Integration of anterior samples information and modeling in repeated survey with complex rotation pattern: an illustration with CPS

Workshop on Data Integration

D. Bonnéry, JPSM, UMD
Based on joint work with
Y. Cheng (Census Bureau) and **P. Lahiri** (JPSM, UMD).

August 6, 2019

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

1. The Current Population Survey (CPS)
2. Model-assisted approach
3. Maximum likelihood estimation
4. Application to CPS

The Current Population Survey (CPS)

- American labor force statistics
- Stratified multi-stage sampling design
- Repeated survey with
  - 8 rotation groups (samples of households) per month
  - rotation pattern: 4-8-4

- m: time indicator (month).
- U_m: the population at time m,
- S_m: the sample at time m,
- $w_{m,k}$: the sampling adjusted weight indivudual k at time m,
- Employment status of individual k at time m is:

$$\mathrm{y}_{k,m} = \begin{cases} \mathrm{status}_1 = (1,0,0) & \text{if } k \text{ is "employed"} \\ \mathrm{status}_2 = (0,1,0) & \text{if } k \text{ is "unemployed"} \\ \mathrm{status}_3 = (0,0,1) & \text{if } k \text{ is "not in the labor force"} \end{cases}$$

- $\mathrm{t}_{\mathrm{y}_m} = \sum_{k \in U_m} \mathbf{y}_{k,m}$.

All rotation groups $G_1, \ldots$ drawn simultenaously.

Table: CPS Rotation chart

	G_1	G_2	G_3	G_4	G_5	G_6	G_7	G_8	G_9	G_{10}	G_{11}	G_{12}	G_{13}	G_{14}	G_{15}	G_{16}	G_{17}	G_{18}	G_{19}	G_{20}
Jan 05	$S_{1,8}$	$S_{1,7}$	$S_{1,6}$	$S_{1,5}$									$S_{1,4}$	$S_{1,3}$	$S_{1,2}$	$S_{1,1}$				
Feb 05		$S_{2,8}$	$S_{2,7}$	$S_{2,6}$	$S_{2,5}$									$S_{2,4}$	$S_{2,3}$	$S_{2,2}$	$S_{2,1}$			
Mar 05			$S_{3,8}$	$S_{3,7}$	$S_{3,6}$	$S_{3,5}$									$S_{3,4}$	$S_{3,3}$	$S_{3,2}$	$S_{3,1}$		
Apr 05				$S_{4,8}$	$S_{4,7}$	$S_{4,6}$	$S_{4,5}$									$S_{4,4}$	$S_{4,3}$	$S_{4,2}$	$S_{4,1}$	
May 05					$S_{5,8}$	$S_{5,7}$	$S_{5,6}$	$S_{5,5}$									$S_{5,4}$	$S_{5,3}$	$S_{5,2}$	$S_{5,1}$
Jun 05						$S_{6,8}$	$S_{6,7}$	$S_{6,6}$	$S_{6,5}$									$S_{6,4}$	$S_{6,3}$	$S_{6,2}$
Jul 05							$S_{7,8}$	$S_{7,7}$	$S_{7,6}$	$S_{7,5}$									$S_{7,4}$	$S_{7,3}$
Aug 05								$S_{8,8}$	$S_{8,7}$	$S_{8,6}$	$S_{8,5}$									$S_{8,4}$
Sep 05									$S_{9,8}$	$S_{9,7}$	$S_{9,6}$	$S_{9,5}$								
Oct 05										$S_{10,8}$	$S_{10,7}$	$S_{10,6}$	$S_{10,5}$							
Nov 05											$S_{11,8}$	$S_{11,7}$	$S_{11,6}$	$S_{11,5}$						
Dec 05												$S_{12,8}$	$S_{12,7}$	$S_{12,6}$	$S_{12,5}$					
Jan 06													$S_{13,8}$	$S_{13,7}$	$S_{13,6}$	$S_{13,5}$				
Feb 06														$S_{14,8}$	$S_{14,7}$	$S_{14,6}$	$S_{14,5}$			
Mar 06															$S_{15,8}$	$S_{15,7}$	$S_{15,6}$	$S_{15,5}$		
Apr 06																$S_{16,8}$	$S_{16,7}$	$S_{16,6}$	$S_{16,5}$	
May 06																	$S_{17,8}$	$S_{17,7}$	$S_{17,6}$	$S_{17,5}$
Jun 06																		$S_{18,8}$	$S_{18,7}$	$S_{18,6}$
Jul 06																			$S_{19,8}$	$S_{19,7}$
Aug 06																				$S_{20,8}$

$S_{m,r}$ be the rth rotation group for month m, $S_m = \bigcup_{r=1}^{8} S_{m,r}$.

- How to integrate the information from the previous months ?
- How to account for the special rotation pattern ?

For $m \in \{1, \dots M\}$, $r \in \{1, \dots, 8\}$, $i \in \{1, \dots, 3\}$,

$$\hat{\mathrm{t}}_{\mathbf{y}m}^{\mathrm{mis},r} = 8 \sum_{k \in S_{m,r}} w_{m,k} \mathsf{y}_{k,m}.$$

When $\mathrm{Cov}\left[\hat{\mathrm{t}}_{\mathbf{y}m,i}^{\mathrm{mis},r}, \hat{\mathrm{t}}_{\mathbf{y}m',i'}^{\mathrm{mis},r'}\right]$ is known, the best linear combinaison of the m.i.s. estimates can be defined.

Currently at the Census, the AK estimator is the empirical best linear estimator in a subclass of linear estimators.

- Unrealistic assumptions were made on the covariance matrix: 10-year stationarity. To our knowledge, the model assumptions on the covariance matrix were not studied or called into question.
- Outdated optimisation techniques were used to compute the optimal coefficients.
- The optimal AK estimator for month to month change is not optimal for level estimation.
- The estimators are sensible to the "rotation bias".

Other linear estimates (Yansaneh-Fuller) suffer from the same problem of estimation of covariance.

Regression composite estimators (Statistics Canada)

- Depend on a tuning parameter α.
- Do provide a compromise optimal estimator.
- Perform well on simulations.
- Why is it working ?

Model Assisted estimation

Consider auxiliary variables $\mathbf{x}$ of known total $\mathrm{t}_{\mathbf{x}}$.
In a **model-assisted** approach, Generalized Regression is associated to a **superpopulation linear model**: $\forall k \in U$, $\mathrm{y}_k, \mathrm{x}_k$ distribution is such

$$\mathrm{E}\left[\mathrm{y}_k \mid \mathrm{x}_k\right] = \mathrm{x}_k \beta.$$

Then the Generalized Regression estimator is

$$\hat{\mathrm{t}}_{\mathbf{y}}^{\mathrm{GREG}} = \hat{\mathrm{t}}_{\mathbf{y}}^{\mathrm{Direct}} + \left(\mathrm{t}_{\mathbf{x}} - \hat{\mathrm{t}}_{\mathbf{x}}^{\mathrm{Direct}}\right) \hat{\beta}.$$

It will be design consistent irrespective of the model and effective if the model fits the data.

The generalized regression estimator is also equal to

$$\hat{\mathrm{t}}_{\mathbf{y}}^{\mathrm{c}} = \sum_{k \in S} w_k^* \mathrm{y}_k,$$

where w_k^* minimizes $\sum_k (w_k^* - w_k)^2 / w_k$ under the constraint:

$$\sum_{k \in S} w_k^* \mathrm{x}_k = \mathrm{t}_{\mathbf{x}}.$$

Since $\mathbf{y}_{m-1}$ is correlated to $\mathbf{y}_m$, why not choose $\mathbf{x}_m = \mathbf{y}_{m-1}$?

Note that the linear assumption holds:

$$\mathrm{E}\left[\mathrm{y}_{k,m} \mid \mathrm{y}_{k,m-1}\right] = \mathrm{y}_{k,m-1}\beta_m,$$

where

$$\beta_m = \mathrm{P}\left(\mathrm{y}_{m,k} = \mathrm{status}_{i'} \mid \mathrm{y}_{m-1,k} = \mathrm{status}_i\right)_{i,i'=1,\ldots,3}.$$

The GREG estimator is

$$\hat{\mathrm{t}}^{\mathrm{GREG}}_{\mathbf{y}_m} = \hat{\mathrm{t}}^{\mathrm{Direct}}_{\mathbf{y}_m} + \left(\mathrm{t}_{\mathbf{y}_{m-1}} - \hat{\mathrm{t}}^{\mathrm{Direct}}_{\mathbf{y}_{m-1}}\right)\hat{\beta}_m.$$

But

- the **characteristic** $\mathrm{y}_{m-1,k}$ is unknown for $k \in S_m \setminus S_{m-1}$,
- the **total** $\mathrm{t}_{\mathbf{y}_{m-1}}$ is unknown.

Singh, A. and Merkouris, P. (1995).
Composite estimation by modified regression for repeated surveys.
In *ASA Proc. Surv. Res. Meth. Sec*, pages 420–5.

Solution:

- replace **characteristic** $\mathrm{y}_{m-1,k}$ by a proxy variable,
- replace **total** $\mathrm{t}_{\mathbf{y}_{m-1}}$ by an estimation.

Proxy variables:

- MR1:

$$\mathrm{x}_{m,k,MR1} = \begin{cases} \mathrm{y}_{m-1,k} & \text{if } k \in S_m \cap S_{m-1}, \\ N_{m-1}^{-1} \hat{\mathrm{t}}_{\mathrm{y},t-1}^{\mathrm{c}} & \text{if } k \in S_m \setminus S_{m-1}, \end{cases}$$

- MR2:

$$\mathrm{x}_{m,k,MR2} = \begin{cases} \left(\tau^{-1}[\mathrm{y}_{m-1,k} - \mathrm{y}_{m,k}] + \mathrm{y}_{m,k}\right) & \text{if } k \in S_m \cap S_{m-1}, \\ \mathrm{y}_{m,k} & \text{if } k \in S_m \setminus S_{m-1}, \end{cases}$$

- MR1 will carry misrepresentativeness of first sample over time.
- MR2 will chain evolutions without taking into account the levels of intrant rotation groups.

Solution: $\mathsf{x}_{m,k,\mathrm{MR3}} = [\mathsf{x}_{m,k,\mathrm{MR1}}, \mathsf{x}_{m,k,\mathrm{MR2}}]$

Fuller, W. A. and Rao, J. (2001).

A regression composite estimator with application to the canadian labour force survey.

Survey Methodology, 27(1):45–52.

Solution: define a new proxy variable. Let $\alpha \in [0, 1]$:

$$\mathsf{x}_{m,k} = \alpha\, \mathsf{x}_{m,k,\mathrm{MR1}} + (1 - \alpha)\, \mathsf{x}_{m,k,\mathrm{MR2}}$$

Why is it working ?

The calibration equation can be rewritten:

$$\left(\sum_{k\in S_m\cap S_{m-1}} w^*_{m,k}\mathsf{y}_{m-1,k} + \sum_{k\in S_m\setminus S_{m-1}} w^*_{m,k}\hat{\mathsf{y}}_{m-1,k}\right) - \hat{\mathrm{t}}^{\mathrm{c}}_{\mathbf{y},m-1}$$

$$= (1-\alpha)\left((\tau^{-1}-1)\left(\sum_{k\in S_m\cap S_{m-1}} w^*_{m,k}\left(\mathsf{y}_{m,k}-\mathsf{y}_{m-1,k}\right)\right)\right.$$

$$\left. - \sum_{k\in S_t\setminus S_{m-1}} w^*_{m,k}\left(\mathsf{y}_{m,k}-\hat{\mathsf{y}}_{m-1,k}\right)\right)$$

where $\hat{\mathsf{y}}_{k,m-1} = N^{-1}_{m-1}\hat{\mathrm{t}}^{\mathrm{c}}_{\mathsf{y},m-1}$, $\tau = \frac{3}{4}$

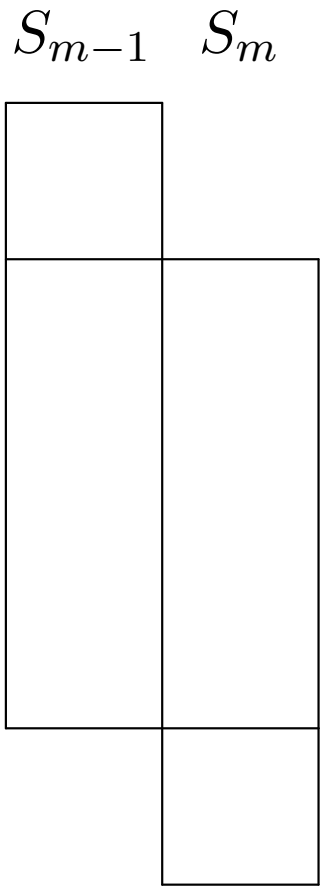
We notice that:

- Composite Regression calibration equation is a balance between two hard constraints: new set of weights will compromise between level and change.
- Does not require revision of $\hat{\mathrm{t}}_{\mathrm{y},m-1}$

Unanswered questions:

- How to compute the variance of this estimator ?
- How to choose the best α ?

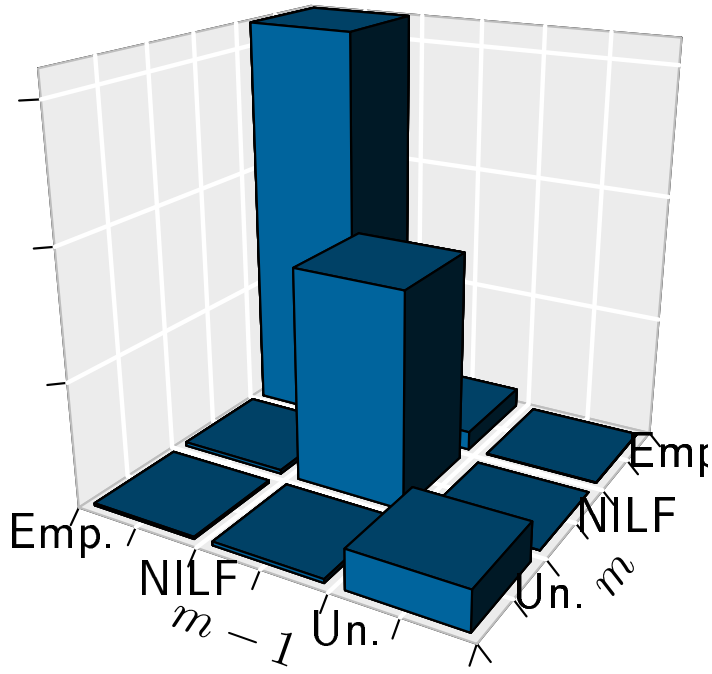
Take a step back:

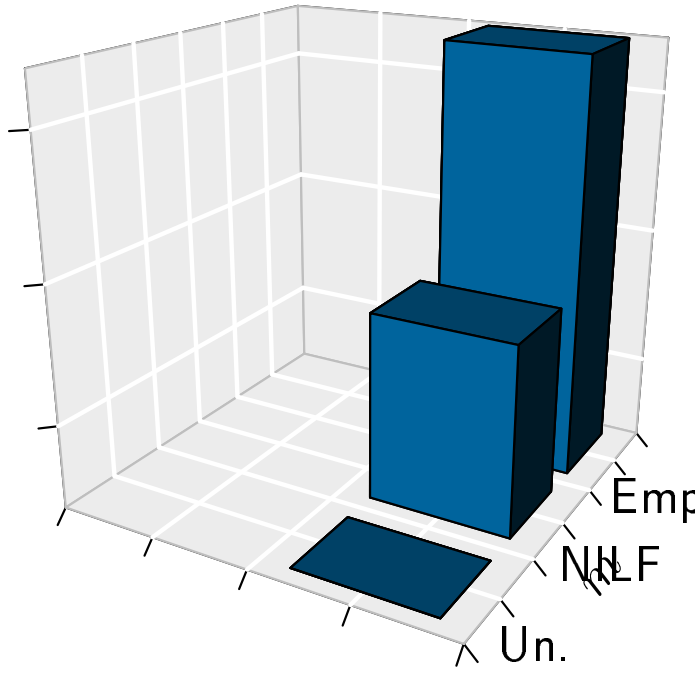


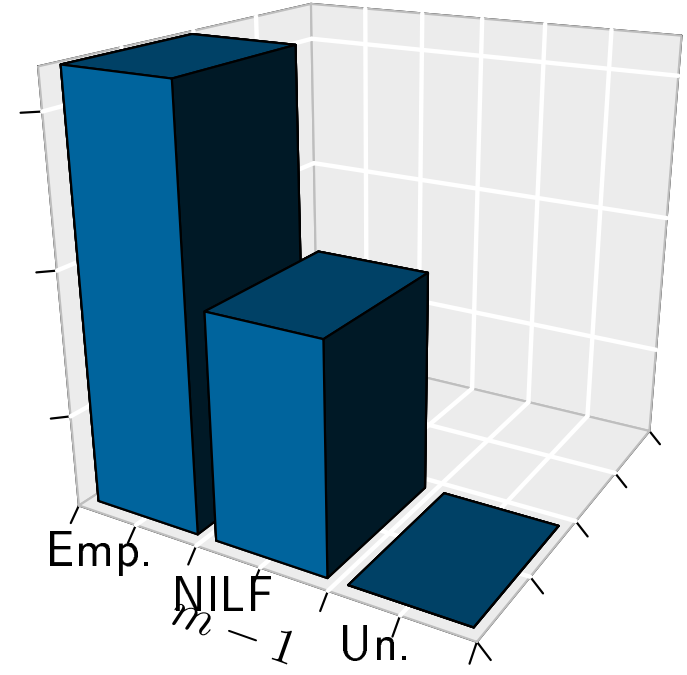
$$(N_{1,.}, N_{2,.}, N_{3,.})$$

$$\begin{pmatrix} N_{1,1} & N_{1,2} & N_{1,3} \\ N_{2,1} & N_{2,2} & N_{2,3} \\ N_{3,1} & N_{3,2} & N_{3,3} \end{pmatrix}$$

$$(N_{.,1}.N_{.,2}, N_{.,3})$$


Emp.
NILF
Un.
$m-1$
Emp
NILF
Un.
m


Emp
NILF
Un.
m


Emp.
NILF
Un.
$m-1$

When

$$N_{i_1,i_2} \sim \text{Multinomial}(\gamma_{i_1,i_2}),$$

the likelihood is

$$\sum_{i_1=1}^{3} N_{i_1,.} \ln\left(\gamma_{i_1,1} + \gamma_{i_1,2} + \gamma_{i_1,3}\right) + \sum_{i_1,i_2=1}^{3} N_{i_1,.} \ln\left(\gamma_{i_1,i_2}\right) + \sum_{i_2=1}^{3} N_{.,i_2} \ln\left(\gamma_{1,i_2} + \gamma_{2,i_2} + \gamma_{3,i_2}\right).$$

We assume the following population model: $\forall k, (\mathsf{y}_{m,k})_{m\in\{1,\dots M\}}$ are the i.i.d realisation of a random variable $Y = (Y_1, \dots, Y_m)$. Let γ a $\underbrace{3 \times 3 \times \dots \times 3}_{M \text{ times}}$ array, such that

$$\gamma_{i_1,\dots,i_M} = \mathrm{P}\left(\bigcap_{m=1}^{M} Y_m = \mathrm{status}_{i_m}\right).$$

Example: for $M = 3$,

$$\gamma_{1,3,3} = \mathrm{P}\left(\{Y_1 = \mathrm{status}_1\} \cap \{Y_2 = \mathrm{status}_3\} \cap \{Y_3 = \mathrm{status}_3\}\right)$$

$$\begin{aligned}
&\mathrm{P}\left(Y_3 = \mathrm{status}_1\right) \\
&\quad = \sum_{i_1=1}^{3} \sum_{i_2=1}^{3} \mathrm{P}\left(\{Y_1 = \mathrm{status}_{i_1}\} \cap \right. \\
&\qquad\qquad \left. \{Y_2 = \mathrm{status}_{i_2}\} \cap \{Y_3 = \mathrm{status}_1\}\right)
\end{aligned}$$

$$\mathrm{P}\left(Y_M = \mathrm{status}_1\right) = \sum_{i_1,\ldots,i_{M-1}=1}^{3} \gamma_{i_1,\ldots,i_{M-1},1}$$

We do not observe y_k, but $y_k|_{A_k}$, for example

$$A_k = \{1, 2, 3, 4, 13, 14, 15, 16\},$$

$$A_k = \{2, 3, 4, 5, 14, 15, 16, 17\}.$$

The contribution to the log-likelihood of $\mathrm{y}_k|_{A_k}$ is:

$$w_k \ln\Big(\mathrm{P}\left(Y|_{A_k} = \mathrm{y}_k|_{A_k}\right)\Big) = w_k \ln\Big(\sum\{\gamma_{i_1,\ldots,i_M} \mid \forall k \in A_k, i_m = y_{k,m}\}\Big) .$$

So the log-likelihood is:

$$\begin{aligned}
&\mathscr{L}\left(\gamma; \left(\mathrm{y}_k|_{A_k}\right)_{k\in\bigcup_{m=1}^{M} S_m}\right) \\
&\quad = \sum_{k\in\bigcup_{m=1}^{M} S_m} w_k \ln\left(\mathrm{P}\left(Y|_{A_k} = \mathrm{y}_k|_{A_k}\right)\right) \\
&\quad = \sum_{k\in\bigcup_{m=1}^{M} S_m} w_k \ln\left(\sum_{i_1,\ldots,i_M|\forall m\in A_k, i_m = y_{k,m}} \gamma_{i_1,\ldots,i_M}\right) .
\end{aligned}$$

If $M = 2$, then $A_k \in \{\{1\}, \{1,2\}, \{2\}\}$. Let

- $N_{i_1,.} = \sum \left\{ w_k \mid k \in \bigcup_{m=1}^{M} S_m,\ A_k = \{1\}, y_1 = i_1 \right\}$
- $N_{i_1,i_2} =$
  $\sum \left\{ w_k \mid k \in \bigcup_{m=1}^{M} S_m,\ A_k = \{1,2\}, y_1 = i_1, y_2 = i_2 \right\}$
- $N_{.,i_2} = \sum \left\{ w_k \mid k \in \bigcup_{m=1}^{M} S_m,\ A_k = \{1\}, y_2 = i_2 \right\}$

The likelihood is

$$\mathscr{L}\left(\gamma; \left(\mathsf{y}_k|_{A_k}\right)_{k \in \bigcup_{m=1}^{M} S_m}\right)$$

$$= \sum_{i_1=1}^{3} N_{i_1,.} \ln\left(\gamma_{i_1,1} + \gamma_{i_1,2} + \gamma_{i_1,3}\right)$$

$$+ \sum_{i_1,i_2=1}^{3} N_{i_1,.} \ln\left(\gamma_{i_1,i_2}\right) + \sum_{i_1=1}^{3} N_{.,i_2} \ln\left(\gamma_{1,i_2} + \gamma_{2,i_2} + \gamma_{3,i_2}\right).$$

If $M = 3$, then $A_k \in \{\{1\}, \{1,2\}, \{1,2,3\}, \{2,3\}, \{3\}\}$.
Let

- $N_{i_1,.,.} = \sum \left\{ w_k \mid k \in \bigcup_{m=1}^{M} S_m,\ A_k = \{1\}, y_1 = i_1 \right\}$
- $N_{i_1,i_2,.} =$
  $\sum \left\{ w_k \mid k \in \bigcup_{m=1}^{M} S_m,\ A_k = \{1,2\}, y_1 = i_1, y_2 = i_2 \right\}$
- $N_{i_1,i_2,i_3} =$
  $\sum \left\{ w_k \mid k \in \bigcup_{m=1}^{M} S_m,\ A_k = \{1,2,3\}, y_1 = i_1, y_2 = i_2, y_3 = i_3 \right\}$
- $N_{.,i_2,i_3} =$
  $\sum \left\{ w_k \mid k \in \bigcup_{m=1}^{M} S_m,\ A_k = \{2,3\}, y_2 = i_2, y_3 = i_3 \right\}$
- $N_{.,.,i_3} = \sum \left\{ w_k \mid k \in \bigcup_{m=1}^{M} S_m,\ A_k = \{3\}, y_3 = i_3 \right\}$

$$\mathscr{L}\left(\gamma;\left(\mathrm{y}_k|_{A_k}\right)_{k\in\bigcup_{m=1}^{M}S_m}\right) = \sum_{i_1=1}^{3} N_{i_1,.,.}\ln\left(\sum_{i_2,i_3=1}^{3}\gamma_{i_1,i_2,i_3}\right) + \sum_{i_1,i_2=1}^{3} N_{i_1,i_2,.}\ln\left(\sum_{i_3=1}^{3}\gamma_{i_1,i_2,i_3}\right) + \sum_{i_1,i_2,i_3=1}^{3} N_{i_1,i_2,i_3}\ln\left(\gamma_{i_1,i_2,i_3}\right) + \sum_{i_2,i_3=1}^{3} N_{.,i_2,i_3}\ln\left(\sum_{i_1=1}^{3}\gamma_{i_1,i_2,i_3}\right) + \sum_{i_3=1}^{3} N_{.,.,i_3}\ln\left(\sum_{i_1,i_2=1}^{3}\gamma_{i_1,i_2,i_3}\right).$$

- No assumption on the stochastic structure.
- Optimality obtained for all function of γ: month to month change, month level for totals, employment rates...

Let

$$\hat{\gamma}_{\mathrm{MLE}} \in \mathrm{argmax}_{\gamma} \left\{ \mathscr{L} \left(\gamma; \left(\mathsf{y}_k|_{A_k} \right)_{k \in \bigcup_{m=1}^{M} S_m} \right) \right\}.$$

Define

$$\hat{\gamma}^{\mathrm{MLE}}|_{m,i} = \sum_{i'_1,\ldots,i'_{m-1},i'_{m+1},i'_M} \gamma_{i'_1,\ldots,i'_{m-1},i,i'_{m+1},i'_M}$$

Define

$$\hat{\mathrm{u}}_m^{\mathrm{MLE}} = \frac{\hat{\gamma}^{\mathrm{MLE}}|_{m,2}}{\hat{\gamma}^{\mathrm{MLE}}|_{m,1} + \hat{\gamma}^{\mathrm{MLE}}|_{m,2}}$$

- Choice of number of lags, for example -12,-3,-2,-1,0,1,2,3,12.
- Define a best compromise if no revision is required.
- Integration into existing adjustment procedure: weight adjustment based on MLE estimates.

`github.com/DanielBonnery/pubBonneryChengLahiri2016`

- Choice of lags -2,-1,0,1,2.
- Generation of 3 synthetic populations.
- CSP design was mimicked.
- MSE was computed with Monte Carlo
- MLE performs (slightly) better than Regression Composite, for level and month to month change.

Why?

- More information is used.
- No need to compromise between level and change.

M	$\dim(\gamma)$
5	243
10	59,049

Proposition:

- For each month, select lags (ex: 12,3,2,1,0,-1,-2,-3,-12)
- Compute MLE.
- Under the no revision constraint, define a criterium for best compromise

Current work.

- Parameter dimension limits the information that can be used.
- How to reduce the model dimenstion when modeling the microdata

Solution

- Use machine learning methods to predict each individual outcume, with a measure of uncertainty.
- Include additional variables.
- Integrate the uncertainty.

Thank you for your attention.