Hierarchical Bayesian Methods for Combining Estimates from Multiple Surveys

Adrijo Chakraborty

NORC at the University of Chicago

January 30, 2015

Joint work with Gauri Sankar Datta and Yang Cheng

Outline

Outline

- Sometimes multiple surveys are conducted to estimate a characteristic.
- Estimates of this characteristic from different surveys could be combined to provide an overall, and hopefully better estimate.
- Estimates obtained from different surveys may not always agree to each other.

In order to estimate the number of occupied housing units (households), many surveys are conducted by the United States Census Bureau. Difference among the survey estimates are noticeable in the following table.

Table: *Estimates of households, obtained in different surveys (numbers in 1000s)*.

Survey	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
CPS/ASEC	111278	112000	113343	114384	116011	116783	117181	117538	119927	121084
HVS	104994	105636	106971	108667	109736	110173	110475	112295	112899	113533
ACS	107367	108420	109902	111091	111617	112378	113101	113616	114567	114992
AHS	.	105842	.	108871	.	110692	.	111806	.	114907

Estimates from CPS/ASEC are consistently high over the years and estimates from Housing Vacancy Survey and American Housing Survey are typically low

Table: *Estimates of households, obtained in different surveys (numbers in 1000s).*

Survey	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
CPS/ASEC	111278	112000	113343	114384	116011	116783	117181	117538	119927	121084
HVS	104994	105636	106971	108667	109736	110173	110475	112295	112899	113533
ACS	107367	108420	109902	111091	111617	112378	113101	113616	114567	114992
AHS	.	105842	.	108871	.	110692	.	111806	.	114907

Table: *Estimates of households (numbers in 1000s)*.

Survey	2007
CPS/ASEC	116783
HVS	110173
ACS	112378
AHS	110692

Table: *Standard errors of the estimates obtained in different surveys (numbers in 1000s).*

Survey	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
CPS/ASEC	260	260	235	234	261	261	261	262	262	262
HVS	185	182	179	204	194	187	181	174	173	171
ACS	.	.	.	144	146	144	147	161	163	180
AHS	.	165	.	218	.	231	.	238	.	396

Outline

Let, h_t= Number of households in year t.

- y_{it} is the estimate of h_t from i^{th} survey, $i = 1, \ldots, 4$.
- s_{it} is the estimated standard error of y_{it}.

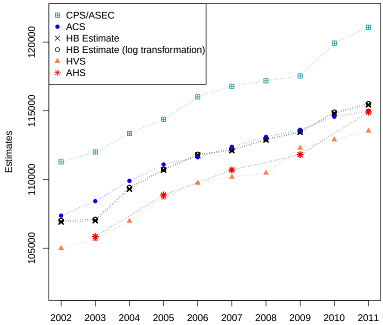
Proposed model M_A:

$$\begin{aligned} y_{it} &= h_t + e_{it} \\ h_t &= h_{t-1} + e_t^*, \end{aligned} \tag{1}$$

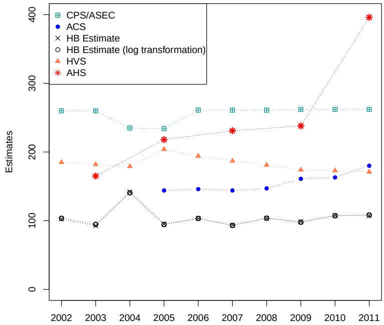
where, e_t^*'s are independently distributed with a truncated normal distribution truncated above 0, with variance $\sigma^2_{e^*}$ and $e_{it} \sim N(0, s_{it}^2)$. We assume uniform priors for the unknown parameters in the model.

Table: *Posterior standard deviations and the standard errors (numbers in 1000s)*.

Year	Proposed method (M_A) Posterior sd	CPS/ASEC s.e	HVS s.e	ACS s.e	AHS s.e
2002	103.48	260	185	.	.
2003	93.73	260	182	.	165
2004	141.15	235	179	.	.
2005	94.65	234	204	144	218
2006	103.65	261	194	146	.
2007	92.93	261	187	144	231
2008	103.76	261	181	147	.
2009	97.42	262	174	161	238
2010	107.55	262	173	163	.
2011	107.08	262	171	180	396



(a)



(b)

Figure: *(a) Point estimates (b) Uncertainty.*

Outline

Proposed model M_B:

$$y_{it} = h_t + \alpha_i + e_{it}$$
$$h_t = h_{t-1} + e_t^*,$$

where, $\sum_{i=1}^{4} \alpha_i = 0$, α_i measures the bias for i^{th} survey. In model M_B, e_t^*'s are independently distributed with a truncated normal distribution truncated above 0, with variance $\sigma^2_{e^*}$ and $e_{it} \sim N(0, s^2_{it})$.

We assume uniform priors for the unknown parameters in the model.

Table: Bayesian inference of the bias contrasts based on M_B.

Parameter	Posterior Mean	Posterior sd	Simulated Quantiles 2.5%	Median	97.5%
$\alpha_1 - \alpha_2$	6388.732	99.34	6190.98	6389.906	6579.72
$\alpha_1 - \alpha_3$	4421.02	105.97	4218.93	4420.899	4630.89
$\alpha_1 - \alpha_4$	6123.44	136.46	5861.28	6122.919	6389.846
$\alpha_2 - \alpha_3$	-1967.72	86.29	-2135.48	-1967.01	-1801.206
$\alpha_2 - \alpha_4$	-265.2917	123.41	-508.40	-264.15	-20.36
$\alpha_3 - \alpha_4$	1702.41	126.35	1459.79	1701.92	1944.453

Proposed model M_C:

$$y_{it} = h_t + \alpha_i + e_{it}$$
$$h_t = \beta_0 + \beta_1 x_t + \eta_t,$$

where $\eta_t \overset{\text{iid}}{\sim} N(0, \sigma^2_\eta)$, x_t is the total population in the United States at year t.

As before, We impose an additive constraint $\sum_{i=1}^{4} \alpha_i = 0$ in the model.

Table: Bayesian inference of the bias contrasts based on M_C.

Parameter	Posterior Mean	Posterior sd	Simulated Quantiles 2.5%	Median	97.5%
$\alpha_1 - \alpha_2$	6390.31	98.44	6200.86	6390.38	6583.22
$\alpha_1 - \alpha_3$	4416.34	102.90	4216.10	4416.27	4618.69
$\alpha_1 - \alpha_4$	6127.05	137.17	5859.32	6127.30	6395.12
$\alpha_2 - \alpha_3$	−1973.98	86.61	−2141.1	−1975.04	−1803.76
$\alpha_2 - \alpha_4$	−263.27	125.39	−505.33	-262.62	−12.56
$\alpha_3 - \alpha_4$	1710.71	128.53	1463.46	1710.32	1962.27

Table: *Bias corrected estimates (rounded) of households from* 2002 – 2011 *for three different surveys based on model M_C (numbers in 1000s)*.

Survey	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
CPS	107044.6	107766.6	109109.6	110150.6	111777.6	112549.6	112947.6	113304.6	115693.6	116850.6
HVS	107150.9	107792.9	109127.9	110823.9	111892.9	112329.9	112631.9	114451.9	115055.9	115689.9
ACS	107549.9	108602.9	110084.9	111273.9	111799.9	112560.9	113283.9	113798.9	114749.9	115174.9
AHS	.	107735.6	.	110764.6	.	112585.6	.	113699.6	.	116800.6

Table: *Bias corrected estimates (numbers in 1000s).*

Survey	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
CPS	107044.6	107766.6	109109.6	110150.6	111777.6	112549.6	112947.6	113304.6	115693.6	116850.6
HVS	107150.9	107792.9	109127.9	110823.9	111892.9	112329.9	112631.9	114451.9	115055.9	115689.9
ACS	107549.9	108602.9	110084.9	111273.9	111799.9	112560.9	113283.9	113798.9	114749.9	115174.9
AHS	.	107735.6	.	110764.6	.	112585.6	.	113699.6	.	116800.6

Table: *Original survey estimates (numbers in 1000s).*

Survey	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
CPS	111278	112000	113343	114384	116011	116783	117181	117538	119927	121084
HVS	104994	105636	106971	108667	109736	110173	110475	112295	112899	113533
ACS	107367	108420	109902	111091	111617	112378	113101	113616	114567	114992
AHS	.	105842	.	108871	.	110692	.	111806	.	114907

Survey	2007
CPS/ASEC	116783
HVS	110173
ACS	112378
AHS	110692

Survey	2007
CPS/ASEC	112549.6
HVS	112329.9
ACS	112560.9
AHS	112585.6

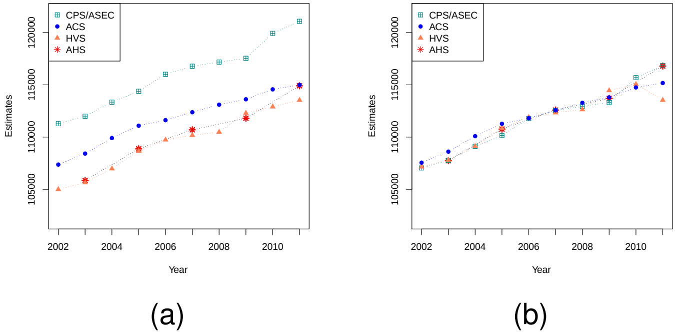


Figure: *(a) Original data (b) After bias correction.*

Table: *HB estimates based on model M_B and M_C (numbers in 1000s)*

Year	Estimate M_B	Estimate M_C	Posterior SD M_B	Posterior SD M_C
2002	107127.72	107136.87	156.28	151.33
2003	107768.22	107786.69	113.87	112.93
2004	109125.61	109129.91	146.15	142.17
2005	110893.35	110869.47	96.35	94.77
2006	111824.62	111796.26	111.76	109.90
2007	112505.36	112495.83	96.71	95.77
2008	113016.36	113024.20	106.61	107.80
2009	113921.37	113934.36	97.85	97.56
2010	115029.86	115032.04	110.01	110.53
2011	115780.69	115788.58	108.23	108.81

Outline

- Number of households estimated by different surveys differ considerably, which may create ambiguity among the researchers and impact decisions of the government organizations.
- We have studied various methods which successfully combine the estimates obtained from different surveys.

- We have achieved considerable gain in precision using our proposed models.
- In future, we would like to develop efficient model selection techniques which select the appropriate model for a given data set. These techniques will be helpful for survey researchers particularly when the scope of external evaluations are restrictive.

Thank you!

- Arthur Cresce Jr, Yang Cheng, and Christopher Grieves (2013). Household Estimates Conundrum: Effort to Develop More Consistent Household Estimates Across Current Survey. *2013 Federal Committee on Statistical Methodology Research Conference.*
- Alan Gelfand and Adrian Smith (1990). Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association,* 85, (410) 398-409.
- Howard R. Hogan (2012). Reducing User Confusion with Joint Data Releases and User Education. *2012 Federal Committee on Statistical Methodology Statistical Policy Seminar.*
- Cheng Y., Chakraborty A., Datta G.S. (2014). Hierarchical Bayesian Methods for Combining Surveys. *2014 Proceedings of the American Statistical Association, Survey Research Methods Section, 4099-4111.*

Collaborators

- Yang Cheng, US Census Bureau.
- Gauri Datta, University of Georgia and US Census Bureau.

Collaborators

- Yang Cheng, US Census Bureau.
- Gauri Datta, University of Georgia and US Census Bureau.