# Analysis of Generalized Variance Function Estimators from Complex Sample Surveys

MoonJung Cho*

**Key Words:** Bias, Confidence interval properties, Degrees of freedom, Equation error, Lognormal model, Simulation study, U.S. Current Employment Statistics (CES) survey

## 1. Introduction

For applied work with generalized variance function (GVF) models for sample survey data, one generally seeks to develop a model that produces variance estimators that are approximately unbiased and relatively stable. Through simulation, we evaluate the bias and variance of model coefficients, and the bias and variance of the GVF estimator. In addition, we compare and contrast confidence interval coverage rates and widths of the GVF estimator to design-based estimators. We study these properties with varying degrees of freedom for the GVF estimators and a refined bias adjustment factor for nonlinear transformations in the lognormal model. Our simulation study is based on the data from the U.S. Current Employment Statistics (CES) survey.

## 2. Variance Function Model

Define $\hat{\theta}_{jt}$ a point estimator of $\theta_{jt}$, a finite population mean or total where $j$ is the domain index at time $t$. For example, in CES survey, domains are the combinations of industries and areas. Define $V_{pjt}$= $V_p(\hat{\theta}_{jt})$ as the design variance of $\hat{\theta}_{jt}$, and $\hat{V}_{pjt} = \hat{V}_p(\hat{\theta}_{jt})$ as an estimator of $V_{pjt}$. 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, $V_{pjt}$, as a function of the estimate and possibly other variables (Wolter, 2007). The common specification is

$$V_{pjt} = f(X_{jt}, \gamma) + q_{jt} \tag{1}$$

where $X_{jt}$ is a vector of predictor variables potentially relevant to estimators of $V_{pjt}$, $q_{jt}$ is a univariate "equation error" with the mean $0$, and $\gamma$ is a vector of variance function parameters which we need to estimate. Note especially that $q_{jt}$ represents the deviation of $V_{pjt}$ from its modeled value $f(X_{jt}, \gamma)$. Furthermore, one needs to supplement model (1) with the decomposition

$$\hat{V}_{pjt} = V_{pjt} + \epsilon_{jt}, \tag{2}$$

where $\epsilon_{jt}$ is a random term that reflects sampling error in the estimator $\hat{V}_{pjt}$. Under the assumption that $\hat{V}_{pjt}$ is design unbiased for $V_{pjt}$, the error term $\epsilon_{jt}$ has design expectation equal to zero.

*Office of Survey Methods Research, U.S. Bureau of Labor Statistics, 2 Massachusetts Ave. N.E. Washington, D.C. 20212. The views expressed in this presentation are those of the author and do not necessarily reflect the policies of the U.S. Bureau of Labor Statistics. The author thanks John Eltinge for many helpful comments on variance function models; and Julie Gershunskaya and Larry Huff on helpful comments on the Current Employment Statistics survey.

We will use a special form of model (1) on the logarithmic scale in our CES applications,

$$\ln(V_{pjt}) \quad = \quad X_{jt}\,\gamma + q^{*}_{jt}, \tag{3}$$

where $q^{*}_{jt}$ is a general error term with mean equal to zero. As Johnson and King (1987) demonstrated in the Young Adult Literacy Survey, prediction can be improved by transforming to the logarithmic scale. The advantages of log transformation are that it converts multiplicative relationships to linear relationships, and reduces the impact of extreme values.

## 3. CES Data and Model Fitting

The CES survey collects data on employment, hours, and earnings from 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, 2011). This paper will focus only on total employment in the reporting establishment.

Using the benchmark data, $x_{j0}$, at the benchmark month $0$ from Quarterly Census of Employment and Wages (QCEW) data, the CES program obtains weighted link relative estimator, $\hat{y}_{jt}$, to estimate the total employment, $x_{jt}$, within the domain $j$ and month $t$,

$$\hat{y}_{jt} = x_{j0}\hat{R}_{jt}$$

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

We used the direct variance estimators $\hat{V}_{pjt}$ from the survey as the dependent variables in GVF models. We assume that $\hat{V}_{pjt}$ is a design unbiased estimator for $V_{pjt}$, i.e., $E_p(\hat{V}_{pjt}) = V_{pjt}$. Our sample consists of Unemployment Insurance (UI) accounts, which report nonzero employment for previous and current months. Let $n_{jt}$ be a number of responding UI accounts within the domain $j$ and month $t$. In fact, $t$ can be considered as the month distance between the reference month $t$ and the benchmark month $0$. In this paper, we consider only domains with at least 12 reporting UI accounts. There are 430 domains (industry-area combinations) in our CES data. Each domain 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

$$\ln(\hat{V}_{jt}) \quad = \quad \gamma_0 + \gamma_1 \ln(x_{j0}) + \gamma_2 \ln(n_{jt}) + \gamma_3 \ln(t) + e. \tag{4}$$

In this model, we assume that both intercepts and slopes are constant across the industries and areas.

## 4. "Degrees of Freedom" Measures for Estimation and Prediction Errors Under Variance Function Models

Let A be a positive random variable with finite positive mean and variance. Then under a standard approach, (e.g., Satterthwaite (1941) and Kendall and Stuart (1968, p. 83)), the random variable $\{E(A)\}^{-1}\,d\,A$ has the same first and second moments as those of a $\chi^2_d$ random variable, where we define "degrees of freedom" term

$$d = \{V(A)\}^{-1}\,2\,\{E(A)\}^{2}. \tag{5}$$

Specifically, for the random variables $V_{pjt}$ and $\hat{V}_{pjt}$ defined in expressions (1) and (2), $\{f(X_{jt},\gamma)\}^{-1}\, d_{q_{jt}} V_{pjt}$ has the same first and second moments as a $\chi^2_{d_{pjt}}$ random variable, where

$$d_{q_{jt}} = \{V(q_{jt})\}^{-1}\; 2\; \{f(X_{jt},\gamma)\}^2 . \tag{6}$$

Similarly, conditional on $V_{pjt}$, $(V_{pjt})^{-1}\, d_{\epsilon_{jt}} \hat{V}_{pjt}$ has the same first and second moments as a $\chi^2_{d_{pjt}}$ random variable, where

$$d_{\epsilon_{jt}} = \{V(\epsilon_{jt}|X_{jt})\}^{-1}\; 2\; (V_{pjt})^2 . \tag{7}$$

## 5. Equation Error and Estimation Error under Lognormal Models

Under the model defined by expressions (2) and (3), define $\epsilon^*_{jt} = \ln(\hat{V}_{jt}) - \ln(V_{jt})$ and assume that

$$\epsilon^*_{jt} \sim N(0, \sigma^2_{\epsilon^*}) \tag{8}$$

and

$$q^*_{jt} \sim N(0, \sigma^2_{q^*}). \tag{9}$$

Under additional regularity conditions, $\hat{\sigma}^2_e$ is a consistent estimator for the sum $\sigma^2_{q^*} + \sigma^2_{\epsilon^*}$.

If one does not have satisfactory information about the estimation-error variance term $\sigma^2_{\epsilon^*}$, then one may consider use of the predictor

$$V^*_{pjt} = \exp\left(X_{jt}\hat{\gamma} + 2^{-1}\hat{\sigma}^2_e\right). \tag{10}$$

The term $d_{\epsilon_{jt}}$ is usually known (up to a reasonable level of approximation) and equals the constant $d_\epsilon$ for all $j$ and $t$. Additional calculations for the moments of the lognormal distribution then show that

$$\sigma^2_{\epsilon^*} = \Psi\left(1, 2^{-1}d_\epsilon\right) \tag{11}$$

where $\Psi(a,b)$ is the $\Psi$ function with arguments $a$ and $b$ (Abramowitz and Stegun 1972, p.258). Similarly, under the lognormal model (9), define $d_q = \{V(q_{jt})\}^{-1}\; 2\; \{E(V_{jt}\}^2$, then

$$\sigma^2_{q^*} = \Psi\left(1, 2^{-1}d_q\right). \tag{12}$$

Finally, based on substitution of $\hat{\sigma}^2_{q^*}$ for $\sigma^2_e$ in expression (10), define the predictor

$$V^{**}_{pjt} = \exp\left(X_{jt}\hat{\gamma} + 2^{-1}\hat{\sigma}^2_{q^*}\right). \tag{13}$$

## 6. Simulation Study

### 6.1 Design of the Study

To evaluate the properties of $\hat{\gamma}$ and $V^{**}_{pjt}$, we carried out a simulation study based on the following variables produced for each of $R = 1000$ replicates.

First, based on the 5160 vectors $\left(\hat{V}_{pjt(r)}, X_{jt}\right)$, where $X_{jt} = (1, \ln(x_{j0}), \ln(n_{jt}), \ln(t))$, we carried out ordinary least squares regression of $\ln(\hat{V}_{pjt(r)})$ on $X_{jt}$ to produce the coefficient vector estimate $\hat{\gamma}_{(r)}$. Table 1 shows coefficient estimates $(\hat{\gamma})$. We then computed the fixed values of $f_{jt}$.

$$f_{jt} \quad = \quad \gamma_0 + \gamma_1 \ln(x_{j0}) + \gamma_2 \ln(n_{jt}) + \gamma_3 \ln(t) \tag{14}$$

based on the numerical values of the coefficient vector $\gamma$ for model $(f)$ presented in the Table 1, for all 5160 combinations of domain $j$ and month $t$ considered in Section 3.

Second, we generated the normal $(0, \sigma^2_{q^*})$ random variables $q^*_{jt(r)}$ for the 5160 cases, and then generated

$$V_{pjt(r)} = exp(f_{jt} + q^*_{jt(r)}).$$

In addition, we generated $\hat{\theta}_{jt(r)}$ as independent normal $(x_{j0}, V_{pjt})$ independent random variables; generated $\epsilon^*_{jt(r)}$ as independent normal $(0, \sigma^2_{\epsilon^*})$ random variables; and generated

$$\hat{V}_{pjt(r)} = V_{pjt(r)} \ \exp(\epsilon^*_{jt(r)}).$$

The term $\hat{\sigma}^2_{(r)}$ equal to the regression mean squared error; the term $\hat{\sigma}^2_{q^*(r)}$ defined by expression (12); and the predicted variances $V^{**}_{pjt(r)}$ defined by expression (13). In addition, we computed the confidence intervals for $\theta_{jt}$ based on the direct variance estimates $\hat{V}_{pjt(r)}$

$$\hat{\theta}_{jt(r)} \pm t_{d_\epsilon, 1-\alpha/2} \left(\hat{V}_{pjt(r)}\right)^{1/2} \tag{15}$$

and based on the GVF predictors $V^{**}_{pjt(r)}$

$$\hat{\theta}_{jt(r)} \pm t_{d_q, 1-\alpha/2} \left(V^{**}_{pjt(r)}\right)^{1/2} \tag{16}$$

where $t_{d,1-\alpha/2}$ is the upper $1-\alpha/2$ quantile of a $t$ distribution on $d$ degrees of freedom. Finally, taking averages over the $R$ replicates, we computed estimates of the biases of the coefficient estimates

$$R^{-1} \sum_{r=1}^{R} \left(\hat{\gamma}_{(r)} - \gamma\right) \tag{17}$$

and the average domain-specific relative bias of $V^{**}_{pjt}$ is

$$\left(n^{-1} R^{-1} \sum_{r=1}^{R} \sum_{t=1}^{12} \sum_{j=1}^{430} V^{-1}_{pjt} \Delta_{pjt(r)}\right) \tag{18}$$

where $\Delta_{pjt(r)} = V^{**}_{pjt(r)} - V_{pjt}$ and $n = J \times T = 430 \times 12 = 5160$. In addition, we computed the coverage rates and mean widths for the confidence intervals of $\hat{V}_{pjt}$ and $V^{**}_{pjt}$, and compared those properties of the GVF estimator to design-based estimators.

We repeated these steps for the 8 values of $d_q = 4, 6, 30$ and $400$. Results are displayed in Table 1.

### 6.2 Numerical Results

Table 2 presents the relative bias of the coefficient estimates as given in the expression (17), with the corresponding simulated standard deviations placed in parentheses. Note that the bias terms are all small relative to the coefficient values in Table 1 and relative to their reported standard deviations. Table 3 presents the selected values of $d_q$, and the corresponding values of $\sigma^2_{q^*}$ based on the expression (12); and the the average domain-specific relative bias values given by the expression (18). Note that the relative bias terms are fairly large for $d_q = 4$, but decline to values close to zero as $d_q$ increases. Table 4 reports the quantiles of the widths of the confidence intervals regarding expressions (15) and (16), respectively. As $d_q$ increases, interquartile range (IQR) value of $V^{**}_{pjt}$ decreases. This reflects the increasing efficiency of $V^{**}_{pjt}$ relative to $\hat{V}_{pjt}$ as $d_q$ increases with $d_\epsilon$ held equal to 6.

We explored possible time trends and employment size effects in the bias and confidence interval values. Since all results were very similar across different $d_q$ values, we arbitrarily selected $d_q = 30$ case. Hence all figures from 1 to 5 are from $d_q = 30$ case.

Figure 1 plots relative bias against month-distance: $month = 1$ means one month away from the benchmark month $0$. We didn't identify any substantial time effects for the relative-bias results. Figure 2 plots relative bias against log of employment size at benchmark month $0$ with loess (locally weighted scatter plot smooth) line of span=0.3 inserted. Again, we did not observe any substantial employment-size effects for the relative-bias results. Figure 3 shows coverage rates of both $V^{**}_{pjt}$ and $\hat{V}_{pjt}$ against month-distance. Note that all coverage rates exceeded the nominal value of $0.95$; coverage rates of $\hat{V}_{pjt}$ is slightly higher and values from $V^{**}_{pjt}$ is slightly lower than than $0.96$. This is due to the fact that $\hat{V}_{pjt}$ has wider confidence width as shown in Figure 4. We didn't identify any substantial time effects in coverage rates of both $V^{**}_{pjt}$ and $\hat{V}_{pjt}$.

As one would expect from the positive coefficient $\gamma_1$ and $\gamma_3$ in Table 1, the widths of the intervals (15) and (16) did increase over month-distance and employment size as shown in Figures 4 and 5.

## 7. Summary

In this paper, we presented simple methods to simulate GVF estimator. Through simulation, we evaluate the bias and variance of model coefficients, and the bias and variance of the GVF estimator. The bias terms of coefficients were small relative to true coefficient values and to their standard deviations. Relative bias of GVF estimator declined as $d_q$ increased and no substantial time effects were observed. Coverage rates for both simulated $\hat{V}$ and $V^*$ exceeded the nominal value of $0.95$ and showed no time effects.

**Table 1: Coefficient Estimates of Model ($f$)**

| | *intercept* $\gamma_0$ | $\ln(x_{j0})$ $\gamma_1$ | $\ln(n_{jt})$ $\gamma_2$ | $\ln(t)$ $\gamma_3$ |
|---|---|---|---|---|
| EST. | -1.43 | 1.16 | 0.22 | 1.17 |
| s.e. | 0.66 | 0.09 | 0.12 | 0.07 |
| $t_\gamma$ | -2.17 | 12.77 | 1.78 | 16.72 |

**Table 2: Bias of Coefficient Estimates of Model $(f)$**

| $d_q$ | *Coefficient Bias (Standard Deviation)* | | | |
|---|---|---|---|---|
| | $\hat{\gamma}_0$ | $\hat{\gamma}_1$ | $\hat{\gamma}_2$ | $\hat{\gamma}_3$ |
| 4 | 0.0010 | -0.0001 | 0.0009 | -0.0010 |
| | (0.247) | (0.026) | (0.034) | (0.061) |
| 6 | 0.0009 | -0.0000 | 0.0006 | -0.0008 |
| | (0.215) | (0.022) | (0.030) | (0.054) |
| 30 | 0.0007 | -0.0000 | 0.0002 | 0.0008 |
| | (0.164) | (0.017) | (0.023) | (0.042) |
| 400 | 0.0006 | 0.0000 | 0.0000 | -0.0001 |
| | (0.152) | (0.016) | (0.021) | (0.038) |

**Table 3: Relative Bias of GVF estimator $V_{pjt}^{**}$**

| $d_q$ | $\sigma_q^2$ | *rel bias* |
|---|---|---|
| 4 | 0.645 | 0.906 |
| 6 | 0.395 | 0.484 |
| 30 | 0.069 | 0.072 |
| 400 | 0.005 | 0.006 |

**Table 4: Quantiles of CI $(f)$**

| $d_q$ | | 0.01 | 0.05 | 0.10 | 0.25 | 0.50 | 0.75 | 0.90 | 0.95 | 0.99 | IQR |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 4 | $\hat{V}_{pjt}$ | 0.20 | 0.31 | 0.37 | 0.50 | 0.74 | 1.13 | 1.64 | 2.10 | 3.63 | 0.63 |
| | $V_{pjt}^{**}$ | 0.23 | 0.36 | 0.43 | 0.58 | 0.87 | 1.32 | 1.92 | 2.46 | 4.25 | 0.74 |
| 6 | $\hat{V}_{pjt}$ | 0.19 | 0.30 | 0.36 | 0.48 | 0.72 | 1.09 | 1.58 | 2.03 | 3.50 | 0.61 |
| | $V_{pjt}^{**}$ | 0.19 | 0.30 | 0.36 | 0.48 | 0.72 | 1.09 | 1.59 | 2.04 | 3.52 | 0.50 |
| 30 | $\hat{V}_{pjt}$ | 0.18 | 0.29 | 0.34 | 0.46 | 0.69 | 1.05 | 1.52 | 1.96 | 3.36 | 0.59 |
| | $V_{pjt}^{**}$ | 0.15 | 0.23 | 0.27 | 0.37 | 0.55 | 0.84 | 1.22 | 1.57 | 2.71 | 0.47 |
| 400 | $\hat{V}_{pjt}$ | 0.18 | 0.28 | 0.34 | 0.46 | 0.69 | 1.04 | 1.51 | 1.94 | 3.35 | 0.58 |
| | $V_{pjt}^{**}$ | 0.14 | 0.22 | 0.26 | 0.35 | 0.53 | 0.80 | 1.16 | 1.48 | 2.57 | 0.45 |

## REFERENCES


Abramowitz, M. and Stegun, I.A. (1972), *Handbook of Mathematical Functions*: New York: Dover Publications, INC.

Binder, D.A. (1983), "On the Variances of Asymptotically Normal Estimators from Complex Surveys," in *International Statistical Review*, 51, 279-292.

Cho, M.J., Eltinge, J.L., Gershunskaya, J. and Huff, L. (2002), "Evaluation of the Predictive Precision of Generalized Variance Functions in the Analysis of Complex Survey Data," In *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.

Karlberg, F. (2000), "Survey Estimation for Highly Skewed Population in the Presence of Zeros," *Journal of Official Statistics*, 16, 229-241.

Kendall, M.G. and Stuart, A. (1968). *The Advanced Theory of Statistics*: Vol.3, New York: Hafner Publishing

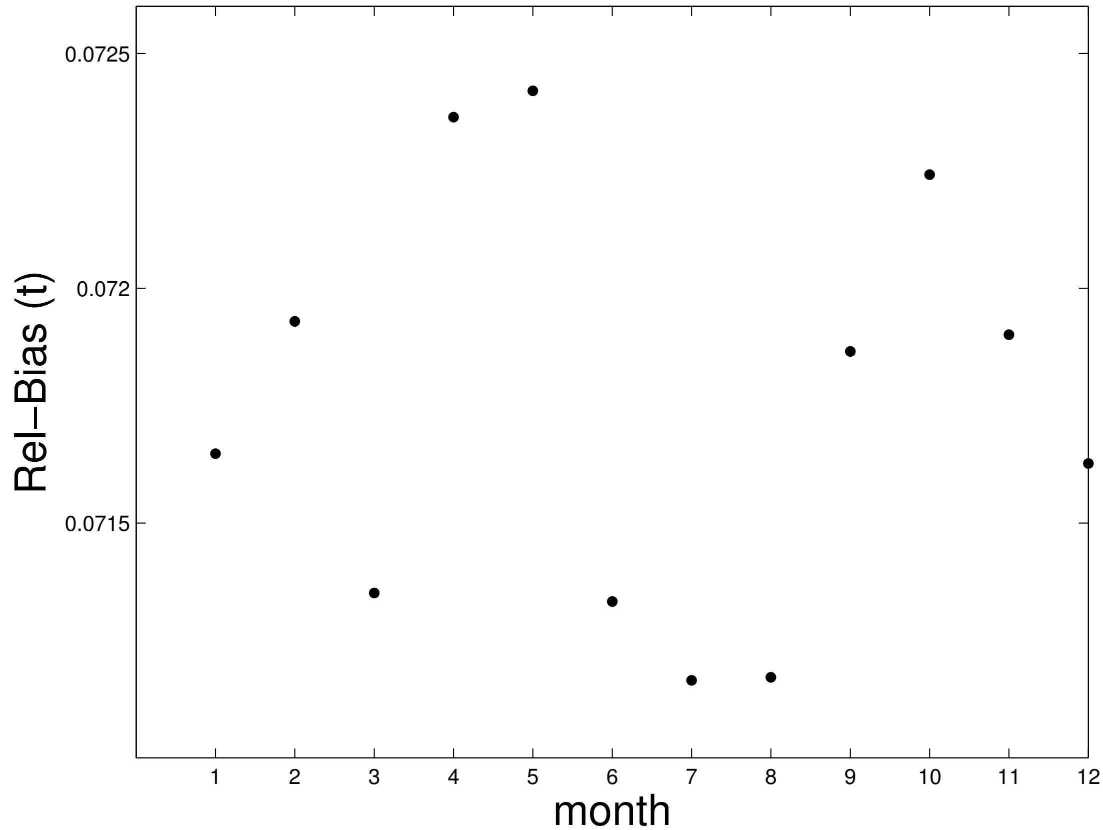


**Figure 1**: *Relative Bias ($V^*$) against Months :* $d_q = 30$

Company.
Satterthwaite, F.E. (1946), "An Approximate Distribution of Estimates of Variance Components," *Biometrics*, Bulletin 2, 110-114.
U.S. Bureau Of Labor Statistics (2011), Employment, Hours, and Earnings from the Establishment Survey. Chapter 2 of *BLS Handbook of Methods*, U.S. Department of Labor.
Available at: URL=http://www.bls.gov/opub/hom/pdf/homch2.pdf (Accessed June 2013)
Valliant, R. (1987), "Generalized Variance Functions in Stratified Two-Stage Sampling," *Journal of American Statistical Association*, 82, 499-508.
Wolter, K.M. (2007), *Introduction to Variance Estimation*: Second Edition, New York: Springer-Verlag.

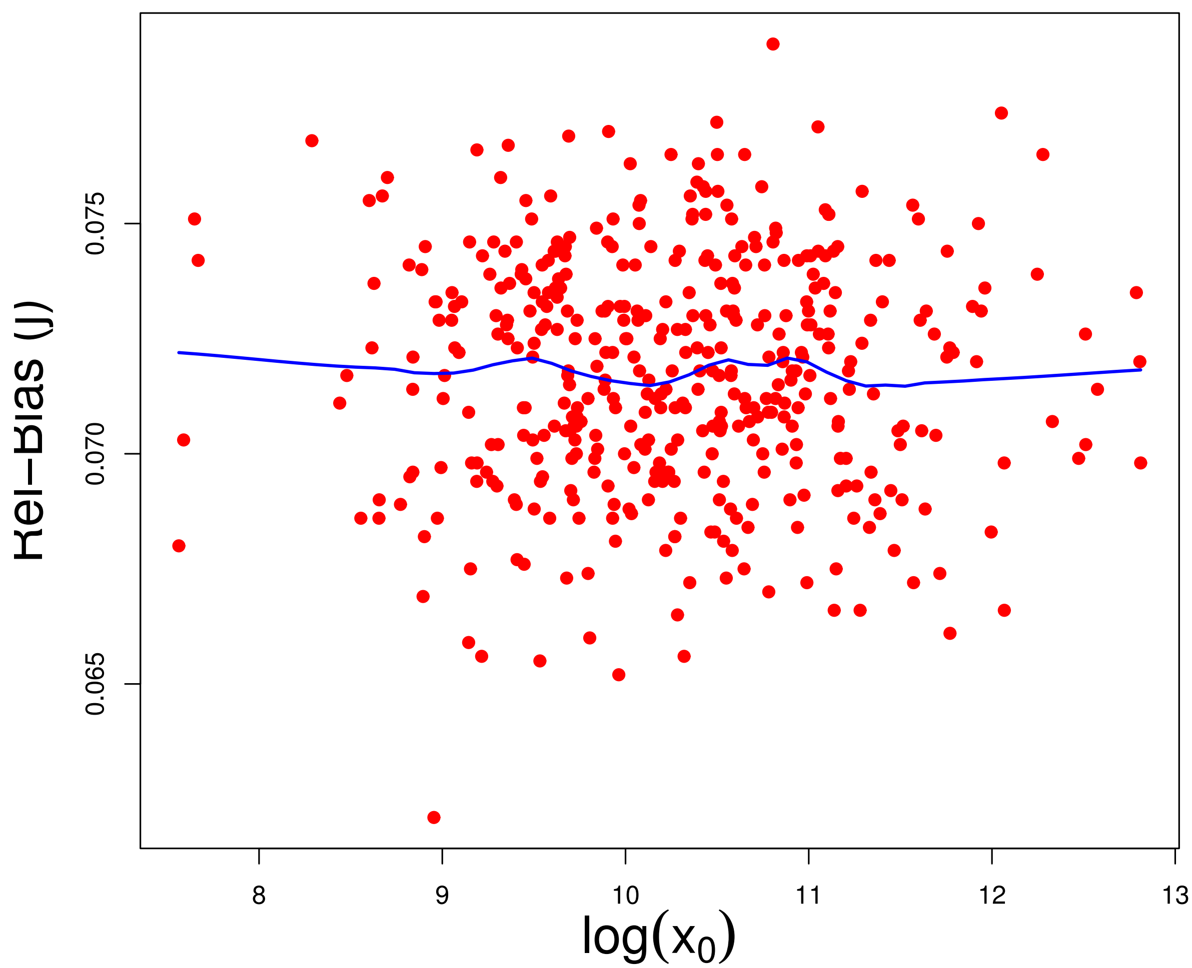


**Figure 2**: *Relative Bias ($V^*$) against Employment Size:* $d_q = 30$

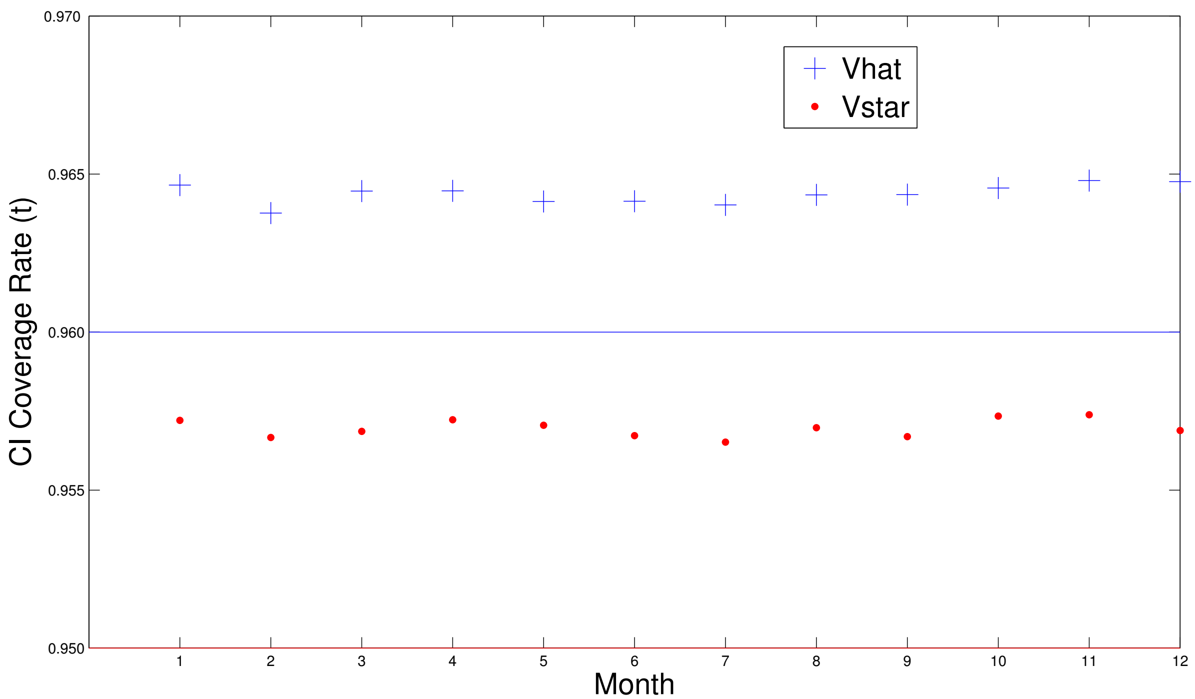


**Figure 3**: *Coverage Rate against Months* $d_q = 30$

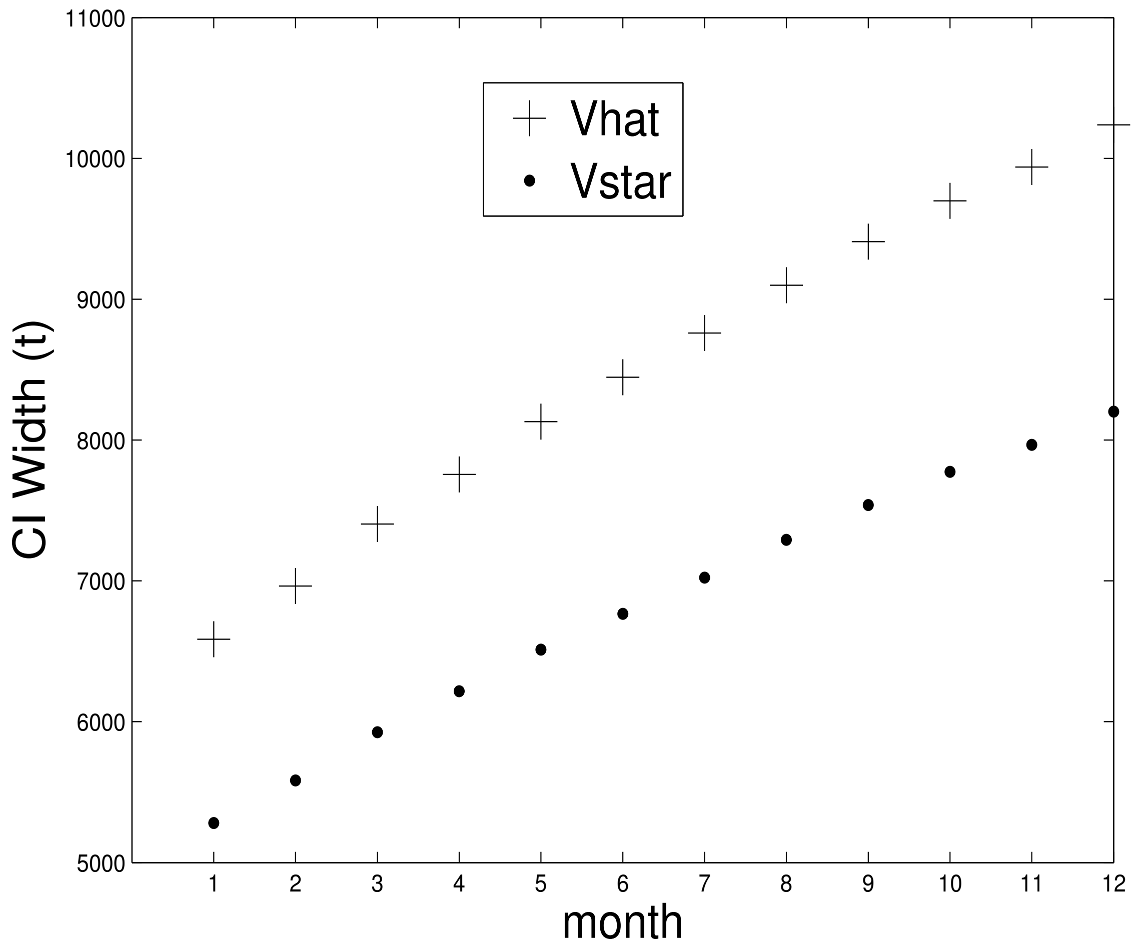


**Figure 4**: *CI Width against Months:* $d_q = 30$

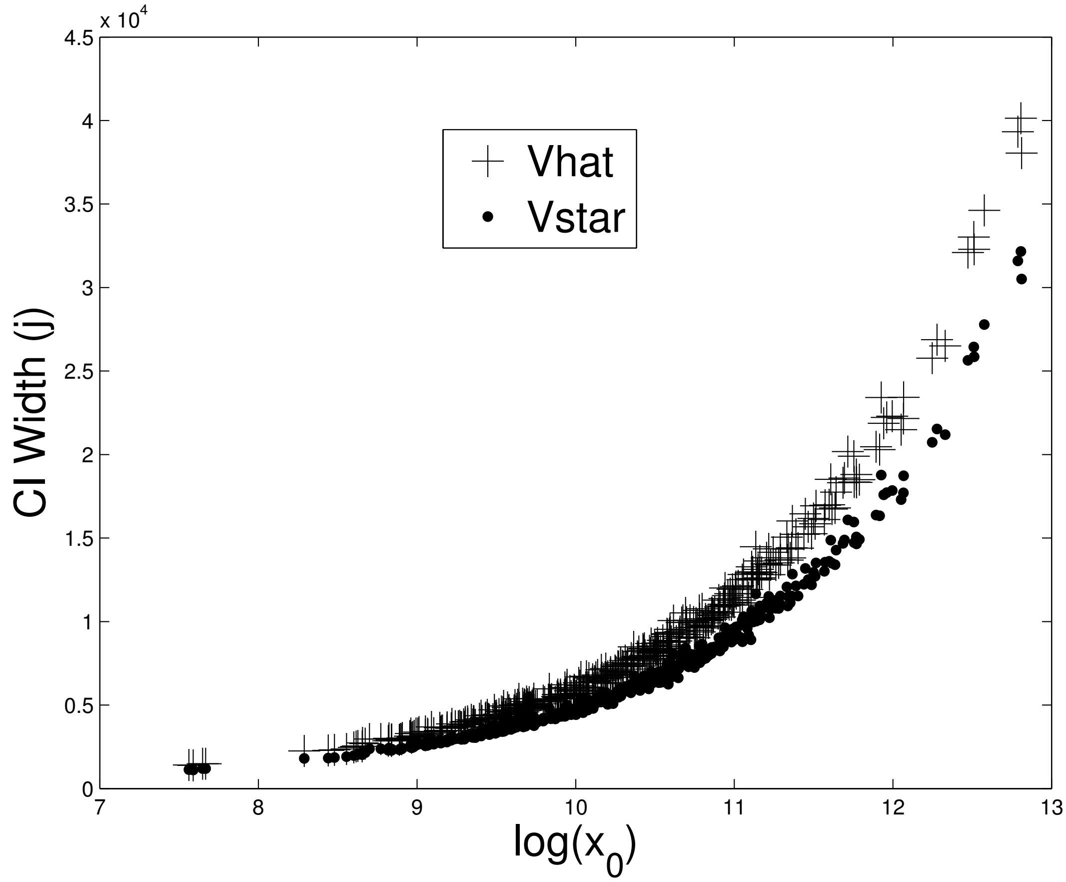


**Figure 5**: *CI Width against Employment Size:* $d_q = 30$