients a (Intercept), b (log_est), and c (log2_est) were highly significantly greater than zero.

Number of Observations Read	1059
Number of Observations Used	1059

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	2	3978.05623	1989.02812	25382.0	<.0001
Error	1056	82.75221	0.07836		
Corrected Total	1058	4060.80844			

Root MSE	0.27994	**R-Square**	0.9796
Dependent Mean	8.45861	**Adj R-Sq**	0.9796
Coeff Var	3.30947		

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	2.12560	0.13127	16.19	<.0001
log_est	1	0.35394	0.02578	13.73	<.0001



Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
log2_est	1	0.02144	0.00123	17.45	<.0001

Fitting models 3 to 5 at vehicle and person level produced the similar results. Model 5 not only had the lowest average absolute relative error, it also had the highest R-square. All estimated coefficients of model 5 were significantly greater than zero and its residual distribution showed better fit (see Appendix A). In addition, using the accident level R^2 estimated from the combined 2016-2017 data in Table 3, the F-tests:

$$F = \frac{\left(R^2_{model\ 5} - R^2_{model\ 3}\right)/1}{\left(1 - R^2_{model\ 5}\right)/(1059 - 3)} = 305 \gg F_{(1;\ 1056;\ 0.05)} = 3.85$$

and

$$F = \frac{\left(R^2_{model\ 5} - R^2_{model\ 4}\right)/1}{\left(1 - R^2_{model\ 5}\right)/(1059 - 3)} = 186 \gg F_{(1;\ 1056;\ 0.05)} = 3.85$$

rejected the null hypotheses that Model 3 and Model 4 fit the data. In summary, Model 5 was the final model.

Table 8: Estimated Coefficients of the Final Model From 2016 to 2019

Final Model $ln(ste(x)) = a + b * ln(x) + c * ln^2(x)$	Accident Level Coefficients	Vehicle Level Coefficients	Person Level Coefficients

2016	a = 1.92772
	b = 0.38750
	c = 0.01947
		a = 1.17146
		b = 0.53866
		c = 0.01425
			a = 1.79032
			b = 0.40622
			c = 0.01930

2017	a = 2.33171
	b = 0.30826
	c = 0.02344
		a = 1.43152
		b = 0.48824
		c = 0.01629
			a = 2.05394
			b = 0.35287
			c = 0.02119

2018	a = 2.33242
	b = 0.31521
	c = 0.02258
		a = 1.69299
		b = 0.44262
		c = 0.01787
			a = 2.02774
			b = 0.35777
			c = 0.02075

2019*	a = 2.19494
	b = 0.33465
	c = 0.02185
		a = 1.70176
		b = 0.43713
		c = 0.01826
			a = 2.14416
			b = 0.32619
			c = 0.02238

*: CRSS 2019 data became available when we were making the revision of this report.

Table 8 presents the estimated coefficients of the final model from 2016-2019. To use the final model to estimate the standard error, first calculate the point estimate X, then use the following formula:

$$ste(X) = e^{a+bX+cX^2}$$

Appendix C provides GVF standard error estimates using this final model for 2016 – 2019 CRSS estimates at crash, vehicle, and person levels.

5 Examples of Using GVF

We use two examples to show how to use the final model and Appendix C to estimate the standard errors of a total estimate and a proportion estimate.

Example 1: Estimate the standard error of a sub-population size estimate.

In the vehicle file, variable HITRUN_IM indicates whether each vehicle is "hit-and-run" vehicle or not. The total number of "hit-and-run" vehicles in 2018 police reported in-transport crashes is estimated as $X = 817{,}573$ (the summation of the weights of all "hit-and-run" vehicles). Using the 2018 vehicle level model coefficients listed in Table 8, the corresponding GVF standard error estimate is:

$$ste(X) = e^{1.69299+0.44262*\ln(817{,}573)+0.01787*(\ln(817{,}573))^2} = 61{,}756$$

compared with the actual variance estimate using SAS PROC SURVEY procedure: 66,812.

Alternatively, we can also use the GVF standard error estimate tables in Appendix C. For estimate $X = 817{,}573$ there is no standard error estimate in the 2018 CRSS GVF Standard Error Estimate table. We need to make approximation by interpolation. The following is an excerpt of the 2018 CRSS GVF Standard Error Estimate table around estimate $X = 817{,}573$ at vehicle level:

2018 CRSS GVF Standard Error Estimates	
Vehicle	
Estimate (X)	Standard Error*
800,000	60,500
900,000	67,500

$X = 817{,}573$ is between $X = 800{,}000$ and $X = 900{,}000$. Therefore, we approximate the standard error for estimate $X = 817{,}573$ by interpolation as the following:

$$ste(X) = 60{,}500 + \frac{817{,}573 - 800{,}000}{900{,}000 - 800{,}000} * (67{,}500 - 60{,}500) = 61{,}730$$

Example 2: Estimate the standard error of a proportion estimate.

The proportion estimate is referred to the ratio of two total estimates:

$$\hat{R} = \frac{\hat{X}_d}{\hat{X}_p}$$

Here $\hat{X}_p$ is the total estimate of X (numeric or categorical) for population p, $\hat{X}_d$ is the total estimate of X for domain d within population p. So, in general $\hat{X}_d \leq \hat{X}_p$.

Our goal is to use the GVF variance estimates for $\hat{X}_p$ and $\hat{X}_d$ to estimate the standard error of $\hat{R}$. When $\hat{R}$ and $\hat{X}_p$ are uncorrelated, by Wolter (2007) approximation (7.2.7):

$$V_{\hat{R}}^2 = V_{\hat{X}_d}^2 - V_{\hat{X}_p}^2$$

here $V_{\hat{R}}^2$, $V_{\hat{X}_d}^2$, and $V_{\hat{X}_p}^2$ are the relative variance of $\hat{R}$, $\hat{X}_d$, and $\hat{X}_p$ respectively. This is equivalent to:

$$var(\hat{R}) = \hat{R}^2 \left[\frac{var(\hat{X}_d)}{\hat{X}_d^2} - \frac{var(\hat{X}_p)}{\hat{X}_p^2}\right]$$

In other words, to approximate the variance of a ratio estimate, we can first find variance estimates: $var(\hat{X}_d)$ and $var(\hat{X}_p)$, then use the above formula to find the variance and the standard error of the ratio estimate $\hat{R}$.

Recall that an important criterion we used for GVF is that the relative variance V^2 is a decreasing function of the magnitude of the expectation X. Because of this and notice $\hat{X}_d \leq \hat{X}_p$ we have:

$$V_{\hat{R}}^2 = V_{\hat{X}_d}^2 - V_{\hat{X}_p}^2 \geq 0$$

Hence,

$$var(\hat{R}) = \hat{R}^2 \left[\frac{var(\hat{X}_d)}{\hat{X}_d^2} - \frac{var(\hat{X}_p)}{\hat{X}_p^2}\right] \geq 0$$

This indicates the above variance estimator for the proportion estimates normally produce non-negative variance estimates.

The GVF for the standard error estimate of the proportion estimate is:

$$ste(\hat{R}) = \hat{R}\sqrt{\frac{var(\hat{X}_d)}{\hat{X}_d^2} - \frac{var(\hat{X}_p)}{\hat{X}_p^2}}$$

In Example 1, it is estimated there were $\hat{X}_d = 817{,}573$ “hit-and-run” vehicles in 2018. This comprises of 6.7854% of total vehicles involved in a police reported crash (X_p=12,049,038 – the summation of the weights of all sampled vehicles). To estimate the associated standard error of this proportion estimate $\hat{R} = 6.7854\%$, notice:

$$var(\hat{X}_d) = ste^2(X_d) = 61{,}756^2$$

$$var(\hat{X}_p) = \left[e^{1.69299+0.44262*\ln(12{,}049{,}038)+0.01787*(\ln(12{,}049{,}038))^2}\right]^2 = 856{,}137^2$$

$$ste(\hat{R}) = 6.7854\% * \sqrt{\left(\frac{61{,}756}{817{,}573}\right)^2 - \left(\frac{856{,}137}{12{,}049{,}038}\right)^2} \approx 0.17\%$$

compared with the actual standard error estimate 0.42% calculated from SAS PROC SURVEY procedure. This example reminds us the GVFs give ballpark estimates.

6 References

National Highway Traffic Safety Administration. (2017). *Traffic safety facts 2015* (Report No. DOT HS 812 384). Available at https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812384

Rosén, B. (1997). On sampling with probability proportional to size. *Journal of Statistical Planning and Inference, Vol. 62*, pp. 159-191.

Wolter, K. (2007). *Introduction to variance estimation*. Springer-Verlag New York, Inc.

Zhang, F., Subramanian, R., Chen, C.-L., & Noh, E. Y. (2018, March). *Crash Report Sampling System: Design overview, analytic guidance, and FAQs* (Report No. DOT HS 812 509). National Highway Traffic Safety Administration. Available at https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812509

Zhang, F., Noh, E. Y., Subramanian, R., & Chen, C.-L. (2019, May). *Crash Report Sampling System: Sample design and weighting* (Report No. DOT HS 812 706). National Highway Traffic Safety Administration. Available at https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812706

APPENDIX A: SAS PROC REG Output for Model Comparison

The following is the SAS PROC REG output for models 3 to 5 using 2016-2017 combined estimates.

Model 3:

The SAS System

MODEL 3: ln[ste(X)]=a + b*ln(X)

The REG Procedure
Model: yhat
Dependent Variable: log_se

Number of Observations Read	1059
Number of Observations Used	1059

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	3954.18733	3954.18733	39200.3	<.0001
Error	1057	106.62111	0.10087		
Corrected Total	1058	4060.80844			

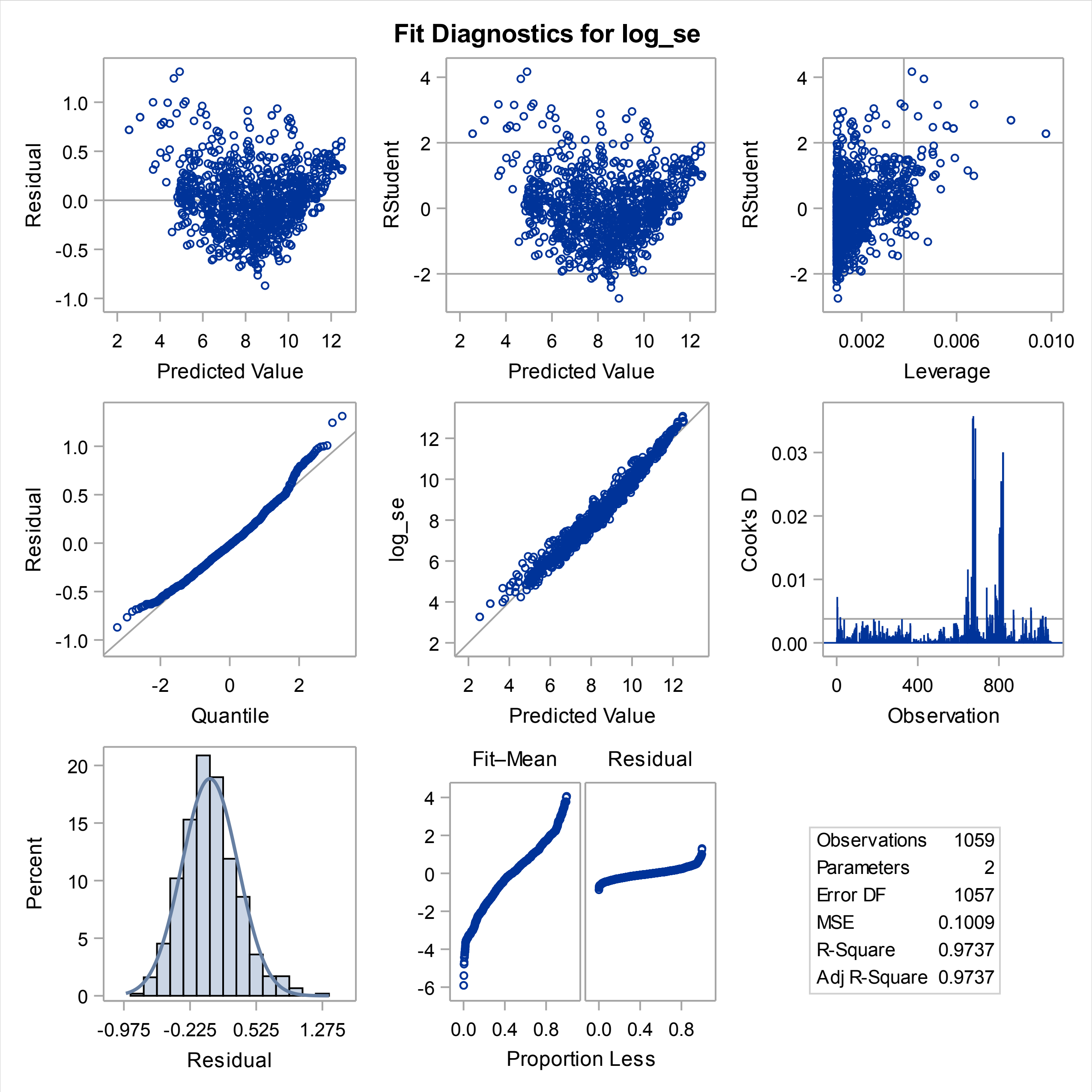
Root MSE	0.31760	**R-Square**	0.9737
Dependent Mean	8.45861	**Adj R-Sq**	0.9737
Coeff Var	3.75478		

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	-0.06256	0.04413	-1.42	0.1566
log_est	1	0.79950	0.00404	197.99	<.0001

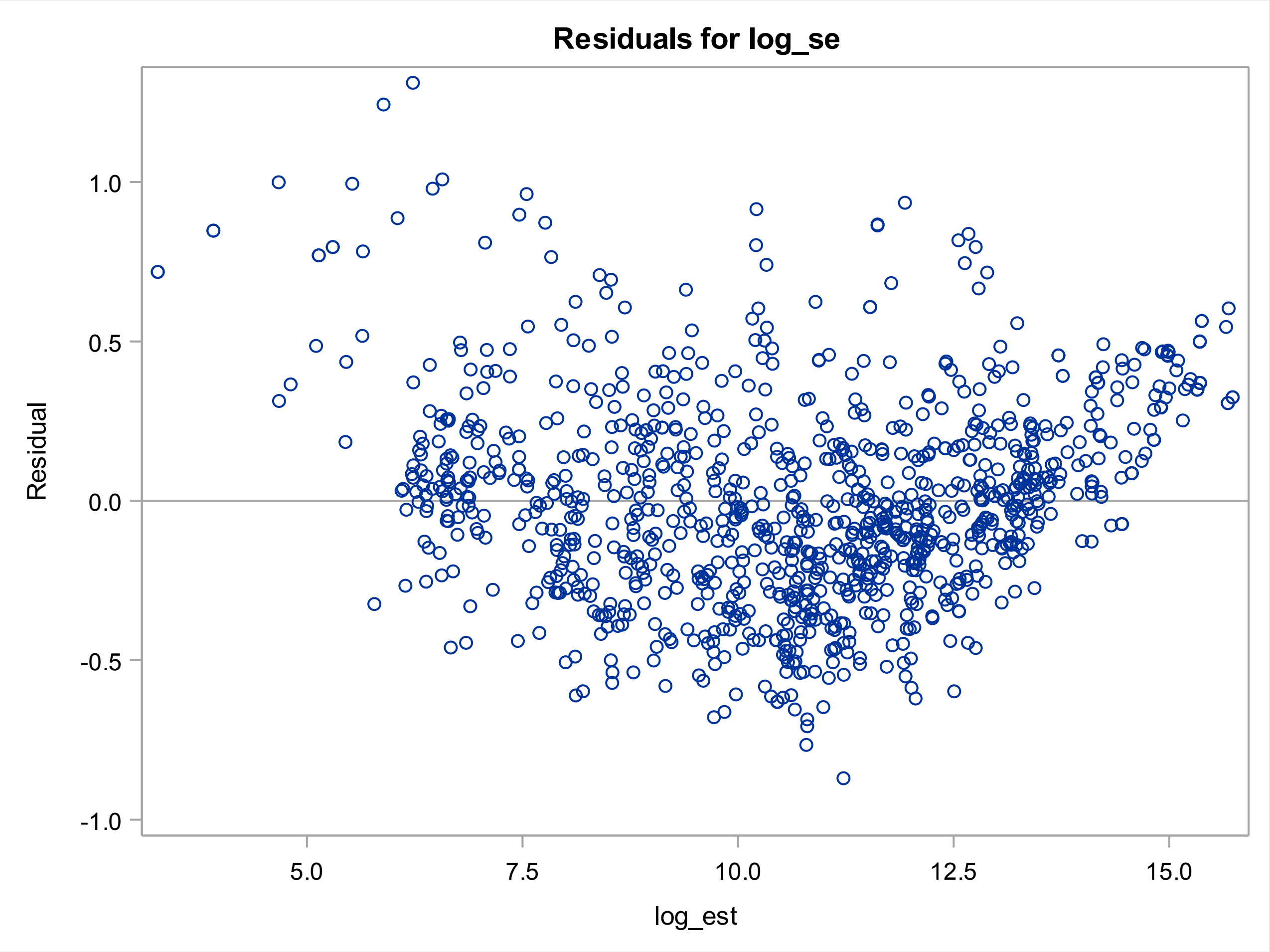
The SAS System

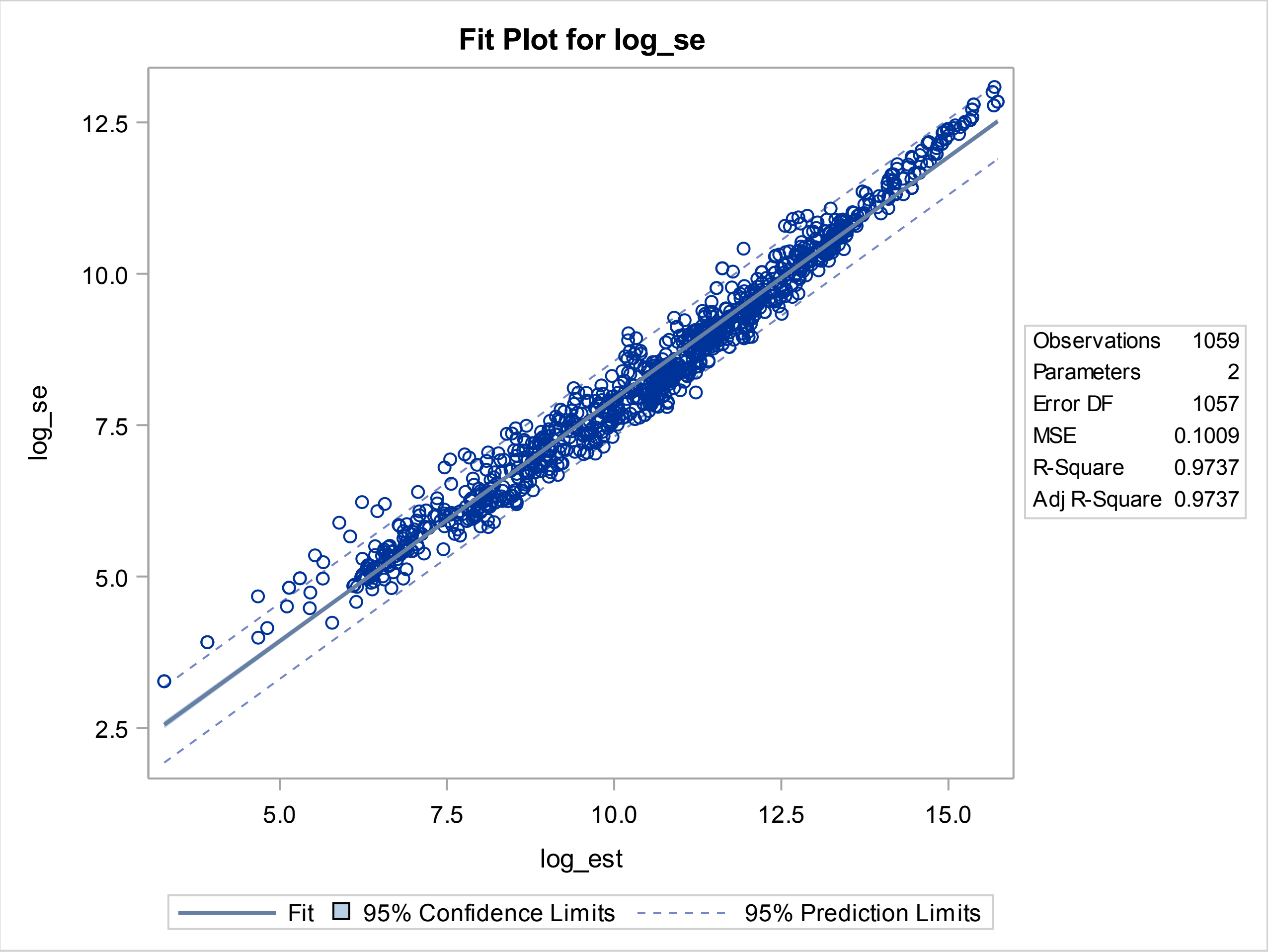
MODEL 3: In[ste(X)]=a + b*In(X)

The REG Procedure
Model: yhat
Dependent Variable: log_se



Observations	1059
Parameters	2
Error DF	1057
MSE	0.1009
R-Square	0.9737
Adj R-Square	0.9737


Residuals for log_se
Residual
1.0
0.5
0.0
-0.5
-1.0
5.0
7.5
10.0
12.5
15.0
log_est


Fit Plot for log_se
log_se
12.5
10.0
7.5
5.0
2.5
5.0
7.5
10.0
12.5
15.0
log_est
Observations 1059
Parameters 2
Error DF 1057
MSE 0.1009
R-Square 0.9737
Adj R-Square 0.9737
Fit
95% Confidence Limits
95% Prediction Limits

Model 4:

The SAS System

MODEL 4: In[ste(X)]=a + b*In(X)^2

The REG Procedure
Model: yhat
Dependent Variable: log_se

Number of Observations Read	1059
Number of Observations Used	1059

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	3963.28185	3963.28185	42954.3	<.0001
Error	1057	97.52660	0.09227		
Corrected Total	1058	4060.80844			

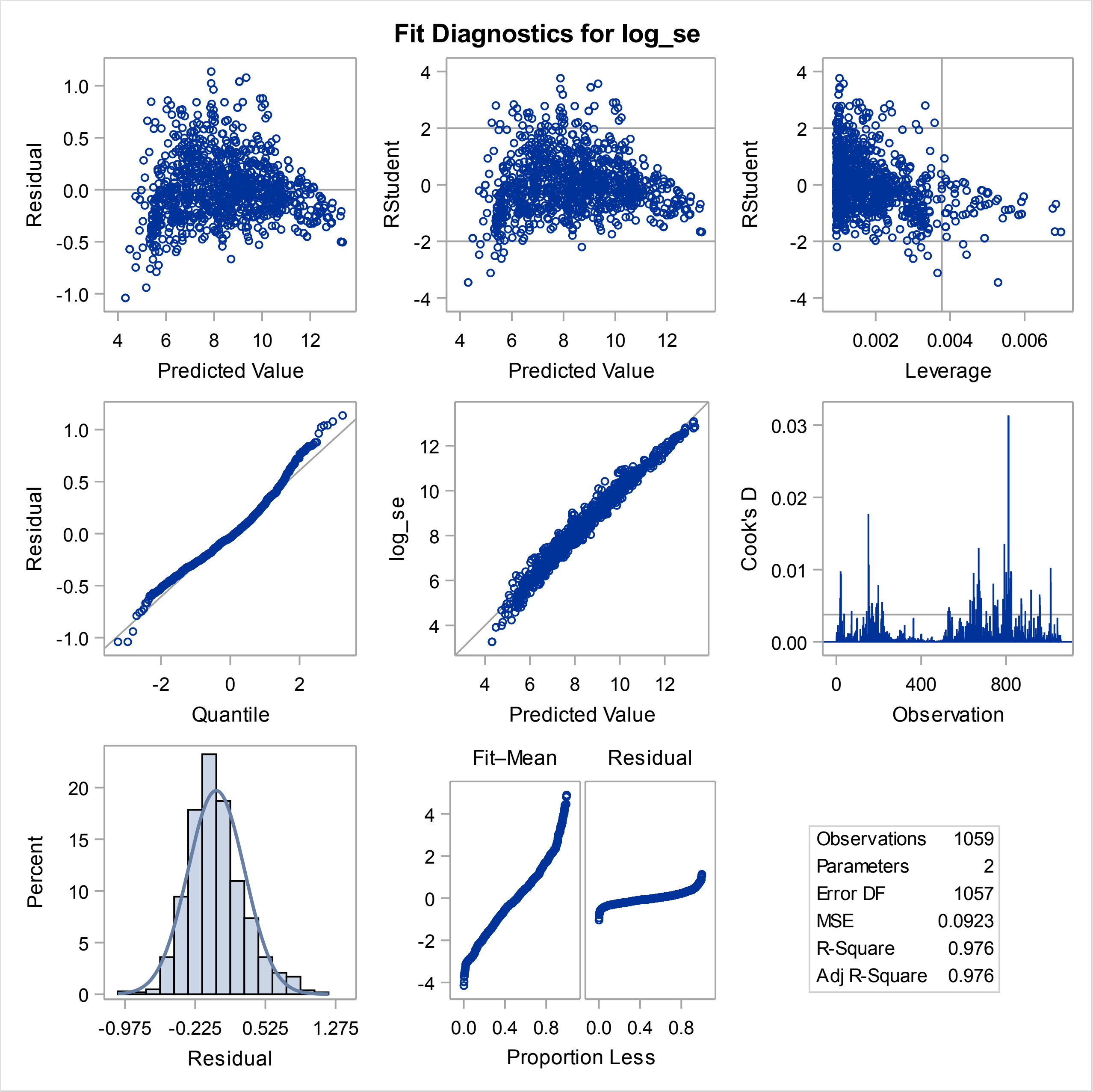
Root MSE	0.30376	**R-Square**	0.9760
Dependent Mean	8.45861	**Adj R-Sq**	0.9760
Coeff Var	3.59108		

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	3.90257	0.02388	163.41	<.0001
log2_est	1	0.03815	0.00018405	207.25	<.0001

The SAS System

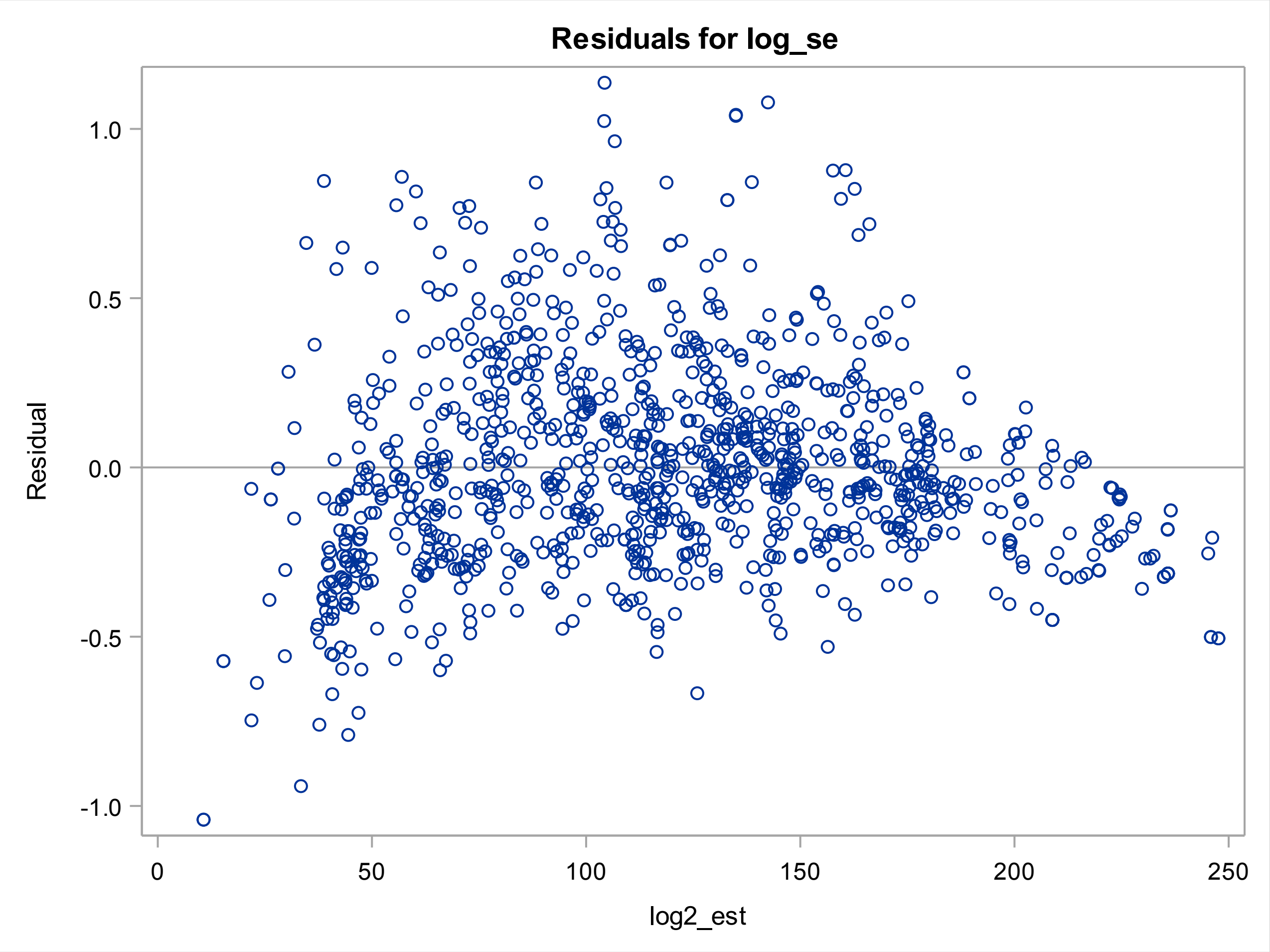
MODEL 4: ln[ste(X)]=a + b*ln(X)^2

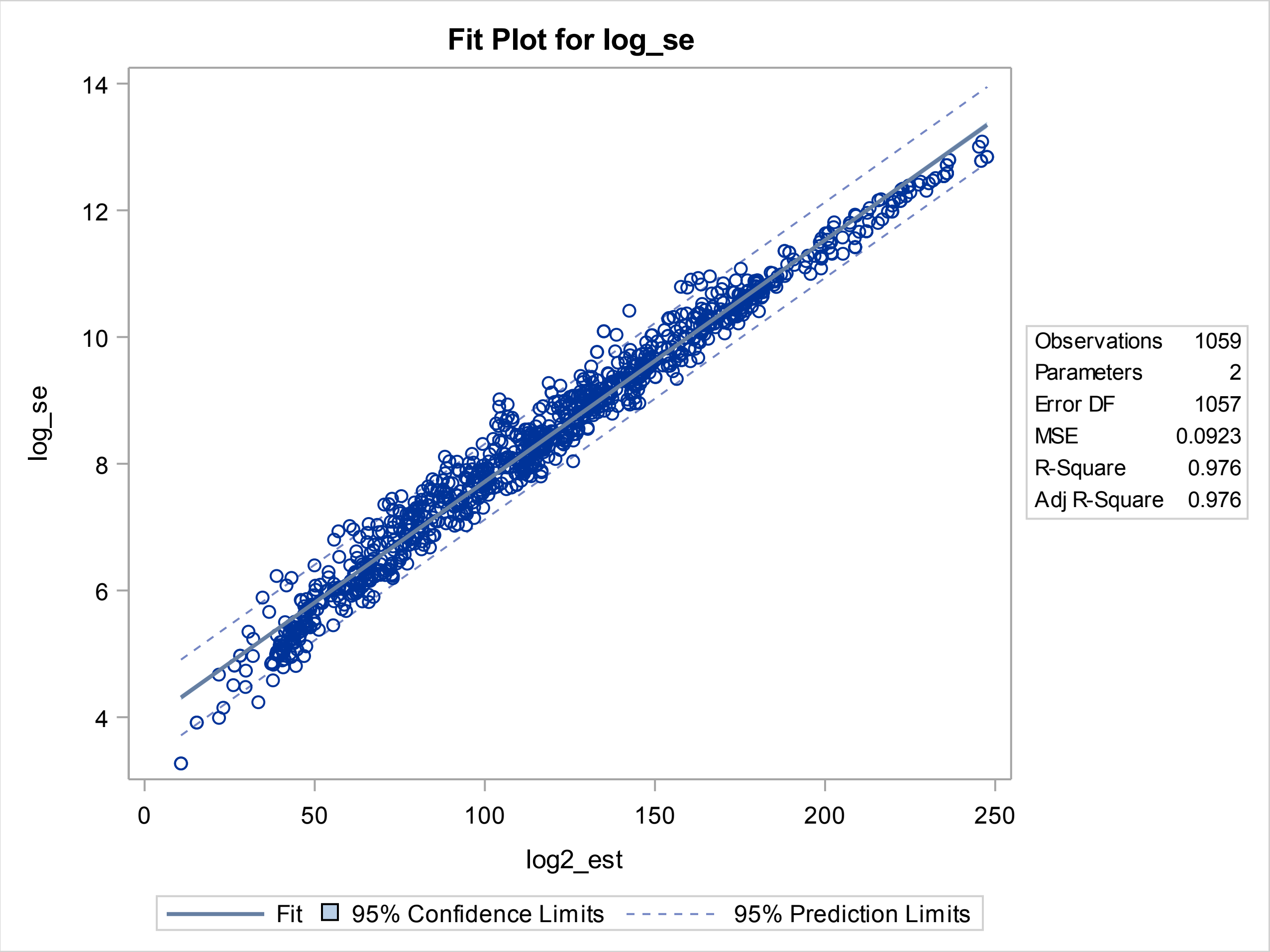
The REG Procedure
Model: yhat
Dependent Variable: log_se



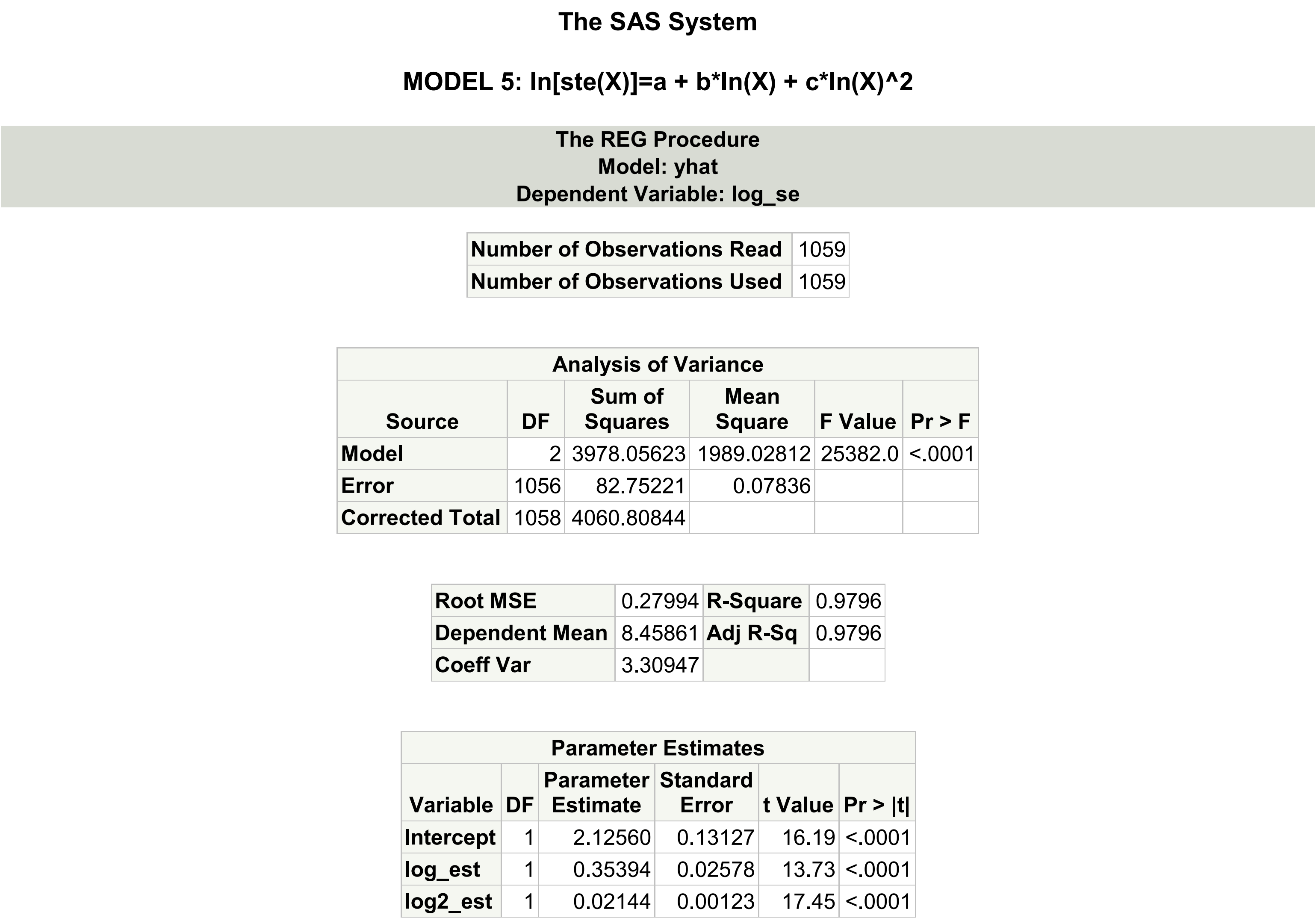
Fit Diagnostics for log_se

Observations	1059
Parameters	2
Error DF	1057
MSE	0.0923
R-Square	0.976
Adj R-Square	0.976


Residuals for log_se
Residual
1.0
0.5
0.0
-0.5
-1.0
0
50
100
150
200
250
log2_est


Fit Plot for log_se
log_se
14
12
10
8
6
4
0
50
100
150
200
250
log2_est
Observations 1059
Parameters 2
Error DF 1057
MSE 0.0923
R-Square 0.976
Adj R-Square 0.976
Fit
□ 95% Confidence Limits
95% Prediction Limits

Model 5:



The SAS System

MODEL 5: ln[ste(X)]=a + b*ln(X) + c*ln(X)^2

The REG Procedure
Model: yhat
Dependent Variable: log_se

Number of Observations Read	1059
Number of Observations Used	1059

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	3978.05623	1989.02812	25382.0	<.0001
Error	1056	82.75221	0.07836		
Corrected Total	1058	4060.80844			

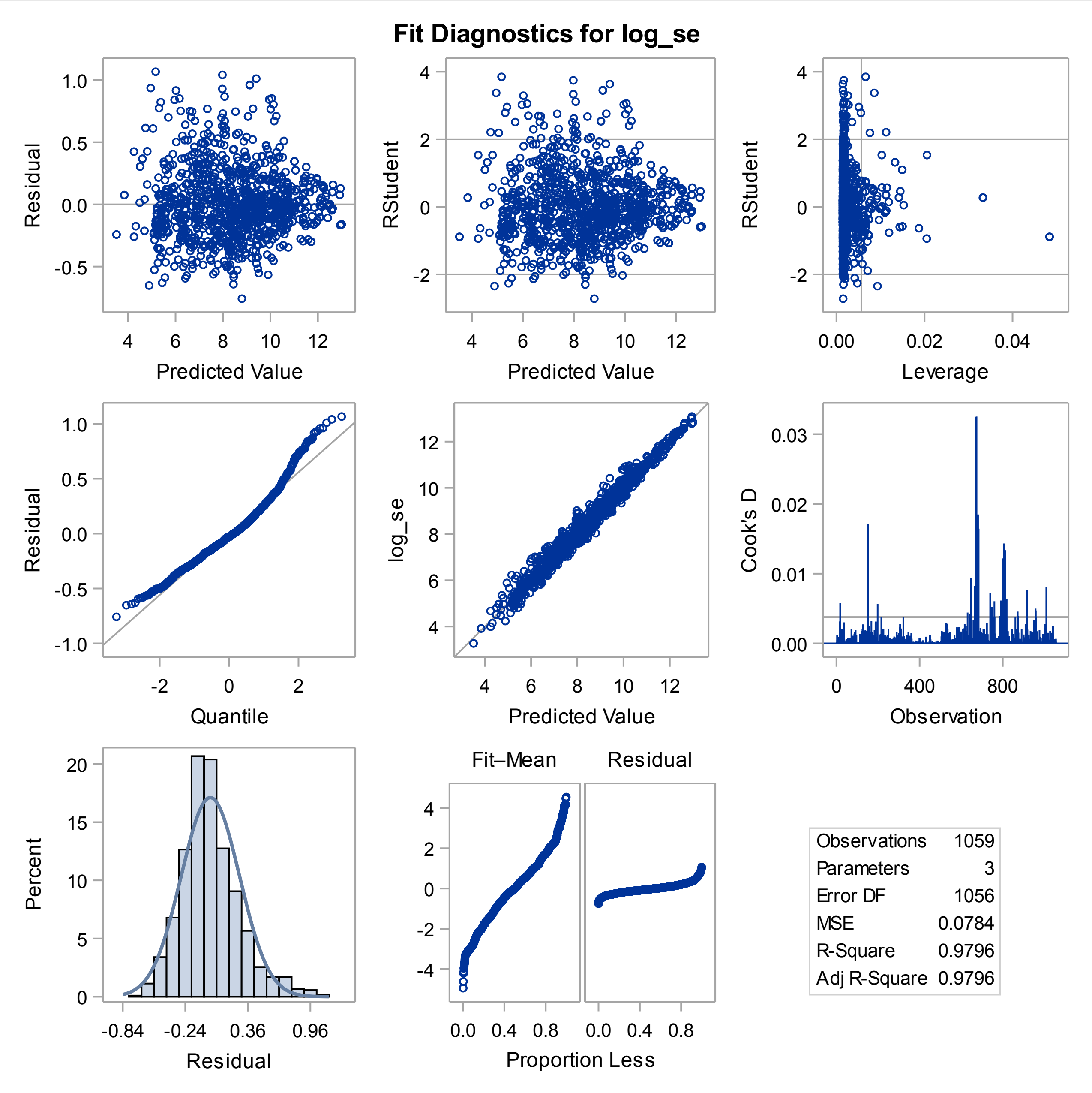
Root MSE	0.27994	**R-Square**	0.9796
Dependent Mean	8.45861	**Adj R-Sq**	0.9796
Coeff Var	3.30947		

Parameter Estimates

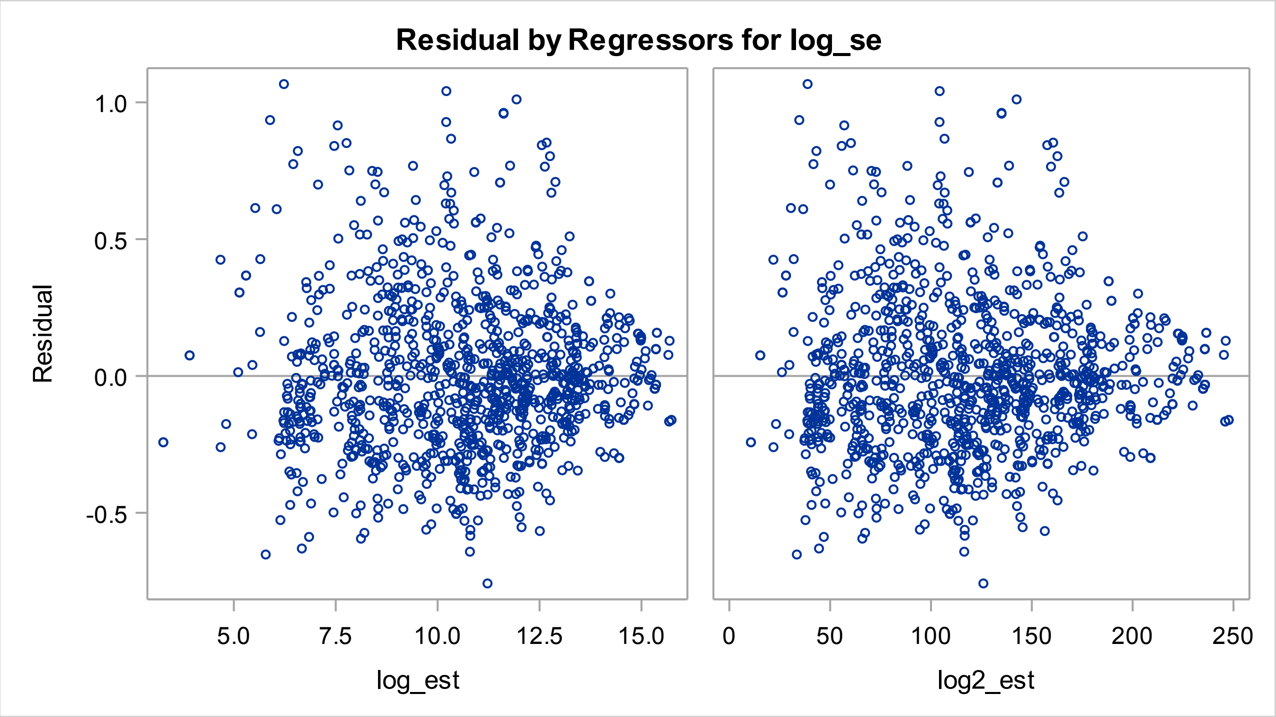
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	2.12560	0.13127	16.19	<.0001
log_est	1	0.35394	0.02578	13.73	<.0001
log2_est	1	0.02144	0.00123	17.45	<.0001

MODEL 5: ln[ste(X)]=a + b*ln(X) + c*ln(X)^2

The REG Procedure
Model: yhat
Dependent Variable: log_se


Fit Diagnostics for log_se

Observations	1059
Parameters	3
Error DF	1056
MSE	0.0784
R-Square	0.9796
Adj R-Square	0.9796


Residual by Regressors for log_se
Residual
1.0
0.5
0.0
-0.5
5.0
7.5
10.0
12.5
15.0
log_est
0
50
100
150
200
250
log2_est

APPENDIX B: CRSS GVF Major SAS Programs

Categorical Variable Point and Variance Estimates

```
PROC SQL;
        CREATE TABLE DATAACC AS SELECT DISTINCT
                A.*,
                B.A_CT
        FROM CRSS2018.ACCIDENT A
        LEFT JOIN CRSS2018.ACC_AUX B
        ON (A.CASENUM = B.CASENUM);
QUIT;

data sudtab28;
  set DATAACC;
        format _ALL_;

        IF A_CT = 1 THEN CRASH_TYPE= 1;                     /* Single Veh  */
        ELSE IF A_CT IN (2,3) THEN CRASH_TYPE= 2;           /* Multi Veh   */
        ELSE CRASH_TYPE = 0;
        IF REL_ROAD IN (1,11) THEN ROADWAY = 1;             /* On Roadway  */
        IF REL_ROAD IN (4,5,6) THEN ROADWAY = 2;            /* Off-roadway */
        IF REL_ROAD IN (2) THEN ROADWAY = 3;                /* Shoulder            */
        IF REL_ROAD IN (3) THEN ROADWAY = 4;                /* Median              */
        IF REL_ROAD IN (7,8,10,98,99) THEN ROADWAY = 9;     /* Other               */

        IF MAXSEV_IM IN (1,2,3,5) THEN LEVEL2 = 1;          /* INJURY CRASHES */
        ELSE IF MAXSEV_IM IN (0,6,8) THEN LEVEL2 = 2;       /* PROPERTY-DAMAGE-ONLY CRASHES */
        ELSE LEVEL2 = 0;                                    /* OTHER CRASHES */
run;

proc sort data= sudtab28;
  by psustrat PSU_VAR casenum;
run;

proc surveyfreq data= sudtab28 VARMETHOD=JK;
        title1 "National Highway Traffic Safety Administration";
        title3 "&DATASET. -- Crashes Table &_tabnum -- Computed by PROC SURVEYFREQ";
        title5 "CRASH BY CRASTH TYPE, RELATION TO ROADWAY, AND CRASH SEVERITY";
        cluster PSU_VAR;
        strata psustrat;
        weight weight;
        tables LEVEL2*CRASH_TYPE*ROADWAY;
        ods output crosstabs=sewgts2;
run;

data stderr28 (keep=by_lvl row_lvl col_lvl tabl_num nsum wsum sewgt log_se log2_est);
  length by_lvl row_lvl col_lvl tabl_num nsum wsum sewgt log_se log2_est 8;
  set sewgts2;

  by_lvl = LEVEL2;
  row_lvl = CRASH_TYPE;
  col_lvl = ROADWAY;
  nsum = frequency;
  wsum = wgtfreq;
  sewgt = stddev;

  if (row_lvl eq .) then row_lvl = 0;
  if (col_lvl eq .) then col_lvl = 0;

  log_se = log(sewgt);
  est = wsum;
  log_est = log(wsum);
  log2_est = (log(wsum))**2;
run;
```

Numerical Variable Point and Variance Estimates

```
proc surveymeans data= sudtab28 SUM SUMWGT VARMETHOD=jk;
        title1 "National Highway Traffic Safety Administration";
```

```
       title3 "&DATASET. -- Crashes Table &_tabnum -- Computed by PROC SURVEYFREQ";
       cluster PSU_VAR;
       strata psustrat;
       weight weight;
       DOMAIN MANCOL_IM;
       VAR VE_FORMS;
       ods output DOMAIN=sewgts2;
run;

data stderr_VE_FORMS (keep= by_lvl row_lvl col_lvl  nsum wsum sewgt log_se log2_est);
  length by_lvl row_lvl col_lvl  nsum wsum sewgt log_se log2_est 8;
  set sewgts2;

  by_lvl = .;
  col_lvl = MANCOL_IM;
  row_lvl = .;
  nsum = sum;
  wsum = sum;
  sewgt = stddev;

  if (row_lvl eq .) then row_lvl = 0;
  if (col_lvl eq .) then col_lvl = 0;

  log_se = log(sewgt);
  est = wsum;
  log_est = log(wsum);
  log2_est = (log(wsum))**2;

run;

proc sort data= stderr_VE_FORMS;
  by by_lvl row_lvl col_lvl;
run;
```

Final Model Fitting (crash level data)

```
data LOG_EM_CRASH_1617_3;
      set GVF2016.stderr24 GVF2016.stderr25 GVF2016.stderr26 GVF2016.stderr28
GVF2016.stderr29 GVF2016.stderr30 GVF2017.stderr24 GVF2017.stderr25
GVF2017.stderr26 GVF2017.stderr28 GVF2017.stderr29 GVF2017.stderr30
GVF2016.stderr_VE_TOTAL GVF2016.stderr_VE_FORMS GVF2016.stderr_PERMVIT
GVF2016.stderr_NO_INJ_IM GVF2016.stderr_PEDS GVF2016.stderr_PERNOTMVIT
GVF2016.stderr_PVH_INVL GVF2017.stderr_VE_TOTAL GVF2017.stderr_VE_FORMS
GVF2017.stderr_PERMVIT GVF2017.stderr_NO_INJ_IM GVF2017.stderr_PEDS
GVF2017.stderr_PERNOTMVIT GVF2017.stderr_PVH_INVL;

  /* DELETE RECORDS WITH SAMPLE SIZE <= 15 AS THESE ESTIMATES ARE UNSTABLE */
      IF NSUM <= 15 THEN DELETE;

      log_se = log(sewgt);
      sqrt_se = sqrt(sewgt);
      inv_wsum = 1/wsum;
      inv_wsum2 = 1/wsum**2;
      log_est = log(wsum);
      log2_est = (log(wsum))**2;

      v2 = sewgt**2/wsum**2;
      v2inv = wsum**2/sewgt**2;
      logv2 = log(v2);

      wsum2 = wsum**2;
      inv_se = 1/sewgt;
      sewgt2=sewgt**2;
run;

proc sort data=LOG_EM_CRASH_1617_3;
```

```
      by tabl_num by_lvl row_lvl col_lvl;
run;

MODEL 5: LOG-LOG-QUADRATIC LOG

PROC REG DATA=LOG_EM_CRASH_1617_3;
      MODEL log_se = log_est log2_est;
RUN;
```

APPENDIX C: CRSS GVF Standard Error Estimates

2016 CRSS Estimates and GVF Standard Error Estimates					
Crash		Vehicle		Person	
Estimate (X)	Standard Error*	Estimate (X)	Standard Error*	Estimate (X)	Standard Error*
1,000	300	1,000	300	1,000	200
5,000	800	5,000	900	5,000	800
6,000	900	10,000	1,500	10,000	1,300
7,000	1,000	20,000	2,700	20,000	2,200
8,000	1,100	30,000	3,800	30,000	3,100
9,000	1,200	40,000	4,800	40,000	3,900
10,000	1,300	50,000	5,800	50,000	4,700
20,000	2,200	60,000	6,800	60,000	5,400
30,000	3,000	70,000	7,700	70,000	6,200
40,000	3,700	80,000	8,700	80,000	6,900
50,000	4,400	90,000	9,600	90,000	7,600
60,000	5,200	100,000	10,500	100,000	8,300
70,000	5,800	200,000	19,300	200,000	15,100
80,000	6,500	300,000	27,800	300,000	21,700
90,000	7,200	400,000	36,000	400,000	28,000
100,000	7,900	500,000	44,100	500,000	34,300
200,000	14,200	600,000	52,100	600,000	40,600
300,000	20,200	700,000	60,000	700,000	46,800
400,000	26,000	800,000	67,900	800,000	53,000
500,000	31,700	900,000	75,700	900,000	59,100
600,000	37,400	1,000,000	83,500	1,000,000	65,300
700,000	43,000	2,000,000	160,500	2,000,000	126,300
800,000	48,600	3,000,000	236,700	3,000,000	187,500
900,000	54,200	4,000,000	312,800	4,000,000	249,100
1,000,000	59,700	5,000,000	388,800	5,000,000	311,200
2,000,000	114,500	6,000,000	464,900	6,000,000	373,800
3,000,000	169,000	7,000,000	541,200	7,000,000	436,900
4,000,000	223,600	8,000,000	617,700	8,000,000	500,500
5,000,000	278,600	9,000,000	694,300	9,000,000	564,500
6,000,000	333,800	10,000,000	771,200	10,000,000	629,000
6,500,000	361,500	11,000,000	848,300	11,000,000	693,800
7,000,000	389,300	12,000,000	925,500	12,000,000	759,200
*: $ste(X) = e^{a+bln(X)+cln(X)^2}$					
a = 1.92772 b = 0.38750 c = 0.01947		a = 1.17146 b = 0.53866 c = 0.01425		a = 1.79032 b = 0.40622 c = 0.01930	

2017 CRSS Estimates and GVF Standard Error Estimates					
Crash		Vehicle		Person	
Estimate (X)	Standard Error*	Estimate (X)	Standard Error*	Estimate (X)	Standard Error*
1,000	300	1,000	300	1,000	200
5,000	800	5,000	900	5,000	700
6,000	900	10,000	1,500	10,000	1,200
7,000	1,000	20,000	2,600	20,000	2,100
8,000	1,100	30,000	3,600	30,000	2,800
9,000	1,200	40,000	4,600	40,000	3,500
10,000	1,300	50,000	5,500	50,000	4,200
20,000	2,200	60,000	6,500	60,000	4,900
30,000	3,000	70,000	7,400	70,000	5,600
40,000	3,800	80,000	8,300	80,000	6,200
50,000	4,500	90,000	9,100	90,000	6,900
60,000	5,200	100,000	10,000	100,000	7,500
70,000	5,900	200,000	18,400	200,000	13,600
80,000	6,600	300,000	26,400	300,000	19,400
90,000	7,300	400,000	34,200	400,000	25,100
100,000	8,000	500,000	41,900	500,000	30,700
200,000	14,600	600,000	49,600	600,000	36,300
300,000	20,900	700,000	57,200	700,000	41,800
400,000	27,100	800,000	64,700	800,000	47,300
500,000	33,300	900,000	72,200	900,000	52,800
600,000	39,400	1,000,000	79,700	1,000,000	58,300
700,000	45,500	2,000,000	153,900	2,000,000	112,900
800,000	51,700	3,000,000	227,900	3,000,000	167,700
900,000	57,800	4,000,000	302,000	4,000,000	223,000
1,000,000	63,900	5,000,000	376,400	5,000,000	278,900
2,000,000	125,300	6,000,000	451,200	6,000,000	335,300
3,000,000	187,800	7,000,000	526,300	7,000,000	392,300
4,000,000	251,400	8,000,000	601,800	8,000,000	449,700
5,000,000	316,100	9,000,000	677,700	9,000,000	507,700
6,000,000	381,700	10,000,000	753,900	10,000,000	566,100
6,500,000	414,900	11,000,000	830,500	11,000,000	625,000
7,000,000	448,400	12,000,000	907,400	12,000,000	684,300
*: $ste(X) = e^{a+bln(X)+cln(X)^2}$					
a = 2.33171 b = 0.30826 c = 0.02344		a = 1.43152 b = 0.48824 c = 0.01629		a = 2.05394 b = 0.35287 c = 0.02119	

2018 CRSS Estimates and GVF Standard Error Estimates					
Crash		Vehicle		Person	
Estimate (X)	Standard Error*	Estimate (X)	Standard Error*	Estimate (X)	Standard Error*
1,000	300	1,000	300	1,000	200
5,000	800	5,000	900	5,000	700
6,000	900	10,000	1,500	10,000	1,200
7,000	1,000	20,000	2,500	20,000	2,000
8,000	1,100	30,000	3,500	30,000	2,800
9,000	1,200	40,000	4,400	40,000	3,500
10,000	1,300	50,000	5,300	50,000	4,100
20,000	2,100	60,000	6,200	60,000	4,800
30,000	2,900	70,000	7,000	70,000	5,400
40,000	3,700	80,000	7,800	80,000	6,100
50,000	4,400	90,000	8,700	90,000	6,700
60,000	5,100	100,000	9,500	100,000	7,300
70,000	5,800	200,000	17,300	200,000	13,200
80,000	6,400	300,000	24,800	300,000	18,800
90,000	7,100	400,000	32,100	400,000	24,200
100,000	7,700	500,000	39,300	500,000	29,600
200,000	14,000	600,000	46,400	600,000	34,900
300,000	19,900	700,000	53,500	700,000	40,200
400,000	25,700	800,000	60,500	800,000	45,400
500,000	31,500	900,000	67,500	900,000	50,700
600,000	37,200	1,000,000	74,500	1,000,000	55,900
700,000	42,800	2,000,000	143,800	2,000,000	107,600
800,000	48,500	3,000,000	213,000	3,000,000	159,400
900,000	54,100	4,000,000	282,500	4,000,000	211,400
1,000,000	59,700	5,000,000	352,300	5,000,000	263,900
2,000,000	115,700	6,000,000	422,500	6,000,000	316,800
3,000,000	172,100	7,000,000	493,200	7,000,000	370,100
4,000,000	229,200	8,000,000	564,300	8,000,000	423,800
5,000,000	286,900	9,000,000	635,700	9,000,000	477,900
6,000,000	345,300	10,000,000	707,600	10,000,000	532,300
6,500,000	374,700	11,000,000	779,900	11,000,000	587,200
7,000,000	404,300	12,000,000	852,600	12,000,000	642,400
*: $ste(X) = e^{a+bln(X)+cln(X)^2}$					
a = 2.33242 b = 0.31521 c = 0.02258		a = 1.69299 b = 0.44262 c = 0.01787		a = 2.02774 b = 0.35777 c = 0.02075	

2019 CRSS Estimates and GVF Standard Error Estimates					
Crash		Vehicle		Person	
Estimate (X)	Standard Error*	Estimate (X)	Standard Error*	Estimate (X)	Standard Error*
1,000	300	1,000	300	1,000	200
5,000	800	5,000	900	5,000	700
6,000	900	10,000	1,400	10,000	1,100
7,000	1,000	20,000	2,500	20,000	1,900
8,000	1,100	30,000	3,500	30,000	2,700
9,000	1,200	40,000	4,400	40,000	3,300
10,000	1,200	50,000	5,300	50,000	4,000
20,000	2,100	60,000	6,100	60,000	4,600
30,000	2,900	70,000	7,000	70,000	5,300
40,000	3,600	80,000	7,800	80,000	5,900
50,000	4,300	90,000	8,600	90,000	6,500
60,000	5,000	100,000	9,500	100,000	7,100
70,000	5,700	200,000	17,300	200,000	12,800
80,000	6,400	300,000	24,800	300,000	18,400
90,000	7,000	400,000	32,200	400,000	23,800
100,000	7,700	500,000	39,400	500,000	29,100
200,000	13,800	600,000	46,600	600,000	34,400
300,000	19,700	700,000	53,800	700,000	39,700
400,000	25,500	800,000	60,900	800,000	44,900
500,000	31,200	900,000	68,000	900,000	50,200
600,000	36,900	1,000,000	75,100	1,000,000	55,400
700,000	42,500	2,000,000	145,500	2,000,000	107,800
800,000	48,100	3,000,000	215,900	3,000,000	160,700
900,000	53,600	4,000,000	286,900	4,000,000	214,200
1,000,000	59,200	5,000,000	358,300	5,000,000	268,500
2,000,000	114,700	6,000,000	430,200	6,000,000	323,400
3,000,000	170,400	7,000,000	502,700	7,000,000	378,900
4,000,000	226,800	8,000,000	575,700	8,000,000	435,100
5,000,000	283,700	9,000,000	649,100	9,000,000	491,800
6,000,000	341,200	10,000,000	723,100	10,000,000	549,000
6,500,000	370,200	11,000,000	797,500	11,000,000	606,800
7,000,000	399,300	12,000,000	872,300	12,000,000	665,100
*: $ste(X) = e^{a+bln(X)+cln(X)^2}$					
a = 2.19494 b = 0.33465 c = 0.02185		a = 1.70176 b = 0.43713 c = 0.01826		a = 2.14416 b = 0.32619 c = 0.02238	

DOT HS 813 041
December 2020

U.S. Department
of Transportation

National Highway
Traffic Safety
Administration

NHTSA