Improved Longitudinal Imputation Method in Survey of Doctorate Recipients

Minsun Riddles[1] Jean Opsomer[1] Wan-Ying Chang[2]
Laura Alvarez-Rojas[1] Shelley Brock Roth[1] Medha Uppala[1]

[1]Westat [2]National Center for Science and Engineering Statistics, U.S. National Science Foundation

FCSM 2023 — Building Toward the Future:
Strengthening and Expanding the Capacity of the Federal Statistical Ecosystem.

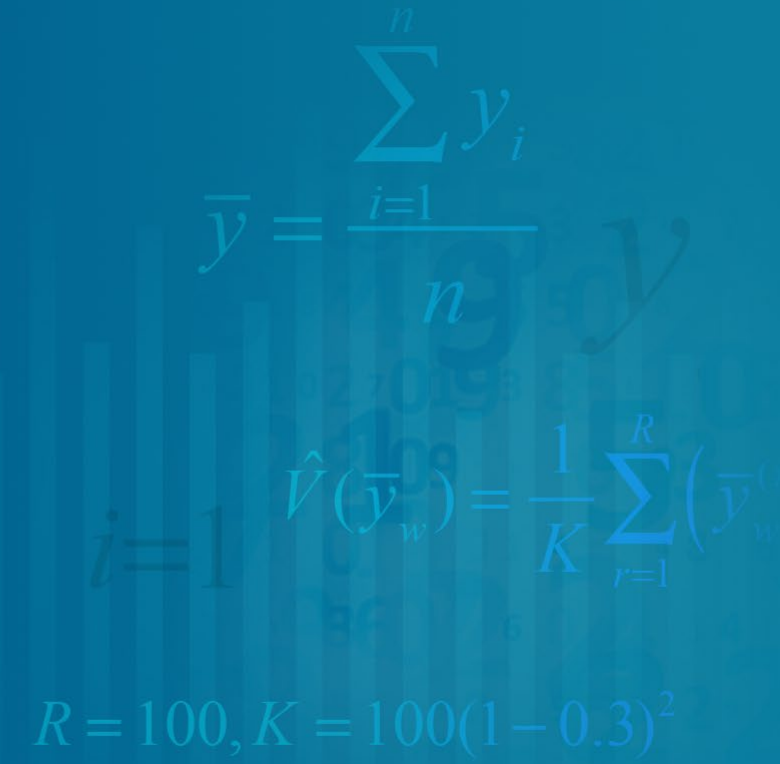
Disclaimer

- This presentation provides preliminary results of exploratory research sponsored in part by the National Center for Science and Engineering Statistics (NCSES) within the National Science Foundation (NSF). This information is being shared to inform interested parties of ongoing activities and to encourage further discussion. Any views expressed are those of the authors and not necessarily those of NCSES or NSF.

Outline

- Background
- Imputation Process
- Evaluation
- Summary

Background


$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$
$\hat{V}(\bar{y}_w) = \frac{1}{K}\sum_{r=1}^{R}$
$R = 100, K = 100(1-0.3)^2$

Survey of Doctorate Recipients

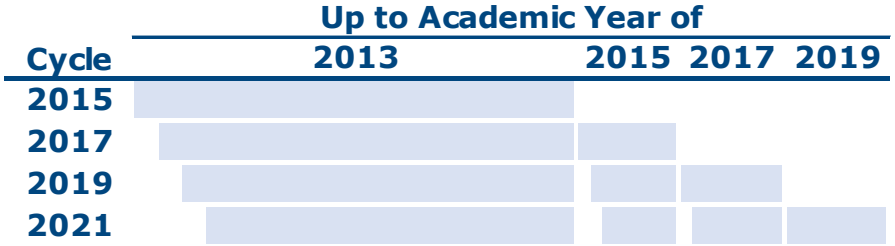
- Biennial survey of U.S. research doctorate holders in science, engineering, and health fields
  - Launched in 1973
  - Conducted by the National Center for Science and Engineering Statistics within National Science Foundation
  - With sponsorship from National Institutes of Health
- Provides demographic, education, and career history information
- Target population contains U.S. research doctorate holders residing in U.S. and out of U.S.

Sample Design

- Refreshed sample for 2015 cycle, extending its coverage and increasing the sample size
- Stratified by field of study, gender, and underrepresented minority status
- Maintains a continuing cohort until aged out
- Replenishes sample with new PhDs in every cycle

→Generating cross-sectional data

and longitudinal data



SDR 2015/2017/2019 Cross-sectional Data

- Sample of 120,000+ doctorate holders
- 80,000+ respondents who responded to critical items
- Weighting compensates for doctorate holders who did not respond
- Imputation addresses item nonresponse

Longitudinal SDR (L-SDR) 2015-2025 Sample

- Sample of ~40,000 doctorate holders
- Drawn from SDR 2015 respondents who were 65 years old or younger as of 2015
- Stratified by
  - Reported employment sector in 2015
  - Age
  - Underrepresented minority status
  - Gender

L-SDR 2015-2019 Data

- ~36,500 who responded in both or either of the 2017 and 2019 cycles
- Weighting compensates for individuals who did not respond in both cycles
- Imputation addresses unit nonresponse in one of 2017 and 2019 cycles
- Limited findings are publicly available[1]

	2015	2017	2019
L-SDR 2015-2019 Sample	Y_{15}	Y_{17}	Y_{19}
	Y_{15}		Y_{19}
	Y_{15}	Y_{17}	
	Y_{15}		

Weighting →

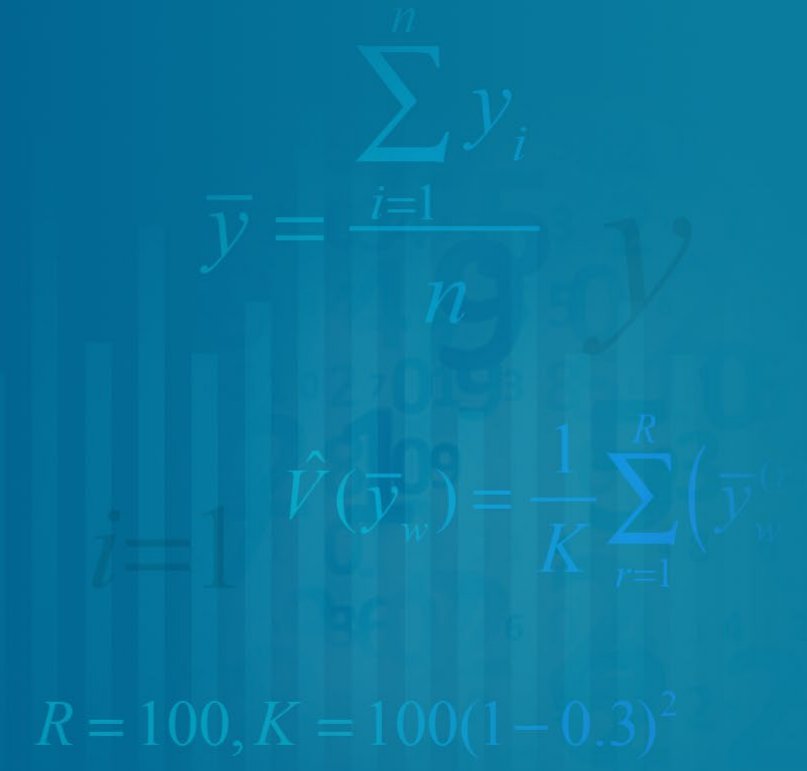
2015	2017	2019
Y_{15}	Y_{17}	Y_{19}
Y_{15}		Y_{19}
Y_{15}	Y_{17}	

Imputation →

Y_{15}	Y_{17}	Y_{19}	L-SDR 2015-2019 Data
Y_{15}	$Y_{17,imp}$	Y_{19}	
Y_{15}	Y_{17}	$Y_{19,imp}$	

[1]https://ncses.nsf.gov/pubs/nsf22326#section11241

Imputation Process



Imputation Implemented | L-SDR 2015-2019 Data

	Phase 1 Imputation	Phase 2 Imputation
Nonresponse (NR) type	Item NR	Unit NR of 1 cycle
Respondents	Cross-sectional	Longitudinal
Imputation Method	Hot-deck	Hot-deck

Information usage	Phase 1 Imputation	Phase 2 Imputation
Demographic	√	√
Cross-sectional	√	--
Longitudinal	√ limited	√ **systematically**

Phase 1 | Imputation for Cross-sectional Data

- Hot-deck imputation replaces each missing value with a randomly selected observed response in its imputation cell containing "similar" individuals
- Imputation cells defined by crossing cross-sectional, demographic, and longitudinal variables closely related to the variable to be imputed

ID	Gender	Age	Employed in 2019			
			Raw	Donor ID	Donor value	Imputed
1	Male	30-45	Yes			Yes
2	Male	30-45		1	Yes	Yes
3	Male	46-60	Yes			Yes
4	Male	46-60	No			No
5	Male	46-60	No			No
6	Male	46-60		5	No	No
7	Female	30-45	Yes			Yes
8	Female	30-45	No			No
9	Female	30-45		7	Yes	Yes
10	Female	46-60	No			No

Phase 2 | Imputation for Longitudinal Data

- Hot-deck imputation with imputation cells defined by longitudinal variables first and selected demographic variables
- Treat imputed data for other cycles as reported

ID	Employed in 2015	Employed in 2017	Gender	Age	Employed in 2019			
					Raw	Donor ID	Donor value	Imputed
1	Yes	Yes	Male	30-45	Yes			Yes
2	Yes	Yes	Male	30-45		1	Yes	Yes
3	Yes	Yes	Male	46-60	Yes			Yes
4	Yes	No	Male	46-60	No			No
5	Yes	No	Male	46-60	No			No
6	No	No	Male	46-60		5	No	No
7	Yes	Yes	Female	30-45	Yes			Yes
8	Yes	No	Female	30-45	No			No
9	No	No	Female	30-45		7	Yes	Yes
10	No	No	Female	46-60	No			No

Motivation

- Phase 1 imputation sometimes created longitudinal transitional patterns that were not reported
  - Employment sectors

2015	2017	2019
4 year university	→ 2 year college	→ self-employed
Non-US government	→ 2 year college	→ state/local government

  - Influential cases with large changes in salary between cycles

→ Improve donor matching in cross-sectional imputation to align two imputation phases through **systematic usage of historically reported data** in Phase 1 imputation for cross-sectional data while keeping Phase 2 imputation the same

Proposed Phase 1 Imputation Approach

- For Phase 1 imputation, form three imputation groups

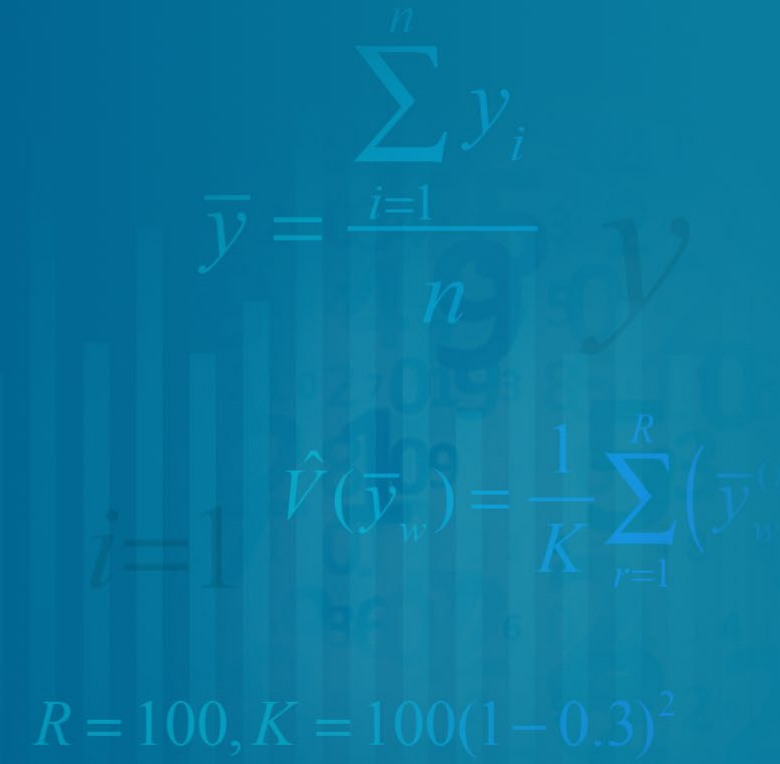
Item reported?			Imputation group 1		Imputation group 2		Imputation group 3	
2015	2017	2019	Recipients	Donors	Recipients	Donors	Recipients	Donors
Yes	Yes	Yes		√		√		√
No	Yes	Yes				√		√
Yes	No	Yes						√
No	No	Yes						√
Yes	Yes	No	√					
No	Yes	No			√			
Yes	No	No					√	
No	No	No					√	

No= Did not respond or was not sampled in the cycle

Proposed Phase 1 Imputation Approach

- Update imputation cell formation by systematically using longitudinal variables in imputation groups 1 & 2
  - **Time of doctorate award**
  - **Reported values in previous cycles (up to 2)**
  - Cross-sectional or demographic variables currently used

Evaluation


$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$
$\hat{V}(\bar{y}_w) = \frac{1}{K}\sum_{r=1}^{R}$
$R = 100, K = 100(1-0.3)^2$

Evaluation | Cross-sectional Results

- Applied to proposed Phase 1 imputation approach to employment section in SDR 2019 data while keeping Phase 2 imputation approach the same
- Compared marginal distributions of variables subject to imputation using goodness-of-fit chi-square tests
  - No significant differences out of ~100 variables
- Compared estimates in cells with 100+ respondents of 20 detailed statistical tables
  - <1% estimates with significant differences

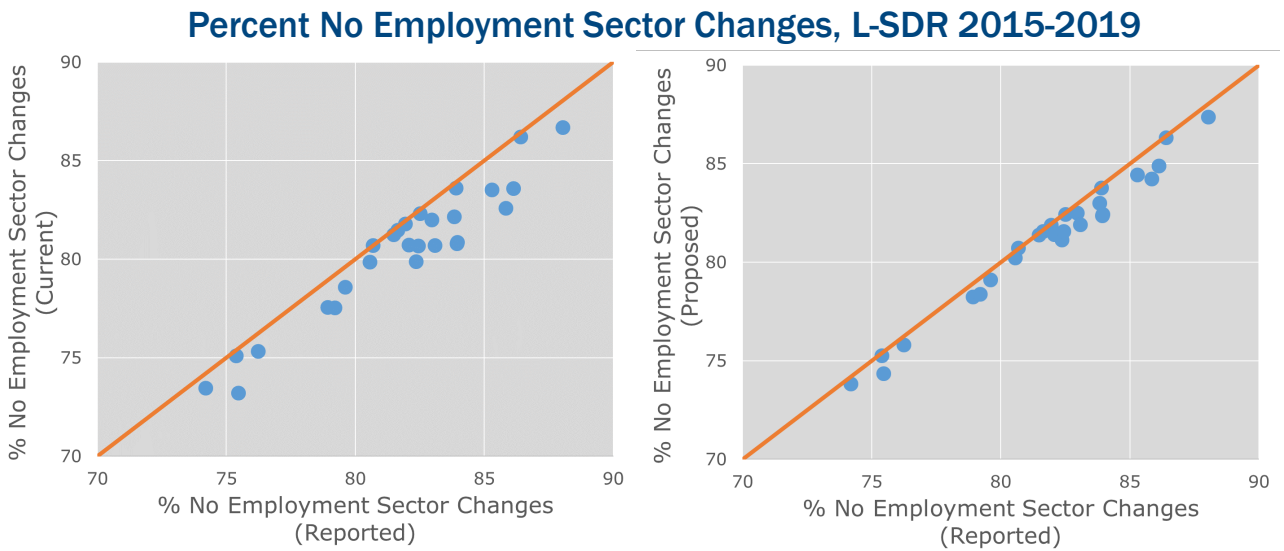
Evaluation | Longitudinal Pattern

- Examined longitudinal patterns
  - Closer to those among respondents who responded in 2015, 2017, 2019
  - New longitudinal patterns rarely introduced by imputation (5% → <1%)
  - Changes in salary more aligned with reported data

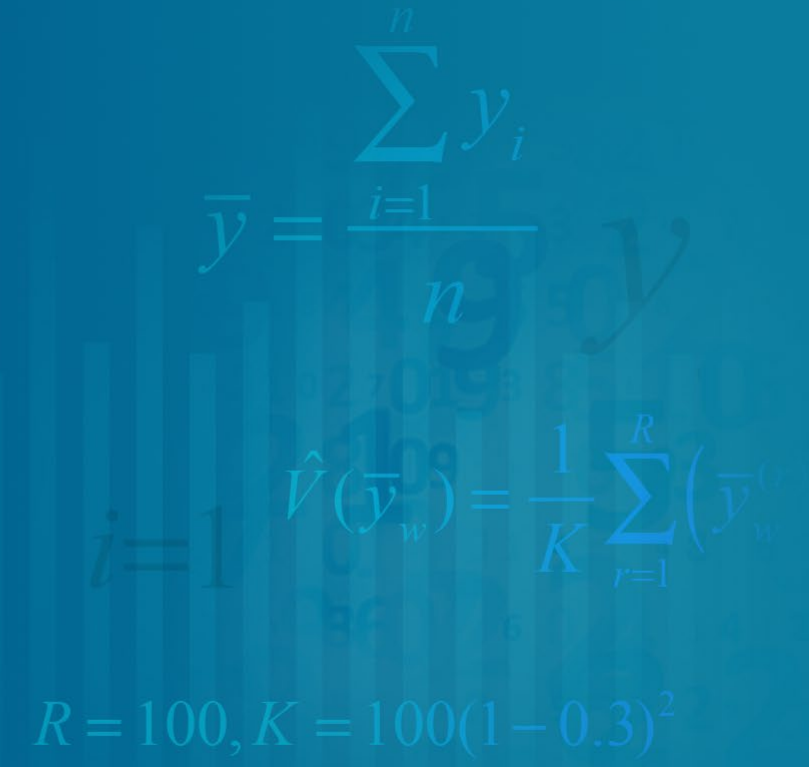
Average ratio of Salary in 2019 to	Reported	Reported + Imputed	
		Current	Proposed
2015	1.12	1.15	1.13
2017	1.07	1.08	1.07

Evaluation | Longitudinal Estimates - Preliminary

- Examined longitudinal estimates by various domain
  - More aligned with estimates based on reported cases



Summary


$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$
$\hat{V}(\bar{y}_w) = \frac{1}{K}\sum_{r=1}^{R}$
$R = 100, K = 100(1-0.3)^2$

Summary

- New imputation approach to improve donor matching process
  - Aligned data matching strategies of both imputation phases
  - Reduced chances of creating erroneous longitudinal patterns
  - Longitudinal estimates closer to estimates among reported cases
  - Minimal impact on cross-sectional estimates

- More to come!

Thank You

For more information, please contact:

MinsunRiddles@westat.com