Estimating Inequality with Missing Incomes

P. Brunori[a,b], P. Salas-Rojo[a] and P. Verme[c]

III - London School of Economics[a]; University of Florence[b]; World Bank[c]

July, 2023

Warm up and stretching

Official income inequality estimates usually rely on survey data.

- Surveys might be affected by missing observations, often concentrated around the tails of distributions (Hlasny et al., 2021).
- Missings may bias inequality measures.

The literature has proposed many methods to deal with the missing data problem.

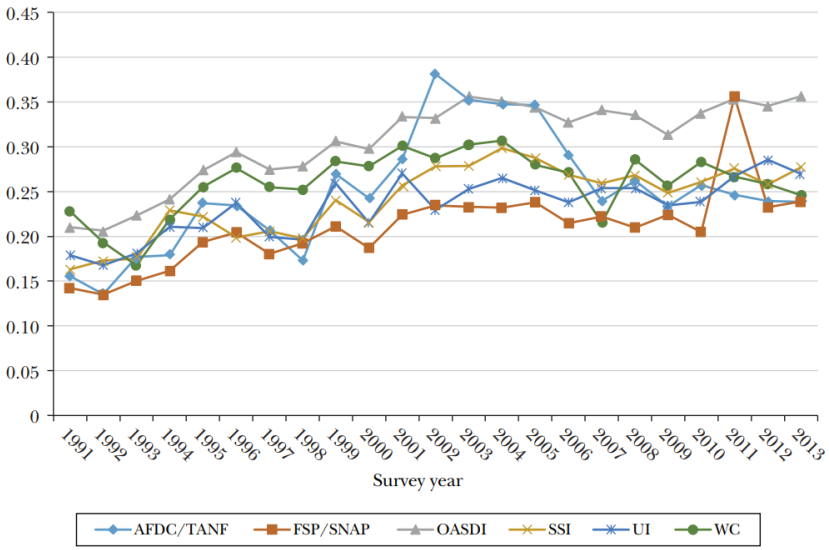
Layout of the Presentation

1. Present the problem of missing incomes.
2. Explain popular solutions.
3. Explain experiment.
4. Main results.

Causes of missing incomes

- Unit non-responses (Not explored here).
- Item non-responses.
- Misreportings.

Up to 65% of datasets in LIS have missing income values (Hlasny, 2020).
Figure from (Meyer et al., 2015)



Missing incomes in developing countries

Issue is salient for top incomes due to lack of alternative sources and weaker survey infrastructures (Alvaredo and Gasparini (2015); Ravallion (2022)).

Similar for bottom and middle incomes, due to seasonality and within-year variability. They can accentuate misreporting.

Consumption and/or expenditure are often used as a proxy. But this might underestimate inequality!

Types of missing (Rubin, 1976)

- Missing Completely At Random (MCAR). Missing probability is homogeneous.
- Missing At Random (MAR). Missing probability is driven by covariates.
- Missing Not At Random (MNAR). Missing probability is correlated with the affected variable.

Usually hard to distinguish between these patterns in the data.

Popular solutions: Deletion and mean

a) Deletion:

Shortcut, erase incomplete observations.

b) Imputation of the mean:

$$\forall i \in M : y_i = \frac{\sum_{j \notin M} y_j}{N_j} \tag{1}$$

Popular solutions: Imputations

c) Single and Multiple imputations (Rubin, 1978):

$$\forall j \notin M : log(y_j) = \beta \mathbf{x}_j + \epsilon_j \tag{2}$$

$$\epsilon_j \sim N(0, \sigma^2) \tag{3}$$

$$\forall i \in M : \hat{y}_i = exp(\hat{\beta} \mathbf{x}_i + \sigma^2/2) \tag{4}$$

d) Matching (PMM):

Similar to parametric imputations but substituting missing incomes with a random y_j drawn from observations with similar $\hat{y}_i$ (Schenker and Taylor, 1996).

Popular solutions: Parametric distributions

e) As in (Jenkins, 2017), where $y_0 > 0$ is a scale parameter defining the threshold over which the Pareto tail is adjusted, and $\theta > 0$ is a data-specific shape parameter.

$$F_\theta(y) = 1 - \left(\frac{y_j}{y_0}\right)^\theta, y_j \geq y_0 \tag{5}$$

Popular solutions: Reweighting

f) As proposed in (Korinek et al., 2006):

$$P_i(x_i, \theta) = \frac{e^{g(x_i,\theta)}}{1 + e^{g(x_i,\theta)}} \tag{6}$$

where $g(x_i, \theta)$ is a stable function of observable characteristics and θ is a vector of parameters estimated with GMM:

$$\hat{\theta} = argmin \sum_{j=1,\ldots,J} \left[(\hat{m}_j - m_j) w_j^{-1} (\hat{m}_j - m_j) \right] \tag{7}$$

Popular solutions: Machine Learning prediction methods

g) LASSO (Tibshirani, 1996)

$$\forall j \notin M : \sum_{j=1}^{n} (log(y_j) - \beta \mathbf{x}_j)^2 - \lambda \sum_{c=1}^{C} |\beta_c| \tag{8}$$

h) Trees and Random Forests

Partition the sample into non-overlapping subgroups based on the regressors' space. Prediction average the realization of y by subgroups (Hothorn et al., 2006).

Our experiment

1. Take a complete dataset.
2. Simulate different missing patterns.
3. Apply correction methods.
4. Evaluate differences in Gini and (out of sample, OOS) Root Mean Squared Error (RMSE).

Warning: trade-off between inequality and RMSE

Imagine we impute using a single regression:

$$\forall j \notin M : log(y_j) = \beta \mathbf{x}_j + \epsilon_j \tag{9}$$

Then regularize with LASSO:

$$\forall j \notin M : \sum_{j=1}^{n} (log(y_j) - \beta \mathbf{x}_j)^2 - \lambda \sum_{c=1}^{C} |\beta_c| \tag{10}$$

Some β parameters would disappear. We would improve OOS RMSE, but necessarily imputation will have less inequality.

Baseline data

5^{th} wave of National Income Dynamics Study (South Africa 2017).

We focus on labor income.

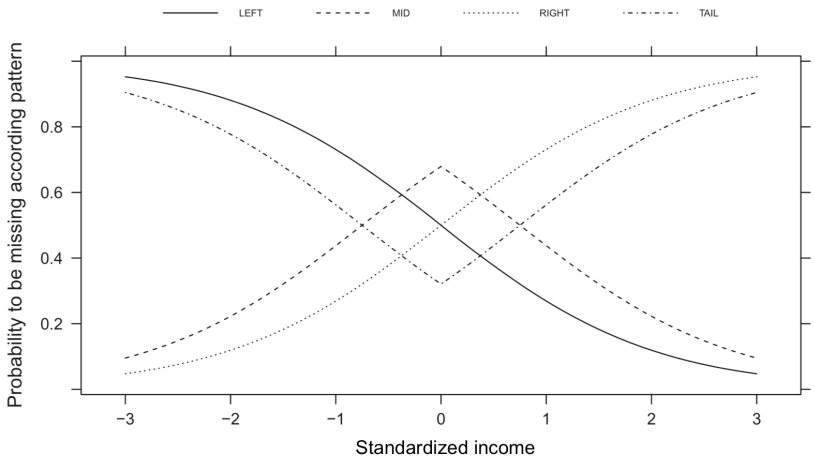
N	Mean	Gini
7199	8057.23	53.46

All results were obtained after 100 bootstrapped repetitions and significance estimated with a difference-in-means test.

Other patterns (MCAR and MNAR)

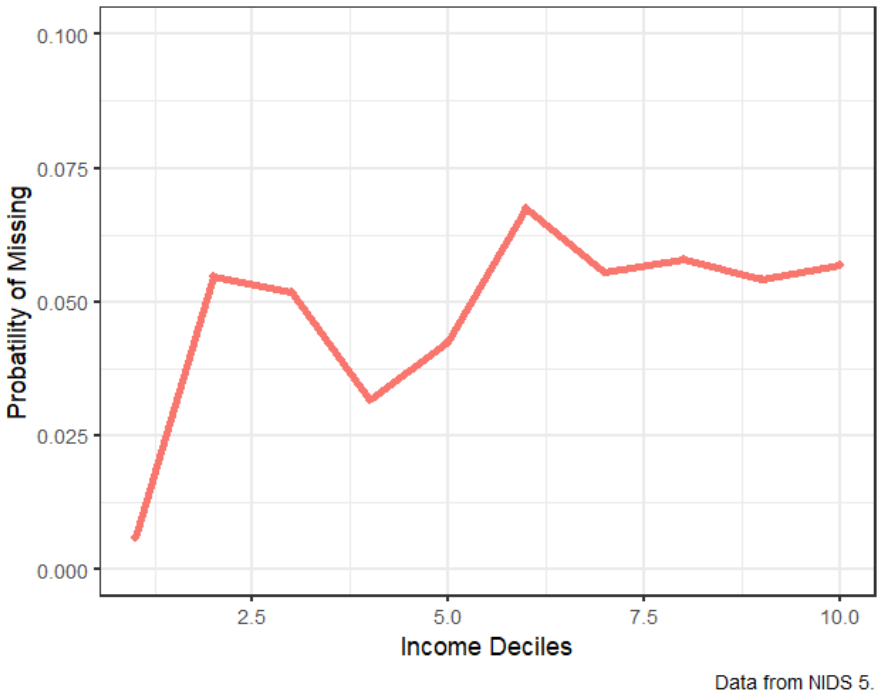
For MCAR, all observations have the same probability.

For MNAR, we use the non-linear functions illustrated in (Schouten et al., 2018)



Our preferred data corruption: MAR

We estimate missing probabilities based on covariates.



Results MAR (Gini I)

N	Share (%)	Deletion	Mean	SP	MP	PMM
6983	3	-0.72	-1.68	-0.49	-0.86	-0.41
6839	5	-1.00	-2.57	-1.47	-1.26	-0.50
6695	7	-1.33	-3.50	-1.58	-1.68	-0.74
6479	10	-1.60	-4.66	-2.04	-2.13	-0.84
6119	15	-2.07	-6.66	-1.86	-2.81	-1.03
5759	20	-2.45	-8.64	-2.44	-3.45	-1.18
5039	30	-3.32	-13.05	-2.75	-4.61	-1.67
4319	40	-3.86	-17.58	-5.04	-5.46	-1.94
3600	50	-4.44	-22.65	-6.71	-6.26	-2.16

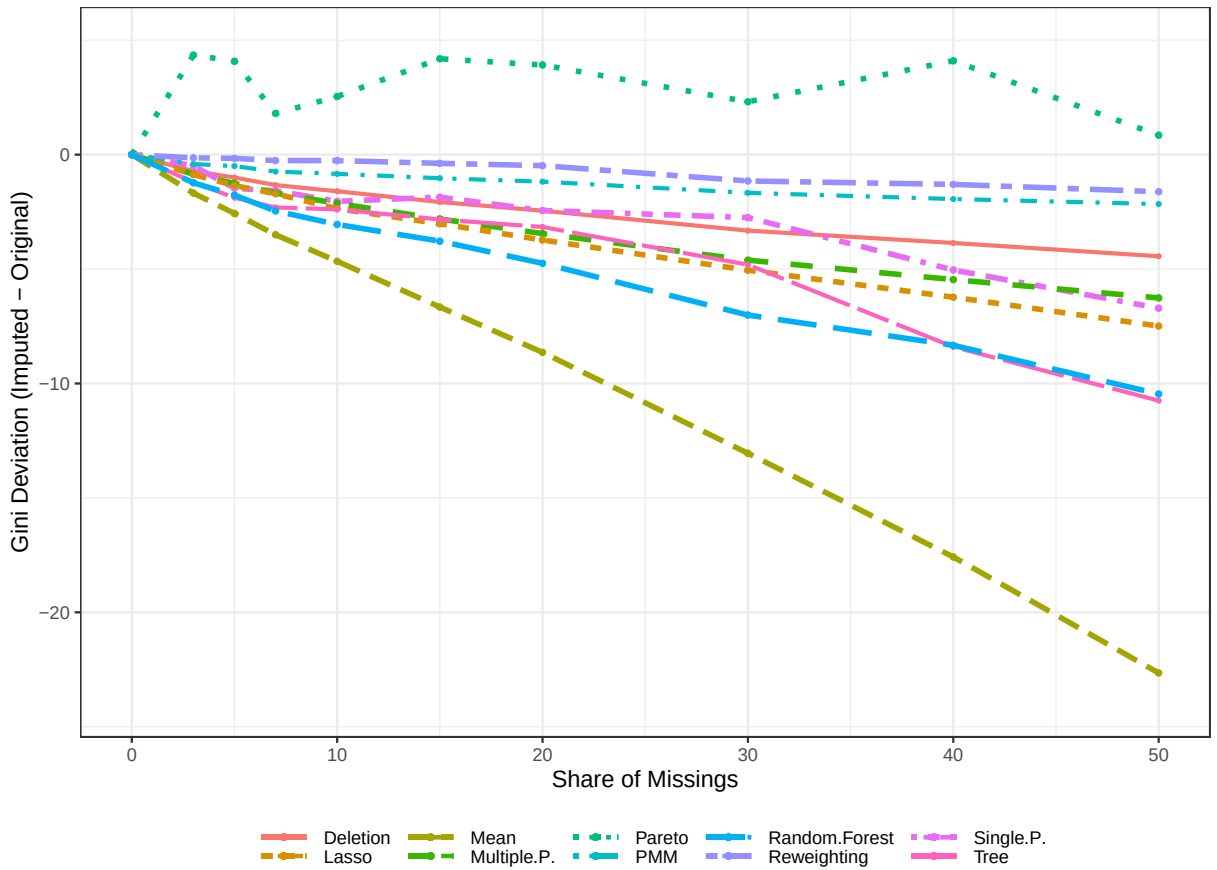
(No *, all different from zero).

Results MAR (Gini II)

N	Share (%)	Pareto	Rwght	LASSO	Tree	RF
6983	3	4.35	-0.14	-0.87	-1.21	-1.23
6839	5	4.08	-0.16	-1.35	-1.88	-1.77
6695	7	1.80	-0.26	-1.71	-2.31	-2.46
6479	10	2.54	-0.26	-2.36	-2.39	-3.05
6119	15	4.19	-0.38	-3.02	-2.83	-3.78
5759	20	3.92	-0.48	-3.73	-3.16	-4.75
5039	30	2.31	-1.15	-5.05	-4.81	-7.01
4319	40	4.10	-1.30	-6.23	-8.40	-8.33
3600	50	0.85	-1.62	-7.49	-10.75	-10.46

(No *, all different from zero).

Gini Summary



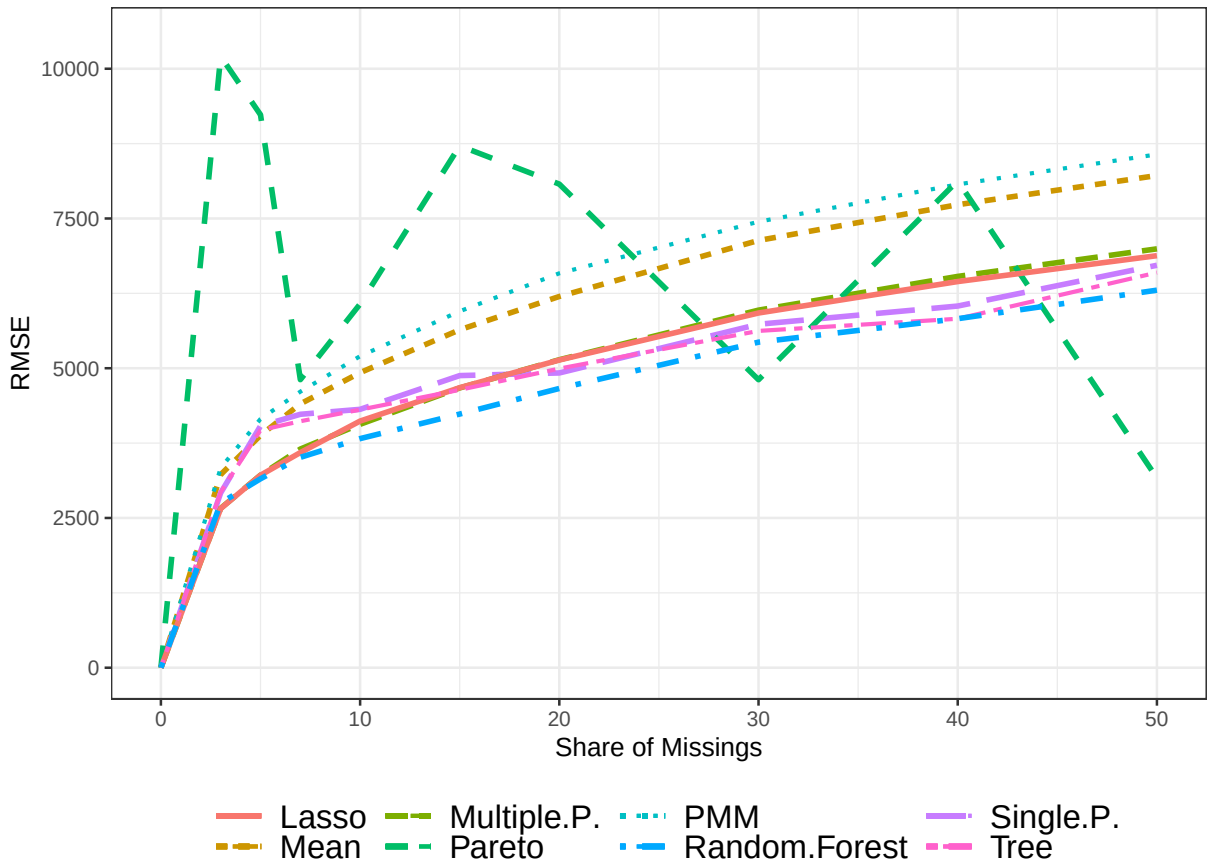
Results MAR (RMSE I)

N	Share (%)	Mean	SP	MP	PMM
6983	3	3224	2918	2661	3334
6839	5	3879	4036	3215	4153
6695	7	4405	4231	3651	4609
6479	10	4926	4314	4064	5201
6119	15	5642	4877	4671	5951
5759	20	6198	4921	5142	6585
5039	30	7133	5734	5970	7446
4319	40	7730	6038	6533	8069
3600	50	8217	6719	6993	8570

Results MAR (RMSE II)

N	Share (%)	Pareto	LASSO	Tree	RF
6983	3	10207	2657	2933	2750
6839	5	9232	3214	3966	3154
6695	7	4809	3592	4115	3512
6479	10	6073	4118	4308	3825
6119	15	8713	4675	4639	4235
5759	20	8074	5137	4994	4662
5039	30	4808	5922	5623	5435
4319	40	8129	6447	5826	5827
3600	50	3166	6879	6597	6303

RMSE


RMSE
10000
7500
5000
2500
0
0
10
20
30
40
50
Share of Missings
Lasso
Mean
Multiple.P.
Pareto
PMM
Random.Forest
Single.P.
Tree

Other results

When we impose MCAR, deletion and reweighting deliver accurate inequality estimates. See ▸ Appendix.

When missings are MNAR/MID: The best methods are the Multiple Parametric and the LASSO (mean bias around 0). See ▸ Appendix.

Other results

When missings are MNAR/LEFT: The best methods are the Deletion and Reweighting (mean bias around 0). See ▸ Appendix.

When missings are MNAR/RIGHT or MNAR/TAIL: The best methods are the Deletion, Pareto Models and Reweighting. Still, the inequality biases persist. See ▸ Right Appendix or ▸ Tail Appendix.

Limitations

Our main concern is external validity.

- Although we tried many specifications, this is only one dataset.
- We selected data from South Africa so the distribution was highly skewed.

Still, these methods have their idiosyncratic strengths and limitations.

RMSE and inequality trade-off prevails. It is unlikely that we find a radical ordering in their performance

Take-home ideas

- We show how biased inequality measures can be when missing incomes are not properly assessed.
- Sample reweighting is the best method to reduce the inequality bias.
- Other methods, such as Deletion, seem to palliate the bias in some patterns (MNAR at the tails).
- If your objective is to minimize OOS RMSE, Machine Learning (especially random forests) is a good solution.

Q&A

Thank you for your attention!

All remaining questions can be addressed to p.salas-rojo@lse.ac.uk

Gini (MCAR)

N	Share of Missing	Deletion	Imputation of the Mean	Single Parametric	Multiple Parametric	PMM
6983	3	0.02*	-1.49	-0.31	-0.23	-0.02
6839	5	-0.01*	-1.25	-0.59	-0.43	-0.08
6695	7	-0.01*	-1.78	-0.84	-0.60	-0.11
6479	10	-0.02*	-2.64	-1.22	-0.88	-0.12
6119	15	-0.01*	-4.11	-1.60	-1.31	-0.20
5759	20	0.03*	-5.70	-2.43	-1.74	-0.17
5039	30	-0.02*	-9.43	-4.59	-2.79	-0.34
4319	40	-0.05*	-13.72	-5.42	-3.84	-0.44
3600	50	0.00*	-18.57	-6.91	-4.94	-0.54

N	Share of Missing	Pareto Model	Reweighting	LASSO	Tree	Random Forest
6983	3	2.08	0.09*	-0.21	-0.19	-0.36
6839	5	2.05	0.01*	-0.37	-0.45	-0.57
6695	7	2.02	-0.02*	-0.55	-0.77	-0.82
6479	10	2.16	0.03*	-0.81	-1.15	-1.11
6119	15	2.12	0.04*	-1.21	-1.22	-1.68
5759	20	2.19	-0.02*	-1.68	-2.34	-2.34
5039	30	2.37	0.05*	-2.71	-4.20	-3.92
4319	40	2.67	-0.04*	-3.79	-5.06	-5.46
3600	50	2.92	0.00*	-4.95	-7.42	-7.23

Return to ▸ Slide.

Gini (MID)

N	Share of Missing	Deletion	Imputation of the Mean	Single Parametric	Multiple Parametric	PMM
6983	3	0.30	-0.60	-0.12	-0.03	-0.19
6839	5	0.49	-1.03	-0.27	-0.06	-0.29
6695	7	0.68	-1.49	-0.38	-0.09	-0.48
6479	10	0.96	-2.23	-0.55	-0.14	-0.62
6119	15	1.42	-3.58	-0.80	-0.23	-0.97
5759	20	1.87	-5.11	-0.79	-0.31	-1.28
5039	30	2.77	-8.61	-1.02	-0.45	-1.71
4319	40	3.64	-12.83	-1.96	-0.61	-2.67
3600	50	4.47	-17.60	-2.07	-0.70	-3.31

N	Share of Missing	Pareto Model	Reweighting	LASSO	Tree	Random Forest
6983	3	4.26	0.32	-0.02	-0.13*	-0.22*
6839	5	5.99	0.51	-0.04	-0.20*	-0.34*
6695	7	7.66	0.70	-0.04	-0.31	-0.45*
6479	10	6.25	0.97	-0.06	-0.41	-0.62
6119	15	5.35	1.44	-0.09	-0.51	-0.89
5759	20	3.36	1.90	-0.12	-0.79	-1.31
5039	30	4.16	2.84	-0.23	-1.13	-2.00
4319	40	7.34	3.82	-0.35	-1.85	-2.65
3600	50	5.65	4.80	-0.58	-1.71	-3.29

Return to ▸ Slide.

Gini (LEFT)

N	Share of Missing	Deletion	Imputation of the Mean	Single Parametric	Multiple Parametric	PMM
6983	3	0.03	-0.50	-0.17	-0.10*	-0.01*
6839	5	0.05	-0.89	-0.26	-0.18	-0.04
6695	7	0.05	-1.32	-0.47	-0.28	-0.06
6479	10	0.06	-2.00	-0.73	-0.43	-0.08
6119	15	0.09	-3.31	-1.18	-0.69	-0.11
5759	20	0.13	-4.79	-1.78	-0.96	-0.10
5039	30	0.20	-8.35	-2.59	-1.57	-0.14
4319	40	0.27	-12.69	-2.88	-2.21	-0.08
3600	50	0.30	-17.87	-3.76	-2.88	-0.10

N	Share of Missing	Pareto Model	Reweighting	LASSO	Tree	Random Forest
6983	3	3.61	0.07	-0.10	-0.15*	-0.22*
6839	5	4.91	0.11	-0.17	-0.29	-0.36
6695	7	6.41	0.14	-0.24	-0.45	-0.52
6479	10	8.31	0.20	-0.36	-0.67	-0.80
6119	15	3.40	0.31	-0.57	-1.01	-1.22
5759	20	5.21	0.45	-0.79	-1.50	-1.69
5039	30	3.70	0.76	-1.33	-2.19	-2.82
4319	40	6.74	1.11	-1.96	-3.00	-3.94
3600	50	3.50	1.46	-2.69	-3.62	-5.12

Return to ▸ Slide.

Gini (RIGHT)

N	Share of Missing	Deletion	Imputation of the Mean	Single Parametric	Multiple Parametric	PMM
6983	3	-4.73	-5.73	-4.70	-4.96	-4.49
6839	5	-5.43	-6.90	-5.20	-5.79	-5.27
6695	7	-6.09	-8.02	-5.68	-6.55	-5.85
6479	10	-6.70	-9.32	-6.53	-7.33	-6.38
6119	15	-7.51	-11.36	-7.84	-8.42	-7.14
5759	20	-8.15	-13.34	-9.09	-9.39	-7.76
5039	30	-9.12	-17.32	-10.62	-11.06	-8.54
4319	40	-9.88	-21.69	-11.75	-12.52	-9.06
3600	50	-10.50	-26.50	-14.46	-13.92	-9.44

N	Share of Missing	Pareto Model	Reweighting	LASSO	Tree	Random Forest
6983	3	2.64	-3.36	-5.03	-5.55	-5.16
6839	5	6.56	-4.22	-5.99	-6.34	-6.14
6695	7	3.68	-4.78	-6.77	-6.79	-6.86
6479	10	5.80	-5.54	-7.69	-8.10	-8.09
6119	15	2.82	-6.55	-8.79	-9.55	-9.63
5759	20	4.21	-7.33	-9.78	-11.42	-10.95
5039	30	0.75	-8.63	-11.50	-13.52	-13.12
4319	40	3.88	-9.75	-13.03	-14.56	-15.29
3600	50	3.15	-10.67	-14.57	-17.56	-17.35

Return to ▸ Slide.

Gini (TAIL)

N	Share of Missing	Deletion	Imputation of the Mean	Single Parametric	Multiple Parametric	PMM
6983	3	-4.08	-4.93	-3.94	-4.24	-3.92
6839	5	-4.82	-6.07	-4.82	-5.07	-4.64
6695	7	-5.31	-6.94	-5.16	-5.64	-5.11
6479	10	-5.96	-8.20	-6.03	-6.41	-5.75
6119	15	-6.83	-10.11	-6.82	-7.48	-6.53
5759	20	-7.53	-11.92	-7.88	-8.36	-7.17
5039	30	-8.63	-15.55	-10.03	-9.94	-8.19
4319	40	-9.65	-19.53	-12.19	-11.51	-9.12
3600	50	-10.50	-23.85	-13.00	-12.88	-9.88

N	Share of Missing	Pareto Model	Reweighting	LASSO	Tree	Random Forest
6983	3	3.95	-3.36	-4.28	-4.52	-4.32
6839	5	2.13	-4.22	-5.15	-5.64	-5.34
6695	7	4.43	-4.78	-5.82	-6.44	-6.21
6479	10	6.06	-5.54	-6.60	-6.94	-7.19
6119	15	3.65	-6.55	-7.78	-8.12	-8.31
5759	20	5.14	-7.33	-8.64	-9.47	-9.71
5039	30	3.41	-8.63	-10.30	-11.81	-12.29
4319	40	1.69	-9.75	-11.85	-14.99	-14.83
3600	50	3.56	-10.67	-13.41	-16.38	-16.97

Return to ▸ Slide.

References I

Alvaredo, F. and Gasparini, L. (2015). Chapter 9 - recent trends in inequality and poverty in developing countries. In Atkinson, A. B. and Bourguignon, F., editors, *Handbook of Income Distribution*, volume 2 of *Handbook of Income Distribution*, pages 697–805. Elsevier.

Hlasny, V. (2020). Nonresponse bias in inequality measurement: Cross-country analysis using luxembourg income study surveys. *Social Science Quarterly*, 101(2):712–731.

Hlasny, V., Ceriani, L., and Verme, P. (2021). Bottom incomes and the measurement of poverty and inequality. *Review of Income and Wealth*.

Hothorn, T., Hornik, K., and Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics*, 15(3):651–674.

References II

Jenkins, S. P. (2017). Pareto models, top incomes and recent trends in uk income inequality. *Economica*, 84(334):261–289.

Korinek, A., Mistiaen, J. A., and Ravallion, M. (2006). Survey nonresponse and the distribution of income. *The Journal of Economic Inequality*, 4(1):33–55.

Meyer, B. D., Mok, W. K., and Sullivan, J. X. (2015). Household surveys in crisis. *Journal of Economic Perspectives*, 29(4):199–226.

Ravallion, M. (2022). Missing top income recipients. *The Journal of Economic Inequality*, 20(1):205–222.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

References III

Rubin, D. B. (1978). Multiple imputations in sample surveys-a phenomenological bayesian approach to nonresponse. In *Proceedings of the survey research methods section of the American Statistical Association*, volume 1, pages 20–34. American Statistical Association.

Schenker, N. and Taylor, J. M. (1996). Partially parametric techniques for multiple imputation. *Computational statistics & data analysis*, 22(4):425–446.

Schouten, R. M., Lugtig, P., and Vink, G. (2018). Generating missing values for simulation purposes: a multivariate amputation procedure. *Journal of Statistical Computation and Simulation*, 88(15):2909–2930.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B*, 58(1):267–88.