Topics in Model-Assisted Point and Variance Estimation in Clustered Samples

By

Timothy Kennel

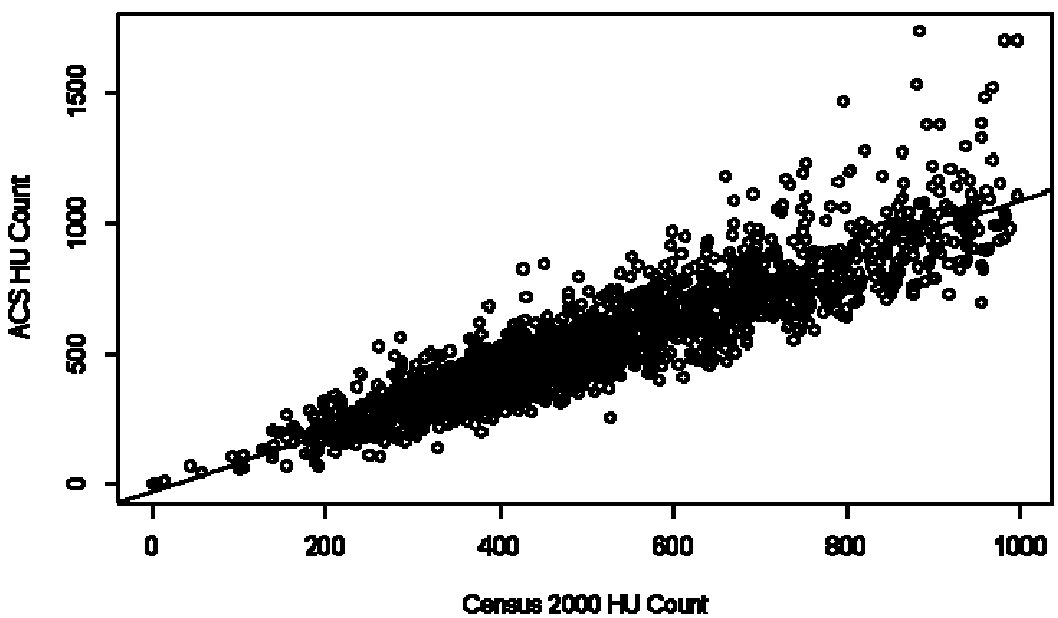
Federal Committee on Statistical Methodology Research Conference

THURSDAY, DECEMBER 3, 2015

Outline

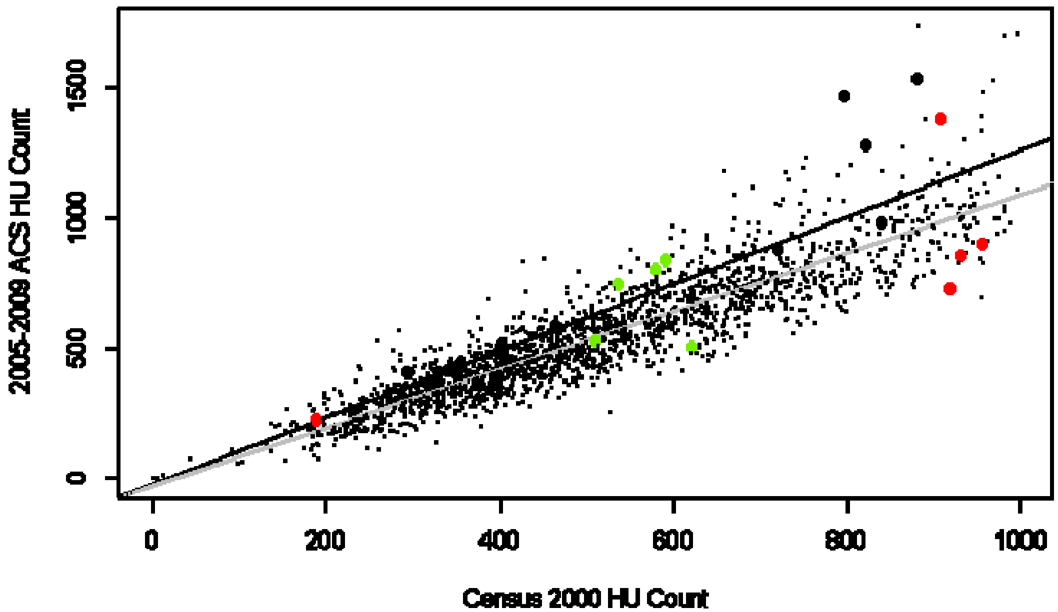
1. Improved Variance Estimators for Generalized Regression Estimators in Cluster Samples
2. Multivariate Logistic-Assisted Estimators of Totals from Clustered Survey Samples in the presence of Complete Auxiliary Information
3. Design-based Inference Assisted by Generalized Linear Models for Cluster Samples

Population



ACS HU Count
0
500
1000
1500
0
200
400
600
800
1000
Census 2000 HU Count

Sample Leverages



Estimator

- Generalized Regression Estimator (GREG)
  - $\hat{t}_y^{gr} = \sum_{\in \ U} \hat{y}_k + \sum_{\in \ s} d_k(y_k - \hat{y}_k)$
  - $var_M\left(\hat{t}_y^{gr}\right) = \sum_{\in \ s} \boldsymbol{g}_i^T \boldsymbol{\Pi}_i^{-1} \psi_i \boldsymbol{\Pi}_i^{-1} \boldsymbol{g}_i$

- Sandwich Variance Estimators
  - $v_R = \sum_{\in \ s} \boldsymbol{g}_i^T \boldsymbol{\Pi}_i^{-1} \boldsymbol{r}_i \boldsymbol{r}_i^T \boldsymbol{\Pi}_i^{-1} \boldsymbol{g}_i$
  - $v_D = \sum_{\in \ s} \boldsymbol{g}_i^T \boldsymbol{\Pi}_i^{-1} (\boldsymbol{I}_n - \boldsymbol{H}_{\boldsymbol{ii}})^{-1} \boldsymbol{r}_i \boldsymbol{r}_i^T \boldsymbol{\Pi}_i^{-1} \boldsymbol{g}_i$
  - $v_J = \sum_{\in \ s} \boldsymbol{g}_i^T \boldsymbol{\Pi}_i^{-1} (\boldsymbol{I}_n - \boldsymbol{H}_{\boldsymbol{ii}})^{-1} \boldsymbol{r}_i \boldsymbol{r}_i^T (\boldsymbol{I}_n - \boldsymbol{H}_{\boldsymbol{ii}})^{-1} \boldsymbol{\Pi}_i^{-1} \boldsymbol{g}_i$

Confidence Interval Coverage

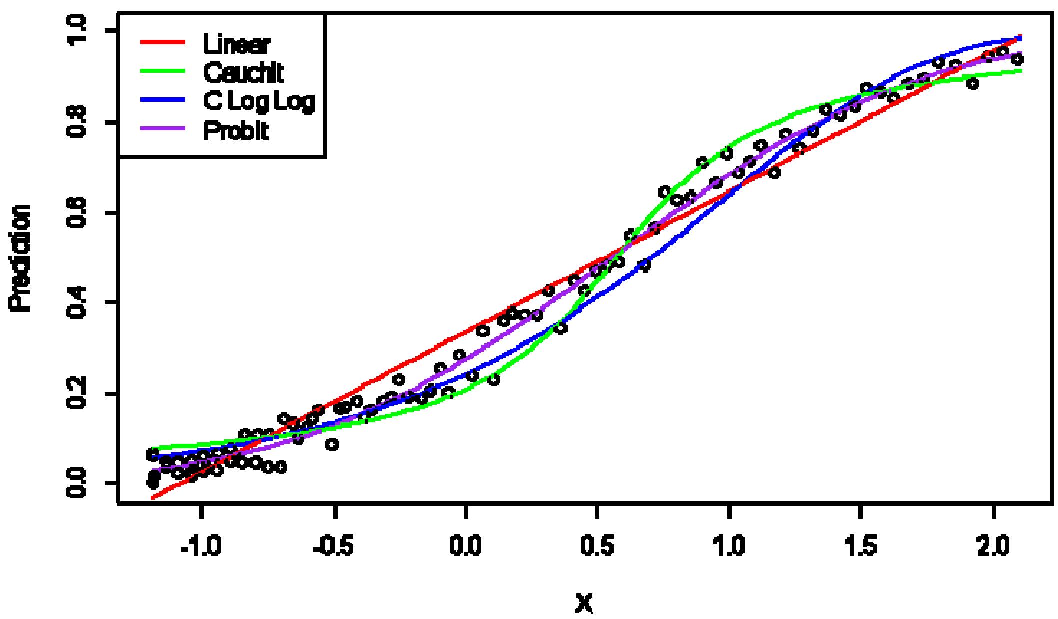
Estimator	**Lower**	**Middle**	**Upper**
Empirical	3.9	95.3	0.8
v_R	18.3	77.2	4.5
v_D	10.8	87.0	2.2
v_J	4.9	94.1	1.0

Conclusion of Leverage Adjusted Variance Estimators

- Small samples
  - Confidence interval coverage is closer to nominal value.
  - Central tendency (median) is closer to true value.
  - Extreme estimates are possible.
  - More variable.
- Large samples
  - Confidence interval coverage is closer to nominal value.
  - Conservative estimates.
  - Asymptotically unbiased.

Design-based Inference Assisted by Generalized Linear Models for Clustered Samples in the Presence of Complete Auxiliary Information

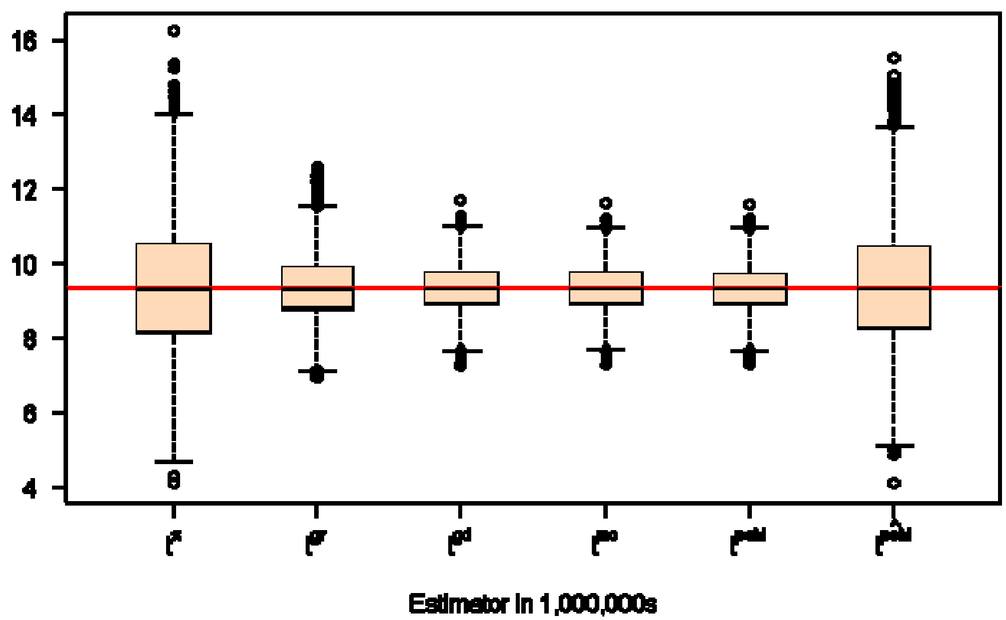
Example of a Binary Response from the 2000 Tract Level Planning Database



Estimators

- $\hat{t}_y^{\pi} = \sum_{\in\ s} d_k y_k$
- $\hat{t}_y^{pr} = \sum_{\in\ U} \hat{\mu}_k$
- $\hat{t}_y^{gr} = \sum_{\in\ U} \hat{y}_k + \sum_{\in\ s} d_k(y_k - \hat{y}_k)$
- $\hat{t}_y^{gd} = \sum_{\in\ U} \hat{\mu}_k + \sum_{\in\ s} d_k(y_k\ \ \mu\,\hat{}_k)$
- $\hat{t}_y^{mc} = \sum_{\in\ s} w_k^{mc} y_k$
- $\hat{t}_y^{peM} = M \sum_{\in\ s} p_k^{pe} y_k$
- $\hat{t}_y^{pe\hat{M}} = \hat{M} \sum_{\in\ s} p_k^{pe} y_k$

Box Plot of Logistic-Assisted Estimators of Renters in Large Samples



Results

- Calibrated estimators are asymptotically unbiased.
- Use canonical ink or calibrated estimators.
- Clear variance reductions of $\hat{t}_y^{gd}$, $\hat{t}_y^{mc}$, and $\hat{t}_y^{peM}$ over established estimators.
- GLM-assisted estimators require complete data.
- Estimators could be unstable in small samples.
- Performance of variance estimators depends on the sample design and sample size.

Contact

- Timothy Kennel
  - Email: Timothy.L.Kennel@census.gov