EVALUATION OF GENERALIZED VARIANCE FUNCTION ESTIMATORS FOR THE U.S. CURRENT EMPLOYMENT SURVEY

Moon J. Cho, John L. Eltinge, Julie Gershunskaya, Larry Huff,
U.S. Bureau of Labor Statistics
Moon J. Cho, 2 Massachusetts Avenue NE, Washington, DC 20212
(Cho_M@bls.gov)

Key Words: Complex sample design, Covered Employment and Wages Program (ES-202), Design-based inference, Generalized least squares, Model-based inference, Superpopulation model, Variance estimator stability.

1. Introduction

In applied work with generalized variance function models for sample survey data, one generally seeks to develop and validate a model that is relatively parsimonious and that produces variance estimators that are approximately unbiased and relatively stable. This development and validation work often begins with regression of initial variance estimators (computed through standard design-based methods) on one or more candidate explanatory variables.

Evaluation of the adequacy of the resulting regression fit depends heavily on the relative magnitudes of error terms associated, respectively, with pure sampling variability of the initial design-based variance estimators; the deterministic lack of fit in the proposed generalized variance function model; and the random equation error associated with the generalized variance function model. This paper presents some simple methods of evaluating the relative magnitudes of the sampling error and equation error terms. Both parametric and nonparametric regression methods are used in producing smoothed estimators of the mean squared equation error in the underlying generalized variance function model fit.

Some of the proposed diagnostics are applied to data from the U.S. Current Employment Survey.

2. Variance Function Model

Define $\hat{\theta}_j$ a point estimator of θ_j, a finite population mean or total. Let $\theta_{\xi\, j}$ be a superpopulation analogue of θ_j where j is the domain index. For example, in CES survey, domains are the combinations of industries, areas and time. Define $V_{p\,j}= V_p(\hat{\theta}_j)$ as the design variance of $\hat{\theta}_j$, and $\hat{V}_{p\,j} = \hat{V}_p(\hat{\theta}_j)$ as an estimator of $V_{p\,j}$. Throughout this paper, the subscript "p" denotes the method to obtain an expectation or variance evaluated with respect to the sample design.

The generalized variance function method models the variance of a survey estimator, $V_{p\,j}$, as a function of the estimate and possibly other variables (Wolter 1985). The common specification is

$$V_{p\,j} \quad = \quad f(\theta_j, X_j, \gamma_j) + q_j \qquad (1)$$

where X_j is a vector of predictor variables potentially relevant to estimators of $V_{p\,j}$, q_j is a univariate "equation error" with the mean 0, and γ_j is a vector of variance function parameters which we need to estimate. Note especially that q_j represents the deviation of $V_{p\,j}$ from its modeled value $f(\theta_j, X_j, \gamma_j)$.

3. Current Employment Survey Data

The CES survey collects data on employment, hours, and earnings from 400,000 nonfarm establishments monthly. Employment is the total number of persons employed full or part time in a nonfarm establishment during a specified payroll period. An establishment, which is an economic unit, is generally located at a single location, and is engaged predominantly in one type of economic activity (BLS Handbook,1997). This paper will focus only on total employment in the reporting establishment.

One important feature of the CES program is that complete universe employment counts of the previous year become available from the Unemployment Insurance tax records on a lagged basis (Butani, Stamas and Brick, 1997). The quarterly unemployment insurance files are generally transmitted five months after the end of the quarter by the states to BLS. It takes BLS an additional 3 months to process these files through various edits as well as perform

The views expressed in this paper are those of the authors and do not necessarily reflect the policies of the U.S. Bureau of Labor Statistics.

record linkage to previous quarters before making it available as the sampling frame (Butani, Stamas and Brick, 1997). This data known as ES202 data are used annually to benchmark the CES sample estimates to these universe counts (Werking, 1997). Using the benchmark data, x_{ia0}, at the base period from ES202 data, the CES program obtains weighted link relative estimator, $\hat{y}_{iat}$, to estimate the total employment, x_{iat}, within the industry i, area a and month t,

$$\hat{y}_{iat} = x_{ia0}\hat{R}_{iat}$$

where $\hat{R}_{iat}$ is the growth ratio estimate from benchmark month 0 to current month t.

The CES sample design uses stratified sampling of Unemployment Insurance (UI) accounts with strata defined by state, industry and employment size class (BLS Handbook, 1997). CES aims primarily at meeting the requirements for the national estimates. As for finer domains which are defined by geographic characteristics, and industrial classifications, effective sample sizes within occupational classifications become so small that the standard design based estimators are not precise enough to satisfy the needs of prospective data users (Eltinge, Fields, Fisher, Gershunskaya, Getz, Huff, Tiller and Waddington, 2001). It is necessary to have stable estimates of $V(\hat{y}_{iat})$ for the finer domains.

4. Model Fitting

We used the direct variance estimators from the survey as the dependent variables in GVF models. In the CES survey, we have direct estimators, $\hat{V}_{p\,j}$ of $V_{p\,j}$, from Fay's method which is a variant of the balanced half-samples replication methods. Each replicate half-sample estimate is formed based on a Hadamard matrix. In the standard balanced half-samples replication methods, only the selected ones are used to estimate the variance, and the weights for the selected units are multiplied by a factor 2 to form the weights for the replicate estimate (Wolter, 1985). However, in Fay's method, one-half of the sample is weighted down by a factor $K(0 \leq K < 1)$ and the remaining half is weighted up by a compensating factor $2 - K$ (Judkins, 1990). In our CES example, $K = 0.5$.

We assume that $\hat{V}_{p\,j}$ is a design unbiased estimator for $V_{p\,j}$, i.e., $E_p(\hat{V}_{p\,j}) = V_{p\,j}$. Our sample consists of Unemployment Insurance accounts, which report nonzero employment for previous and current months. Let n_{iat} be a number of responding UI accounts within the industry i, area a and month t. In this paper, we consider only domains with at least 12 reporting UI accounts. There are 430 industry-area combinations in our CES data. Each industry-area combination has data from January to December of the year 2000. Hence we have 5160 industry-area-time combinations. For the current analysis, we considered data from the following six industries: Mining, Construction and Mining, Construction, Manufacturing Durable Goods, Manufacturing Nondurable Goods, Wholesale Trade. Consider the GVF model

$$\begin{aligned} log(\hat{V}_{iat}) &= \gamma_0 + \gamma_1 log(x_{ia0}) + \gamma_2 log(n_{iat}) \\ &\quad + \gamma_3 log(t_{ia0}) + e \ . \end{aligned} \qquad (2)$$

In this model, we assume that both intercepts and slopes are constant across the industries and areas. Various modified models can be considered from (2). For example, we may allow the intercepts to vary across industries. Further, we may allow the intercepts and the slopes to vary across industries.

5. Residual Decomposition

Suppose that a model fitting method (e.g., ordinary least squares perhaps on a transformed scale; or nonlinear least squares) leads to the point estimator $\hat{\gamma}_j$. This in turn leads to the estimated variances,

$$V^*_{p\,j} \overset{def}{=} f(\hat{\theta}_j, \hat{X}_j, \hat{\gamma}_j) \ . \qquad (3)$$

Note that $V^*_{p\,j}$ is the variance estimator based on the model, which is transformed back onto original variance scale.

From the definition of the direct variance estimator

$$\hat{V}_{p\,j} = V_{p\,j} + \epsilon_j \qquad (4)$$

where ϵ_j has a mean 0 and a constant variance. Recall the variance function model in (1),

$$V_{p\,j} = f(\theta_j, X_j, \gamma_j) + q_j$$

Then the resulting residuals are

$$\begin{aligned} &\hat{V}_{p\,j} - V^*_{p\,j} \\ &\quad = (\hat{V}_{p\,j} - V_{p\,j}) - (V^*_{p\,j} - V_{p\,j}) \\ &\quad = \epsilon_j - \{f(\hat{\theta}_j, \hat{X}_j, \hat{\gamma}_j) - f(\theta_j, X_j, \gamma_j) - q_j\} \quad (5) \\ &\quad = \epsilon_j + \{q_j - E(q_j)\} + E(q_j) \\ &\quad\quad - \{f(\hat{\theta}_j, \hat{X}_j, \hat{\gamma}_j) - f(\theta_j, X_j, \gamma_j)\} \ . \quad (6) \end{aligned}$$

In the equation above, ϵ_j is a pure estimation error in the original $\hat{V}_{p\,j}$ estimates with $E(\epsilon_j) = 0$,

$\{q_j - E(q_j)\}$ is random equation error, and $E(q_j)$ is deterministic lack-of-fit in our model attributable e.g., to omitted regressors or misspecified functional form. $\{f(\hat{\theta}_j, \hat{X}_j, \hat{\gamma}_j) - f(\theta_j, X_j, \gamma_j)\}$, the last term in (6), is a parameter estimation error attributable to errors $\{(\hat{\theta}_j, X_j, \hat{\gamma}_j) - (\theta_j, X_j, \gamma_j)\}$.

Exploratory analysis of the adequacy of our estimated values, V^*_{pj}, may focus on the magnitude of the prediction errors, $\left(V^*_{pj} - V_{pj}\right)$, relative to the errors, $\left(\hat{V}_{pj} - V_{pj}\right)$, in the original estimators $\hat{V}_{pj}$. If $E(V^*_{pj} - V_{pj})^2$ is smaller than the variance of $\hat{V}_{pj}$, then we would prefer V^*_{pj}. In some cases, we may find that

$$\delta(\theta_j, X_j, \gamma_j) \overset{def}{=} E\left[\left\{f(\hat{\theta}_j, \hat{X}_j, \hat{\gamma}_j) - V_{pj}\right\}^2 \mid \theta_j, X_j, \gamma\right]$$

varies across values of θ_j or X_j with $\delta(\theta_j, X_j, \gamma_j) << V_p(\hat{V}_{pj} - V_{pj})$ only in some cases. In this case, we might prefer V^*_{pj} for some, but not all values of X_j. This is one case in which we need to consider a variance function model for the equation errors q_j as well as possibly non constant values of the mean squared estimation errors, $E\left[\left\{f(\hat{\theta}_j, \hat{X}_j, \hat{\gamma}_j) - f(\theta_j, X_j, \gamma_j)\right\}^2 \mid \theta_j, X_j, \gamma_j\right]$.

6. Conditional Expected Squared Error

We evaluate error sizes in terms of related general measures and conditional expected squared error. We may assume that for some known d,

$$ {V_{pj}}^{-1} d \hat{V}_{pj} \overset{p}{\sim} \chi^2(d) \tag{7}$$

where $\overset{p}{\sim}$ refers to the distribution induced by the random sampling design, conditional on the finite population. Thus, $E_p(\hat{V}_{pj}) = V_{pj}$, $V_p(\hat{V}_{pj}) = 2V^2_{pj}/d$, and

$$\begin{aligned} E_p(\hat{V}^2_{pj}) &= \{E_p(\hat{V}_{pj})\}^2 + V(\hat{V}_{pj}) \\ &= V^2_{pj} + 2V^2_{pj}/d \\ &= d^{-1}(d+2)V^2_{pj} \ . \end{aligned}$$

Consequently, an unbiased estimator of $V_p(\hat{V}_{pj})$ is:

$$\hat{V}_p(\hat{V}_{pj}) = (d+2)^{-1}\, 2\hat{V}^2_{pj} \ . \tag{8}$$

For our CES survey example, six employment size classes were used for stratification. Hence we use $d = 6$.

In this section, we obtain and model the conditional squared error, $E\{(V^*_{pj} - V_{pj})^2 | X_j\}$. Consider

$$\begin{aligned} & E\{(\hat{V}_{pj} - V^*_{pj})^2 | X_j\} \\ & \quad = \ E\left[\{(\hat{V}_{pj} - V_{pj}) + (V_{pj} - V^*_{pj})\}^2 | X_j\right] . \end{aligned}$$

Recall the equation (5)

$$\hat{V}_{pj} - V^*_{pj} = \epsilon_j + q_j - \{f(\hat{\theta}_j, \hat{X}_j, \hat{\gamma}_j) - f(\theta_j, X_j, \gamma_j)\}.$$

From the variance function model in (1), and the definition of V^*_{pj} in (3), we have

$$V_{pj} - V^*_{pj} = q_j - \{f(\hat{\theta}_j, \hat{X}_j, \hat{\gamma}_j) - f(\theta_j, X_j, \gamma_j)\}.$$

We are now assuming that for all X_j, $E\left(\epsilon_j \left[q_j - \{f(\theta_j, X_j, \gamma_j) - f(\hat{\theta}_j, \hat{X}_j, \hat{\gamma}_j)\}\right] | X_j\right)$ is much smaller than $E(\epsilon^2_j | X_j)$, and $E\{(V_{pj} - V^*_{pj})^2 | X_j\}$. Generally this will be true provided that the number of domains is relatively large.

Under the distribution assumption of $\hat{V}_{pj}$ in (7) and the assumptions above, an approximately unbiased estimator of $E\{(V^*_{pj} - V_{pj})^2 | X_j\}$ is:

$$r_j \overset{def}{=} (\hat{V}_{pj} - V^*_{pj})^2 - (d+2)^{-1}\, 2\hat{V}^2_{pj}.$$

7. Model Fitting for Conditional Expected Squared Error

For a given function $f(\theta_j, X_j, \gamma_j)$, we may consider a model to produce a smooth version, $h_f(\theta_j, X_j, \omega)$, of the "nonparametric" estimator, $E\{(V^*_{pj} - V_{pj})^2 | X_j\}$ such that:

$$E\{(V^*_{pj} - V_{pj})^2 | X_j\} = h_f(\theta_j, X_j, \omega) + c_j$$

where $E(c_j) = 0$. To develop possible appropriate models, consider the case in which

$$E(q^2_j) >> E[\{f(\hat{\theta}_j, \hat{X}_j, \hat{\gamma}_j) - f(\theta_j, X_j, \gamma_j)\}^2].$$

If we believed that $V(q_j)$ is approximately proportional to V^2_{pj}, then we could consider a regression model fit for r_j such that

$$r_j \ = \ \omega_0 + \omega_1 V^*_{pj} + \omega_2 {V^*_{pj}}^2 \ . \tag{9}$$

We consider the following model for r_j to evaluate the adequacy of the GVF model (2) for V_{pj}. When we divide each term in (9) with ${V^*_{pj}}^2$, we have

$${V^*_{pj}}^{-2} r_j = {V^*_{pj}}^{-2}\omega_0 + {V^*_{pj}}^{-1}\omega_1 + \omega_2 \ . \tag{10}$$

We used the following ordinary least squared method to compute ω where ω is the vector of $(\omega_0, \omega_1, \omega_2)$. Define $\mathbf{X} = [V^{*}_{pj}{}^{-2}, V^{*}_{pj}{}^{-1}, \mathbf{1}]$ a $j \times k$ matrix where j is the number of domains, and k is the number of coefficients in (10). Since we have 5160 industry-area-time combinations, and have three coefficients in (10), $j = 5160$ and $k = 3$. Hence $\hat{\omega} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(V^{*}_{pj}{}^{-2} r_j)$. Hhat $(\hat{h})$, the smooth version of r_j, is the product of the corresponding elements of $\mathbf{X}\,\hat{\omega}$ and the weight $V^{*}_{pj}{}^{2}$.

8. Data Analysis

Figure 1 shows the plot of r_j, an approximately unbiased estimator of the mean squared error (MSE) of V^{*}_{pj}, against $log(V^{*}_{pj})$. Hhat, the smooth version of r_j, is also plotted against $log(V^{*}_{pj})$. This figure demonstrates the sensitivity of Hhat to outstanding values of r_j.

Figures 2 and 3 present results from two nonparametric regression methods known as locally weighted regression (loess) and a supersmoothed estimator (supsm). For some general background on these smoothing methods, see the section of MathSoft (1995, section 7.11). Figure 2 displays the locally weighted regression smoothing predictors (loess). In locally weighted regression smoothing, the nearest neighbors of each point are used for regression, and the number of neighbors is specified as a percentage of the total number of points. This percentage is called the span. Figure 2 shows the loess with two different span sizes, 0.1 and 0.5 respectively. Not surprisingly, a loess predictor with a larger span size shows less sensitivity toward outstanding data values. A loess predictor with span size 0.5 didn't fit the data as well as the loess predictor with span size 0.1 at the tail. Figure 3 shows the result of supersmoothing r_j as a function of $log(V^{*}_{pj})$ adjunct to a loess predictor. With loess, the span is constant over the entire range of predictor values. Supersmoother, however, chooses a span for the predictor values r_j based on only the leave-one-out residuals whose predictor values r_i are in the neighborhood of r_j. A supersmoother fit the data better than the loess predictor with span size 0.1 in the tail.

In Figure 4, we plotted several estimators of the standard error of $\hat{V}_{pj}$ against V^{*}_{pj}. $(2\hat{V}^{2}_{pj})/(d+2)$ is denoted as SE2 which is the unbiased variance estimator of $\hat{V}_{pj}$. $(2V^{*}_{pj})/d$ is denoted as SE1 which could be a reasonable variance estimator of $\hat{V}_{pj}$ if the error, $V^{*}_{pj} - V_{pj}$ is small compared to the error, $\hat{V}_{pj} - V_{pj}$. As seen previously, Hhat is the smooth version of r_j.

9. Discussion

In closing, we note several possible extensions of the current work. First, we have focused on modeling of the variance of sampling error alone. In some work with small domain estimation, there is also interest in modeling of the variances of prediction errors, which may include components of both sampling error and model error. Second, one may complement the current evaluation of predictive precision of a GVF with formal significance testing for specific coefficients of a given GVF model. Third, one may develop additional diagnostics that are specifically focused on evaluation of the effect of GVF lack of fit on specific statistics e.g., confidence intervals for finite population means or variance-based weights in construction of weighted least squares estimators. These issues will be considered in other papers.

10. References

Binder, D.A.(1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.

Bureau Of Labor Statistics (1997). *BLS Handbook of Methods.* U.S. Department of Labor.

Butani, S., Stamas, G. and Brick, M.(1997). Sample Redesign for the Current Employment Statistics Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 517-522.

Eltinge, J.L., Fields, R.C., Gershunskaya, J., Getz, P., Huff, L., Tiller, R., and Waddington, D.(2001). Small Domain Estimation in the Current Employment Statistics Program *Unpublished Background Material for the FESAC Session on Small Domain Estimation at the Bureau of Labor Statistics.*

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

Judkins, D.R.(1990). Fay's Method for Variance Estimation. *Journal of Official Statistics*, 6, 223-239.

MathSoft, Inc.(1995), S-PLUS Guide to Statistical and Mathematical Analysis, Seattle, WA

Werking, G.(1997). Overview of the CES Redesign. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 512-516.

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

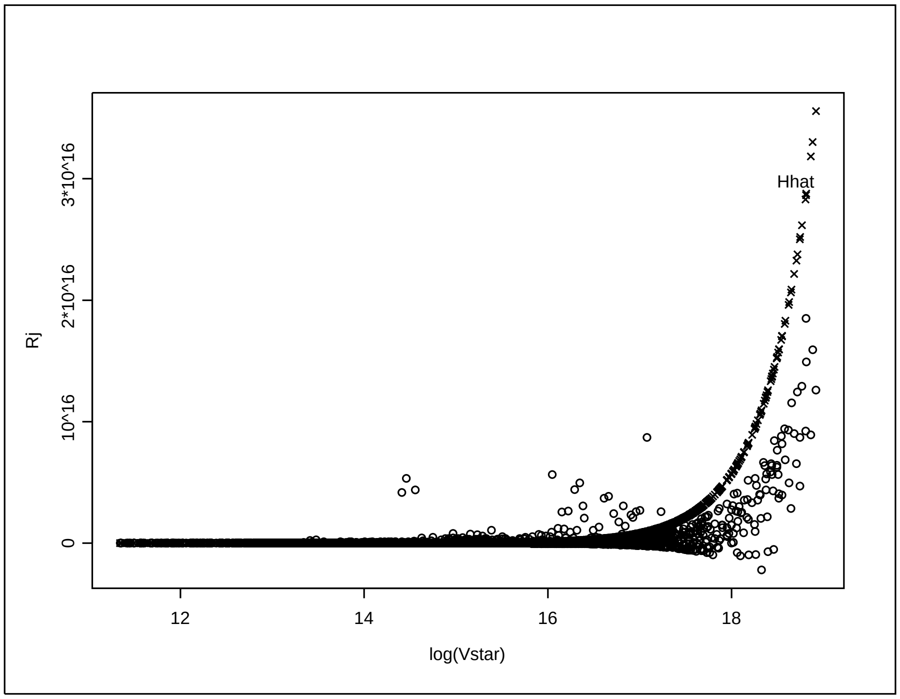

Figure 1: Rj on log(Vstar)

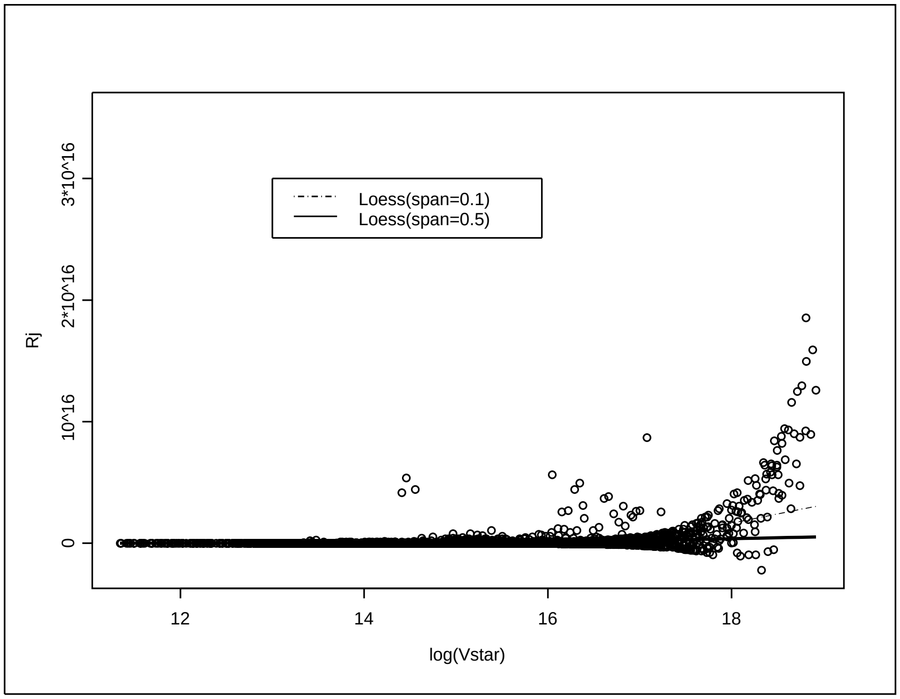

Figure 2: Loess-smoothed Rj on log(Vstar) with different span widths

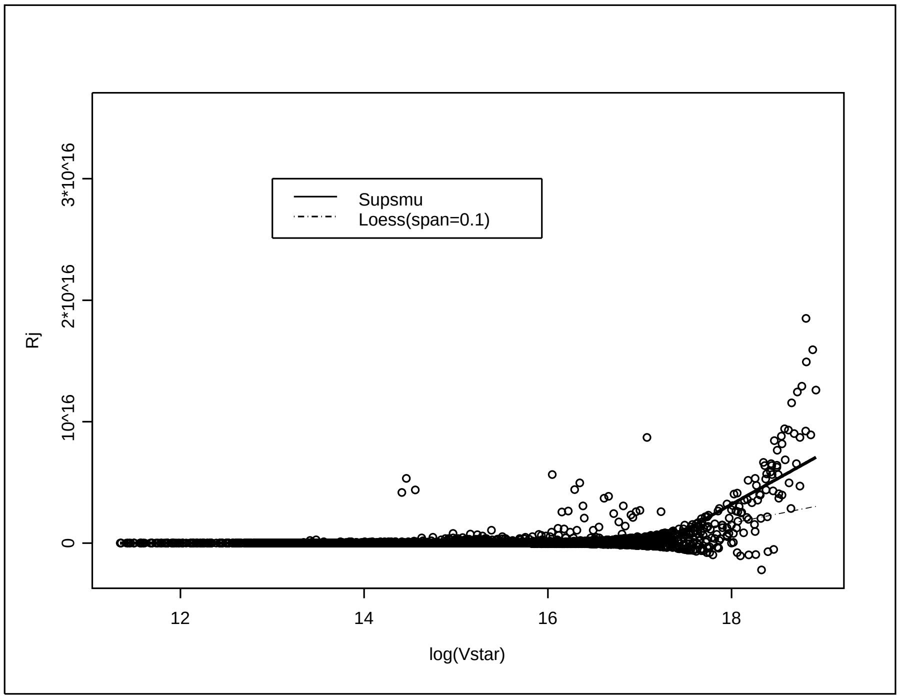

Figure 3: Supersmoothed and Loess-smoothed Rj on log(Vstar)

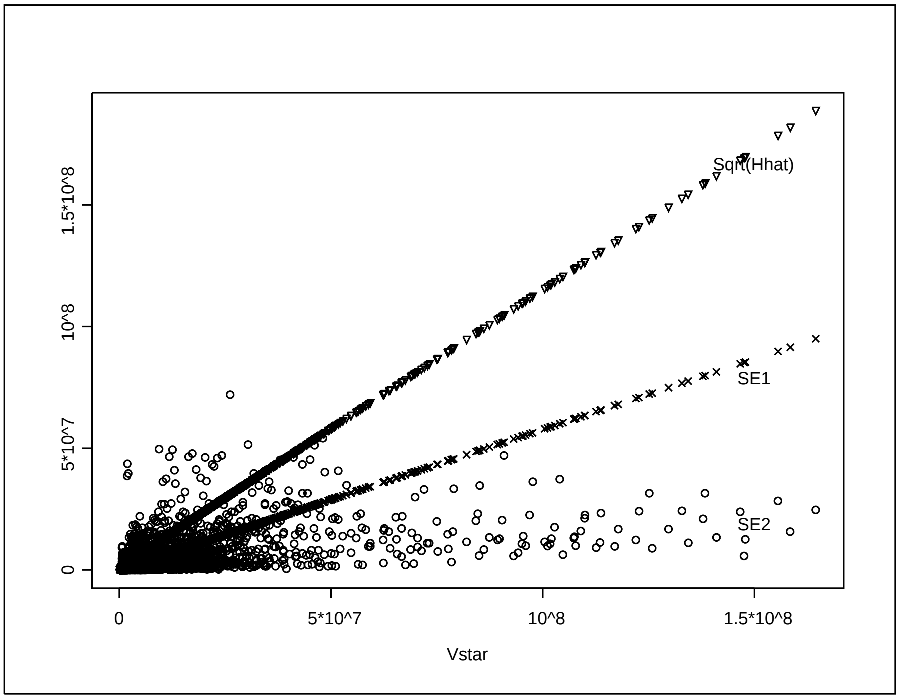

Figure 4: Sqrt(Hhat), SE1 and SE2 on Vstar