

Online Appendix to: Deng & Lindeboom (2021), “Early-life Famine Exposure, Hunger Recall and Later-life Health”

Zichen Deng Maarten Lindeboom*

Appendix A Theoretical Results

A.1 Model

Suppose we are interested in estimating the treatment effect of a binary treatment D on outcome Y in a primary population of interest, which is confounded by measured covariates X and unmeasured ones U , with the aid of an instrumental variable Z . However, we only observe $(Y_i, Z_i, X_i), i = 1, \dots, N_1$ from this population F_p . As a remedy, suppose an additional sample $(D_j, Z_j, X_j), j = 1, \dots, N_0$ is available from an auxiliary population F_a , possibly different from the primary population. Let R be an indicator variable equal to 1 if drawn from the primary population and 0 otherwise. We use the notation $D(0)$ to represent the latent D in the primary sample. The following assumptions give a formal definition of the data combination model.

Assumption A1 (Random Sampling). *With probability $Q \in (\xi, 1 - \xi)$ for $0 < \xi < 0.5$, we draw a unit at random from F_p and record its realizations of Y , Z , and X , otherwise we*

*This version: Tuesday 28th September, 2021. Deng: NHH Norwegian School of Economics, zichen.deng@nhh.no; Lindeboom: Vrije Universiteit Amsterdam, Centre for Health Economics, Monash University, Tinbergen Institute and IZA Bonn, m.lindeboom@vu.nl. We gratefully acknowledge the valuable comments from the co-editor Thierry Magnac and three anonymous reviewers.

draw a unit at random from F_a and record its realizations of D , Z , and X .

Assumption A2 (Weak Overlap). Let $\mathcal{X}_p = \{x : f_p(x) > 0\}$ and $\mathcal{X}_a = \{x : f_a(x) > 0\}$, then $\mathcal{X}_p \subseteq \mathcal{X}_a$.

Assumption A3 (Conditional Distributional Equality). $F_p(D(0)|Z, X) = F_a(D|Z, X)$,
 $F_p(Y|Z, X) = F_a(Y|Z, X)$

Similar to [Graham, Pinto, and Egel \(2016\)](#); [Shu and Tan \(2020\)](#), Assumption [A1](#) defines how the data are generated. Assumption [A2](#) states that the support of the common variables (Z, X) in the primary sample is contained within the support of the auxiliary sample. This ensures that for each unit in the study (primary) sample, there will be matching units with similar values of X in the auxiliary sample.¹ Assumption [A3](#) requires predictive invariance for the treatment between the two heterogeneous populations. The distributions of (Y, Z, X) and (D, X, Z) in the two populations differ only in terms of their marginal distributions for the always measured variable, (Z, X) . This assumption is similar to the idea of selection-on-observables.

Let P^* be the matched sample generated by matching each unit in the primary sample, i , to the auxiliary sample, $J(i)$ with replacement. We only consider one-to-one matching, since the auxiliary sample in our empirical application is only slightly larger than the primary sample. We choose the sets of matches $J(i)$ to minimize the sum of the matching discrepancies, $\sum_{i=1}^{N_1} d(X_i, X_{J(i)})$, where $d(\cdot)$ is the distance metric to measure the matching discrepancies. The commonly used distance metrics include, for example, the Mahalanobis distance. Similar to the matching literature, we assume that the sum of matching discrepancies vanishes (i.e., $\frac{1}{\sqrt{N_1}} \sum_{i=1}^{N_1} d(X_i, X_{J(i)}) \xrightarrow{p} 0$) quickly enough to allow asymptotic unbiasedness as $N_0, N_1 \rightarrow \infty$ with $N_0 > N_1$.

We now describe the population distribution targeted by the matched sample, P^* . Since $F_p(\cdot)$ and $F_a(\cdot)$ are the cumulative distribution functions from the primary and auxiliary

¹In the empirical application this is also verified: for less than 5% of the individuals in the main sample, we can't find a match from the auxiliary sample.

samples, we define $E_p[\cdot]$ and $E_a[\cdot]$ as the corresponding expectation operators. We define a matching target distribution, F_p^* , as

$$E_p^*[(D, Z, X) \in A | R = 1] = E_p[(D, Z, X) \in A | R = 1] \text{ and}$$

$$E_p^*[(D, Z, X) \in A | R = 0] = E_p[E_a[(D, Z, X) \in A | Z, X, R = 0] | R = 1],$$

where E_p^* represents the corresponding expectation operators on matched targeting distribution and R an indicator that equals 1 for the primary sample and 0 otherwise. The first expression holds because the primary sample is our matched targeting distribution. The second expression, the distribution of (D, Z, X) in the auxiliary sample, is generated by integrating the conditional distribution of (D, Z, X) given Z, X over the distribution of Z, X in the primary sample.

Assumptions A1, A2, and A3 allow researchers to balance the primary and the auxiliary sample. To proceed, let $K = g(Z, X)$ be a $(k \times 1)$ vector of functions of (Z, X) , and let $\tilde{\beta}$ be the vector of regression coefficients obtained from regressing D on K in the matched sample. The choice of K can be but is not limited to (Z, X) . To ensure that matching is working, we also need to assume that conditional expectations are well-behaved and $H = E(KK')$ is invertible. Other assumptions can be found in [Abadie and Imbens \(2012\)](#).

The following Proposition [A1](#) formalizes the idea that the first-stage estimates of the matched sample recover the parameters of the matching target distribution (i.e., the distribution of the primary sample).

Proposition A1. *Under regularity conditions, the regression coefficients ($\tilde{\beta}$) of D on K in the matched sample, P^* , are unbiased estimates of the analogous regression coefficients (β) in the population of the primary sample.*

Proof. We use the notation $D(0)$ to represent the latent D in the primary sample. Therefore, the regression coefficient in the primary (target) sample is defined by $E_p[(D(0) - K'b)^2]$.

$$E_p[(D(0) - K'b)^2] = E_p[E_p[(D(0) - K'b)^2|Z, X]] \quad (1)$$

$$= E_p[E_a[(D - K'b)^2|Z, X]] \quad (2)$$

$$= E_p[E_a[(D - K'b)^2|Z, X, R = 0]|R = 1] \quad (3)$$

$$= E^*[(D - K'b)^2|R = 0] \quad (4)$$

The equality in (6) follows from the law of iterated expectations; the equality in (7) follows from propensity score equality (Assumption 3). Equations (8) and (9) follow from the definition of the matching target distribution. Until here, we have shown that matching under propensity score equality allows us to reproduce the first stage setting for the primary sample. Therefore, the regression coefficient in the primary sample is recovered using the matched sample.

To further establish the large sample property of the estimator, let $\tilde{\beta}$ be the vector of the sample regression coefficients obtained from regressing D on K in the matched sample,

$$\tilde{\beta} = \operatorname{argmin}_{b \in R^k} \frac{1}{N_1} \sum_{i \in P^*} (D - K'b)^2 = \left(\frac{1}{N_1} \sum_{i \in P^*} K K' \right)^{-1} \frac{1}{N_1} \sum_{i \in P^*} K D. \quad (5)$$

From 6-9, the matching procedure makes sure that $\frac{1}{N_1} \sum_{i \in P^*} K K' \xrightarrow{p} H$. $H = E(K K')$ is the Hessian, which is invertible by assumption.

$$\tilde{\beta} - \beta = \left(\frac{1}{N_1} \sum_{i \in P^*} K K' \right)^{-1} \frac{1}{N_1} \sum_{i \in P^*} (K D - K K' \beta) \xrightarrow{p} 0 \quad (6)$$

□

A.2 Nonlinear Models

Above, we used a linear model for the second step after balancing the primary and the auxiliary sample. The results also hold for more complex (nonlinear) models. For example, we can consider the following moment condition proposed in [Graham, Pinto, and Egel \(2016\)](#)

$$\mathbb{E}_p[\psi_p(Y; \beta) - \psi_a(D, Z_1; \beta)e(Z)] = 0 \quad (7)$$

with $Z = (Z'_0, Z'_1)'$. $E_p[\cdot]$ denotes expectations taken with respect to the primary population. β is the parameter of interest. There exist identification results for moment condition 7, when D and Y are observed in two different samples ([Chen, Hong, Tarozzi et al., 2008](#)). Note that both TSIV and TSTSLs methods can be seen as a special case of the moment condition 7. For example, we have the linear model of [Angrist and Krueger \(1992\)](#) if we take in 7 $e(Z) = Z$, $\psi_p(Y; \beta) = Y$ and $\psi_a(D, Z_1; \beta) = D'\gamma_1 + Z'_1\gamma_2$ with $\beta = (\gamma_1, \gamma'_2)'$.

A.3 Simulation Results

We perform simulation results for the two classical methods (TSIV and TSTSLs), the Inverse Probability Tilting method (IPT), and the two-step-TSTSLs estimator proposed in this paper. In each of our experiments, we assume that X in both the primary sample and the auxiliary sample is distributed according to a truncated normal distribution, with support $[0, 2]$. The location and scale parameters of both distributions, (μ_p, ω_p^2) and (μ_a, ω_a^2) , may differ. We assume a multinomial sampling scheme: with probability $Q_0 = 1/2$ a draw of (Y, Z, X) is taken at random from the population to constitute the primary sample; otherwise, a draw of (D, Z, X) is taken from the population to constitute the auxiliary sample. We set $\mu_p = 1.5$ and $\mu_a = 0.5$. We vary ω_p and ω_a to reflect the overlap between the primary sample and the auxiliary sample. In case 1, we take $\omega_p = \omega_a = 1$. Alternatively, in case 2, we take $\omega_p = \omega_a = 0.3$. In case 1, there is much overlap, which means in practice that the distribution of X does not differ too much in both samples. In case 2, there is little

overlap, implying that the distribution of X in both samples differs a lot. Finally, we assume that Y and D are generated according to the following data generating process:

For the primary sample we generate data according to

$$Y = 0.5D + U, \tag{8}$$

and the endogenous variable D in the auxiliary sample is generated by

$$D = 0.5Z + \theta XZ + V, \tag{9}$$

where Z is distributed as $N(0, 1)$ and (U, V) are distributed independently of (Z, X) as

$$\begin{pmatrix} U \\ V \end{pmatrix} \sim N \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}. \tag{10}$$

For each simulation, we generated an i.i.d. sample of size $N_0 = 1000$ of (Y, Z, X) from the population (the primary sample) and an i.i.d. sample of size $N_1 = 1000$ of (D, Z, X) from the population (the auxiliary sample). We then merge the two samples. With $\theta = 0$, the setting is simplified to the classical two sample models. With $\theta \neq 0$, we simulate misspecification.

Table [A1](#) presents the results for four different scenarios. In scenario 1 the model is correctly specified, the overlap is good ($\omega_p = \omega_a = 1$) and there is no misspecification ($\theta = 0$). All four methods (see the first four rows) perform well with a very small bias and a small Root Mean Squared Error (RMSE). However, as expected, the two-step estimator performs worse than the other three methods on efficiency. The two-step-TSTSLS does not use the information of all data points, which results in a larger RMSE. In scenario 2, the overlap is good, but the model is misspecified ($\theta = 0.3$). The IPT estimator performs best with the smallest bias and RMSE. We next turn to situations where the overlap in the distribution of X in both samples is poor. In scenario 3, we take ($\theta = 0$) (i.e., the model is correctly specified), while in scenario 4, we assume that both the model is misspecified

($\theta = 0.3$) and the overlap is poor. In scenario 3, the two-step-TSTOLS estimator performs best on the bias, but the RMSE (like the IPT estimator) is less efficient. In scenario 4, all estimators are biased, but the two-step-TSTOLS estimator performs best.

Table A1. Monte Carlo Results

	(1)	(2)	(3)	(4)
	N	Asym. Bias	Std dev.	RMSE
Scenario 1: Good overlap and correct specification				
TSIV	1000	0.005	0.080	0.080
TSTOLS	1000	0.004	0.080	0.080
IPT	1000	0.006	0.084	0.084
TWO-STEP-TSTOLS	1000	0.006	0.085	0.085
Scenario 2: Good overlap and incorrect specification				
TSIV	1000	0.060	0.054	0.081
TSTOLS	1000	0.059	0.054	0.080
IPT	1000	0.003	0.054	0.054
TWO-STEP-TSTOLS	1000	0.028	0.053	0.060
Scenario 3: Bad overlap and correct specification				
TSIV	1000	0.002	0.084	0.084
TSTOLS	1000	0.001	0.084	0.084
IPT	1000	-0.862	26.666	26.666
TWO-STEP-TSTOLS	1000	0.016	0.117	0.119
Scenario 4: Bad overlap and incorrect specification				
TSIV	1000	0.202	0.065	0.212
TSTOLS	1000	0.200	0.064	0.210
IPT	1000	0.108	1.910	1.912
TWO-STEP-TSTOLS	1000	0.088	0.078	0.117

In a second set of simulations, we vary the trade-off between efficiency (RMSE) and bias when the model is misspecified with a varying degree of misspecification. For these simulations we fix the overlap parameter $\omega_p = 0.3$ (bad overlap). We subsequently take 1,000 repeated simulations under four scenarios where we vary the degree of misspecification, i.e. we vary θ . In scenario 1, we take $\theta = 0$, i.e. the model is specified correctly. This scenario is equal to the scenario 3 of Table A1. In scenario 2 to 4, we gradually increase the degree of misspecification, with steps of 0.1, i.e. we take $\theta = 0.1, 0.2, 0.3$ for scenarios 2, 3, 4, respectively. We report the results of these simulations in Table A2.

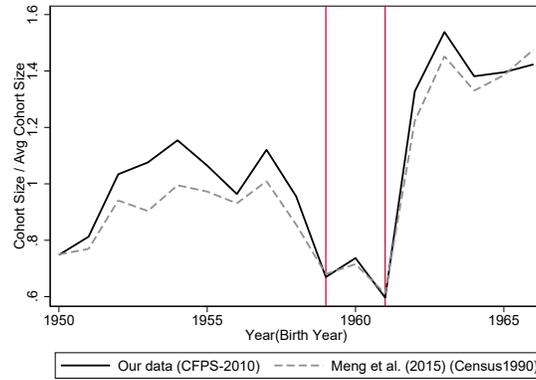
In scenario 1 The bias of the TSIV, TSTSLS and the two-step-TSTSLS are similar and outperform the IPT estimator. However, the two-step-TSTSLS (and the IPT) is less efficient than the TSIV and TSTSLS estimators. When we gradually increase the degree of misspecification (scenarios 2 to 4), the performance of the TSIV, the TSTSLS and the IPT estimators deteriorate quickly. The two-step-TSTSLS estimator performs very well in comparison with the other estimators. The IV estimate can be understood as the ratio of the intention to treat estimate and the first stage estimate. Therefore, misspecifications in the first stage regression translate in relatively large biases. Similarly, misspecifications in the first stage will increase the bias of the TSTSLS. The two-step-TSTSLS is robust against misspecifications and therefore we suggest that this method can be used as a robustness check in empirical applications.

Table A2. Additional Monte Carlo Results

	(1)	(2)	(3)	(4)
	N	Asym. Bias	Std dev.	RMSE
Scenario 1: Bad overlap and correct specification ($\theta = 0.0$)				
TSIV	1000	0.002	0.084	0.084
TSTSLS	1000	0.001	0.084	0.084
IPT	1000	-0.862	26.666	26.666
TWO-STEP-TSTSLS	1000	0.016	0.117	0.119
Scenario 2: Bad overlap and incorrect specification ($\theta = 0.1$)				
TSIV	1000	0.080	0.071	0.107
TSTSLS	1000	0.079	0.071	0.106
IPT	1000	0.492	3.905	3.934
TWO-STEP-TSTSLS	1000	0.045	0.095	0.105
Scenario 3: Bad overlap and incorrect specification ($\theta = 0.2$)				
TSIV	1000	0.150	0.068	0.165
TSTSLS	1000	0.148	0.068	0.163
IPT	1000	0.419	10.174	10.178
TWO-STEP-TSTSLS	1000	0.071	0.085	0.111
Scenario 4: Bad overlap and incorrect specification ($\theta = 0.3$)				
TSIV	1000	0.202	0.065	0.212
TSTSLS	1000	0.200	0.064	0.210
IPT	1000	0.108	1.910	1.912
TWO-STEP-TSTSLS	1000	0.088	0.078	0.117

Appendix B Additional Figures and Tables

Figure B1. Cohort Loss in CFPS



Notes: The figure compares the relative survivor birth cohort sizes in our data set (CFPS-2010, the solid line) with the relative cohort sizes in [Meng, Qian, and Yared \(2015\)](#) (Census1990, the dashed line).

Table B1. Effects on Separate Components

	(1)	(2)	(3)	(4)
		Components		
	Metabolic syndrome (index)	Diabetes	High blood pressure	Obesity
<i>Panel A: Female</i>				
Hunger before age 5	0.38*** (0.15)	0.034 (0.021)	0.089* (0.052)	0.058 (0.039)
Mother literate		0.0011 (0.0054)		-0.0023 (0.0069)
Age		0.041 (0.038)		0.12* (0.070)
Age squared(/100)		-0.039 (0.037)		-0.12* (0.069)
Observations	2517	2517	2517	2517
<i>Panel B: Male</i>				
Hunger before age 5	0.032 (3.93)	-0.031 (0.059)	-0.038 (0.30)	0.13 (0.093)
Mother literate		-0.0062 (0.0044)		0.022** (0.011)
Age		-0.0031 (0.041)		0.064 (0.096)
Age squared(/100)		0.0017 (0.040)		-0.059 (0.094)
Observations	2612	2612	2612	2612

Notes: Each coefficient is from a separate regression. All regressions use the log(EDR) as the instrumental variable. The sample contains all individuals born between 1957 and 1962 in three waves of CFPS. Three components, diabetes, hypertension, and obesity, are dummy indicators constructed from CFPS. Standard errors clustered by province based on matched bootstrap (Abadie and Spiess, 2019) with 999 replications appear in parentheses. *, **, *** indicates significance at the 10%, 5% and 1% level respectively.

Table B2. Reduced-form Estimates at Age 0-5

	(1)	(2)	(3)	(4)
	Metabolic syndrome (index)			
	Female	Female	Male	Male
log(EDR)	0.033*** (0.013)	0.034*** (0.012)	0.0021 (0.013)	0.0032 (0.013)
Mother literate		0.033 (0.030)		0.030 (0.028)
Age		0.43*** (0.16)		0.14 (0.087)
Age squared(/100)		-0.40** (0.16)		-0.13 (0.084)
Observations	2517	2517	2612	2612

Notes: The results are based on reduced-form estimates from separate regressions. All regressions use the (matched) primary sample of individuals born between 1957 and 1962 from three waves of the China Family Panel Survey (CFPS). Standard errors clustered by province based on bootstrap with 999 replications appear in parentheses. *, **, *** indicates significance at the 10%, 5% and 1% level, respectively.

Table B3. Effects of Hunger at Age 0-5

	(1)	(2)	(3)	(4)
	Metabolic syndrome (index)			
	Female	Female	Male	Male
<i>Panel A: log(EDR) as the instrumental variable</i>				
Hunger before age 5	0.38*** (0.14)	0.40*** (0.14)	0.032 (0.46)	0.048 (0.36)
Mother literate		0.025 (0.030)		0.032 (0.032)
Age		0.65* (0.35)		0.13 (0.21)
Age squared(/100)		-0.61* (0.34)		-0.12 (0.21)
Observations	2517	2517	2612	2612
<i>Panel B: EDR as the instrumental variable</i>				
Hunger before age 5	0.40 (0.41)	0.43 (0.39)	0.031 (2.57)	0.060 (0.55)
Mother literate		0.024 (0.035)		0.032 (0.034)
Age		0.68 (0.57)		0.13 (0.31)
Age squared(/100)		-0.65 (0.56)		-0.12 (0.31)
Observations	2517	2517	2612	2612

Notes: The results are based on TSIV estimates from separate regressions. All regressions are based on the (matched) primary sample of individuals born between 1957 and 1962 from three waves of the China Family Panel Survey (CFPS). Panel A uses the log(EDR) as the instrumental variable. Panel B uses the EDR as the instrumental variable. Standard errors clustered by province based on matched bootstrap (Abadie and Spiess, 2019) with 999 replications appear in parentheses. *, **, *** indicates significance at the 10%, 5% and 1% level, respectively.

Table B4. OLS Estimates at Age 0-5

	(1)	(2)	(3)	(4)
	Metabolic syndrome (index)			
	Female	Female	Male	Male
<i>Panel A: Matched recall</i>				
Hunger before age 5	-0.0017 (0.037)	-0.00034 (0.035)	0.055* (0.033)	0.057 (0.035)
Mother literate		0.032 (0.029)		0.032 (0.028)
Age		0.41*** (0.15)		0.13 (0.089)
Age squared(/100)		-0.38*** (0.15)		-0.12 (0.087)
Observations	2517	2517	2612	2612
<i>Panel B: Own recall</i>				
Hunger before age 5	-0.022 (0.079)	-0.039 (0.074)	-0.019 (0.040)	-0.023 (0.042)
Mother literate		0.032 (0.029)		0.029 (0.027)
Age		0.41** (0.16)		0.13 (0.083)
Age squared(/100)		-0.38** (0.16)		-0.12 (0.080)
Observations	2517	2517	2612	2612

Notes: The results are based on simple OLS regressions. All regressions use the (matched) primary sample of individuals born between 1957 and 1962 from three waves of the China Family Panel Survey (CFPS). Standard errors clustered by province based on bootstrap with 999 replications appear in parentheses. *, **, *** indicates significance at the 10%, 5% and 1% level, respectively.

Table B5. Effects of Hunger at Age 0-5

	(1)	(2)	(3)	(4)
	Hospital Visits			
	Female	Female	Male	Male
Hunger before age 5	0.21 (0.14)	0.21 (0.16)	0.077 (0.18)	0.078 (0.38)
Mother literate		0.0033 (0.017)		-0.025* (0.014)
Age		0.19 (0.23)		0.11 (0.18)
Age squared(/100)		-0.18 (0.22)		-0.11 (0.17)
Observations	2517	2517	2612	2612

Notes: The results are based on simple OLS regressions. All regressions use the (matched) primary sample of individuals born between 1957 and 1962 from three waves of the China Family Panel Survey (CFPS). Standard errors clustered by province based on matched bootstrap (Abadie and Spiess, 2019) with 999 replications appear in parentheses. *, **, *** indicates significance at the 10%, 5% and 1% level, respectively.

Table B6. Metabolic Syndrome and Hospital Visits

	(1)	(2)	(3)	(4)
	Hospital Visits			
	Female	Female	Male	Male
Metabolic syndrome(index)	0.069*** (0.020)	0.070*** (0.019)	0.027 (0.027)	0.028 (0.027)
Mother literate		0.012 (0.018)		-0.027* (0.015)
Age		0.049 (0.12)		0.13* (0.072)
Age squared(/100)		-0.044 (0.11)		-0.12* (0.068)
Observations	2517	2517	2612	2612

Notes: The results are based on simple OLS regressions. All regressions use the (matched) primary sample of individuals born between 1957 and 1962 from three waves of the China Family Panel Survey (CFPS). Standard errors clustered by province based on bootstrap with 999 replications appear in parentheses. *, **, *** indicates significance at the 10%, 5% and 1% level, respectively.

References

- Abadie, Alberto and Guido W Imbens. 2012. “A martingale representation for matching estimators.” *Journal of the American Statistical Association* 107 (498):833–843.
- Abadie, Alberto and Jann Spiess. 2019. “Robust Post-Matching Inference.” *Working paper* .
- Angrist, Joshua and Alan Krueger. 1992. “The Effect of Age at School Entry on Educational Attainment: An Application of Instrumental Variables with Moments from Two Samples.” *Journal of the American Statistical Association* 87 (418):328–336.
- Chen, Xiaohong, Han Hong, Alessandro Tarozzi et al. 2008. “Semiparametric Efficiency in GMM Models with Auxiliary Data.” *Annals of statistics* 36 (2):808–843.
- Graham, Bryan S, Cristine Campos de Xavier Pinto, and Daniel Egel. 2016. “Efficient Estimation of Data Combination Models by the Method of Auxiliary-to-Study Tilting (AST).” *Journal of Business & Economic Statistics* 34 (2):288–301.
- Meng, Xin, Nancy Qian, and Pierre Yared. 2015. “The Institutional Causes of China’s Great Famine, 1959–1961.” *The Review of Economic Studies* 82 (4):1568–1611.
- Shu, Heng and Zhiqiang Tan. 2020. “Improved methods for moment restriction models with data combination and an application to two-sample instrumental variable estimation.” *Canadian Journal of Statistics* .