Analytic Tools for Evaluating Variability of Standard Errors in Large-Scale Establishment Surveys

MJ. Cho, J. Eltinge, J. Gershunskaya and L. Huff

U.S. Bureau of Labor Statistics
cho.moon@bls.gov

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

Outline

Large-scale **establishment surveys** exhibit temporal or cross-sectional **variability** in their published **standard errors**.

Use **generalized variance function** framework to provide tools to evaluate these patterns of variability.

Outline (continued)

- Establishment Survey
- Generalized Variance Functions (GVFs)
- Current Employment Statistics Program (CES)
- Numerical Results

Establishment Survey

- Many survey variables are continuous, heavily **skewed population distribution**.
  In our example, individual employment counts range from single digits to tens of thousands. Most units have counts in the single or double digits.
- Initiation of new sample units can be expensive, and time consuming.
  Slow initiation and **attrition** may lead to increased variability.

Sources of Variability

1. Changes in factors **controllable** (e.g. realized sample size)
2. Changes in factors **observable** but not controllable (e.g. the true population parameter)
3. Changes in factors neither observable nor controllable (e.g. short-term local changes in economic conditions)
4. Sampling variablity of the variance estimator

explore sources of variability using GVF models

BLS

Generalized Variance Function Model

Johnson and King (1987, JOS), Valliant (1987, JASA), Wolter (2007, Ch 7)

Mathematical model describing the relationship between variance of a survey estimator and predictors

Generalized Variance Function Model

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

Given a domain j,

$V_{p\,j}$: true design-based variance
θ_j: a finite population mean or total
X_j: a vector of predictor variables
γ: a vector of function parameters
q_j: a random error with the mean 0

Current Employment Statistics Program

collects data on employment, hours and earnings of nonfarm establishments

- Active CES sample includes approximately one third of all nonfarm payroll employees
- When firms are sampled, they are retained for two years or more

BLS

Sample Design and Special Features

- Sample Unemployment Insurance (**UI**) accounts
- **Stratification** by state, industry and employment size class
- Complete **universe employment counts** of the previous year become available from the **Quarterly Census of Employment and Wages** employment total on a lagged basis
- **Benchmark** the sample estimates annually

BLS

Point Estimators

Given a domain j and a month t,

$$\hat{\theta}_{jt,total} = x_{j0}\,\hat{R}_{jt}$$

$\hat{\theta}_{jt,total}$:	estimator of total employment
x_{j0}:	known total at benchmark month 0
$\hat{R}_{jt}$:	growth ratio estimator from 0 to month t

BLS

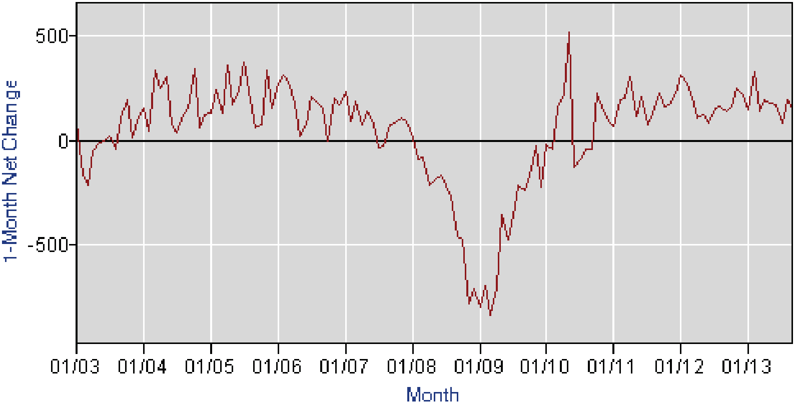
Point Estimators (continued)

$$\hat{\theta}_{jt,change} = \hat{\theta}_{jt} - \hat{\theta}_{j,t-1}$$

$$\hat{\theta}_{jt,ratio} = \hat{\theta}_{jt} \,/\, \hat{\theta}_{j,t-1}$$

BLS

One-Month Employment Change



THOUSANDS; Seasonally Adjusted; http://www.bls.gov/

BLS

Use **generalized variance function** framework
to evaluate **temporal**
or **cross-sectional variability**
in design variance of the CES

Common Group

Find groups of domains with similar GVF coefficients γ
(Wolter, 2007, Section 7.3)

- In CES application, group by years or industries
- Empirical evidence of equality or inequality of coefficients across groups
- Need satisfactory estimator of $V(\hat{\gamma})$

Prospective Models (f)

$$log(V_{jt}) = \gamma_0 + \gamma_1 log(x_{j0}) + \gamma_2 log(t) + \gamma_3 log(n_{jt}) + q_{jt}$$

x_{j0}: known total employment at benchmark month 0

t: distance from 0 to reference month t

n_{jt}: sample size

q_{jt}: a random univariate "equation error" reflecting lack of model fit

estimated γ by OLS regression

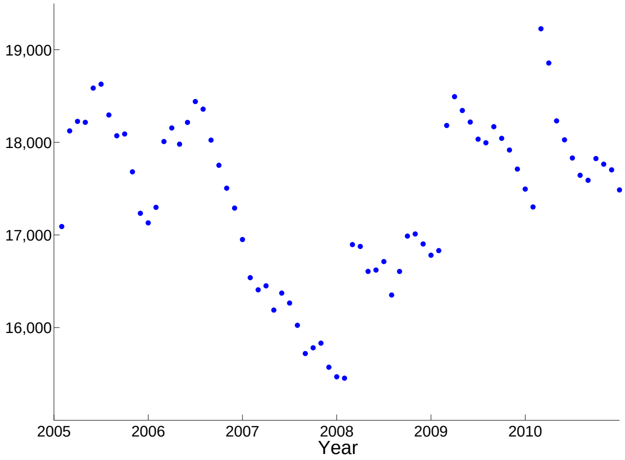
BLS

Observed Patterns in n_{jt}

number of responding sample units

- Substantial variability across industries
- "Saw tooth" patterns due to periodic initiation of new units and continuing attrition of current units

Number of Responding Sample Units across Years: Construction



Coefficient Estimates for Model (f)

$$log(V_{jt}) \quad = \quad \gamma_0 + \gamma_1 log(x_{j0}) + \gamma_2 log(t) + \gamma_3 log(n_{jt}) + q_{jt}$$

	intercept γ_0	$\log(x_{j0})$ γ_1	$\log(t)$ γ_2	$\log(n_{jt})$ γ_3
EST.	**-1.43**	**1.16**	**1.17**	**0.22**
s.e.	0.66	0.09	0.07	0.12
t_γ	-2.17	12.77	16.72	1.78

confounding of x_{j0} with n_{jt}

$\log(n_{jt})$ *provided very limited additional value*

BLS

Final Model

$$log(V_{jt}) \;=\; \gamma_0 + \gamma_1 log(x_{j0}) + \gamma_2 log(t) + q_{jt}$$

x_{j0}: known total employment at benchmark month
t: distance from 0 to reference month t
q_{jt}: a random univariate “equation error” reflecting lack of model fit

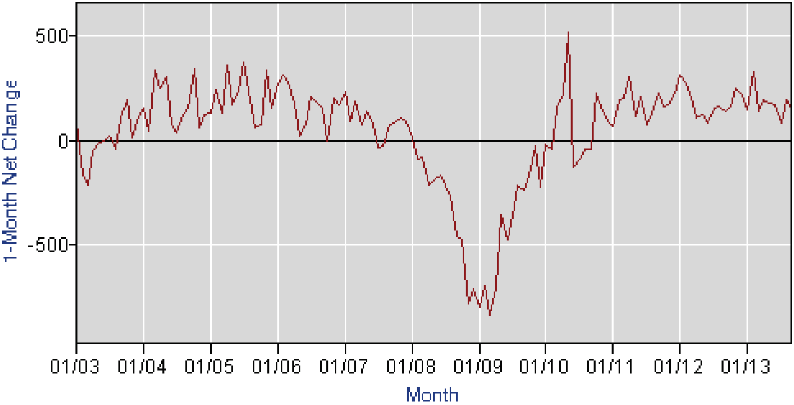
Testing Homogeneity of Coefficient γ: Estimating Equation Approach (cf. Binder, 1983)

- **Sample design** and **estimation** features are important
- Dependent variables $\hat{V}_{jt}$ may be strongly **correlated** across months, due to the **form of the estimators** as well as the use of a **rotation** sample design
- Sampling is essentially **independent across domains**
- Thus, decompose estimating equation into sum of terms across independent domains

test temporal and cross sectional homogeneity in the CES

BLS

One-Month Employment Change



THOUSANDS; Seasonally Adjusted; http://www.bls.gov/

BLS

Temporal homogeneity

$$log(V_{jt}) = \begin{cases} \gamma_{10} + \gamma_{11} log(x_{j0}) + \gamma_{12} log(t) + q_{jt} & \text{if 2005-2007} \\ \gamma_{20} + \gamma_{21} log(x_{j0}) + \gamma_{22} log(t) + q_{jt} & \text{if 2008-2010} \end{cases}$$

Test homogeneity of coefficients across year groups:

$$H_0 : (\gamma_{10}, \gamma_{11}, \gamma_{12}) = (\gamma_{20}, \gamma_{21}, \gamma_{22})$$

$$W = (A\,\hat{\gamma})' \, [(A\,V(\hat{\gamma})\,A')']^{-1} \, (A\,\hat{\gamma})$$

$$\text{where } A = \begin{pmatrix} 1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 & 0 & -1 \end{pmatrix}, \quad \hat{\gamma} = \begin{pmatrix} \gamma_{10} \\ \gamma_{11} \\ \gamma_{12} \\ \gamma_{20} \\ \gamma_{21} \\ \gamma_{22} \end{pmatrix}$$

BLS

Grouping by Years: cutoff=9.69

Year Group 1 : 2005-2007
Year Group 2 : 2008-2010

Estimator	$\gamma_{1,0}$	$\gamma_{1,1}$	$\gamma_{1,2}$	$\gamma_{2,0}$	$\gamma_{2,1}$	$\gamma_{2,2}$	*W*
Total	0.26	1.08	1.33	0.38	1.15	0.87	11.14
(s.e.)	(4.46)	(0.27)	(0.12)	(2.50)	(0.16)	(0.09)	
Change	-2.18	1.27	0.32	-1.45	1.29	-0.12	6.95
(s.e.)	(2.65)	(0.17)	(0.13)	(2.84)	(0.18)	(0.11)	
Ratio	-2.27	-0.72	0.30	-1.60	-0.71	-0.09	5.73
(s.e.)	(2.59)	(0.17)	(0.13)	(2.75)	(0.17)	(0.10)	

significant coefficients for $\log(x_{j0})$
test statistic for total is larger than cutoff point at $\alpha = 0.05$

BLS

Description of Industries

Industry	Description	Classification
1	Mining and logging	Goods-producing
2	Construction	Goods
3	Durable goods manufacturing	Goods
4	Non-durable goods manufacturing	Goods
5	Wholesale trade	Service-providing
6	Retail trade	Service
7	Transportation and warehousing	Service
8	Utilities	Service
9	Information	Service
10	Financial activities	Service
11	Professional and business services	Service
12	Education and health services	Service
13	Leisure and hospitality	Service
14	Other services	Service

Cross-Sectional Homogeneity (cutoff=12.72)

$$log(V_{jt}) = \begin{cases} \gamma_{10} + \gamma_{11} log(x_{j0}) + \gamma_{12} log(t) + q_{jt} & \text{if Goods} \\ \gamma_{20} + \gamma_{21} log(x_{j0}) + \gamma_{22} log(t) + q_{jt} & \text{if Service} \end{cases}$$

Test statistic similar to year-group case

Estimator	$\gamma_{1,0}$	$\gamma_{1,1}$	$\gamma_{1,2}$	$\gamma_{2,0}$	$\gamma_{2,1}$	$\gamma_{2,2}$	*W*
Total	6.69	0.69	1.15	-2.35	1.29	1.08	15.94
(s.e.)	(2.17)	(0.16)	(0.10)	(3.02)	(0.18)	(0.12)	
Change	4.87	0.90	-0.25	-4.73	1.44	0.23	65.33
(s.e).	(1.61)	(0.12)	(0.07)	(1.86)	(0.13)	(0.12)	
Ratio	4.27	-1.07	-0.21	-4.68	-0.56	0.23	53.52
(s.e.)	(1.64)	(0.12)	(0.08)	(1.90)	(0.13)	(0.12)	

strong indication of differences in the Goods and Services coefficients

BLS

Quantiles of Residuals from Two Groups of Industries

(for total employment)

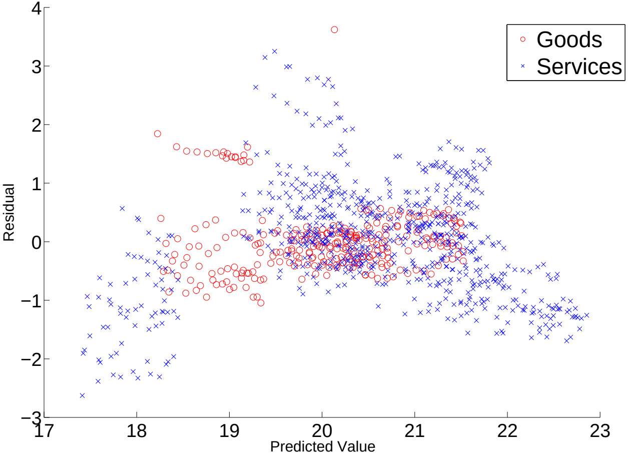
$$\hat{q}_{jt} = log(\hat{V}_{jt}) - X_{jt}\,\hat{\gamma}$$

Group	0.01	0.10	0.25	0.50	0.75	0.90	0.99	IQR
Goods	-0.98	-0.69	-0.52	-0.23	0.65	1.07	1.52	1.17
Services	-1.63	-0.79	-0.42	-0.02	0.38	0.84	2.26	0.80

Goods-producing industries have a wider IQR

BLS

Log-Scale Residuals against Predicted Values $(X_{jt}\hat{\gamma})$



Summary

- Presented tools to evaluate patterns of variability using generalized variance function framework
- Evaluated temporal and cross-sectional variability by examining GVF coefficients across groups
- For GVF coefficients, estimated their variance estimators using estimating-equation to take into account for clustering

THANK YOU.

MoonJung Cho
Cho.Moon@BLS.GOV