Regression Composite Estimation for Current Population Survey

Tim Trudell
U.S. Census Bureau

Joint work with
Yang Cheng, formerly U.S. Census Bureau
Daniel Bonnery, University of Maryland
Partha Lahiri, University of Maryland

2019 Joint Statistical Meetings, Denver

The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.
Outline

1. Background
2. Current Population Survey (CPS) sample design
3. CPS direct estimator
4. AK composite estimation
5. Composite regression estimation
6. Simulation study
7. Application to CPS data
8. Conclusion
CPS background

- American labor force statistics
- Oldest national survey - since 1940
- A household sample survey sponsored by Bureau of Labor Statistics and U.S. Census Bureau
- Monthly sample of 72,000 Households
- Primary source of labor force data:
  - Monthly national unemployment rate
- Labor force data for people aged 16 and over
CPS sample design

Stratified multi-stage sampling design:

• Primary Sampling Unit (PSU)
  • Consist of county or a group of counties
  • 1,987 PSUs with 506 self-representing (SR) PSUs, roughly 70% of total population (2,205 PSUs with 446 SR PSU for 2000)
  • Group non-self-representing (NSR) PSUs into stratum

• Stratified two-stage design on NSR groups
  • First stage: select one PSU per stratum by probability proportional to size method restricted to population 16+
  • Second stage: Systematic sampling on the clusters of 4 sampled housing units
CPS rotation panel design

• Repeated rotating sample design with 8 rotation panels
• 4-8-4 Sample Pattern: a housing unit selected is interviewed for 4 consecutive months, out for 8 months, and then interview for another 4 months
• Survey modes: personal visit interviews and telephone interviews based on sampling in the different panels
• Self weighting at the state level: all sampled units have equal weights. Self weighting samples often yield smaller variance, and sample statistics are more robust
### CPS/SCHIP ROTATION CHART

**January 2012 - March 2014**

**Sample and Rotation**

<table>
<thead>
<tr>
<th>Year/Month</th>
<th>A88/B88</th>
<th>A89/B89</th>
<th>A90/B90</th>
<th>A91/B91</th>
<th>A92/B92</th>
<th>A93/B93</th>
</tr>
</thead>
<tbody>
<tr>
<td>2012</td>
<td>JAN</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>. .</td>
</tr>
<tr>
<td></td>
<td>FEB</td>
<td>. .</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>MAR</td>
<td>. .</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>APR</td>
<td>. .</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>MAY</td>
<td>. .</td>
<td>7</td>
<td>8</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>JUNE</td>
<td>. .</td>
<td>8</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>JULY</td>
<td>. .</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>AUG</td>
<td>. .</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>2013</td>
<td>JAN</td>
<td>. .</td>
<td>7</td>
<td>8</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>FEB</td>
<td>. .</td>
<td>8</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>MAR</td>
<td>. .</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>APR</td>
<td>. .</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>MAY</td>
<td>. .</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>JUNE</td>
<td>. .</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>JULY</td>
<td>. .</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>AUG</td>
<td>. .</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>SEPT</td>
<td>. .</td>
<td>7</td>
<td>8</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>OCT</td>
<td>. .</td>
<td>8</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>NOV</td>
<td>. .</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>DEC</td>
<td>. .</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>2014</td>
<td>JAN</td>
<td>. .</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>FEB</td>
<td>. .</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>MAR</td>
<td>. .</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
</tbody>
</table>
Notations

- \( t \): time indicator (month)
- \( U_t \): the population at the time \( t \), of size \( N_t \)
- \( S_{t,i} \): the \( i \)th rotation group for month \( t \)
- \( S_t = U_{i=1}^8 S_{t,i} \)
- \( w_{t,k} \): the weight after adjustment of individual \( k \) at month \( t \)
- \( y_t \): a vector of study variables \( (y_{t,k})_{k \in U_t} \)
- \( Y_t = \sum_{k \in U_t} y_{t,k} \): the unknown total to be estimated at month \( t \)
Direct estimator

The weights $w_{t,k}$ are obtained after different adjustment procedures:
- 2 procedures at the household level
- 5 procedures at the individual level

The direct estimator of $Y_t$ is

$$\hat{Y}_t = \frac{1}{8} \sum_{i=1}^{8} \hat{Y}_{t,i}$$

where $\hat{Y}_{t,i} = \sum_{k \in s_{t,i}} w_{t,k} y_{t,k}$, the estimator of $Y_t$ based on data in the $i$th rotation panel at month $t$, $i = 1, \ldots, 8$. 
Composite estimator before 1985

Let $\Delta_t = Y_t - Y_{t-1}$ be the month-to-month change. Then

$$Y_t = Y_{t-1} + \Delta_t$$

which suggests that we may construct more efficient estimator by incorporating historical information properly. Because of the 4-8-4 sample rotation scheme, six of eight rotation panels in the sample for month $t-1$ remains in sample for month $t$. The month-to-month change can be estimated as

$$\hat{\Delta}_t = \frac{1}{6} \sum_{i \in S} (\hat{Y}_{t,i} - \hat{Y}_{t-1,i-1})$$

where $S = \{2, 3, 4, 6, 7, 8\}$. The composite estimator before 1985 is defined as

$$\hat{Y}_t^C = (1 - K)\hat{Y}_t + K(\hat{Y}_{t-1}^C + \hat{\Delta}_t), \quad 0 \leq K \leq 1$$
AK composite estimator

• In 1985, a different composite estimator was introduced by adding another term to the previously described composite estimator $\hat{Y}_t^C$, a term that is the estimator of the net difference between the incoming and continuing parts of the current month’s sample:

$$\hat{\beta}_t = \frac{1}{8} \left( \sum_{i \notin S} \hat{Y}_{t,i} - \frac{1}{3} \sum_{i \in S} \hat{Y}_{t,i} \right)$$

• Then the AK composite estimator is defined as

$$\hat{Y}_t^{AK} = (1 - K)\hat{Y}_t + K(\hat{Y}_{t-1}^{AK} + \hat{\Delta}_t) + A\hat{\beta}_t, \ 0 \leq K \leq 1.$$ 

Note: constants A and K are determined empirically.
Composite regression estimation

  Composite estimation by modified regression for repeated surveys
  *ASA Proceedings Survey Research Methods Section*, page 400-425

• Fuller, W. A. and Rao, J. (2001)
  A regression composite estimator with application to the Canadian labor force survey
  *Survey Methodology*, Vol. 27, No. 1, page 45-51
Composite regression estimation (cont.)

• \( x_{t,k} \): a vector of auxiliary variables for unit \( k \) at time \( t \)

• Definition of \( z_{t,k} \) and \( \hat{t}^C_{y,t} \):

\[
\begin{align*}
z_{t,k} &= \begin{cases} 
\alpha \left( \tau^{-1}[y_{t-1,k} - y_{t,k}] + y_{t,k} \right) + (1 - \alpha)y_{t-1,k} & \text{if } k \in S_t \cap S_{t-1} \\
\alpha y_{t,k} + (1 - \alpha) N_{t-1} t^C_{y,t-1} & \text{if } k \in S_t \setminus S_{t-1}
\end{cases}
\end{align*}
\]

and

\[
\hat{t}^C_{y,t} = \sum_{k \in S_t} w^C_{t,k} y_{t,k}
\]

• \( \tau \) is a fixed number, which closes to \( \left( \sum_{k \in S_t} w_{t,k} \right)^{-1} \sum_{k \in S_t \cap S_{t-1}} w_{t,k} \)
Composite regression estimation  (cont.)

• \( w_{t,k}^C \) minimizes \( \sum_k (w_{t,k}^C - w_{t,k})^2 / w_{t,k} \) under the constraints:
  \[
  \sum_{k \in S_t} w_{t,k}^C x_{t,k} = X_t
  \]
  and
  \[
  \sum_{k \in S_t} w_{t,k}^C z_{t,k} = \hat{y}_{y,t}
  \]
  Then, the regression composite estimator of \( Y_t \) is
  \[
  \hat{\hat{y}}_{y,t}^C = \sum_{k \in S_t} w_{t,k}^C y_{t,k}
  \]
Simulation study

• Create population of size 100,000
• Rotation group size: 100
• Employment status constrained at the population level to match CPS estimates
• Change of status affects the smallest number of people
• Number of replications of the sample selection process: 1,000
Figure: Month-to-month change estimates

(a) Mean

(b) Bias

(g) MSE

Methods:
- Direct
- AK
- R.C., α = 0.95

Data range: 2005-2011

U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov
Figure: Level estimates
Figure: Month-to-month change estimates
Application to CPS data
Figure: Level estimates: difference with direct estimates

- AK
- Direct
- RC

2005 2006 2007 2008 2009 2010 2011 2012
Figure: Month-to-month: difference with direct estimates

AK
Direct
RC

2005 2006 2007 2008 2009 2010 2011

5e-04 0e+00 -5e-04 -1e-03
Conclusion

• Regression Composite performs better with respect to month-to-month change estimation on simulated population
• Regression Composite estimate is closer to the direct estimate
• Weights ratio dispersion after Regression Composite adjustment is smaller
Thank you! Questions?

Contact information:

Tim Trudell: tim.trudell@census.gov