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
Sample Leverages

![Graph showing the relationship between Census 2000 HU Count and 2005-2009 ACS HU Count. The data points form a trend line indicating a positive correlation.]
Estimator

• Generalized Regression Estimator (GREG)
  • \( \hat{t}_y^{gr} = \sum_{\epsilon U} \hat{y}_k + \sum_{\epsilon s} d_k (y_k - \hat{y}_k) \)
  • \( \text{var}_M(\hat{t}_y^{gr}) = \sum_{\epsilon s} g_i^T \Pi_i^{-1} \psi_i \Pi_i^{-1} g_i \)

• Sandwich Variance Estimators
  • \( \nu_R = \sum_{\epsilon s} g_i^T \Pi_i^{-1} r_i r_i^T \Pi_i^{-1} g_i \)
  • \( \nu_D = \sum_{\epsilon s} g_i^T \Pi_i^{-1} (I_n - H_{ii})^{-1} r_i r_i^T \Pi_i^{-1} g_i \)
  • \( \nu_J = \sum_{\epsilon s} g_i^T \Pi_i^{-1} (I_n - H_{ii})^{-1} r_i r_i^T (I_n - H_{ii})^{-1} \Pi_i^{-1} g_i \)
## Confidence Interval Coverage

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Lower</th>
<th>Middle</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Empirical</td>
<td>3.9</td>
<td>95.3</td>
<td>0.8</td>
</tr>
<tr>
<td>$\nu_R$</td>
<td>18.3</td>
<td>77.2</td>
<td>4.5</td>
</tr>
<tr>
<td>$\nu_D$</td>
<td>10.8</td>
<td>87.0</td>
<td>2.2</td>
</tr>
<tr>
<td>$\nu_J$</td>
<td>4.9</td>
<td>94.1</td>
<td>1.0</td>
</tr>
</tbody>
</table>
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

- $\hat{t}_y^\pi = \sum_{s} s \ d_k y_k$
- $\hat{t}_y^{pr} = \sum_{u} u \ \hat{\mu}_k$
- $\hat{t}_y^{gr} = \sum_{u} u \ \hat{y}_k + \sum_{s} s \ d_k (y_k - \hat{y}_k)$
- $\hat{t}_y^{gd} = \sum_{u} u \ \hat{\mu}_k + \sum_{s} s \ d_k (y_k \ \mu \hat{\mu}_k)$

- $\hat{t}_y^{mc} = \sum_{s} w_{mc} y_k$
- $\hat{t}_y^{peM} = M \sum_{s} p_{k}^{pe} y_k$
- $\hat{t}_y^{pe\hat{M}} = \hat{M} \sum_{s} p_{k}^{pe} y_k$
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