Proceedings of the Annual Meeting of the American Statistical Association, August 5-9, 2001

ON PROCEDURES TO SUMMARIZE VARIANCES FOR SURVEY ESTIMATES

Don Jang
Mathematica Policy Research, Inc., 600 Maryland Ave. S.W., Suite 550, Washington, DC 20024-2512

Key Words: Design effect, Errors-in-variables, Generalized variance function

Variance estimation is an important research area for sampling statisticians; consequently, if appropriate design variables are provided in the final dataset, nowadays data analysts can calculate variance estimates for most survey statistics via public or custom software. However, some data users are not be familiar with variance estimation procedures or may not have access to software to compute variance estimates under a complex survey design. For this reason, a simple, but useful, approach is to develop approximate variances for statistics of interest. One such technique is called the generalized variance function (GVF) technique. It has been customary to use GVF methods based on (1) design effects or (2) regression models.

In this paper, I compare two customary techniques to develop generalized variance estimates. Moreover, I expand the regression model approach to allow an errors-in-variables approach to see whether the resultant GVF estimates perform better than traditional regression-based GVFs.

1. Need to Summarize Variance Estimates

Large surveys produce many survey estimates in their reports. These estimates with their standard errors result in a substantial number of pages in survey reports. So it has been customary to produce point estimates for survey characteristics in the report while (1) standard errors are put in an appendix of the report; (2) average design effects are reported to users can approximate variance estimates especially for proportions; or (3) parameter estimates for GVFs are reported. With (2) or (3), users summarize standard error estimates for the reported survey estimates.

For sophisticated analysts, a public-use dataset is a good source for details about the survey and obtaining unpublished estimates. In particular, survey data usually provide appropriate analysis weights so users may calculate variance estimates using Taylor Series linearization or the replication method. In many cases, developing customized software to implement a variance estimation method is very time consuming, and thus many users tend to use variance estimation software. That is, data users can calculate estimates directly from the data set using generalized software; however, not all data-users have appropriate software. Or even if users have some facility to estimate variances using existing software, the costs of publishing variances for many (all) items in a report may be excessive. Moreover, public-use datasets available do not always provide full design information needed for appropriate variance estimation - usually because of reasons of confidentiality. In that sense, it is helpful and cost-effective for data-users to have a simple tool to obtain variance estimates.

Nowadays, it is becoming more common to release a database to the public through the internet. This allows users to specify variables and obtain estimates interactively. Providing design-based variance estimates for user-specified estimates may be expensive and time-consuming. Consequently, it is helpful to have a tool to approximate variance estimates for the analysts' specific estimates in a less computationally intensive fashion.

Summarized variances can be also used for optimal sample design. Sample size is usually determined to meet certain precision goals quantified from summarized variance estimates of similar or past surveys. A successfully developed variance calculation mechanism for one survey can be used for another survey. Finally, variance estimates obtained from a summary method can be more stable than directly calculated variance estimates, because they are based on a group of variables rather than an individual variable.

GVFs can provide a simple tool to calculate variance estimates for similar characteristics in a quick and simple manner. Statistics considered for GVF estimation are totals, proportions, averages, and ratios. In the next section, I review empirical studies in which GVFs were used for variance approximation.

2. Empirical Studies

There have been many empirical studies that calculate GVFs for selected statistics for many domains of interest in complex surveys. See for example, Salvucci et al. 1993. In particular, Salvucci et al. present a simple procedure that users can follow to calculate variances using GVFs. Even without a thorough theoretical justification, one issue is the choice of a fitting procedure, such as ordinary least squares, weighted nonlinear, unweighted nonlinear, and iteratively reweighted, etc. One of the fitting algorithms used is the Gauss-Newton algorithm.

The GVF approach can provide secondary data analysts variance estimates, even though it lacks theoretical justification (Bieler and Williams, 1990). John and King (1987) used modified goodness of model diagnostics to evaluate GVF-derived variance estimates, since they believe underestimation is more severe than overestimation.

There have been two approaches to approximating or summarizing variance estimates: average design effect and regression models. These two approaches were evaluated for proportion estimates (see, for example, John and King, 1987; and Bieler and Williams, 1990). Empirical studies have found there is no gain to use detailed models over the simple design effect approach. GVFs can be used for other estimates not used in developing GVFs in the same survey or/and similar surveys. Without a theoretical justification thus far, GVFs are believed to be more stable than directly calculated variance estimates. Valliant (1992) tried to develop GVFs for a price index using nonparametric smoothing. Valliant (1987) showed theoretical justification that GVFs are consistent and more stable under some regularity conditions.

In the following section, I present the functional forms for two customary procedures: average design effect and regression based approach. Moreover, I introduce an errors-in-variables model to allow for the variability of survey estimates used as predictors. Then, I compare simple average design effect approaches with regression-based GVF approaches. These comparisons are specifically implemented into the 1999 Science and Engineers Statistical Data System (SESTAT) database.

3. Average Design Effect (Adeff) Approach

In this section, I discuss the average design effect technique. The design effect is usually used to measure the impact on variability of survey estimates due to complex sampling procedures from the hypothesized simple random sampling. The design effect can be obtained as the ratio of the design-based variance to the variance based on simple random sampling:

$$Deff(\hat{q}) = \frac{Var_D(\hat{q})}{Var_S(\hat{q})} \tag{3.1}$$

where $Var_D(\hat{q})$ is the design-based variance and $Var_S(\hat{q})$ is the variance from a simple random sampling (SRS). One can then calculate a sample-based design effect as the ratio of two directly calculated variance estimators:

$$\widehat{Deff}(\hat{q}) = \frac{\widehat{Var}_D(\hat{q})}{\widehat{Var}_S(\hat{q})} \tag{3.2}$$

Therefore the design-based variance estimator can be obtained by multiplying the design effect and the simple random sample based variance estimator.

Consider estimates of proportions. Suppose $\hat{P}$ is the proportion of persons with a certain attribute. Then, from (3.2), the design-based variance estimator can be obtained with the proportion estimate, its design effect, and the sample size:

$$\widehat{Var}_D(\hat{P}) = \frac{\hat{P}(1-\hat{P})}{n} \times \widehat{Deff}(\hat{P}) \tag{3.3}$$

Given point estimates and sample sizes, if the design effect of a group of variables can be summarized, then so can variance estimates. This conjecture can be formulized in the following way:

$$Deff_i = ADeff + e_i \tag{3.4}$$

If this relation (3.4) holds for a set of variables, then the predicted average design effect estimates can be used to summarize variance estimates for those variables.

Given average design effects, variance estimates can be approximated from a relatively simple formula. For example, given survey estimates of proportions, users only need to know the sample size and the average design effect:

$$\widehat{Var}_{Adeff}(\hat{P}) = \frac{\hat{P}(1-\hat{P})}{n} \times \widehat{Adeff}$$

Or equivalently, effective sample sizes can be provided instead of average design effects and initial sample sizes:

$$\widehat{Var}_{Adeff}(\hat{P}) = \frac{\hat{P}(1-\hat{P})}{n_{\overline{eff}}}$$

where $n_{\overline{eff}} = \widehat{Adeff}^{-1} n$ is called the effective sample size and can be interpreted as the sample size that provides the same precision as a simple random sample. In this case, the usual SRS–based formula can be used to obtain variance estimates.

The average design effect approach can be also extended to the estimates of totals. Suppose X denotes the total number of units with a certain attribute and T is the total population size; that is, $X=TP$. Then, a design based variance estimator is the product of a design effect and the variance based on a simple random sample. This is the product of the square of population size estimate $\hat{T}$ and the simple random sample based variance estimate of the proportion:

$$\hat{V}_D(\hat{X}) = \widehat{Deff}(\hat{X}) \times \hat{V}_S(\hat{X}) \doteq \widehat{Deff}(\hat{X}) \times \frac{\hat{T}^2 \hat{P}(1-\hat{P})}{n}$$

Then, given both total population size and sample size, one can obtain variance estimates for totals using the average design effect approach.

There is a relationship between design effects for totals and proportions, if these estimates are nearly uncorrelated; that is, if $COV(\hat{P},\hat{T}) \simeq 0$. Under this assumption, the relative variance of proportions can be expressed as the difference of relative variances of $\hat{X}$ and $\hat{T}$: $P^{-2}\mathrm{Var}(\hat{P}) \approx X^{-2}\,\mathrm{Var}(\hat{X}) - T^{-2}\,\mathrm{Var}(\hat{T})$. Then, the design effect for estimates of totals can be expressed as a factor of the design effect for proportions:

$$\widehat{Deff}(\hat{X}) \simeq \widehat{Deff}(\hat{P})\left[1 + \frac{CV^2(\hat{T})}{CV^2(\hat{P})}\right]$$

Since the factor above is greater than or equal to 1, the design effect for total is also larger than that for proportion. However, if T is known, then these two design effects would be same.

4. Generalized Variance Function (GVF) Approach

This section discusses another popular approach, the generalized variance function technique, that can help data users calculate variance estimates for many variables in a relatively short time period. The idea behind this technique is that given point estimates and coefficient of variation estimates, variance estimates or standard errors can be obtained from the following relationship:

$$CV = \frac{SE(\hat{q})}{q} \Leftrightarrow se(\hat{q}) = \widehat{CV}\hat{q}$$

If the coefficient of variation can be summarized for a set of items, so can the standard errors. Algebraically, the GVF can be expressed as:

$$g\{CV(\hat{\mathbf{q}})\} = f(\mathbf{q};\mathbf{b}) \qquad (4.1)$$

where $\mathbf{q}$ is a vector of population parameters (e.g., totals) to be estimated; $\mathbf{b}$ is a vector of model parameters; and $\hat{\mathbf{q}}$ is a survey estimator where $\mathrm{E}(\hat{\mathbf{q}})$ Ñ $\mathbf{q}$. That is, a function of the coefficient of variation or equivalent measures can be expressed as a function of population characteristics $\mathbf{q}$ and model parameters $\mathbf{b}$. Here the final functional form for g and f are usually determined based on empirical data plots.

To use GVF functional relationship (4.1), directly calculated variance estimates and point estimates are needed in place of the unknown values. Then, through standard procedure, the model can be fitted. The resultant GVF-derived estimates can be expressed as:

$$\sqrt{\hat{V}_{\mathrm{GVF}}(\hat{\mathbf{q}})} = \hat{\mathbf{q}}g^{-1}\{f(\hat{\mathbf{q}};\hat{\mathbf{b}})\}$$

where $\hat{\mathbf{b}}$ is the estimated parameter. Consequently, once model parameter estimates $\hat{\mathbf{b}}$ are obtained, users can simply put the survey estimates $\hat{\mathbf{q}}$ into the GVF functional form to derive variance estimates.

For the GVF approach, let us start with totals. For estimates of totals, variance approximation can begin with the conjecture that the square of the coefficient of variation of the estimator is a decreasing function of the expectation of the estimator (see, for example, Valliant, 1987 and references cited therein): $CV^2(\hat{X}) \propto E(\hat{X})^{-1}$. That is, the square of the coefficient of variation is linearly related to the inverse of the total or equivalently the variance is a quadratic function of the total:

$$CV^2(\hat{X}) = a + b\,X^{-1} \qquad (4.2)$$

$$Var(\hat{X}) = a\,X^2 + b\,X \qquad (4.3)$$

where CV is the coefficient of variation of $\hat{X}$, an estimator of a characteristic X, and a and b are unknown parameters. While acknowledging little theoretical justification of the above models, they have been supported theoretically under some conditions. In fact, this relationship makes sense since estimates of totals might be regarded as binomial distribution under simple random sampling. Suppose X is distributed as a Bernoulli with a probability P to have a value 1 and 1-P a value of zero. Then, the total of the characteristic, $\hat{X}$, can be regarded as a Binomial distribution with a probability of success, P. The total number of population characteristics can be guessed as X=NP. Since its variance under the binomial assumption is $NP(1-P)$, it can be shown that the square of the coefficient of variation is a decreasing function of the total. So this provides a rough justification for model (4.2) or (4.3) for totals. Usually, a direct model fit to this relationship results in heteroscadiscity of error terms. A simple log-transformation on both response and predictor variables would stabilize the error variance:

$$\log[Var(\hat{X})] = a + b\log(X) \qquad (4.4)$$

I use this model (4.4) with the SESTAT data sets presented later. Then, the GVF-derived standard error estimates based on (4.4) are a function of estimated model coefficients and total estimates:

$$se_{GVF}(\hat{X}) = \exp(\hat{a}/2)\,\hat{X}^{(\hat{b}/2)} \qquad (4.5)$$

With good GVF models for totals, one can also calculate GVF-derived variance estimates for proportions if the same assumptions are made. That is, if the covariance terms between $\hat{P}$ and $\hat{T}$ are negligible, then the relative variance of estimates of proportions can be expressed as the simple difference of two relative variances of $\hat{X}$ and $\hat{T}$:

$$P^{-2}\mathrm{Var}(\hat{P}) \approx X^{-2}\,\mathrm{Var}(\hat{X}) - T^{-2}\,\mathrm{Var}(\hat{T}) \text{ if } \mathrm{Cov}(\hat{P},\hat{T}) \approx 0$$

Then, the final form of GVF-derived standard error estimates for proportion estimates would be another function of model parameter estimates, estimated proportions, and population totals:

$$se_{GVF}(\hat{P}) = \hat{P}\exp(\hat{a}/2)\hat{T}^{\left(\frac{\hat{b}}{2}-1\right)}\left[\hat{P}^{\left(\hat{b}-2\right)}-1\right]^{\frac{1}{2}}$$

So given the model parameter estimates, users only need to insert estimates into the GVF functional form to obtain standard errors for totals or percentage values.

5. Errors-in-Variables Model

Given the model in (4.4), Section 4 discusses to fit the following model

$$Log[\widehat{\mathrm{Var}(\hat{X})}] = a + b\,\log(\hat{X})$$

where $\log(\hat{X})$ and $Log[\widehat{\mathrm{Var}(\hat{X})}]$ are estimated from the data. Usual regression approach assumes that $\log(\hat{X})$ values are constant. However, they are also subject to sampling errors with the following first and second moments:

$$\log(\hat{X}) \stackrel{\cdot}{\sim} [\log(X), CV^2(\hat{X})].$$

By allowing the variance of $\log(\hat{X})$, a direct application of Fuller (1987, p.187) with equal weights gives the estimators,

$$\tilde{a} = k^{-1}\sum_{i=1}^{k}\log[\widehat{\mathrm{Var}(\hat{X}_i)}] - \tilde{b}\,k^{-1}\sum_{i=1}^{k}\log(\hat{X}_i) \text{ and}$$

$$\tilde{b} = \frac{\sum_{i=1}^{k}\left\{\log(\hat{X}_i) - k^{-1}\sum\log(\hat{X}_i)\right\}\log\{\mathrm{Var}(\hat{X}_i)\}}{\sum_{i=1}^{k}\left\{\log(\hat{X}_i) - k^{-1}\sum\log(\hat{X}_i)\right\}^2 - \hat{s}_u}$$

where $\hat{s}_u = \sum_{i=1}^{k} CV^2(\hat{X}_i)$.

In the application presented later, this errors-in-variables approach did not make much difference in terms of predictability.

6. Application to Scientist and Engineers Statistical Data System (SESTAT)

So far, I have discussed two approaches to approximate or summarize variances, particularly for proportions and totals. In this section, I provide examples of the application of these methods. The dataset considered for this study comes from the Scientist and Engineers Statistical Data System (SESTAT). The population of this data system consists of all residents of the United States with Bachelor's degree or higher who are noninstitutionalized, age 75 or less, and either trained as or working as a scientist or engineer. This data set has very complex survey design features in that it consists of three independent survey components, the National Survey of College Graduates, the National Survey of Recent College Graduates, and the Survey of Doctorate Recipients. Because of its complexity and lack of design information available for public data, GVFs have been made available to the public. For detailed information, see the homepage for SESTAT (http://sestat.nsf.gov). A total of 12 domains and 97 variables were considered for this study.

Models can be fit by domains and/or by type of outcomes. Figure 1 shows design effect distributions within domain and across domains. Before calculating average design effects, this plot would be used to help decide whether to use one value for all variables or domain-specific values. As seen in the plot, there is nontrivial variation within and across domains. Consequently, with distributional variation across domains, domain-specific average design effects were used. From the twelve domains, I chose two domains for further investigation: total number of scientists and engineers and the total number of scientists and engineers with bachelor's degrees.

Goodness of fit measures need to be considered. For the regression type GVF approach, R^2 can be used as a quick check for model validity. For both approaches, evaluating standard errors calculated using the variance summarization techniques, a simple but useful diagnostic statistic called the relative standard error was used:

$$RSE = \frac{se_D(\hat{q}) - se_{Approx}(\hat{q})}{se_D(\hat{q})}$$

This is the relative difference between directly calculated standard error and approximated standard errors. As an ad-hoc criterion, 20% *RSE* is frequently used as a cut point to determine whether summary-based variance estimates are acceptable.

Figure 2 shows proportions of relative standard errors with less than 20% for the average design effect approach for the 12 domains considered. A little over 80% of variables have less than 20% of RSE when using the average design effect based standard errors as presented in the blue line. I also separated variables into two groups based on the estimated proportion; one group is for moderate *P* values between 0.1 and 0.9 and the other is out of range. There is clear distinction between these two categories. For moderate P values, the average design effect approach works pretty well. However, not surprisingly, there is substantial variation for small P values. From this result, one might say users ought to be cautious about using the average design effect procedure for small P. Since the variance of small P values is unstable, it appears to be better to use aggregated variance measures based on average design effects.

For the regression approach, a log-transformed model was fitted for the 12 domains and all models were fit reasonably well with less systematic error patterns and large R^2. All models have greater than 80% of R^2. Using the formula in (4.5), standard errors can be predicted.

Figures 3 and 4 show plots of relative standard errors for the two predicted standard error models: one is from average design effect approach, the other one is for the GVF approach. Figure 3 illustrates the proportion case and indicates the average design effects approach gives an overestimation, while the GVF approach shows a little better prediction though it tends to have a little underestimation. For total estimates (Figure 4), since we used the same average design effects obtained from proportion estimates, the magnitude of overestimation of standard errors based average design effects seems to be reduced, and for some domains, it even underestimates. The GVF approach seems to work better with less relative standard errors. Unlike the proportion case, it gives conservative variances, and this is good.

7. Discussion

For large surveys, it is better to attempt separate variance summarization for the key domains. In general, the two methods considered perform well for most proportions and totals with P between 0.1 and 0.9. Conversely, estimates with extreme P values show wild patterns; so it is inappropriate to use summarized variance estimates blindly. It is more useful to use aggregated ones that produce more reliable standard errors. A close relationship exists between totals and proportions. So with almost the same information, users can obtain variance estimates for totals and proportions. Finally, in fitting the GVF, I also tried to account for the variability of predictors, which are estimated values as opposed to constants. However, with a large sample size, we have seen this made little difference.

REFERENCES

Bieler, G.S. and Williams, R.L. (1990). "Generalized Standard Error Models for Proportions in Complex Design Surveys." *Proceedings of Section on Survey Research Methods of the American Statistical Association*, pp. 272-277.

Fuller, W.A. (1987). *Measurement Error Models.* New York: John Wiley.

Johnson, E.G. and King, B.F. (1987). "Generalized Variance Functions for a Complex Sample Surveys." *Journal of Official Statistics*, Vol. 3, pp. 235-250.

Salvucci, S., Galfond, G., and Kaufman, S. (1993). "Generalized Variance Functions for the Schools and Staffing Surveys." *Proceedings of Section on Survey Research Methods of the American Statistical Association*, pp. 669-674.

Valliant, R. (1987). "Generalized Variance Functions in Stratified Two-Stage Sampling." *Journal of the American Statistical Association*, Vol. 82, pp. 499-508.

Valliant, R. (1992). "Smoothing Variance Estimates for Price Indexes Over Time." *Journal of Official Statistics*, Vol. 8, pp .433-444.

Figure 1: Design Effect Distributions Within and Across Domains

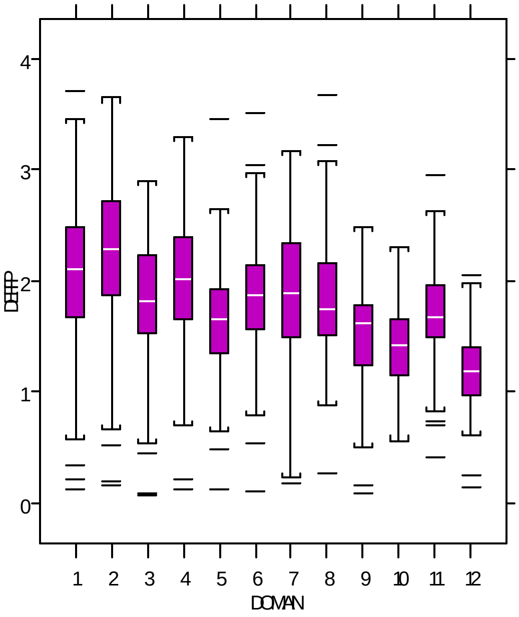


Figure 2: Proportions of RSE less than 20%

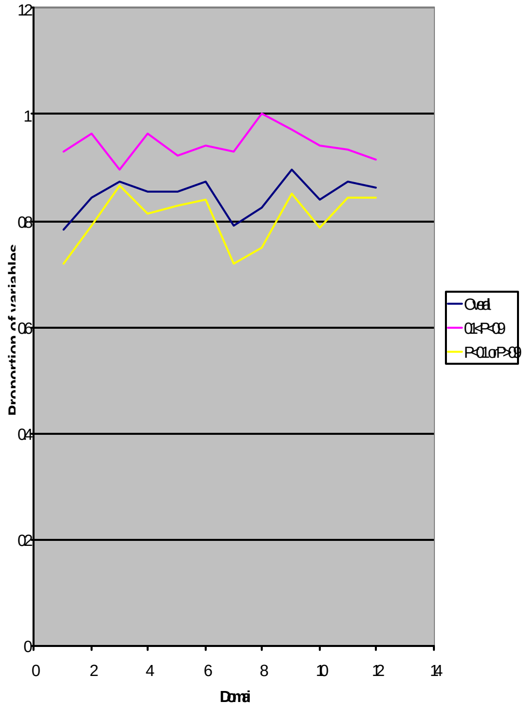


Figure 3: Average RSEs for Proportions from Three Approaches

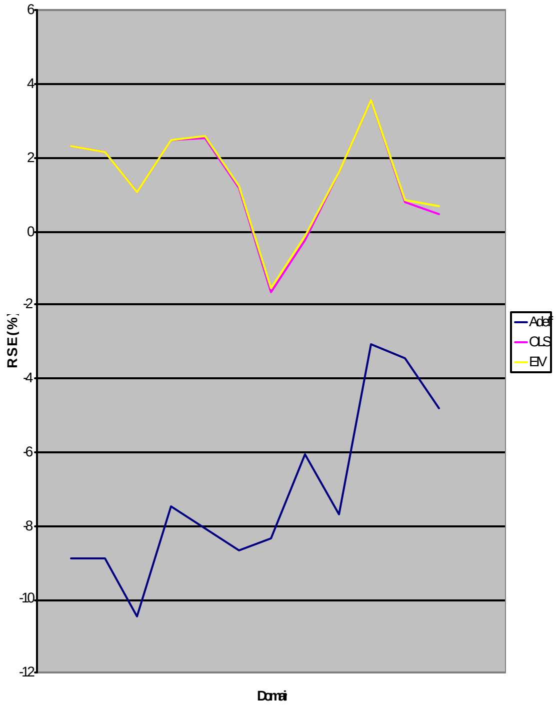
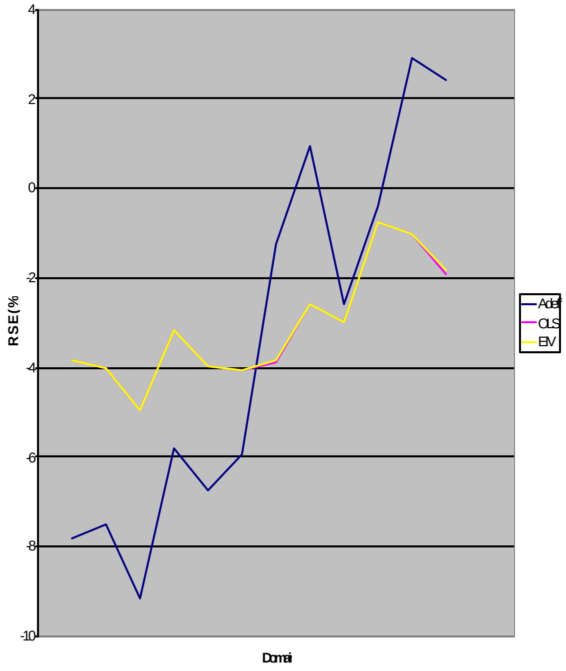


Figure 4: Average RSEs for Totals from Three Approaches


RSE(%
Domai
Adef
OLS
EIV