

Demonstrating Annual Dual-System Estimates using Administrative Records and the American Community Survey

Brandon P. Pipher

Disclaimer: Any views expressed in this presentation are those of the authors and not necessarily those of the U.S. Census Bureau

The Census Bureau has reviewed this data product to ensure appropriate access, use, and disclosure avoidance protection of the confidential source data (Project No. 7531903, Sub-Project No. 7532179
Disclosure Review Board (DRB) approval number: DRB-ID: 2026-0204)

The Challenge

The Challenge

Statistical Agencies are under pressure to produce statistics with:

- Utmost Quality
- Greater Frequency
- More Granularity and Greater Detail
- Reduced Costs and Greater Efficiency

Simultaneously, for Surveys:

- Response Rates are Declining
- Fieldwork Costs are Rising

What about Administrative Records (AR) estimates?

- **Pro:** Lower marginal cost vs traditional surveys
- **Pro:** High frequency / near real-time availability
- **Pro:** Reduced respondent burden
- **Pro:** Improved coverage over select populations (vs surveys)
- **Pro:** Census-like level of detail
- **Con:** Systematic coverage biases
- **Con:** Dependent on data quality and record linkage

Hybrid Solution: Combining AR and Surveys

Hybrid Solution – Combining AR and Surveys

Dual-System Estimation (DSE):

- Uses two incomplete lists of a population to better estimate the total population
- Allows estimation of people missed by both lists

Why use AR and Survey?

- Administrative Records provide greater granularity and coverage at a marginal cost
- Survey data improves accuracy of Administrative Record estimates

Implementing the Hybrid Solution: Log-Linear Modeling

Log-Linear Modeling allows creating dual-system estimates that simultaneously addresses:

- Overall unit missingness (people missed by both lists)
- Item missingness (Due to either unit missingness or characteristic non-response)

Log-Linear Models are fit by Expectation-Maximization Algorithm

- Missing At Random (MAR) is assumed

Details can be found in *Van Der Heijden et al 2021*

Applying the Hybrid Solution: Survey + AR

The Data

Dual-System Estimation requires two lists.

For our example, we use the following:

- **List A:** Demographic Frame (aka Administrative Records)
- **List B:** American Community Survey (ACS)

List A: Demographic Frame (DF)

What is it?

- Administrative Record (AR) based
- Derived from administrative records and third-party data
- Comprehensive database of person-level data
- Contains demographic characteristics and addresses associated to individuals
- Can be produced yearly and with respect to varying reference dates

List A: Demographic Frame (DF) – Continued

How is it created?

- Capture Administrative Records within some period near a reference date
 - Examples include IRS, Medicaid, Medicare, HUD, ...
- Apply record linkage to person records and addresses to assign unique keys
- Filter out-of-scope cases like deaths, unborn children, or out-of-country cases
- People are often captured at multiple addresses
 - Models are used to predict a person's most-likely address for a given reference date
 - These models result in a frame with one person at one place

List B: American Community Survey (ACS)

What is it?

- Annual survey conducted by Census Bureau
- Probabilistic sample
- Designed to be representative of the resident U.S. population
- ACS 1-year reported estimates only cover larger areas (> 65k population)
 - Survey responses are collected over the course of a calendar year
 - Estimates are very current
- ACS 5-year estimates are reported for all areas
 - Requires 60 months of data collection
 - Estimates are more like a 'mid-point' estimate, and less current

Data Tables

Contingency Table

Data is sourced using:

- 2023 1-Year ACS Weighted Responses
- 2023 v3 Demographic Frame
- Rounded to nearest Million

Table: Dual-System using Demographic Frame and Non-Interview Adjusted 2023 ACS (by Hispanic Origin)					
		On ACS			All
		On		Off	
		ACS Hispanic Origin		ACS Hispanic Origin	
		Hispanic	Non-Hispanic	Missing	
On DF	DF Hispanic Origin				
On	Hispanic	32,000,000	3,000,000	-	48,000,000
	Non-Hispanic	3,000,000	195,000,000	-	222,000,000
	Missing	11,000,000	32,000,000	-	62,000,000
Off	Missing	2,000,000	7,000,000	-	-
All		48,000,000	237,000,000	-	-

Contingency Table

Data is sourced using:

- 2023 1-Year ACS Weighted Responses
- 2023 v3 Demographic Frame
- Rounded to nearest Million

Table: Dual-System using Demographic Frame and Non-Interview Adjusted 2023 ACS (by Hispanic Origin)					
		On ACS			All
		On		Off	
		ACS Hispanic Origin		ACS Hispanic Origin	
		Hispanic	Non-Hispanic	Missing	
On DF	DF Hispanic Origin				
On	Hispanic	32,000,000	3,000,000	-	48,000,000
	Non-Hispanic	3,000,000	195,000,000	-	222,000,000
	Missing	11,000,000	32,000,000	-	62,000,000
Off	Missing	2,000,000	7,000,000	-	-
All		48,000,000	237,000,000	-	-

Values determined using **Survey Weights** for people found on ACS

Contingency Table

Data is sourced using:

- 2023 1-Year ACS Weighted Responses
- 2023 v3 Demographic Frame
- Rounded to nearest Million

Table: Dual-System using Demographic Frame and Non-Interview Adjusted 2023 ACS (by Hispanic Origin)					
		On ACS			All
		On		Off	
		ACS Hispanic Origin		ACS Hispanic Origin	
		Hispanic	Non-Hispanic	Missing	
On DF	DF Hispanic Origin				
On	Hispanic	32,000,000	3,000,000	-	48,000,000
	Non-Hispanic	3,000,000	195,000,000	-	222,000,000
	Missing	11,000,000	32,000,000	-	62,000,000
Off	Missing	2,000,000	7,000,000	-	-
All		48,000,000	237,000,000	-	-

Values determined by observed **AR Total Counts** from DF

Contingency Table

Data is sourced using:

- 2023 1-Year ACS Weighted Responses
- 2023 v3 Demographic Frame
- Rounded to nearest Million

Table: Dual-System using Demographic Frame and Non-Interview Adjusted 2023 ACS (by Hispanic Origin)					
		On ACS			All
		On		Off	
		ACS Hispanic Origin		ACS Hispanic Origin	
		Hispanic	Non-Hispanic	Missing	
On DF	DF Hispanic Origin				
On	Hispanic	32,000,000	3,000,000	-	48,000,000
	Non-Hispanic	3,000,000	195,000,000	-	222,000,000
	Missing	11,000,000	32,000,000	-	62,000,000
Off	Missing	2,000,000	7,000,000	-	-
All		48,000,000	237,000,000	-	-

Problem:

These values that are unknown, but necessary to produce a DSE
 These represent AR-only people (and unlinked ACS cases)

Contingency Table

Data is sourced using:

- 2023 1-Year ACS Weighted Responses
- 2023 v3 Demographic Frame
- Rounded to nearest Million

Table: Dual-System using Demographic Frame and Non-Interview Adjusted 2023 ACS (by Hispanic Origin)					
		On ACS			All
		On		Off	
		ACS Hispanic Origin		ACS Hispanic Origin	
		Hispanic	Non-Hispanic	Missing	
On DF	DF Hispanic Origin				
On	Hispanic	32,000,000	3,000,000	13,000,000	48,000,000
	Non-Hispanic	3,000,000	195,000,000	24,000,000	222,000,000
	Missing	11,000,000	32,000,000	19,000,000	62,000,000
Off	Missing	2,000,000	7,000,000	-	-
All		48,000,000	237,000,000	-	-

One Potential Solution:

Subtract the difference between **Observed Counts** and **Survey Weights**
(Luckily, there are no negatives!)

Contingency Table

Data is sourced using:

- 2023 1-Year ACS Weighted Responses
- 2023 v3 Demographic Frame
- Rounded to nearest Million

Table: Dual-System using Demographic Frame and Non-Interview Adjusted 2023 ACS (by Hispanic Origin)					
		On ACS			All
		On		Off	
		ACS Hispanic Origin		ACS Hispanic Origin	
		Hispanic	Non-Hispanic	Missing	
On DF	DF Hispanic Origin				
On	Hispanic	32,000,000	3,000,000	13,000,000	48,000,000
	Non-Hispanic	3,000,000	195,000,000	24,000,000	222,000,000
	Missing	11,000,000	32,000,000	19,000,000	62,000,000
Off	Missing	2,000,000	7,000,000	-	-
All		48,000,000	237,000,000	-	-

Notice:

The unobserved cell (and associated marginal totals) are all unobserved and will be estimated by Dual-System Estimation

Contingency Table

Data is sourced using:

- 2023 1-Year ACS Weighted Responses
- 2023 v3 Demographic Frame
- Rounded to nearest Million

Table: Dual-System using Demographic Frame and Non-Interview Adjusted 2023 ACS (by Hispanic Origin)					
		On ACS			All
		On		Off	
		ACS Hispanic Origin		ACS Hispanic Origin	
		Hispanic	Non-Hispanic	Missing	
On DF	DF Hispanic Origin				
On	Hispanic	32,000,000	3,000,000	13,000,000	48,000,000
	Non-Hispanic	3,000,000	195,000,000	24,000,000	222,000,000
	Missing	11,000,000	32,000,000	19,000,000	62,000,000
Off	Missing	2,000,000	7,000,000	-	-
All		48,000,000	237,000,000	-	-

Cells marked in **yellow** represent Measurement Error:

- *Why is this problematic?* Survey and AR differ in their Response
- *One Potential Solution:* Log-Linear Modeling

Contingency Table

Data is sourced using:

- 2023 1-Year ACS Weighted Responses
- 2023 v3 Demographic Frame
- Rounded to nearest Million

Table: Dual-System using Demographic Frame and Non-Interview Adjusted 2023 ACS (by Hispanic Origin)					
		On ACS			All
		On		Off	
		ACS Hispanic Origin		ACS Hispanic Origin	
		Hispanic	Non-Hispanic	Missing	
On DF	DF Hispanic Origin				
On	Hispanic	32,000,000	3,000,000	13,000,000	48,000,000
	Non-Hispanic	3,000,000	195,000,000	24,000,000	222,000,000
	Missing	11,000,000	32,000,000	19,000,000	62,000,000
Off	Missing	2,000,000	7,000,000	-	-
All		48,000,000	237,000,000	-	-

Cells marked in **yellow** represent Item Missingness:

- *Why is this problematic?* Captured by list, but lack a reported characteristic
- *One Potential Solution:* Log-Linear Modeling

Contingency Table

Data is sourced using:

- 2023 1-Year ACS Weighted Responses
- 2023 v3 Demographic Frame
- Rounded to nearest Million

Table: Dual-System using Demographic Frame and Non-Interview Adjusted 2023 ACS (by Hispanic Origin)					
		On ACS			All
		On		Off	
		ACS Hispanic Origin		ACS Hispanic Origin	
		Hispanic	Non-Hispanic	Missing	
On DF	DF Hispanic Origin				
On	Hispanic	32,000,000	3,000,000	13,000,000	48,000,000
	Non-Hispanic	3,000,000	195,000,000	24,000,000	222,000,000
	Missing	11,000,000	32,000,000	19,000,000	62,000,000
Off	Missing	2,000,000	7,000,000	-	-
All		48,000,000	237,000,000	-	-

Cells marked in **yellow** also represent Item Missingness, but a different kind

- This is due to Unit Missingness (Non-Response for a given list)
- Also resolved by Log-Linear modeling

Fitted Estimates

Fitted Estimates from Log-Linear Modeling

Table: Dual-System using Demographic Frame and Non-Interview Adjusted 2023 ACS (by Hispanic Origin)						
		On ACS				All
		On		Off		
		ACS Hispanic Origin		ACS Hispanic Origin		
		Hispanic	Non-Hispanic	Hispanic	Non-Hispanic	
On DF	DF Hispanic Origin					
On	Hispanic	42,160,000	3,288,000	18,890,000	1,473,000	65,810,000
	Non-Hispanic	4,020,000	227,000,000	609,900	34,440,000	266,100,000
Off	Hispanic	1,712,000	97,870	766,900	43,840	2,621,000
	Non-Hispanic	163,200	6,756,000	24,770	1,025,000	7,969,000
All		48,060,000	237,100,000	20,290,000	36,980,000	342,500,000

Fitted Estimates from Log-Linear Modeling

Table: Dual-System using Demographic Frame and Non-Interview Adjusted 2023 ACS (by Hispanic Origin)						
		On ACS				All
		On		Off		
		ACS Hispanic Origin		ACS Hispanic Origin		
		Hispanic	Non-Hispanic	Hispanic	Non-Hispanic	
On DF	DF Hispanic Origin					
On	Hispanic	42,160,000	3,288,000	18,890,000	1,473,000	65,810,000
	Non-Hispanic	4,020,000	227,000,000	609,900	34,440,000	266,100,000
Off	Hispanic	1,712,000	97,870	766,900	43,840	2,621,000
	Non-Hispanic	163,200	6,756,000	24,770	1,025,000	7,969,000
All		48,060,000	237,100,000	20,290,000	36,980,000	342,500,000

Notice that all cells are filled, and no characteristics are missing!

Fitted Estimates from Log-Linear Modeling

Table: Dual-System using Demographic Frame and Non-Interview Adjusted 2023 ACS (by Hispanic Origin)						
		On ACS				All
		On		Off		
		ACS Hispanic Origin		ACS Hispanic Origin		
		Hispanic	Non-Hispanic	Hispanic	Non-Hispanic	
On DF	DF Hispanic Origin					
On	Hispanic	42,160,000	3,288,000	18,890,000	1,473,000	65,810,000
	Non-Hispanic	4,020,000	227,000,000	609,900	34,440,000	266,100,000
Off	Hispanic	1,712,000	97,870	766,900	43,840	2,621,000
	Non-Hispanic	163,200	6,756,000	24,770	1,025,000	7,969,000
All		48,060,000	237,100,000	20,290,000	36,980,000	342,500,000

This is our unobserved cells estimated using Dual-System Estimation

Fitted Estimates from Log-Linear Modeling

Table: Dual-System using Demographic Frame and Non-Interview Adjusted 2023 ACS (by Hispanic Origin)

		On ACS				All
		On		Off		
		ACS Hispanic Origin		ACS Hispanic Origin		
		Hispanic	Non-Hispanic	Hispanic	Non-Hispanic	
On DF	DF Hispanic Origin					
On	Hispanic	42,160,000	3,288,000	18,890,000	1,473,000	65,810,000
	Non-Hispanic	4,020,000	227,000,000	609,900	34,440,000	266,100,000
Off	Hispanic	1,712,000	97,870	766,900	43,840	2,621,000
	Non-Hispanic	163,200	6,756,000	24,770	1,025,000	7,969,000
All		48,060,000	237,100,000	20,290,000	36,980,000	342,500,000

Aggregate by ACS-definition of Hispanic Origin

- Lower Measurement Error

Comparing Estimates

Estimates

	Total Resident Population (% Difference from Baseline)	Non-Hispanic	Hispanic (% of Total)
Dual-System Estimate (ACS with AR)	342,500,000 (+1.7%)	274,080,000	68,350,000 (20.0%)
Population Estimates Program (Vintage 2024)	336,800,000 (baseline)	270,700,000	66,140,000 (19.6%)
American Community Survey (2023 1-Year ACS Estimates)	334,900,000 (-0.6%)	269,800,000	65,140,000 (19.5%)

Note that:

- ACS 1-year underestimates the Population Estimates Program (PEP)
- Hispanic population is historically undercounted (net undercount of 5% in 2020 Census)

This experimental estimate in comparison produces an over-estimate, potentially closer to truth

- Hispanic DSE is $\approx$ 5% greater than the 1-Year ACS Hispanic Population Estimate

Conclusion

Conclusion

- Potential advantages of DSE Methodology combining AR and Surveys:
 - Uses multiple partial lists to better estimate the full population size
 - Can potentially produce lower-level and more detailed estimates
 - Estimates can potentially be produced in a timelier manner than otherwise available
 - Administrative Records can incorporate characteristics not measured by surveys
 - Administrative Record estimates are improved by the accuracy of survey data
- Note: These are experimental estimates still in early research



THANK YOU!



BRANDON.PIPHER@CENSUS.GOV

Citations

- Heijden, P. G. M. van der, Smith, P. A., Cruyff, M., & Bakker, B. (2018). An Overview of Population Size Estimation where Linking Registers Results in Incomplete Covariates, with an Application to Mode of Transport of Serious Road Casualties. *Journal of Official Statistics*, 34(1), 239-263.
 - <https://doi.org/10.1515/jos-2018-0011>
- Peter G. M. van der Heijden et al, Multiple System Estimation using Covariates having Missing Values and Measurement Error: Estimating the Size of the Māori Population in New Zealand, *Journal of the Royal Statistical Society Series A: Statistics in Society*, Volume 185, Issue 1, January 2022, Pages 156–177
 - <https://doi.org/10.1111/rssa.12731>