Accepted to *Information Systems Research*

Online Appendix for "What, Why, and How: An Empiricist's Guide to Double/Debiased Machine Learning"

Bowen Shi
School of Economics and Management, Tsinghua University, sbw22@mails.tsinghua.edu.cn

Xiaojie Mao
School of Economics and Management, Tsinghua University; Research Center for Contemporary Management, Tsinghua University, maoxj@sem.tsinghua.edu.cn

Mochen Yang
Carlson School of Management, University of Minnesota, yang3653@umn.edu

Bo Li
School of Economics and Management, Tsinghua University, libo@sem.tsinghua.edu.cn

Table A1 Estimation results for $\theta_0 = 1$ from different estimators across 200 experiment replications. In each cell, the three numbers stand for the average of the estimates, the standard deviation of the estimates, and the coverage of the associated empirical confidence intervals, all based on the 200 replications

Scenario 1	$p=2$		$p=5$		$p=10$	
	$N=100$	$N=1000$	$N=100$	$N=1000$	$N=100$	$N=1000$
Linear regression	0.5747	0.5300	-0.5673	-0.5019	-2.2739	-2.3294
	(0.5756)	(0.1801)	(0.9851)	(0.3214)	(2.3283)	(0.7784)
	88.5%	26.0%	69.5%	0.5%	60.0%	0.5%
Traditional Semiparametrics						
Kernel (Backfitting)	1.5421	1.2568	1.4049	1.1795	-	-
	(0.5400)	(0.1553)	(0.8487)	(0.2612)	-	-
	-	-	-	-	-	-
Kernel (Speckman)	1.0619	0.9785	1.0895	1.0072	-1.4930	1.3096
	(0.5392)	(0.1558)	(1.4277)	(0.3865)	(44.2363)	(4.8680)
	92.0%	95.0%	77.0%	90.0%	37.5%	30.5%
DML						
Kernel	1.0641	0.9767	1.1741	1.0476	1.4430	1.2811
	(0.5449)	(0.1573)	(0.7812)	(0.2175)	(1.3071)	(0.3433)
	92.5%	95.0%	85.5%	85.5%	84.0%	76.5%
Decision tree	0.8142	0.8903	0.5595	0.6616	0.7495	0.7290
	(0.5682)	(0.1533)	(0.7748)	(0.2262)	(1.4747)	(0.3801)
	89.7%	90.5%	92.2%	69.5%	94.0%	91.5%
Random forest	0.9899	0.9607	0.9101	0.8795	1.3879	0.8926
	(0.5653)	(0.1550)	(0.6845)	(0.1854)	(1.3712)	(0.3215)
	91.5%	94.5%	98.0%	94.0%	95.5%	97.5%
XGBoost	1.0011	0.9557	0.9065	0.9802	0.8236	0.8687
	(0.5418)	(0.1533)	(0.7016)	(0.2137)	(1.3885)	(0.3506)
	92.0%	94.5%	96.0%	93.5%	93.0%	94.5%

Scenario 2	$p=2$		$p=5$		$p=10$	
	$N=100$	$N=1000$	$N=100$	$N=1000$	$N=100$	$N=1000$
Linear regression	0.2470	0.1856	-0.0331	-0.0052	-0.3620	-0.1272
	(1.0102)	(0.3034)	(1.6702)	(0.4610)	(2.7224)	(0.7502)
	86.5%	21.5%	90.0%	49.5%	86.5%	70.5%
Traditional Semiparametrics						
Kernel (Backfitting)	1.9049	1.5643	1.3023	1.1603	-	-
	(0.5708)	(0.1593)	(0.8919)	(0.2695)	-	-
	-	-	-	-	-	-
Kernel (Speckman)	0.9997	0.9656	0.9977	0.9672	-5.7326	0.7569
	(0.5461)	(0.1561)	(1.4302)	(0.3946)	(57.5675)	(3.9324)
	92.5%	93.5%	81.0%	89.5%	35.5%	34.0%
DML						
Kernel	0.9384	0.9537	0.7016	0.8702	0.0788	0.4595
	(0.6558)	(0.1693)	(1.2860)	(0.3274)	(2.1188)	(0.5604)
	94.0%	95.0%	88.0%	90.5%	84.5%	81.5%
Decision tree	0.8799	0.8989	0.7991	0.7221	0.6944	0.5243
	(0.6945)	(0.1703)	(1.0604)	(0.3028)	(1.5685)	(0.4589)
	90.3%	92.0%	92.8%	80.0%	92.2%	78.2%
Random forest	0.9806	0.9669	1.0186	0.9978	0.9883	0.9457
	(0.6718)	(0.1616)	(0.9985)	(0.2778)	(1.6294)	(0.3869)
	96.5%	94.0%	97.5%	95.0%	96.5%	98.0%
XGBoost	0.9692	0.9484	0.8390	0.9656	0.5351	0.8818
	(0.6647)	(0.1636)	(0.9649)	(0.2867)	(1.6505)	(0.3761)
	94.0%	96.0%	96.0%	94.0%	91.0%	92.5%

Table A2 Estimation results of the DML estimators under orthogonal and their counterparts using non-orthogonal estimating equations, with different cross-fitting folds in Scenario 1 (across 200 experiment replications). In each cell, the three numbers stand for the average of the estimates, the standard deviation of the estimates, and the coverage of the associated empirical confidence intervals, all based on the 200 replications

$p=2$	Orthogonal				Non-orthogonal			
	$K=1$	$K=2$	$K=5$	$K=10$	$K=1$	$K=2$	$K=5$	$K=10$
DML								
Kernel	0.9785	0.9754	0.9766	0.9767	1.2568	1.3432	1.2956	1.2837
	(0.1558)	(0.1649)	(0.1608)	(0.1573)	(0.1553)	(0.1603)	(0.1564)	(0.1563)
	95.0%	92.5%	94.0%	95.0%	-	-	-	-
Decision tree	0.9176	0.8595	0.8761	0.8903	0.9382	0.9423	0.9366	0.9461
	(0.1557)	(0.1724)	(0.1591)	(0.1533)	(0.1542)	(0.1715)	(0.1610)	(0.1574)
	93.0%	85.5%	86.0%	90.5%	61.0%	70.0%	68.0%	66.0%
Random forest	0.9835	0.9656	0.9607	0.9607	0.9667	1.0094	1.0121	1.0080
	(0.1700)	(0.1583)	(0.1534)	(0.1550)	(0.1697)	(0.1728)	(0.1701)	(0.1707)
	91.5%	94.0%	94.5%	94.5%	59.5%	67.5%	67.5%	65.0%
XGBoost	0.9554	0.9455	0.9500	0.9557	0.6287	0.9444	0.9671	0.9727
	(0.1620)	(0.1606)	(0.1589)	(0.1533)	(0.1264)	(0.1804)	(0.1827)	(0.1767)
	93.0%	94.0%	95.0%	94.5%	3.0%	61.0%	61.5%	62.5%

$p=5$	Orthogonal				Non-orthogonal			
	$K=1$	$K=2$	$K=5$	$K=10$	$K=1$	$K=2$	$K=5$	$K=10$
DML								
Kernel	1.0072	1.0569	1.0523	1.0476	1.1795	1.6463	1.5542	1.5357
	(0.3865)	(0.2385)	(0.2242)	(0.2175)	(0.2612)	(0.2402)	(0.2430)	(0.2455)
	90.0%	84.5%	86.0%	85.5%	-	-	-	-
Decision tree	0.7261	0.6358	0.6727	0.6616	0.8118	0.9448	0.9720	0.9579
	(0.2321)	(0.2875)	(0.2404)	(0.2262)	(0.2497)	(0.2994)	(0.2589)	(0.2464)
	72.0%	63.0%	71.5%	69.5%	42.2%	76.5%	80.4%	83.4%
Random forest	0.6177	0.8516	0.8672	0.8795	0.9678	1.0803	1.0864	1.1037
	(0.2272)	(0.2200)	(0.1893)	(0.1854)	(0.1865)	(0.2249)	(0.1875)	(0.1908)
	21.0%	87.0%	92.5%	94.0%	37.5%	70.5%	75.0%	71.0%
XGBoost	0.7550	0.9514	0.9790	0.9802	0.6095	0.9435	0.9556	0.9477
	(0.6149)	(0.2331)	(0.1992)	(0.2137)	(0.1598)	(0.2367)	(0.1962)	(0.2066)
	32.0%	93.5%	95.0%	93.5%	3.0%	65.5%	72.0%	68.0%

$p=10$	Orthogonal				Non-orthogonal			
	$K=1$	$K=2$	$K=5$	$K=10$	$K=1$	$K=2$	$K=5$	$K=10$
DML								
Kernel	1.3096	1.3028	1.2844	1.2811	-	-	-	-
	(4.8680)	(0.3482)	(0.3433)	(0.3433)	-	-	-	-
	30.5%	76.5%	74.5%	76.5%	-	-	-	-
Decision tree	1.6620	0.6634	0.6873	0.7290	1.7857	1.4056	1.4292	1.4893
	(0.8605)	(0.4480)	(0.3794)	(0.3801)	(0.6852)	(0.5169)	(0.4491)	(0.4743)
	80.0%	81.5%	90.5%	91.5%	28.8%	69.8%	74.0%	67.2%
Random forest	1.9529	0.8884	0.9046	0.8926	0.8629	1.0828	1.0625	1.0537
	(0.5066)	(0.3691)	(0.3310)	(0.3215)	(0.3229)	(0.4008)	(0.3268)	(0.3073)
	56.0%	97.5%	98.5%	97.5%	39.5%	82.5%	87.5%	90.5%
XGBoost	0.5523	0.9113	0.8738	0.8687	0.5331	0.9745	0.9655	0.9679
	(0.7304)	(0.4064)	(0.3494)	(0.3506)	(0.2335)	(0.4335)	(0.3406)	(0.3190)
	12.5%	93.5%	94.5%	96.5%	2.5%	78.5%	86.0%	83.0%

Table A3 Estimation results of the DML estimators under orthogonal and their counterparts using non-orthogonal estimating equations, with different cross-fitting folds in Scenario 2 (across 200 experiment replications). In each cell, the three numbers stand for the average of the estimates, the standard deviation of the estimates, and the coverage of the associated empirical confidence intervals, all based on the 200 replications

$p=2$	Orthogonal				Non-orthogonal			
	$K=1$	$K=2$	$K=5$	$K=10$	$K=1$	$K=2$	$K=5$	$K=10$
DML								
Kernel	0.9656	0.9461	0.9537	0.9537	1.5643	1.7868	1.6914	1.6670
	(0.1561)	(0.1918)	(0.1749)	(0.1693)	(0.1593)	(0.1733)	(0.1655)	(0.1638)
	93.5%	91.0%	95.0%	95.0%	-	-	-	-
Decision tree	0.8495	0.8588	0.8875	0.8989	0.8652	0.9358	0.9429	0.9415
	(0.1677)	(0.2036)	(0.1725)	(0.1703)	(0.1658)	(0.1910)	(0.1761)	(0.1829)
	83.0%	86.0%	90.0%	92.0%	41.5%	68.5%	73.5%	69.5%
Random forest	0.9298	0.9643	0.9617	0.9669	0.9678	1.0401	1.0366	1.0326
	(0.1963)	(0.1562)	(0.1627)	(0.1616)	(0.1735)	(0.1918)	(0.1781)	(0.1780)
	90.0%	97.0%	94.5%	94.0%	59.5%	70.5%	73.0%	70.5%
XGBoost	0.9311	0.9316	0.9423	0.9484	0.6441	0.9416	0.9600	0.9715
	(0.1637)	(0.1829)	(0.1687)	(0.1636)	(0.1340)	(0.1869)	(0.1965)	(0.1885)
	93.5%	94.5%	94.0%	96.0%	4.5%	64.5%	66.5%	63.0%

$p=5$	Orthogonal				Non-orthogonal			
	$K=1$	$K=2$	$K=5$	$K=10$	$K=1$	$K=2$	$K=5$	$K=10$
DML								
Kernel	0.9672	0.8206	0.8625	0.8702	1.1603	2.5535	2.3857	2.3459
	(0.3946)	(0.3453)	(0.3277)	(0.3274)	(0.2694)	(0.2590)	(0.2663)	(0.2724)
	89.5%	89.0%	90.0%	90.0%	-	-	-	-
Decision tree	0.6290	0.7207	0.7340	0.7221	0.6774	0.9222	0.8943	0.9009
	(0.2415)	(0.3509)	(0.3056)	(0.3028)	(0.2409)	(0.3593)	(0.3094)	(0.3409)
	62.0%	82.0%	86.5%	80.0%	22.6%	68.3%	69.3%	65.3%
Random forest	0.8103	0.9944	0.9977	0.9978	1.0551	1.2950	1.2553	1.2557
	(0.2628)	(0.3330)	(0.2711)	(0.2778)	(0.2718)	(0.3234)	(0.2895)	(0.2794)
	70.0%	94.5%	96.5%	95.0%	34.0%	40.0%	43.5%	45.5%
XGBoost	1.1345	0.9554	0.9576	0.9656	0.7166	0.9673	0.9364	0.9347
	(1.0763)	(0.2893)	(0.2601)	(0.2867)	(0.1893)	(0.2864)	(0.2616)	(0.2884)
	87.0%	91.0%	94.0%	94.0%	20.5%	63.0%	64.5%	58.5%

$p=10$	Orthogonal				Non-orthogonal			
	$K=1$	$K=2$	$K=5$	$K=10$	$K=1$	$K=2$	$K=5$	$K=10$
DML								
Kernel	0.7569	0.3910	0.4700	0.4595	-	-	-	-
	(3.9324)	(0.6041)	(0.5586)	(0.5604)	-	-	-	-
	34.0%	76.5%	83.0%	81.5%	-	-	-	-
Decision tree	0.6824	0.5324	0.5243	0.5243	0.6285	0.8618	0.7852	0.8424
	(0.7768)	(0.5294)	(0.4760)	(0.4589)	(0.9641)	(0.5997)	(0.5265)	(0.5164)
	81.5%	75.3%	80.0%	78.2%	34.5%	68.6%	64.8%	69.6%
Random forest	0.9153	0.9643	0.9447	0.9457	1.0503	1.3866	1.3891	1.3605
	(0.4430)	(0.4426)	(0.3830)	(0.3869)	(0.3911)	(0.4779)	(0.4109)	(0.4034)
	93.5%	96.5%	98.5%	98.0%	41.0%	53.0%	47.5%	52.5%
XGBoost	1.6055	0.8317	0.8532	0.8818	0.7508	2.0195	1.7329	1.6984
	(2.5499)	(0.4180)	(0.3569)	(0.3761)	(0.2056)	(0.3390)	(0.3070)	(0.2988)
	73.5%	92.0%	94.0%	92.5%	22.5%	2.5%	5.0%	9.5%

Table A4 Estimation results for $\theta_0 = 0.8$ using various estimators under different scaling factors α on DGP2 of ACIC 2019. In each cell, the three numbers stand for the average of the estimates, the standard deviation of the estimates, and the coverage of the associated empirical confidence intervals, all based on 100 replications

	$\alpha = 1$	$\alpha = 4$	$\alpha = 16$	$\alpha = 64$
Linear regression	0.8320	0.9409	1.3766	3.1216
	(0.0319)	(0.0422)	(0.1287)	(0.5348)
	81.0%	4.0%	0.0%	0.0%
DML				
Decision tree	0.8194	0.7785	0.7761	0.7724
	(0.0344)	(0.0343)	(0.0337)	(0.0509)
	88.0%	90.0%	90.0%	93.0%
Random forest	0.8032	0.8049	0.8122	0.8287
	(0.0352)	(0.0351)	(0.0348)	(0.0377)
	94.0%	92.0%	92.0%	93.0%
XGBoost	0.8161	0.8173	0.8152	0.8103
	(0.0314)	(0.0324)	(0.0327)	(0.0332)
	93.0%	92.0%	92.0%	94.0%

Table A5 Estimation results for $\theta_0 = 0.8$ with standardized coefficient values (DGP2 of ACIC 2019)

	Point estimate	Standard deviation	Coverage rate
Linear regression	0.5554	0.1921	82.0%
DML			
Decision tree	0.7708	0.0374	93.0%
Random forest	0.7977	0.0331	96.0%
XGBoost	0.8033	0.0347	97.0%
AutoML	0.7938	0.0359	93.0%

Appendix A: Implementation Details

In this section, we discuss implementation aspects of DML and our numerical experiments. In Section A.1, we briefly survey implementations of DML in existing softwares. Then we discuss the details of our implementations in the numerical experiments, such as softwares, model training and hyperparameter tuning, in Section A.2.

A.1. DML Software Implementations

Researchers interested in applying Double Machine Learning (DML) in their studies can find implementations across several software packages in both R and Python. For instance, the R package `npcausal`, `ddml` and its Stata version (Ahrens et al. 2024a), the Python package `EconML`, `CausalML` and the package `DoubleML` (Bach et al. 2021, 2022) all have dedicated functions for DML estimation of a wide range of parameters. These packages support the use of cross-fitting and incorporate multiple machine learning (ML) techniques.

Table A6 Impact of orthogonal vs. non-orthogonal estimating equations and different cross-fitting folds on DML estimation results (DGP2 of ACIC 2019 with standardized coefficient values). The truth is $\theta_0 = 0.8$. In each cell, the three numbers stand for the average of the estimates, the standard deviation of the estimates, and the coverage of the associated empirical confidence intervals, all based on the 100 replications

	Orthogonal				Non-orthogonal			
	$K=1$	$K=2$	$K=5$	$K=10$	$K=1$	$K=2$	$K=5$	$K=10$
DML								
Decision tree	0.6219	0.7613	0.7705	0.7708	0.7924	0.7798	0.7961	0.7936
	(0.0268)	(0.0810)	(0.0399)	(0.0374)	(0.0331)	(0.0909)	(0.0434)	(0.0350)
	0.0%	88.0%	92.0%	93.0%	73.0%	78.0%	85.0%	88.0%
Random forest	0.6645	0.7947	0.7984	0.7977	0.7946	0.7775	0.7937	0.7929
	(0.0273)	(0.0639)	(0.0384)	(0.0331)	(0.0335)	(0.0624)	(0.0402)	(0.0348)
	0.0%	88.0%	94.0%	96.0%	71.0%	70.0%	76.0%	82.0%
XGBoost	0.6798	0.8169	0.8053	0.8033	0.7952	0.7913	0.7981	0.7973
	(0.0322)	(0.0518)	(0.0376)	(0.0347)	(0.0332)	(0.0527)	(0.0365)	(0.0342)
	2.0%	88.0%	93.0%	97.0%	73.0%	75.0%	86.0%	81.0%
AutoML	0.6555	0.7735	0.7946	0.7938	0.7944	0.7413	0.7945	0.7967
	(0.0343)	(0.0957)	(0.0395)	(0.0359)	(0.0334)	(0.0816)	(0.0400)	(0.0343)
	1.0%	85.0%	93.0%	93.0%	72.0%	68.0%	80.0%	83.0%

Table A7 Estimation results for $\theta_0 = 2.1$ with standardized coefficient values (DGP4 of ACIC 2019)

	Point estimate	Standard deviation	Coverage rate
Linear regression	1.8109	0.0552	0.0%
DML			
Decision tree	2.0712	0.0188	72.0%
Random forest	2.0939	0.0191	97.0%
XGBoost	2.0889	0.0187	94.0%
AutoML	2.0929	0.0205	96.0%

In particular, the `DoubleML` implements DML estimation for several models and effects. It includes implementations for the partially linear regression model (Section 2), the average treatment effect for discrete treatments under a general interactive regression (Section B.4 eq. (11)), the partially linear instrumental variables model (Section B.1), and the local average treatment effect (Section B.1 Remark A1). Additionally, it addresses Difference in Differences (DID) models for both repeated outcomes and cross-sections (Section B.2). Recent extensions of the `DoubleML` package accommodate further applications, such as quantile treatment effect estimation, heterogeneous treatment effect estimation, sensitivity analysis in the presence of omitted variables, and the learning of treatment decision rules. The `DoubleML` package builds on the `scikit-learn` library in Python and the `mlr3` ecosystem in R, allowing for the integration of a diverse array of ML algorithms in the estimation of nuisance functions. `ddml` is another package for DML estimation based on Chernozhukov et al. (2018). Compared to the `DoubleML` package, `ddml` focuses specifically on the

Table A8 Impact of orthogonal vs. non-orthogonal estimating equations and different cross-fitting folds on DML estimation results (DGP4 of ACIC 2019 with standardized coefficient values). The truth is $\theta_0 = 2.1$. In each cell, the three numbers stand for the average of the estimates, the standard deviation of the estimates, and the coverage of the associated empirical confidence intervals, all based on the 100 replications

	Orthogonal				Non-orthogonal			
	$K=1$	$K=2$	$K=5$	$K=10$	$K=1$	$K=2$	$K=5$	$K=10$
DML								
Decision tree	1.9086	2.0685	2.0712	2.0712	2.0968	2.0940	2.0967	2.0967
	(0.0157)	(0.0226)	(0.0195)	(0.0188)	(0.0170)	(0.0229)	(0.0185)	(0.0172)
	0.0%	68.0%	69.0%	72.0%	79.0%	79.0%	85.0%	85.0%
Random forest	1.9106	2.1062	2.0964	2.0939	2.0953	2.0898	2.0950	2.0955
	(0.0207)	(0.0242)	(0.0202)	(0.0191)	(0.0169)	(0.0209)	(0.0184)	(0.0173)
	0.0%	88.0%	94.0%	97.0%	78.0%	77.0%	86.0%	85.0%
XGBoost	1.9851	2.0933	2.0892	2.0889	2.0909	2.0858	2.0886	2.0893
	(0.0189)	(0.0215)	(0.0188)	(0.0187)	(0.0170)	(0.0197)	(0.0179)	(0.0172)
	0.0%	92.0%	93.0%	94.0%	79.0%	68.0%	81.0%	81.0%
AutoML	1.9209	2.0878	2.0896	2.0929	2.0928	2.0959	2.0949	2.0944
	(0.0250)	(0.0279)	(0.0212)	(0.0205)	(0.0169)	(0.0209)	(0.0181)	(0.0169)
	0.0%	90.0%	93.0%	96.0%	81.0%	79.0%	85.0%	88.0%

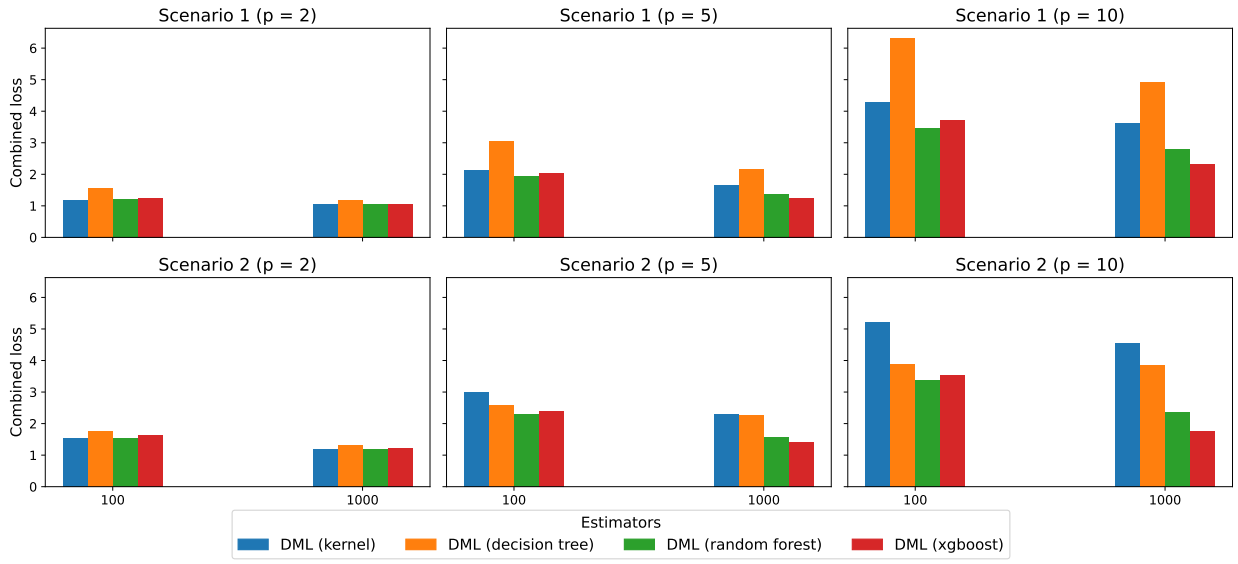


Figure A1 Combined loss of different nuisance estimation methods under orthogonal estimating equation in Scenario 1 and 2 (the top and bottom panels correspond to Scenario 1 and Scenario 2, respectively)

PLR model, the doubly robust estimation of ATE, the PLIV model, and the LATE. In the PLIV part, it also enables the approximation of optimal instruments using nonparametric estimation. In parallel, the `EconML` and `CausalML` package provides extended capabilities for estimating heterogeneous treatment effects

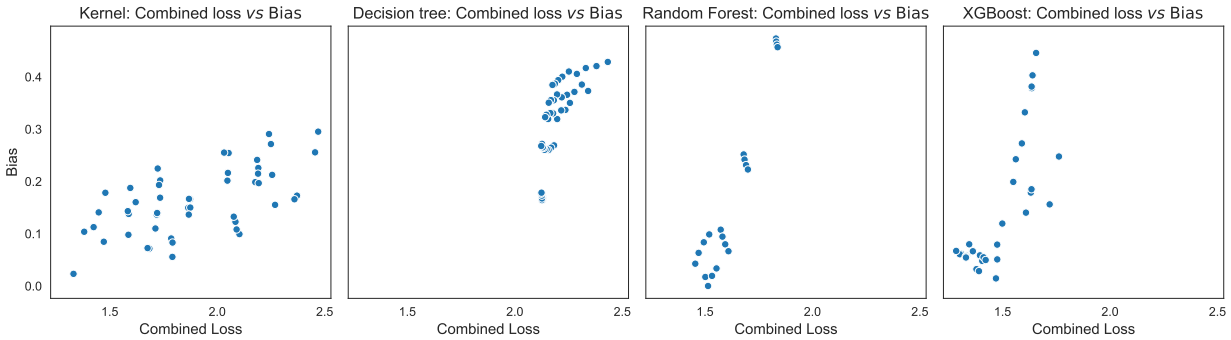


Figure A2 **Relationship between ML nuisance prediction quality (combined loss) and the DML estimation quality (bias) for scenario 1 with $N = 1000$ and $p = 5$.**

and policy learning. The `npcausal` package further extends DML by estimating the average effect curve for continuous treatment. All numerical experiments in this paper use the `DoubleML` package in R or Python.

A.2. Implementation Details for Experiments in Sections 2 and 3

Experiments in Section 2.3. The experiments in Section 2.3 were conducted in the R programming environment. We compare the estimation results of the DML estimator with three benchmarks: linear regression, Bias-Corrected PSM, and weighted regression. Both linear regression and weighted regression are implemented using the `lm` function, with the weighted regression requiring additional estimation of inverse propensity using the `glm` function. For Bias-Corrected PSM, we use the `glm` function for propensity score estimation, the `Matching` package for PSM with and without bias correction.

The DML estimator is implemented using the `DoubleML` package in R (Bach et al. 2021). For cross-fitting, we use the default number of folds $K = 5$. Regarding the estimation of nuisance functions, we use two machine learning techniques: random forest and XGBoost. Since the simulation setting in Section 2.3 is relatively simple, here we directly use the default hyperparameters provided by the functions for fitting random forest and XGBoost. The process of hyperparameter tuning will be discussed in more detail for experiments in Section 2.4.

Experiments in Section 2.4. The experiments in Section 2.4 were conducted in the Python programming environment and include three estimation methods: linear regression, kernel smoothing, and DML. Linear regression is directly implemented using the `LinearRegression` function from the `sklearn` package. Kernel smoothing is implemented using the `KernelReg` function from the `statsmodels` package, where the bandwidths for all kernel smoothing are determined using the standard Scott's Rule of Thumb (Scott 2015).

The DML estimator is implemented using the `DoubleML` package in Python (Bach et al. 2021). Our experiments consider a range of different number of folds in cross-fitting, including $K = 1$ (i.e., no cross-fitting) and $K = 2, 5, 10$. We also tried both orthogonal and non-orthogonal estimating equations. For hyperparameter tuning, we use the "full sample" cross-validation described in Bach et al. (2024) for experiments in this paper. Specifically, we first perform 5-fold cross-validation on the entire sample to select the best hyperparameter configuration for a nuisance ML estimator that minimizes MSE. After selecting the optimal hyperparameters, we implement DML using the same nuisance estimator with the chosen hyperparameter on the entire sample, where the cross-fitting splits are independent of the initial cross-validation splits. This sample splitting scheme is shown to outperform other common schemes in terms of estimation quality or computation speed (Bach et al. 2024).

Specifically, for decision trees, we tune the hyperparameters `max_depth` and `min_samples_split`, with the main selection ranges being $[4, 12]$ and $[2, 20]$, respectively. For random forests, we consider an additional hyperparameter, `n_estimators`, in addition to the decision tree hyperparameters, with the main selection range for `n_estimators` being $[20, 400]$. For xgboost, we similarly focus on the parameters `n_estimators` and `max_depth`, with selection ranges similar to those for random forests. Additionally, we include `learning_rate` in the tuning process, with the main selection range being $[0.001, 1]$.

Experiments in Section 2.5. The experiments on the ACIC 2019 datasets presented in Section 2.5 were conducted in the Python programming environment and included two estimation methods: linear regression and DML. Kernel-based methods were excluded, as they are not applicable in high-dimensional settings (i.e., $p = 200$). For nuisance function estimation in the DML procedure, in addition to the decision tree, random forest, and XGBoost models used in the simulation study, we also incorporated an AutoML-based approach to demonstrate that the nuisance functions m_0 and l_0 can be estimated using different types of ML models. Specifically, the AutoML method was implemented using the `AutoML` class from the `flaml` package, with a `time_budget` of 60 seconds and the default `estimator_list`, which includes a range of commonly used ML techniques such as random forest, XGBoost, LightGBM, and ExtraTrees.

Regarding the data, we adopt two datasets from ACIC 2019—corresponding to the 2nd and 4th data generating processes (i.e., DGP2 and DGP4). Both datasets exhibit two key issues, which motivate us to modify the original data accordingly. First, in both DGP2 and DGP4, the coefficients associated with the nonlinear components are typically 1 to 3 orders of magnitude smaller than those of the linear terms. As a

result, the data generating processes nearly degenerate into linear models, leading to unbiased point estimates even from simple linear regression. To verify the impact of the nonlinear term coefficients on estimation performance, we take DGP2 as an example and multiply the coefficients of the nonlinear terms by a scaling factor $\alpha \in \{1, 4, 16, 64\}$, and compare the estimation results of different methods across these values of α. After confirming the influence of these coefficients, we standardize all coefficient values across terms to be equal while keeping the functional form of the data generating process unchanged.

Second, compared to the high-dimensional covariate setting ($p = 200$), the sample sizes provided in DGP2 and DGP4 ($n = 1000$ or 2000) are insufficient to enable high-quality ML nuisance estimation, so the performance of DML cannot be guaranteed. To better facilitate high-quality ML nuisance estimation, we increase the sample sizes of DGP2 to 5 times its original size and DGP4 to 10 times its original size, respectively, without changing the underlying data generating mechanisms. Specifically, the data generating processes of DGP2 and DGP4 are presented in eq. (1) and eq. (2), respectively.

$$
\begin{aligned}
p &= \alpha_0 + \alpha_1\sqrt{X_1} + \alpha_2 X_5 + \alpha_3 X_{32} + \alpha_4 X_5 X_{32} + \alpha_5 X_{70} + \alpha_6 X_{70}^2 + \alpha_7 \mathbf{1}\left[X_{101} > 2.5\right] + \alpha_8 X_{150} \\
&\quad + \alpha_9 \mathbf{1}\left[X_{179} < -0.5\right](X_{179} + 1.5) \\
D &\sim \text{Bin}(1, 1/(1+\exp(-p))) \\
Y &= D\theta_0 + \beta_1\sqrt{X_1} + \beta_2 X_5 + \beta_3 X_{23} + \beta_4 X_{32} + \beta_5 X_5 X_{32} + \beta_6 X_{70} + \beta_7 X_{70}^2 + \beta_8 \mathbf{1}\left[X_{101} > 2.5\right] \\
&\quad + \beta_9 X_{150} + \beta_{10}\mathbf{1}\left[X_{179} < -0.5\right](X_{179} + 1.5) + \beta_{11} X_{199} + \varepsilon \\
\varepsilon &\sim \mathcal{N}(0, 1)
\end{aligned}
\tag{1}
$$

$$
\begin{aligned}
p &= \alpha_0 + \alpha_1 X_2 + \alpha_2 X_5 + \alpha_3 X_2 X_5 + \alpha_4 X_{12} + \alpha_5 X_{23} + \alpha_6 X_{23}^2 + \alpha_7 X_{12} * X_{23}^2 + \alpha_8 \sqrt{X_{67}} \\
&\quad + \alpha_9 X_{77} + \alpha_{10} \mathbf{1}\left[X_{89} > 19\right] + \alpha_{11} \mathbf{1}\left[X_{95} > 5\right](X_{95} - 3) + \alpha_{12} \exp\left(X_{106}\right) + \alpha_{13} X_{122} \\
&\quad + \alpha_{14} X_{146} + \alpha_{15} X_{122} X_{146} + \alpha_{16} X_{150} + \alpha_{17} X_{168} + \alpha_{18} \mathbf{1}\left[X_{199} > 1\right] \\
D &\sim \text{Bin}(1, 1/(1 + \exp(-p))) \\
Y &= D\theta_0 + \beta_1 + \beta_2 X_2 + \beta_3 X_5 + \beta_4 X_2 X_5 + \beta_5 X_{12} + \beta_6 X_{23} + \beta_7 X_{23}^2 + \beta_8 X_{12} X_{23}^2 + \beta_9 X_{40} \\
&\quad + \beta_{10} \sqrt{X_{67}} + \beta_{11} X_{77} + \beta_{12} \mathbf{1}\left[X_{89} > 19\right] + \beta_{13} \mathbf{1}\left[X_{95} > 5\right](X_{95} - 3) + \beta_{14} \exp\left(X_{106}\right) \\
&\quad + \beta_{15} X_{122} + \beta_{16} X_{133} + \beta_{17} X_{146} + \beta_{18} X_{122} X_{146} + \beta_{19} X_{150} + \beta_{20} X_{168} + \beta_{21} X_{198} \\
&\quad + \beta_{22} \mathbf{1}\left[X_{199} > 1\right] + \varepsilon \\
\varepsilon &\sim t(5)
\end{aligned} \tag{2}
$$

Experiments in Section 3. The Facebook experiments in Section 3 were also conducted in the Python programming environment and include three estimation methods: linear regression, kernel smoothing, and DML. The implementation details for these three methods are similar to those in Section 2.4. The main difference is that we additionally take into account the additional randomness introduced by cross-fitting. Specifically, we repeat the cross-fitting process and DML estimation for 100 times and use the medain of the 100 estimates as the final estimate. The standard error estimation and confidence interval construction also incorporate the variations of the 100 estimates (see Section 3.4 in Chernozhukov et al. (2018)). This can be also implemented by the `DoubleML` package. Fuhr et al. (2024) advocate this approach and argue that "in smaller samples, using a larger number of repetitions can increase the robustness of the estimates considerably".

Appendix B: DML beyond Partially Linear Regression

In this section, we present extensions of DML to other empirical settings. Specifically, we discuss instrumental variable (IV) regression in Section B.1 and Difference-in-Differences (DID) regression in Section B.2. These two serve as examples of common empirical settings widely used in IS empirical research. In Section B.3, we propose a new DML approach to correct the bias in linear regression with ML-generated covariates. This part illustrates the potential of extending DML to new settings relevant for IS empirical research. Finally, we provide a comprehensive survey of DML extensions to other empirical settings in Section B.4.

B.1. DML for IV Regression

IV regression is one of the most popular approaches to address endogeneity in empirical research. When the treatment variable D is endogenous (e.g., due to unobserved confounders), researchers can utilize the two-stage least squares (2SLS) method to estimate the treatment effect of the treatment variable D on the outcome variable Y, given an instrument Z and other control covariates $\boldsymbol{X}$ (e.g., Imbens 2014). Finding IVs that are *relevant* (i.e., correlated with D) and *excluded* (i.e., uncorrelated with the source of endogeneity) can be a challenge in practice. However, even with "perfect" IVs, the widely used 2SLS method still relies on correct model specification. In particular, 2SLS uses a linear IV model that assumes a linear relationship between Y, D and $\boldsymbol{X}$. If this assumption is violated, then 2SLS might lead to highly biased estimates.

To enhance model flexibility and accommodate potential nonlinearities, the partially linear IV (PLIV) model serves as an alternative (Chernozhukov et al. 2018):

$$\begin{aligned} Y &= D\theta_0 + g_0(\boldsymbol{X}) + \varepsilon, \quad E[\varepsilon \mid \boldsymbol{X}, Z] = 0, \\ Z &= r_0(\boldsymbol{X}) + \nu, \quad E[\nu \mid \boldsymbol{X}] = 0. \end{aligned} \tag{3}$$

Similar to our discussions for the PLR model in the previous section, the nonparametric nuisance function g_0 in the PLIV model allows for a general nonlinear relationship between Y and $\boldsymbol{X}$, while the coefficient θ_0 in the parametric component is of the primary interest. The parameter θ_0 can still be interpreted as the average treatment effect of D on Y under standard IV assumptions.

DML Estimation. To apply the general DML method to the PLIV model, Chernozhukov et al. (2018) consider the following Neyman orthogonal estimating equation:

$$\begin{aligned} &E\left[\psi_{\mathrm{PLIV}}\left(\boldsymbol{W}; \theta_0, l_0(\boldsymbol{X}), m_0(\boldsymbol{X}), r_0(\boldsymbol{X})\right)\right] = 0, \\ &\text{where } \psi_{\mathrm{PLIV}}\left(\boldsymbol{W}; \theta, l(\boldsymbol{X}), m(\boldsymbol{X}), r(\boldsymbol{X})\right) = [Y - l(\boldsymbol{X}) - \theta(D - m(\boldsymbol{X}))](Z - r(\boldsymbol{X})), \\ &\text{and } l_0(\boldsymbol{X}) = E(Y \mid \boldsymbol{X}),\ m_0(\boldsymbol{X}) = E(D \mid \boldsymbol{X}),\ r_0(\boldsymbol{X}) = E(Z \mid \boldsymbol{X}). \end{aligned} \tag{4}$$

This estimating equation satisfies the Neyman Orthogonality condition (see Appendix F.2 for a rigorous technical exposition). Then we can use this estimating equation, together with the cross-fitting strategy, to construct a DML estimator for the coefficient θ_0 in the PLIV model. The corresponding procedure is readily available by replacing the generic estimating equation in algorithm 1 with ψ_{PLIV} and the ML-estimated nuisances $\eta_0 = (l_0, m_0, r_0)$. Specifically, the nuisances (l_0, m_0, r_0) can be estimated by regressing Y, D, Z on the covariates $\boldsymbol{X}$ using a wide array of regression or classification ML algorithms. Moreover, in Appendix

Table A9 Estimation results of treatment effect in PLIV model (the ground-truth coefficient value is 1)

	$N=100$	$N=500$	$N=1000$	$N=5000$	$N=10000$
2SLS	-4.2841	-3.8743	-3.7835	-3.6391	-3.6171
	(4.6819)	(2.2113)	(1.5191)	(0.6485)	(0.4549)
DML(Random forest)	**1.2784**	**1.0356**	**0.9739**	**0.9574**	**0.9489**
	(2.5703)	**(0.5247)**	**(0.2741)**	**(0.0972)**	**(0.0625)**
DML(XGBoost)	**1.2322**	**1.1174**	**1.1123**	**0.9870**	**0.9687**
	(5.7904)	**(0.5392)**	**(0.4153)**	**(0.1190)**	**(0.0800)**

F.2 Remark A3, we show that the DML estimator can be interpreted as a nonlinear version of 2SLS: first run ML regressions of the treatment D on instrument Z and covariates $\boldsymbol{X}$ to obtain the fitted value $\hat{D}$, and then run a partially linear regression of Y with respect to $\hat{D}$ and $\boldsymbol{X}$. The resulting DML estimator under the PLIV model also has strong theoretical guarantees such as $\sqrt{N}$-consistency and asymptotic normality, under high-level error rate conditions on the ML nuisance estimators (Appendix G.2). These guarantees provide rigorous underpinnings for the downstream confidence interval or hypothesis testing analyses.

Empirical Study. Next, we compare the performance of the DML estimator with 2SLS estimators through simulation studies on synthetic data. To introduce endogeneity, we perturb the true binary treatment D^* to create a misclassified version D, so measurement error serves as the source of endogeneity. We also simulate another imperfect measurement Z of D^* with another misclassification matrix as a valid IV. We generate data according to the following model with true coefficient $\theta_0 = 1$:

$$X_1 \sim N(0,2^2),\ X_2 \sim N(0,2^2),\ P(D^*=1|X_1,X_2) = \frac{1}{1+e^{-(X_1+X_2+1)}},$$

$$P(D=0 \mid D^*=1) = P(D=1 \mid D^*=0) = 0.1, \quad P(Z=0 \mid D^*=1) = P(Z=1 \mid D^*=0) = 0.1,$$

$$Y = D^* + X_1 + X_2 + X_1X_2 + X_1^2 + X_2^2 + 1 + \varepsilon,\ \varepsilon \sim N\left(0,1^2\right).$$

The configuration involving treatment D^*, covariates $\{X_1, X_2\}$, and outcome Y parallels the setup in the PLR model in Section 2.3. Observations are available for the mismeasured treatment D and the instrument Z, but not for the true treatment D^*. Same as before, we consider sample sizes $N \in \{100, 500, 1000, 5000, 10000\}$. The DML estimator employs random forest and XGBoost algorithms to estimate the nuisance functions. We compute the mean and standard deviation of the parameter estimates over 100 replications and present the results in Table A9.

We observe that the 2SLS estimator, due to the misspecification of the linear IV model, incurs significant bias and often yields estimates with signs contrary to the true coefficient. In contrast, the DML estimator, irrespective of the machine learning technique employed for nuisance function estimation, consistently

achieves lower bias and variance. Notably, the bias and variance decrease as the sample size increases, corroborating the estimator's reliability. This offers strong empirical evidence for the advantages of the DML estimator under the PLIV model, in terms of robustness and estimation accuracy.

Remark A1 (Applications of DML in Other IV models). The PLIV model posits a constant treatment effect and rules out interactions between the treatment and covariates. A more general model could be $Y = \mu_0(Z, \boldsymbol{X}) + \varepsilon$ where $E[\varepsilon \mid \boldsymbol{X}, Z] = 0$. Under this model, the local average treatment effect (LATE) for a binary treatment D, i.e., the average treatment effect for the subpopulation of compliers, can be identified by a binary instrument Z under standard IV assumptions (Imbens 2014). DML estimators tailored for LATE estimation have been provided in Chernozhukov et al. (2018). Although the complier subpopulation is latent and unobserved, certain aggregate properties (e.g., expected value of certain covariates) can be still evaluated from data. Such tasks can also be handled by DML methods (Singh and Sun 2024). Furthermore, recent research has introduced DML approaches for IV estimation of quantile treatment effects (Kallus et al. 2024) and fully nonparametric models that accommodate continuous treatments (Bennett et al. 2023a,b).

B.2. DML for DID Regression

The difference-in-differences (DID) model is another widely used method in empirical research to estimate treatment effects in observational studies (Card and Krueger 1993). In this section, we outline the classic two-group, two-period DID framework (Angrist and Pischke 2009). This setup involves a pre-treatment period ($t = 0$) and a post-treatment period ($t = 1$), during which a subset of units receives treatment (e.g., a policy intervention) while the remainder serves as a control group. Each unit i has a binary treatment indicator D_i and covariates $\boldsymbol{X}_i$. We use variables $Y_{i,t}(1), Y_{i,t}(0)$ to denote the potential outcomes that would occur for a unit i at time t if this unit were treated and were not treated respectively. The standard causal effect parameter in DID analyses is *the average treatment effect on the treated* (ATT) parameter defined as $\theta_0 = E\left[Y_{i,1}(1) - Y_{i,1}(0) \mid D_i = 1\right]$. It measures the treatment's average effect on outcomes at $t = 1$ for the treated units that actually experience the treatment. A core challenge in causal inference, known as the "fundamental problem," is the inability to observe both potential outcomes for any unit simultaneously (Holland 1986, Imbens and Rubin 2015). Researchers can only measure the actual outcome $Y_{i,1} = Y_{i,1}(D_i)$, necessitating assumptions to estimate treatment effects.

DID analysis is particularly relevant when treatment assignment is not random, and the treated and control groups may differ due to unobserved confounders. The DID approach leverages pre-treatment outcomes to

account for these differences, relying on the *parallel trends* assumption, which posits that, in the absence of treatment, the outcome trajectories of the treated and control groups would have been the same (Angrist and Pischke 2009). The pre-treatment outcome for every unit i is denoted as $Y_{i,0} = Y_{i,0}(0)$, since no units experience the treatment yet in period $t = 0$.

DID analysis is typically applied in two scenarios: panel data and repeated cross-section data (Wooldridge 2010). Panel data allows for tracking the same units over time, where the observed data is $\{\boldsymbol{W}_i\}_{i=1}^N = \{Y_{i,0}, Y_{i,1}, D_i, \boldsymbol{X}_i\}_{i=1}^N$. The repeated cross-section data involves different units at each time point, so we observe the outcome for each unit i only in a single period denoted by $T_i \in \{0, 1\}$. The corresponding data is $\{\boldsymbol{W}_i\}_{i=1}^N = \{Y_i, T_i, D_i, \boldsymbol{X}_i\}_{i=1}^N$, with $Y_i = T_i Y_{i,1} + (1 - T_i) Y_{i,0}$. For example, consider a social media platform that introduces a new feature to some users (as the intervention of interest). This corresponds to a panel data setting at the user level because each user can have multiple observations over time, but can amount to a repeated cross-section setting at the user-content level if each piece of content is only observed during one period. Here, we focus on the repeated cross-section data, and defer discussions on panel data to Appendices F.3 and G.3.

In the repeated cross-section context, the standard DID model is a two-way fixed effect regression:

$$Y_i = \mu + \theta_0 T_i D_i + \tau D_i + \delta T_i + \boldsymbol{X}_i^\top \pi + \varepsilon_i, \quad E\left[\varepsilon_i \mid \boldsymbol{X}_i, T_i, D_i\right] = 0, \tag{5}$$

where the coefficient θ_0 captures the treatment effect of interest, and the terms τD_i and δT_i capture the group and time fixed effects, respectively. The coefficients can be easily estimated by running a standard OLS regression. The two-way fixed effect model imposes the (conditional) parallel trends assumption: after controlling for the covariates, the temporal changes in the outcomes in the absence of treatment are identical between the treatment and control groups, or more formally,

$$E\left[Y_{i,1}(0) - Y_{i,0}(0) \mid \boldsymbol{X}_i, D_i = 1\right] = E\left[Y_{i,1}(0) - Y_{i,0}(0) \mid \boldsymbol{X}_i, D_i = 0\right]. \tag{6}$$

However, the two-way fixed effect model does not only require the parallel-trend assumption, but also impose strong functional form restrictions. For example, it posits that the control covariates $\boldsymbol{X}_i$ relate linearly to the outcome Y_i. It also imposes a homogeneous treatment effect, i.e., the treatment effect for every unit i is a constant θ_0. While researchers can include additional terms to capture nonlinearity and heterogeneity, selecting the correct model specification remains a challenge. Again, DML can mitigate this challenge by incorporating ML techniques for more flexible modeling.

DML Estimation. The DML approaches to DID analyses are based on the semiparametric DID model in Abadie (2005). This model only requires the parallel trends assumption in Equation (6), without imposing any restrictions on the functional form of the relationships among the outcomes, the treatment variable, and the control covariates. Thus it is much more robust to model misspecification. While Abadie (2005) originally uses classic nonparametric regressions in the semiparametric inference of the ATT parameter, recent advancements advocate for the use of DML.

In this part, we discuss the DMLDID ATT estimator proposed in Chang (2020) for the repeated cross-section (RCS) setting. This estimator is constructed from the following estimating equation:

$$E\left[\psi_{\mathrm{RCS}}\left(\boldsymbol{W};\theta_0,p_0,\lambda_0,h_0(\boldsymbol{X}),m_0(\boldsymbol{X})\right)\right]=0,$$
$$\text{where } \psi_{\mathrm{RCS}}(\boldsymbol{W};\theta,p,\lambda,h(\boldsymbol{X}),m(\boldsymbol{X}))=\frac{(T-\lambda)Y-h(\boldsymbol{X})}{\lambda\,(1-\lambda)\,p}\frac{D-m(\boldsymbol{X})}{1-m(\boldsymbol{X})}-\theta, \tag{7}$$
$$\text{and } p_0=P\left(D=1\right),\ \lambda_0=P\left(T=1\right),\ m_0(\boldsymbol{X})=E\left[D\mid\boldsymbol{X}\right],\ h_0=E\left[(T-\lambda_0)Y\mid\boldsymbol{X},D=0\right].$$

Section F.3 verifies that this estimating equation satisfies the Neyman Orthogonality condition with respect to the nuisances p_0,λ_0,m_0,h_0. Notably, p_0,λ_0 can be directly estimated by the proportion of the treated group's observations and the proportion of the post-treatment period observations respectively. The nuisance function m_0 can be estimated by a binary regression of D on $\boldsymbol{X}$, and h_0 can be estimated by a regression of $(T-\hat{\lambda})Y$ on $\boldsymbol{X}$ within only the control group corresponding to $D=0$, where $\hat{\lambda}$ is a sample estimate of λ_0. The estimation of the latter two functions can be easily implemented by ML algorithms. We can then construct the DMLDID estimator via Algorithm 1, using the estimating equation ψ_{RCS} with cross-fitted ML nuisance estimators. The resulting DMLDID estimator is again $\sqrt{N}$-consistent and asymptotically normal under fairly general conditions, which enables rigorous statistical inference on the ATT parameter θ_0. See Section G.3 for the theoretical justification.

Simulation Study. To evaluate the performance of the DMLDID estimator, we employ synthetic data and contrast it with the traditional two-way fixed effect regression. Specifically, we simulate covariates $\boldsymbol{X}_i=(X_{i,1},X_{i,2})$ and unobserved confounders U_i, each independently drawn from a standard normal distribution for unit i. We also generate the binary treatment assignment $D_i\in\{0,1\}$, the observed time period $T_i\in\{0,1\}$, and potential outcomes $Y_{i,t}(d)$ for $t,d\in\{0,1\}$ as follows:

$$\Pr\left(D_i=1\mid\boldsymbol{X}_i\right)=\frac{1}{1+e^{U_i+X_{i,1}+X_{i,2}+X_{i,1}X_{i,2}}},\quad P(T_i=1)=0.5,$$
$$Y_{i,0}(0)=Y_{i,0}(1)=f(\boldsymbol{X}_i)+\nu(\boldsymbol{X}_i,U_i)+\varepsilon_{i,0},\ Y_{i,1}(d)=d+4f(\boldsymbol{X}_i)+\nu(\boldsymbol{X}_i,U_i)+\varepsilon_{i,1}(d),\ d=0,1,$$

Table A10 Estimation results of treatment effect in DID model (the ground-truth of ATT is 1)

	$N = 500$	$N = 1000$	$N = 5000$	$N = 10000$
Two-way fixed effect regression	2.0864	2.0903	2.0920	2.0924
	(0.4215)	(0.2980)	(0.1357)	(0.0950)
Semiparametric DID (Kernel)	2.1223	2.0195	1.7904	1.7006
	(0.5698)	(0.4305)	(0.2292)	(0.1723)
Semiparametric DID (Spline)	1.9731	1.9627	1.9470	1.9432
	(0.7277)	(0.4536)	(0.1848)	(0.1266)
DMLDID(Random forest)	**1.1948**	**1.1494**	**1.1176**	**1.1123**
	(1.0086)	**(0.7552)**	**(0.3571)**	**(0.2569)**
DMLDID(XGBoost)	**1.1977**	**1.1110**	**1.0744**	**1.0658**
	(1.0290)	**(0.7895)**	**(0.4052)**	**(0.2942)**

where $f(\boldsymbol{X}_i) = (X_{i,1} + X_{i,2} + X_{i,1}^2 + X_{i,2}^2 + X_{i,1}X_{i,2})/4$, $\nu(\boldsymbol{X}_i, U_i)$ is a normal random variable with mean $U_i f(\boldsymbol{X}_i)$ and variance one, and $\varepsilon_{i,0}, \varepsilon_{i,1}(0), \varepsilon_{i,1}(1)$ are independent standard normal random variables. The unobserved confounders U_i affect treatment assignment and the outcomes but remain time-invariant, thereby justifying the use of the DID model. The true ATT is thus $\theta_0 = E[Y_{i,1}(1) - Y_{i,1}(0) \mid D_i = 1] = 1 + E[\varepsilon_{i,1}(1) - \varepsilon_{i,1}(0) \mid D_i = 1] = 1$.

We assess the performance of the traditional two-way fixed effect regression estimator, semiparametric DID estimator (Abadie 2005) and the DMLDID estimator (with nuisance functions fitted by random forests and XGBoost respectively) across sample sizes $N \in \{500, 1000, 5000, 10000\}$. We repeat the simulations for 10000 times, and report the mean and the standard deviation of the estimates from different methods in Table A10. We find that the two-way fixed effect regression, relying on a misspecified linear model, consistently exhibits significant bias for all sample sizes, falsely indicating a positive ATT. Semiparametric DID estimators, which use a non-orthogonal estimating equation and traditional nonparametric methods (e.g., kernel, spline), also suffer from severe bias. In contrast, the DMLDID estimators, leveraging flexible ML techniques, display substantially reduced bias, which further diminishes as the sample size increases. This demonstrates the advantages of the DMLDID estimator in terms of estimation accuracy and robustness.

Remark A2 (Applications of DML in DID and Panel Data Analyses). Besides the DMLDID estimator of Sant'Anna and Zhao (2020), Zimmert (2020) have also developed similar DML estimators for the canonical two-group and two-period semiparametric DID model in Abadie (2005). The literature continues to evolve, with studies extending DID to multiple periods, groups, continuous treatment and heterogeneous effects (see reviews by De Chaisemartin and d'Haultfoeuille 2023, Roth et al. 2023). In particular, Callaway

and Sant'Anna (2021) introduce an estimator for group-time average effects that incorporates multiple periods and groups with heterogeneous effects, which can also be implemented using DML. Moreover, Zhang (2024) extends DID to settings involving continuous treatments, and Caetano et al. (2022), Caetano and Callaway (2024), Haddad et al. (2024) further develop DML estimators for time-varying treatments and covariates. Beyond DID analyses, some recent literature also study the applications of DML to panel data regressions with unobserved heterogeneity characterized by fixed effects. We refer to Clarke and Polselli (2023), Fuhr and Papies (2024) for a varity of different proposals.

There are also studies that incorporate elements resembling those in DML, yet differ fundamentally from it. For instance, Arkhangelsky et al. (2021), Ben-Michael et al. (2021), Arkhangelsky and Imbens (2022) investigate causal inference in panel data settings, representing recent advances in panel data analyses developed concurrently with DML. Specifically, Arkhangelsky et al. (2021), Ben-Michael et al. (2021) consider the interactive fixed effects model, which accommodates time-varying confounding and generally violates the parallel trends assumption. Arkhangelsky and Imbens (2022) introduce a distinct identification strategy that relies on an additional assumption not necessarily satisfied in DID settings. Hence, the methods they employ differ substantially from DMLDID. Moreover, although these studies and DML share certain "robustness" properties, the underlying notions of "robustness" are fundamentally distinct.

B.3. Correcting Estimation Bias from ML-Generated Covariates

So far, we have discussed how the DML method can enhance estimation robustness to model misspecification for three common empirical models, namely linear regression, IV regression, and DID regression. In all three cases, handling control covariates in a nonparametric manner and applying the DML apparatus (including Neyman Orthogonal estimating equation and cross-fitting) yields more precise and robust estimates than traditional linear models. This section shifts attention to the distinct challenge of statistical inference with ML-generated covariates, illustrating how DML can enable bias correction. Unlike previous sections that primarily review established methodologies, here we offer a novel conceptualization of the issue and develop a new DML estimator.

In many empirical settings, some covariates of interest are not directly observable and must be "mined" from data due to their latent nature (e.g., deriving text sentiment from raw texts or measuring image quality based on raw images in Goh et al. 2013, Zhang et al. 2022) or legal constraints on sensitive information

collection (e.g., Imai and Khanna 2016). A common strategy involves manually labeling a small dataset, utilizing ML models to predict the unobserved covariates, and running regressions based on these predictions. However, prediction errors in ML-generated covariates can result in biased and inconsistent estimations of the target parameter and compromised statistical inference. While Yang et al. (2018), Qiao and Huang (2021) treat ML prediction errors as conventional measurement errors and apply corresponding corrections, ML prediction errors may not align perfectly with the independent, zero-mean, constant-variance errors typically assumed in measurement error literature. Fong and Tyler (2021) consider ML predictions as instrumental variables for the missing covariates, imposing additional exclusion restrictions. Yang et al. (2022) heuristically choose certain decision trees within random forests as instrumental variables for the predictions of other trees, yet they do not provide rigorous theoretical support for statistical inference tasks such as confidence interval construction or hypothesis testing.

We formulate the issue of ML-generated covariates as a missing data problem, where some variables are only observed within a limited subset (the manually labeled data) and are absent elsewhere. Here, the ML models serve as an "imputation" technique to fill the missing values, with the downstream regression's estimation bias originating from imputation errors. In this section, we use the linear regression model as an exemple to formalize the missing data approach and present a DML estimator that offers robust statistical assurances despite imperfect ML imputations. Our DML methodology is more principled and theoretically sound than existing approaches.

The linear regression model with missing covariates can be formulated as:

$$Y = \boldsymbol{X}^\top \boldsymbol{\theta}_0 + \boldsymbol{Z}^\top \boldsymbol{\gamma}_0 + \varepsilon,\ E(\varepsilon \mid \boldsymbol{X}, \boldsymbol{Z}) = 0, \tag{8}$$

where Y denotes the dependent variable, $\boldsymbol{X}$ represents independent variables with potential missingness,[1] $\boldsymbol{Z}$ stands for fully observed independent control variables. In particular, $\boldsymbol{X}$ is observed only within the labeled dataset—used for ML model training and validation (denoted by $R=1$)—and is missing from the unlabeled dataset (denoted by $R=0$). Variables Y and $\boldsymbol{Z}$ are always observed. The primary interest lies in estimating the coefficients $\boldsymbol{\theta}_0$.

[1] Our proposed DML method accommodates both a single partially missing variable or multiple such variables. Here we represent $\boldsymbol{X}$ as a vector of multiple variables for the sake of generality.

If variables $\boldsymbol{X}$ were completely observed, then an ordinary least squares (OLS) regression could be implemented, which solves a sample approximation of the following least squares problem:

$$(\boldsymbol{\theta}_0, \boldsymbol{\gamma}_0) = \arg\min_{\theta,\gamma} E(Y - \boldsymbol{X}^\top \boldsymbol{\theta} - \boldsymbol{Z}^\top \boldsymbol{\gamma})^2.$$

The first-order condition of this optimization yields the OLS estimating equation:

$$E\left[\psi_{\text{OLS}}(\boldsymbol{W}; \boldsymbol{\theta}_0, \boldsymbol{\gamma}_0)\right] = 0, \text{ where } \psi_{\text{OLS}}(\boldsymbol{W}; \boldsymbol{\theta}, \boldsymbol{\gamma}) = \left(Y - \boldsymbol{X}^\top \boldsymbol{\theta} - \boldsymbol{Z}^\top \boldsymbol{\gamma}\right) \begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{Z} \end{pmatrix}. \tag{9}$$

When $\boldsymbol{X}$ is not fully observable, however, averaging the estimating equation over the entire dataset, as implied by eq. (9), is infeasible. Instead, we need to analyze the mechanism behind $\boldsymbol{X}$'s missingness. Prior studies often implicitly or explicitly assume the *missing completely at random* (MCAR) mechanism, where the missingness of $\boldsymbol{X}$ is entirely independent of any observed or unobserved data, and justify this assumption by the fact that labeled and unlabeled data are typically random samples from the same population. In comparison, we adopt a weaker *missing at random* (MAR) assumption, which is common in the missing data literature (Little and Rubin 2019): $R \perp \boldsymbol{X} \mid Y, \boldsymbol{Z}, \boldsymbol{F}$ and $P(R = 1 | Y, \boldsymbol{Z}, \boldsymbol{F}) > 0$. where $\boldsymbol{F}$ represents features used to predict $\boldsymbol{X}$, which are fully observed (features $\boldsymbol{F}$ may or may not overlap with control variables $\boldsymbol{Z}$, depending on specific feature engineering / selection decisions). The MAR assumption posits that the probability of $\boldsymbol{X}$ being missing is conditionally independent of its values given the observed data Y, $\boldsymbol{Z}$ and $\boldsymbol{F}$, and that there is always a non-zero probability of observing $\boldsymbol{X}$.[2] Consequently, the conditional distribution of $\boldsymbol{X}$ given Y, $\boldsymbol{Z}$, $\boldsymbol{F}$ and $R = 0$ is identical to its conditional distribution given Y, $\boldsymbol{Z}$, $\boldsymbol{F}$ and $R = 1$, allowing for the inference of the former from the latter. The MAR assumption is robust unless there are fundamental disparities between the generation or selection of labeled vs. unlabeled datasets.

We may directly fit ML models to impute the values of the missing variables $\boldsymbol{X}$ from other observed features (e.g., variables $\boldsymbol{Z}$, $\boldsymbol{F}$). Then we can run an OLS regression of Y against the imputed values of $\boldsymbol{X}$ and variables $\boldsymbol{Z}$. This amounts to solving the following estimating equation:

$$\begin{aligned} &E\left[\psi_{\text{impute}}(\boldsymbol{W}; \boldsymbol{\theta}_0, \boldsymbol{\gamma}_0, q_0(\boldsymbol{Z}, \boldsymbol{F}))\right] = 0, \text{ where } q_0(\boldsymbol{Z}, \boldsymbol{F}) = E[\boldsymbol{X} \mid \boldsymbol{Z}, \boldsymbol{F}, R = 1], \\ &\text{and } \psi_{\text{impute}}(\boldsymbol{W}; \boldsymbol{\theta}, \boldsymbol{\gamma}, q(\boldsymbol{Z}, \boldsymbol{F})) = \left(Y - q(\boldsymbol{Z}, \boldsymbol{F})^\top \boldsymbol{\theta} - \boldsymbol{Z}^\top \boldsymbol{\gamma}\right) \begin{pmatrix} q(\boldsymbol{Z}, \boldsymbol{F}) \\ \boldsymbol{Z} \end{pmatrix}. \end{aligned} \tag{10}$$

[2] This positive probability condition is needed so that the distribution $\boldsymbol{X} \mid Y, \boldsymbol{Z}, \boldsymbol{F}, R = 1$ would be well-defined.

However, this estimating equation does not satisfy the Neyman Orthogonality condition, leading to estimates that are highly sensitive to biases in the ML imputation of $q_0(\boldsymbol{Z},\boldsymbol{F})$. Consequently, OLS estimators derived from imputed $\boldsymbol{X}$ values are typically biased (Yang et al. 2018).

DML Estimation. While heuristic approaches have been suggested to mitigate ML-induced bias, we advocate for a principled DML approach to this challenge. Drawing from the semiparametric efficiency theory for missing data (Tsiatis 2006, Robins and Rotnitzky 1995), we consider a *doubly robust* (DR) estimating function:

$$\psi_{\mathrm{DR}}(\boldsymbol{W};\boldsymbol{\theta}_0,\boldsymbol{\gamma}_0,\eta_0(Y,\boldsymbol{Z},\boldsymbol{F})) = \frac{R}{\rho_0(Y,\boldsymbol{Z},\boldsymbol{F})}\psi_{\mathrm{OLS}}(\boldsymbol{W};\boldsymbol{\theta}_0,\boldsymbol{\gamma}_0) - \frac{R-\rho_0(Y,\boldsymbol{Z},\boldsymbol{F})}{\rho_0(Y,\boldsymbol{Z},\boldsymbol{F})} E\left[\psi_{\mathrm{OLS}}(\boldsymbol{W};\boldsymbol{\theta}_0,\boldsymbol{\gamma}_0) \mid Y,\boldsymbol{Z},\boldsymbol{F},R=1\right]$$

$$= \begin{pmatrix} \left(\frac{R}{\rho_0(Y,\boldsymbol{Z},\boldsymbol{F})}X - \frac{R-\rho_0(Y,\boldsymbol{Z},\boldsymbol{F})}{\rho_0(Y,\boldsymbol{Z},\boldsymbol{F})}\mu_{10}(Y,\boldsymbol{Z},\boldsymbol{F})\right)\left(Y-\boldsymbol{Z}^\top\boldsymbol{\gamma}_0\right) - \left(\frac{R}{\rho_0(Y,\boldsymbol{Z},\boldsymbol{F})}\boldsymbol{X}\boldsymbol{X}^\top - \frac{R-\rho_0(Y,\boldsymbol{Z},\boldsymbol{F})}{\rho_0(Y,\boldsymbol{Z},\boldsymbol{F})}\mu_{20}(Y,\boldsymbol{Z},\boldsymbol{F})\right)\boldsymbol{\theta}_0 \\ \boldsymbol{Z}\left[Y - \left(\frac{R}{\rho_0(Y,\boldsymbol{Z},\boldsymbol{F})}\boldsymbol{X} - \frac{R-\rho_0(Y,\boldsymbol{Z},\boldsymbol{F})}{\rho_0(Y,\boldsymbol{Z},\boldsymbol{F})}\mu_{10}(Y,\boldsymbol{Z},\boldsymbol{F})\right)^\top\boldsymbol{\theta}_0 - \boldsymbol{Z}^\top\boldsymbol{\gamma}_0\right] \end{pmatrix},$$

where the nuisance functions $\eta_0(Y,\boldsymbol{Z},\boldsymbol{F}) = (\rho_0(Y,\boldsymbol{Z},\boldsymbol{F}),\mu_{10}(Y,\boldsymbol{Z}),\mu_{20}(Y,\boldsymbol{Z},\boldsymbol{F}))$ are given by

$$\rho_0(Y,\boldsymbol{Z},\boldsymbol{F}) = P(R=1\mid Y,\boldsymbol{Z},\boldsymbol{F}),\ \mu_{10}(Y,\boldsymbol{Z},\boldsymbol{F}) = E[\boldsymbol{X}\mid Y,\boldsymbol{Z},\boldsymbol{F},R=1],\ \mu_{20}(Y,\boldsymbol{Z},\boldsymbol{F}) = E[\boldsymbol{X}\boldsymbol{X}^\top\mid Y,\boldsymbol{Z},\boldsymbol{F},R=1].$$

We formally prove that the DR estimating equation $E[\psi_{\mathrm{DR}}(\boldsymbol{W};\boldsymbol{\theta}_0,\boldsymbol{\gamma}_0,\eta_0(Y,\boldsymbol{Z},\boldsymbol{F}))]=0$ adheres to the Neyman Orthogonality condition with respect to the nuisance functions $\eta_0(Y,\boldsymbol{Z},\boldsymbol{F})$ (see Appendix F.4). Thus the DR estimating equation is relatively insensitive to ML-induced bias, and can be used in DML to estimate the coefficients $\boldsymbol{\theta}_0$. Specifically, the nuisance function ρ_0 can be estimated by regressing the labeling indicator R on Y, $\boldsymbol{Z}$ and $\boldsymbol{F}$. In the special MCAR setting where R is completely randomized, ρ_0 is a constant and can be very easily estimated by the proportion of labeled data. The nuisance function μ_{10} can be estimated by regressing $\boldsymbol{X}$ on Y, $\boldsymbol{Z}$ and $\boldsymbol{F}$, and the (j,k)-th entry of μ_{20} can be estimated by regressing the product of the jth and kth variables in $\boldsymbol{X}$, i.e., X_jX_k, on Y, $\boldsymbol{Z}$ and $\boldsymbol{F}$, both only within the labeled data corresponding to $R=1$. These are all amenable to ML estimation. Then we can again apply Algorithm 1 to this doubly robust estimating equation, with nuisance functions $\eta_0(Y,\boldsymbol{Z},\boldsymbol{F})$ estimated by proper ML algorithms.

In Appendix G.4, we also establish the $\sqrt{N}$-consistency and asymptotic normality of the DML estimator of $\boldsymbol{\theta}_0$. These theoretical guarantees enable provably valid confidence intervals and hypothesis testing on $\boldsymbol{\theta}_0$. This significantly improves upon recent approaches, such as ForestIV (Yang et al. 2022), that heuristically use bootstrap for inference, with limited theoretical guarantee. This advantage is ultimately attributed to the principled theoretical framework behind DML.

Empirical Study. In this part, we evaluate the performance of the proposed DML estimator using the real-world dataset collected by Yang et al. (2019) for Facebook business pages. In this evaluation, we hope to estimate the impact of the positive or negative sentiment of user-generated posts on content engagement, specifically measured by the number of comments received by the post (after a logarithmic transformation to reduce skewness). The dataset consists of 429,015 user-generated posts created in 2012 on the business pages of 41 Fortune-500 companies. A completely random sample of 10,157 posts (2.4% of the whole data) had been manually labeled by 5 independent workers on Amazon Mechanical Turk, and the true sentiment label is determined via majority voting. The user-generated posts are numerically encoded by either a traditional bag-of-words (BoW) representation through a term-frequency inverse document-frequency (TF-IDF) matrix, or a 768-dimension embedding vector returned by the BERT pre-trained model (Devlin et al. 2018).

In this problem, the dependent variable Y is the log-transformed number of comments received by each post, the missing variable $\boldsymbol{X}$ is the sentiment of each cost, a binary variable that equals 1 for a positive sentiment and 0 otherwise. The additional control variables $\boldsymbol{Z}$ include the number of words in each post and the content type of the post (one of photo, status, video, or link), and the features $\boldsymbol{F}$ are the BoW or BERT representation of the post. We evaluate four different estimation methods: unbiased regression (estimating the regression using only the labeled data), biased regression (estimating the regression using the unlabeled data and predicted sentiment, without any bias correction), ForestIV (a bias correction methods recently proposed by Yang et al. (2022)), and DML (our approach). For the ML imputation of the missing sentiment variable and the estimation additional nuisance functions, we all use random forests comprised of 100 classification trees. The standard errors of the unbiased and biased regressions are directly from the output of the lm() function in R. The standard error of the ForestIV method is heuristically estimated by a bootstrap procedure suggested by Yang et al. (2022). The standard error for the DML estimator is derived from its asymptotic distribution. The results are summarized in Table A11, with estimates on control variables omitted for brevity.

Table A11 Inference results on Facebook data (*$p < 0.001$, **$p < 0.01$, *$p < 0.05$)**

		BoW+RF			BERT+RF		
	Unbiased	Biased	ForestIV	DML	Biased	ForestIV	DML
sentiment	**-0.202*****	**-0.130*****	**-0.194*****	**-0.203*****	**-0.099*****	**-0.237*****	**-0.222*****
	(0.016)	**(0.007)**	**(0.016)**	**(0.011)**	**(0.011)**	**(0.022)**	**(0.013)**

Compared to coefficient estimates from the unbiased regression, simply plugging in the random forest imputations in the biased regression leads to severe bias.[3] Both ForestIV and DML can clearly mitigate the estimation bias, with DML achieving lower standard errors (and thus higher estimation efficiency). This evaluation shows DML's advantage over the existing heuristic-based ForestIV approach in producing more precise estimates that also have solid theoretical guarantees.

B.4. Applications of DML in estimating other ATE-like parameter

Besides the three settings mentioned above, DML has been also applied to a wide variety of other empirical settings. This subsection provides a comprehensive survey for these applications.

DML has found extensive applications in causal inference. One canonical setting is when the unconfoundedness assumption holds, so that the average effect of a certain treatment on the outcome of interest can be identified by properly controlling covariates to eliminate confounding (Angrist and Pischke 2009). In this setting, one can use the partially linear regression model and estimate the average treatment effect using DML following the introductions in Section 2. However, the partially linear regression model posits an additive linear effect of the treatment without interactions with the control covariates. Therefore, it may not be a good choice if we suspect very complex treatment nonlinearity or treatment-covariate interaction. In that case, we may specify a more general interactive regression model

$$Y = h_0(D, \boldsymbol{X}) + \varepsilon, \quad \mathbb{E}[\varepsilon \mid \boldsymbol{X}, D] = 0, \tag{11}$$

where h_0 is a general nonparametric function. Then we can consider some related parameters of interest such as the average effect $\theta_0 = \mathbb{E}[h_0(1, \boldsymbol{X}) - h_0(0, \boldsymbol{X})]$ when the treatment is binary. Chernozhukov et al. (2018) proposes a DML estimator for binary average treatment effect (ATE) based on the classical doubly robust estimating equation (Bang and Robins 2005). Farrell et al. (2021b) provides theory for this estimator when the nuisance estimation uses deep neural networks, and Farrell et al. (2021a) further study semiparametric inference of parameters that summarize the heterogeneity of treatment effects and propose DML methods based on deep learning. Liu et al. (2021b) extend the use of DML for ATE estimation in logistic partially linear models, and Vansteelandt and Dukes (2022) study generalized partially linear models. Moreover, the growing literature on DML has significantly expanded its scope to estimate ATE-like parameters in the

[3] Moreover, the default OLS standard error tends to underestimate the true standard error of the biased regression, since it ignores the additional uncertainty introduced by the ML imputation of the missing sentiment variable.

context of continuous treatments under unconfoundedness, including dose-response functions (e.g., Kennedy et al. 2017, Colangelo and Lee 2020, Chernozhukov et al. 2022e), average derivative effects (Hines et al. 2023, Klyne and Shah 2023), effects of incremental policies or modified treatment policies (e.g., Kennedy 2019, Schindl et al. 2024), and so on. Moreover, recent authors have also extended DML to parameters beyond average effects, such as the quantile treatment effects (Kallus et al. 2024, Xu et al. 2022b).

When unconfoundedness does not hold, i.e., there exist some unobserved confounders for the treatment and outcome, inferring the treatment effect becomes more difficult. The IV and DID approaches in Sections B.1 and B.2 are two popular ways to address unobserved confounding. As discussed extensively above, DML has already been extended to these two settings (see also Remarks A1 and A2 in Sections B.1 and B.2). Recently, some literature proposes an alternative proximal causal inference approach closely related to the IV approach (Tchetgen et al. 2024). This approach leverages proxy variables that are strongly dependent with the unobserved confounders[4] and satisfy certain exclusion restrictions. DML methods have also been extended to proximal causal inference (Cui et al. 2024, Bennett et al. 2023a,b, Kallus et al. 2021, Ghassami et al. 2022). However, these approaches all need additional auxiliary information (e.g., IV, proxies, repeated cross-section or panel data) to address unmeasured confounding. When there is no such additional information, one may take a sensitivity analysis approach to infer the plausible range of values that the parameter of interest might take for different hypothetical degrees of unobserved confounding (Molinari 2020, Tamer 2010). DML approaches to sensitivity analyses[5] for unobserved confounding have been proposed in Dorn et al. (2021), Chernozhukov et al. (2022a).

DML has been also extended to mediationa analysis that studies the direct effect of a treatment and its indirect effect through some mediators. For example, Farbmacher et al. (2022) applies DML to causal mediation analysis to achieve robust inference of direct and indirect effects. Some recent literature on causal mediation analysis and DML has further expanded to multiple mediators and unmeasured confounding (e.g., Rudolph et al. 2024a,b, Liu et al. 2024b, Shan et al. 2024, Yang et al. 2025, Xu et al. 2022a, Benkeser and Ran 2021). Moreover, DML methods for treatment effect estimation in mediation analysis also exist (Hsu et al. 2023).

[4] In contrast, IV should not be dependent with unobserved confounders after conditioning observed covariates.

[5] We note that there also exist DML solutions to sensitivity analysis in other settings, such as Semenova (2023a, 2024, 2023b).

Furthermore, recent years have seen the application of DML to more complex data settings, especially longitudinal or dynamic data settings. For example, a series of recent works (e.g., Lewis and Syrgkanis 2020, Bodory et al. 2022, Bradic et al. 2024, Chernozhukov et al. 2022d) propose DML methods to infer the dynamic effects of time-varying treatments in a variety of settings. Chen and Ritzwoller (2023), Kallus and Mao (2024), Imbens et al. (2024), Battocchi et al. (2021) propose DML methods to estimate the treatment effects on unobserved or partially observed long-term outcomes in presence of short-term surrogates. Meza and Singh (2021) provides a general framework to study the identification and DML estimation of long-term, mediated, and time-varying treatment effects in a unified manner. It is worth noting that DML has been also applied to survial analyses with censored data (e.g., Lee et al. 2023, Morenz et al. 2024, Vansteelandt et al. 2024).

Beyond the aforementioned settings, DML has also been extended to a variety of other settings. For example, Parikh et al. (2023) and Jung et al. (2024) propose double machine learning approaches to combine experimental and observational data for causal effect estimation. Bia et al. (2024), Chernozhukov et al. (2025), Kato (2023) consider fusing two datasets with different distributions to estimate ATE-like parameters or regression parameters, and propose DML methods to handle the sample selection or distribution shift problem. Li and Owen (2024), Lin et al. (2023) study DML inference with data collected from adaptive experiments where one unit's treatment assignment can adaptively depend on other units' observations. Chiang et al. (2021) extend the DML framework to data with cluster structure where the observations of units within the same cluster can have dependence. They propose robust standard error for statistical inference in this setting. Jung et al. (2021) presents a general algorithm for DML estimation of identifiable causal within the causal graph framework (Pearl 2009).

Finally, we note that DML has been applied in empirical research across many different fields, including information system (e.g., Manzoor et al. 2023, Xu et al. 2024, Zhou et al. 2024, Liu and Huang 2024), agriculture (e.g., Yang et al. 2024, Wang et al. 2024, Yin et al. 2024), energy (e.g., Zhao et al. 2024, Zou et al. 2024), advertising (e.g., Gordon et al. 2023), transportation (e.g., Ling et al. 2024), finance (e.g., Giraldo et al. 2024, Xie et al. 2025), economics (e.g., Beraja et al. 2023, Ellickson et al. 2023, Bonaccolto-Töpfer and Satlukal 2024, Wang and Cheng 2024, Feng et al. 2025), public administration (e.g., Liu et al. 2024a, Pang and Hua 2024), epidemiology (e.g., Chernozhukov et al. 2021a, Moccia et al. 2024) and so on. The motivations for using DML can be broadly categorized into three types. The first is to validate the robustness

of their research conclusions (Zou et al. 2024); the second is to flexibly control for a large number of covariates and address the curse of dimensionality (Zhou et al. 2024, Zhao et al. 2024, Yin et al. 2024, Yang et al. 2024, Wang et al. 2024, Wang and Cheng 2024, Pang and Hua 2024, Manzoor et al. 2023, Xie et al. 2025); and the third is to protect against bias arising from model misspecification (Zhou et al. 2024, Zhao et al. 2024, Yin et al. 2024, Yang et al. 2024, Wang et al. 2024, Wang and Cheng 2024, Pang and Hua 2024). As the above citations suggest, the latter two motivations appear to be the main drivers behind researchers' application of the DML method in practice. Most of these empirical studies consider the PLR model, while a few consider IV or DID models. This motivates us to use the PLR model as the leading example to introduce the DML framework, while additionally explaining the DML methods for IV and DID models in this online appendix.

B.5. Application of DML in HTE estimation

In this paper, our discussion of the DML method is limited to estimation of low-dimensional causal parameters such as average treatment effect, as these are often the target estimands in empirical research literature. Nevertheless, the DML framework is readily applicable to the estimation of heterogeneous treatment effect (HTE) as well. Below we provide a brief discussion of HTE estimation in the DML framework.

Compared to the ATE, which targets the overall causal effect θ_0 on the population, the HTE focuses on the estimand $\theta_0(\boldsymbol{X})$—the causal effect for each subpopulation defined by covariate $\boldsymbol{X}$. Thus, while the ATE captures the average effect, the HTE additionally accounts for heterogeneity in causal effects across individuals. Taking the PLR model as an example, the formulation introduced in Section 2 can be extended to a setting with an HTE parameter $\theta_0(\boldsymbol{X})$ as follows:

$$Y = D\theta_0(\boldsymbol{X}) + g_0(\boldsymbol{X}) + \varepsilon, E\left(\varepsilon \mid D, \boldsymbol{X}\right) = 0 \tag{12}$$

In the estimation of $\theta_0(\boldsymbol{X})$, the Neyman orthogonality property and the cross-fitting strategy of DML remain applicable. Accordingly, following the steps in Algorithm 1, the first stage still involves estimating the nuisance functions $m_0(\boldsymbol{X})$ and $l_0(\boldsymbol{X})$. Once the nuisance functions are obtained, the key difference from ATE estimation is that, instead of solving an estimating equation to obtain a single point estimate of θ_0, the target parameter $\theta_0(\boldsymbol{X})$ in HTE estimation is a function. Estimating such a function requires additional methods, with common choices including meta-learners (e.g., S-learner, T-learner, X-learner) and tree-based models (e.g., causal forest). Implementation of these methods is available in Python packages such as *EconML* and *CausalML*, as well as in R packages such as *grf*.

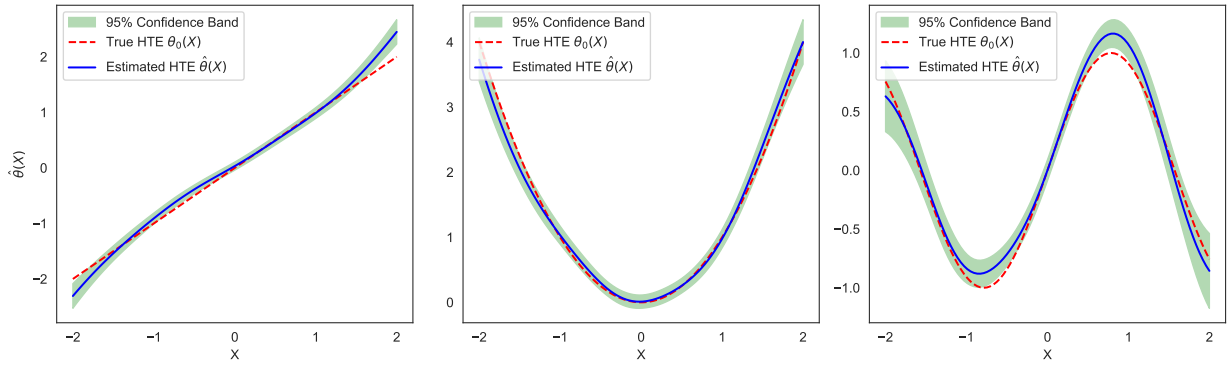


Figure A3 Comparison between the DML-based HTE $\hat{\theta}(X)$ and the ground truth $\theta_0(X)$

To further illustrate this, we design three simulation settings based on eq. (12), each corresponding to a different functional form of the HTE $\theta_0(X)$, including (1) $\theta_0(X) = X$, (2) $\theta_0(X) = X^2$, and (3) $\theta_0(X) = \sin(2X)$. In the implementation, we adopt the CausalForestDML function from the Python package EconML, which uses tree ensembles to estimate the HTE in the second stage after the first-stage of nuisance estimation and construction of Neyman orthogonal estimating functions. The corresponding estimation results are presented in Figure A3. As shown in the figure, in all three simulation settings, the estimated HTE $\hat{\theta}(X)$ closely aligns with the ground truth $\theta_0(X)$, demonstrating that the DML method also performs well in HTE estimation.

Appendix C: Related Literature on Semiparametric Statistics and DML

In Section B, we provided a comprehensive survey of the applications of DML to a wide variety of empirical settings. In this section, we further review related literature on semiparametric statistics that forms the foundation of DML. Then we also review recent studies on the key elements of the DML framework and some proposals to further improve DML.

Semiparametric Statistics

A semiparametric model (Bickel et al. 1993, Tsiatis 2006) is a general statistical model with both parametric components (i.e., the target parameter) and nonparametric components (i.e., nuisance functions). Compared to traditional parametric models (e.g., linear models, generalized linear models), semiparametric models offer more flexibility by allowing nonparametric functional forms for the nuisance components, which reduces the risk of model misspecification. Additionally, this flexibility enables the definition of richer formulations for the target parameter; see Kennedy (2022) for a collection of examples.

In semiparametric models, researchers primarily focus on the estimation and inference of the finite-dimensional target parameter. Extensive research (Levit 1976, Ibragimov et al. 1981, Bickel 1982, Bickel et al. 1993, Robinson 1988, Van der Vaart 1991, 2000, Robins and Rotnitzky 1995, Newey 1990, 1994, Newey et al. 2004, Tsiatis 2006) has been conducted to construct $\sqrt{N}$-consistent and asymptotically normal estimators of the target parameter, with nuisance functions estimated using traditional nonparametric methods, such as kernels and sieves. Some key elements of DML already appear in these semiparametric statistics literature. For example, the Neyman Orthogonal estimating equation for PLR is proposed by Robinson (1988), and the doubly robust estimating equation used for ATE estimation orginates from Robins and Rotnitzky (1995). Newey (1994) also provides quite early discussion on orthogonality. However, these traditional literature typically uses kernels or sieves for nonparametric estimation of nuisance functions. These nonparametric estimators have limited approximation capacity, and do not perform well when the covariate dimension is moderate or high. Later, some literature starts adopting high dimensional LASSO regression for the nuisance estimation (e.g., Belloni et al. 2014, Chernozhukov et al. 2015a,b, Farrell 2015, Belloni et al. 2017). These estimators also leverage Neyman Orthogonality, and they can achieve $\sqrt{N}$-consistency and asymptotic normality under various (approximate) sparsity conditions on the nuisance functions. They do not necessarily need cross-fitting to achieve these desirable properties, as the nuisance estimators using high dimensional sparse linear regressions are still tractable and amenable to refined mathematical analyses.

DML Framework

Building on the existing literature, Chernozhukov et al. (2018) proposes a general DML framework. This framework improves on the existing semiparametric estimators in that it allows for the use of general machine learning methods for more flexible nuisance estimation.

Since its introduction, DML has attracted tremendous attention in the literature. There exist a number of survey papers or exposition papers for DML or semiparametric inference based on machine learning. For example, Kennedy (2022), Hines et al. (2022) review the DML framework from a statistical theory perspective, based on analyses of semiparametric efficiency and influence functions (Fisher and Kennedy 2021), which are very useful tools for constructing Neyman Orthogonal estimating equations. Knaus (2022) review DML methods for causal inference under unconfoundedness. Feyzollahi and Rafizadeh (2024) introduce DML to economists and provide a crash course on basics of machine learning. They focus on the partially linear

regression model without considering other applications. Fuhr et al. (2024) is most similar to our paper, which also provides an introduction to DML in PLR, uses simulations to evaluate the performance of DML, and offers some practical guidelines. However, we note that our paper and Fuhr et al. (2024) are contemporary. With different simulation settings, real-data demonstrations, practical advice, the breadth of empirical settings, and a special focus on IS research, our paper and Fuhr et al. (2024) complement each other.

As we discussed in Section 2, this framework has three key elements: Neyman Orthogonality, cross-fitting, and high-quality ML estimation. Below, we further review some recent literature that study each of these elements respectively.

Neyman Orthogonality. One important prerequisite for the DML estimation is to construct Neyman Orthogonal estimating equations for target parameters under models of interest. Although orthogonal estimating equations are already derived for many standard parameters and models, they may still be unavailable for less-explored settings. Rather than relying on case-by-case derivation, Chernozhukov et al. (2022d) propose an automatic DML framework, which automatically constructs orthogonal estimating equations for a very large class of parameters identified by regression functions. It also proposes to debiases the ML regression estimators by introducing additional high-dimensional sparse nuisance estimators for the so-called "Reisz representer" nuisances. This framework is then extended to accommodate ML estimation of the "Reisz representer" nuisances (Chernozhukov et al. 2022c, 2021b, Hines and Hines 2024), and also applied in a range of empirical settings (e.g., Chernozhukov et al. 2023, Klosin 2021, Chernozhukov et al. 2022d). Moreover, the Neyman Orthogonality considered in the standard DML framework is a "first-order" orthogonality. Some literature has also considered higher-order orthogonality that can further reduce the bias from ML estimation (e.g., Mackey et al. 2018, Liu et al. 2021a, 2017).

Cross-fitting. In Section 2.4, we show the importance of cross-fitting for valid inference with DML using flexible ML techniques. In fact, there are also some exceptions: Chen et al. (2022) shows that cross-fitting is not necessary for some ensemble ML estimators that satisfy certain stability property. But in general cross-fitting is needed. There also exist some variants of cross-fitting, such as the double cross-fitting proposed in Newey and Robins (2018). Double cross-fitting uses different folds of data to estimate different nuisance functions. The simulation study of Zivich and Breskin (2021) demonstrates some benefits of double cross-fitting in estimation and inference. However, with double cross-fitting, each nuisance estimator is trained on fewer folds of data so its finite-sample performance may be also undermined.

ML Estimation. In Section 2.4, we also show the strong positive association between ML nuisance estimation quality and DML estimation qaulity. Therefore, it is very important to use proper ML nuisance estimators and carefully tune the hyperparameters. Bach et al. (2024) use simulations to illustrate the importance of hyperparameter tuning in DML estimation. They show that while the combined loss is strongly associated with the DML error, greedily minimizing the combined loss to select ML models may not always be effective. They also find that using AutoML techniques (Wang and Wu 2019) can also be an effective option. Instead of selecting a single ML model, some other literature also considers ensembles of multiple different models (e.g., Hoffmann 2024, Naimi et al. 2023, Ahrens et al. 2024b, Zivich and Breskin 2021). For hyperparameter tuning, most of the existing literature uses cross-validation with standard error metrics like RMSE. Bach et al. (2024) also evaluates a few different schemes of cross-validation together with cross-fitting. Cui and Tchetgen Tchetgen (2024), Sun et al. (2022) exploit the double robustness property of certain DML estimators and propose some special metrics to guide hyperparameter tuning, but these new metrics have not been widely used yet.

Finally, we note that some literature also has some special considerations in the training of ML models. For example, Chernozhukov et al. (2022c) use a multi-task neural network architecture with shared components for different nuisance estimators to improve efficiency. Klaassen et al. (2024) proposes a neural network architecture for ML nuisance estimation using unstructured, multimodal data (text and images). Fingerhut et al. (2022) proposes to train the two nuisance models in PLR in a coordinated manner, i.e., by minimizing a new combined loss for the two nuisance prediction errors as an proxy for the DML estimation bias.

TMLE Framework

We remark that there is a closely related framework called targeted maximum likelihood estimation (TMLE) developed in the biostatistics literature (Van Der Laan and Rubin 2006). TMLE also leverages Neyman Orthogonality and cross-fitting to use flexible ML techniques in statistical inference, but its procedures are different from DML. Moreover, TMLE is mostly applied in biomedical problems. Díaz (2020) provide a high-level introduction to TMLE and DML as two general approaches for causal inference with machine learning. In principle, most of TMLE methods also have their DML counterparts. In this sense, most of empirical problems that are already solved by TMLE can also have DML solutions. For example, Díaz (2019) propose TMLE method for the inference of treatment effect in survival analysis, and Díaz et al. (2023) propose TMLE

method to infer the effect of modified treatment policies in longitudinal data analysis. Using the orthogonal estimating equation constructed from the "efficient influence function" derived in these papers, we can also construct DML estimators for the corresponding target parameters. Ellul et al. (2024) compares DML and TMLE for ATE estimation in some simulations, finding that they often perform similarly but sometimes TMLE might be more stable. We refer to the monographs Van der Laan et al. (2011), Van der Laan and Rose (2018) for an introduction to the TMLE framework and many biostatistics applications of TMLE. We also refer to the R package `tlverse` for implementations of TMLE methods.

Appendix D: Simulation Studies of DML under Violations of Essential Identification Assumptions

To illustrate the impact of violating these assumptions on DML estimation in PLR models, we provide three simulation examples in this section. These three experiments correspond to (1) violation of unconfoundedness; (2) measurement error; (3) violation of IV exclusion restriction.

Violation of Unconfoundedness. The data generation process for this experiment is based on the setup in Section 2.3, but it incorporates an additional confounder U that affects both the treatment variable D and the outcome Y:

$$X_1 \sim N\left(0, 2^2\right), X_2 \sim N\left(0, 2^2\right), P\left(D=1 \mid X_1, X_2\right) = \frac{1}{1+e^{-(X_1+X_2+U+1)}}$$
$$Y = D + X_1 + X_2 + U + X_1 X_2 + X_1^2 + X_2^2 + 1 + \varepsilon, \varepsilon \sim N\left(0, 1^2\right)$$

Here U is an unobserved variable to which we do not have access in the analysis, so we cannot control for U to eliminate confounding. Instead, we only control for the observed covariates X_1, X_2 in the DML estimators. The estimation results of DML are presented in Table A12. We observe that the DML estimators have huge bias because of the unobserved confounder.

Table A12 Estimation results of the ATE $\theta_0 = 1$ with unobserved variable added to D and Y

	$N=100$	$N=500$	$N=1000$	$N=5000$	$N=10000$
DML(Random forest)	3.4288	3.4828	3.5098	3.5317	3.5138
	(1.0185)	(0.2892)	(0.1925)	(0.0849)	(0.0510)
DML(XGBoost)	3.3492	3.3274	3.3936	3.4994	3.4972
	(1.0529)	(0.3202)	(0.1981)	(0.0770)	(0.0505)

Measurement Error. The data generation process for this experiment is also based on the setup in Section 2.3, but it considers a mismeasured treatment variable:

$$X_1 \sim N\left(0, 2^2\right), X_2 \sim N\left(0, 2^2\right), P\left(D^* = 1 \mid X_1, X_2\right) = \frac{1}{1 + e^{-(X_1 + X_2 + 1)}}$$

$$Y = D^* + X_1 + X_2 + X_1 X_2 + X_1^2 + X_2^2 + 1 + \varepsilon, \varepsilon \sim N\left(0, 1^2\right)$$

$$P(D = 0 \mid D^* = 1) = 0.1, P(D = 1 \mid D^* = 0) = 0.1.$$

Here, variable D^* is the true treatment variable while variable D is a perturbed version of D^* with measurement error. We hide the true treatment D^* in our numerical experiments but only use the mismeasured version D. The estimation results of DML are reported in Table A13. We again observe that DML estimates are highly biased because of the measurement error.

Table A13 Estimation results of the ATE $\theta_0 = 1$ with noise added to treatment selection

	$N = 100$	$N = 500$	$N = 1000$	$N = 5000$	$N = 10000$
DML(Random forest)	0.4489	0.5561	0.5700	0.5608	0.5582
	(0.9242)	(0.3061)	(0.1490)	(0.0535)	(0.0289)
DML(XGBoost)	0.5543	0.5647	0.5527	0.5451	0.5466
	(1.0018)	(0.3065)	(0.1549)	(0.0602)	(0.0384)

Violation of IV Exclusion Restriction. The data generation process is based on the setup in Section B.1, but the IV Z is generated in a different way:

$$X_1 \sim N(0, 2^2),\ X_2 \sim N(0, 2^2),\ P(D^* = 1 | X_1, X_2) = \frac{1}{1 + e^{-(X_1 + X_2 + 1)}},$$

$$P(D = 0 \mid D^* = 1) = P(D = 1 \mid D^* = 0) = 0.1, \quad P(Z = 0 \mid D = 1) = P(Z = 1 \mid D = 0) = 0.1,$$

$$Y = D^* + X_1 + X_2 + X_1 X_2 + X_1^2 + X_2^2 + 1 + \varepsilon,\ \varepsilon \sim N\left(0, 1^2\right).$$

Here, the IV Z directly depends on the mismeasured treatment variable D, violating the exclusion restriction condition. The estimation results of DML and 2SLS are reported in Table A14. We observe that both 2SLS and DML methods have significant bias due to the violation of IV exclusion restriction.

Appendix E: Additional Simulation Results for Propensity Score Matching

Table A15 supplements Table 3 in Section 2.3, providing additional results for PSM. Besides the PSM with linear bias correction presented in Section 2.3 Table 3 (reproduced in the "Bias-Corrected PSM (Linear)" panel in Table A15), Table A15 additionally reports the vanilla PSM without any bias correction ("Vanilla

Table A14 Estimation results of treatment effect in PLIV model (the ground-truth coefficient value is 1)

	$N=100$	$N=500$	$N=1000$	$N=5000$	$N=10000$
2SLS	-1.8584	-1.8628	-1.7383	-1.7748	-1.7532
	(2.8809)	(1.5416)	(1.0254)	(0.4062)	(0.2932)
DML(Random forest)	0.5521	0.4723	0.4732	0.4711	0.4635
	(1.5606)	(0.4092)	(0.2124)	(0.0686)	(0.0366)
DML(XGBoost)	0.5919	0.5217	0.4460	0.4234	0.3946
	(1.8032)	(0.3931)	(0.2184)	(0.0674)	(0.0417)

PSM") and bias-corrected PSM with nonlinear bias correction through correctly specified second-order polynomial regression ("Bias-Corrected PSM (Nonlinear)"). We observe that the linear bias correction indeed sightly reduces the bias of vanilla PSM, but the remaining bias is still quite large. In contrast, when using correctly specified bias correction, the bias becomes small when the sample size is large enough. This means that the large bias of the bias-corrected PSM estimator in Section 2.3 Table 3 is due to the misspecification in the bias correction.

Although the well-specified bias correction achieves very good performance, in practice researchers do not know the true functional form so they may not be able to achieve well-specification. Instead, PSM with linear bias correction is usually the default choice, which may result in unsatisfactory performance under model misspecification. These experiments show that the specification challenge remains within the matching framework. In contrast, DML methods with flexible ML estimation provide an effective alternative approach.

Appendix F: Technical Exposition of Pathwise Derivative and Neyman Orthogonality

Recall that we informally introduce the Neyman Orthogonality condition in Section 2, which is one of the two key elements of the DML method. By constructing Neyman Orthogonal estimation equations, the DML method addresses the bias introduced by the regularization and model selection in ML techniques. To provide a rigorous definition of the Neyman Orthogonality condition, we first introduce the concept of pathwise derivative (Gateaux derivative) in the functional analysis (Giaquinta and Hildebrandt 2004).

Definition A1 (Pathwise derivative). Let $\mathcal{X}$ be a complete normed space (Banach space), and denote by $\mathcal{X}^*$ the space of bounded linear functionals on $\mathcal{X}$. Assume that $F:\mathcal{U}\to\mathbb{R}$ is a functional defined on some open subset $\mathcal{U}$ of $\mathcal{X}$, and let u_0 be some point of $\mathcal{U}$. F is called *pathwise differentiable* at u_0 if there exists some $G\in\mathcal{X}^*$ such that for any $\zeta\in\mathcal{X}$ and r in a small neighborhood of 0, we have

$$F(u_0+r\zeta)=F(u_0)+rG(\zeta)+o(r) \quad \text{as } r\to 0$$

One calls $\partial_r F(u_0+r\zeta)|_{r=0}=G(\zeta)$ the *pathwise derivative* of F at u_0 along the direction ζ.

Table A15 PSM and Bias-Corrected PSM estimation results of the ATE $\theta_0 = 1$.

	$N = 100$	$N = 500$	$N = 1000$	$N = 5000$	$N = 10000$
Vanilla PSM					
Logit, 1:1	2.8563	1.9630	1.8505	1.2833	1.3337
	(1.9518)	(1.2651)	(0.9419)	(0.6186)	(0.4922)
Probit, 1:1	2.7121	1.9392	1.8347	1.2691	1.3404
	(2.1167)	(1.2678)	(0.9738)	(0.6401)	(0.5310)
Logit, 1:5	3.1369	2.3766	2.1747	1.5554	1.5553
	(1.5053)	(0.9618)	(0.7145)	(0.4520)	(0.3361)
Probit, 1:5	3.1584	2.3872	2.1735	1.5289	1.5351
	(1.4823)	(0.9573)	(0.7169)	(0.4750)	(0.3621)
Bias-Corrected PSM (Linear)					
Logit, 1:1	2.8434	1.8668	1.7159	1.2391	1.2612
	(2.7095)	(1.1874)	(0.7924)	(0.4549)	(0.3670)
Probit, 1:1	2.9794	1.8494	1.7081	1.2320	1.2636
	(2.7078)	(1.2087)	(0.8083)	(0.4643)	(0.3835)
Logit, 1:5	3.1901	2.0762	1.8853	1.3797	1.3864
	(2.5335)	(1.1294)	(0.7327)	(0.3920)	(0.2992)
Probit, 1:5	3.1332	2.0771	1.8828	1.3609	1.3734
	(2.5658)	(1.1322)	(0.7388)	(0.4043)	(0.3150)
Bias-Corrected PSM (Nonlinear)					
Logit, 1:1	2.1508	1.0431	0.9515	0.9973	0.9970
	(5.7844)	(0.4402)	(0.3703)	(0.1479)	(0.1101)
Probit, 1:1	1.8586	1.0400	0.9520	0.9996	0.9969
	(4.5050)	(0.4402)	(0.3712)	(0.1466)	(0.1113)
Logit, 1:5	1.1172	1.0185	0.9693	0.9895	1.0084
	(2.4163)	(0.3945)	(0.3088)	(0.1015)	(0.0917)
Probit, 1:5	1.1314	1.0207	0.9724	0.9895	1.0081
	(2.5076)	(0.3928)	(0.3083)	(0.1018)	(0.0921)

Given a semiparametric estimating function $\psi(W;\theta,\eta(\boldsymbol{W}))$ with a finite-dimensional parameter θ and an potentially infinite-dimensional function η, the pathwise derivative of $E[\psi(W;\theta_0,\eta(\boldsymbol{W}))]$ with respect to η at the point η_0 along the direction $\eta-\eta_0$ is given by the following:

$$\partial_\eta E\left[\psi\left(W;\theta_0,\eta_0(\boldsymbol{W})\right)\right]\left[\eta-\eta_0\right]=\partial_r\left\{E\left[\psi\left(W;\theta_0,\eta_0(\boldsymbol{W})+r\left(\eta(\boldsymbol{W})-\eta_0(\boldsymbol{W})\right)\right)\right]\right\}|_{r=0} \tag{13}$$

Based on the pathwise derivative (13), we provide a rigorous definition of the Neyman orthogonality.

Definition A2 (Neyman Orthogonality). An estimating equation $E\left[\psi\left(\boldsymbol{W};\theta_0,\eta_0(\boldsymbol{W})\right)\right]=0$ for the target parameter θ_0 with nuisance functions η_0 satisfies the Neyman Orthogonality condition if the functional $\eta\mapsto E\left[\psi\left(\boldsymbol{W};\theta_0,\eta(\boldsymbol{W})\right)\right]$ is pathwise differentiable at η_0, and its pathwise derivative at η_0 along any direction vanishes:

$$\partial_\eta E\left[\psi\left(W;\theta_0,\eta_0(\boldsymbol{W})\right)\right]\left[\eta-\eta_0\right]=0,\ \forall\eta \tag{14}$$

Intuitively, the pathwise derivative in eq. (13) characterizes how the estimating equation reacts to the perturbation of the nuisance function η in the local region around the truth η_0. Thus, the Neyman Orthogonality condition (14) implies that the value of $E\left[\psi\left(\boldsymbol{W};\theta,\eta(\boldsymbol{W})\right)\right]$ is, on average, insensitive to local perturbations in the nuisance functions. This insensitivity makes the estimation process resilient to bias introduced by machine learning techniques.

In the following, we respectively provide the technical exposition of Neyman orthogonality for PLR model, PLIV model, DID model and the linear regression model with ML-generated covariates.

F.1. PLR Model

In the context of the PLR model, recall the OLS estimating function $\psi_{\mathrm{PLR}}(\boldsymbol{W};\theta,g(\boldsymbol{X}))=D(Y-D\theta-g(\boldsymbol{X}))$ with $\eta=g$. The pathwise derivative of $E\left[\psi_{\mathrm{PLR}}(\boldsymbol{W};\theta_0,g(\boldsymbol{X}))\right]$ at the point g_0 is given by:

$$\partial_\eta E\left[\psi_{\mathrm{PLR}}(\boldsymbol{W};\theta_0,g_0(\boldsymbol{X}))\right][\eta-\eta_0]=-E\left\{D(g(\boldsymbol{X})-g_0(\boldsymbol{X}))\right\},$$

which is in general not equal to 0 at least for some functions g.

Now consider the new estimating function $\psi^{\mathrm{DML}}_{\mathrm{PLR}}(\boldsymbol{W};\theta,l(\boldsymbol{X}),m(\boldsymbol{X}))=[Y-l(\boldsymbol{X})-\theta(D-m(\boldsymbol{X}))](D-m(\boldsymbol{X}))$ with $\eta=(l,m)$. The corresponding pathwise derivative of $E\left[\psi^{\mathrm{DML}}_{\mathrm{PLR}}(\boldsymbol{W};\theta_0,l(\boldsymbol{X}),m(\boldsymbol{X}))\right]$ at the point $\eta_0=(l_0,m_0)$ is given by:

$$\begin{aligned}
&\partial_\eta E\left[\psi^{\mathrm{DML}}_{\mathrm{PLR}}(\boldsymbol{W};\theta_0,l_0(\boldsymbol{X}),m_0(\boldsymbol{X}))\right][\eta-\eta_0]\\
=&-E\left\{\left(D-m_0(\boldsymbol{X})\right)\left(l(\boldsymbol{X})-l_0(\boldsymbol{X})\right)\right\}-E\left\{\left[Y-l_0(\boldsymbol{X})-2\theta_0\left(D-m_0(\boldsymbol{X})\right)\right]\left(m(\boldsymbol{X})-m_0(\boldsymbol{X})\right)\right\}\\
=&-E\left\{E\left[D-m_0(\boldsymbol{X})\mid\boldsymbol{X}\right]\left(l(\boldsymbol{X})-l_0(\boldsymbol{X})\right)\right\}-E\left\{E\left[Y-l_0(\boldsymbol{X})-2\theta_0\left(D-m_0(\boldsymbol{X})\right)\mid\boldsymbol{X}\right]\left(m(\boldsymbol{X})-m_0(\boldsymbol{X})\right)\right\}.
\end{aligned}$$

This pathwise derivative is identical to 0 for any l and m because $E\left[D-m_0(\boldsymbol{X})\mid\boldsymbol{X}\right]=0$ and $E\left[Y-l_0(\boldsymbol{X})\mid\boldsymbol{X}\right]=0$ according to the definitions of l_0,m_0.

Therefore, the naive estimating equation based on ψ_{PLR} does not satisfy the Neyman Orthogonality, but the DML estimating equation based on $\psi^{\mathrm{DML}}_{\mathrm{PLR}}$ does satisfy the Neyman Orthogonality.

F.2. PLIV Model

Recall the PLIV model:

$$Y=D\theta_0+g_0(\boldsymbol{X})+U,\quad E[U\mid\boldsymbol{X},Z]=0,$$

$$Z=r_0(\boldsymbol{X})+V,\quad E[V\mid\boldsymbol{X}]=0,$$

and the estimating equation proposed in Chernozhukov et al. (2018):

$$E\left[\psi_{\text{PLIV}}\left(\boldsymbol{W};\theta_0,l_0(\boldsymbol{X}),m_0(\boldsymbol{X}),r_0(\boldsymbol{X})\right)\right]=0,$$

$$\text{where } \psi_{\text{PLIV}}\left(\boldsymbol{W};\theta,l(\boldsymbol{X}),m(\boldsymbol{X}),r(\boldsymbol{X})\right)=[Y-l(\boldsymbol{X})-\theta(D-m(\boldsymbol{X}))](Z-r(\boldsymbol{X})), \tag{15}$$

$$\text{and } \eta=(l,m,r),\ l_0(\boldsymbol{X})=E(Y\mid\boldsymbol{X}),m_0(\boldsymbol{X})=E(D\mid\boldsymbol{X}),r_0(\boldsymbol{X})=E(Z\mid\boldsymbol{X}).$$

Correspondingly, the pathwise derivative of $E\left[\psi_{\text{PLIV}}\left(\boldsymbol{W};\theta_0,l(\boldsymbol{X}),m(\boldsymbol{X}),r(\boldsymbol{X})\right)\right]$ at the point $\eta_0=(l_0,\, m_0,r_0)$ vanishes: for any $\eta=(l,m,r)$,

$$\partial_\eta E\left[\psi_{\text{PLIV}}\left(\boldsymbol{W};\theta_0,l_0(\boldsymbol{X}),m_0(\boldsymbol{X}),r_0(\boldsymbol{X})\right)\right][\eta-\eta_0]=\theta_0 E\left\{\left(Z-r_0(\boldsymbol{X})\right)\left(m(\boldsymbol{X})-m_0(\boldsymbol{X})\right)\right\}$$

$$-E\left\{\left(Z-r_0(\boldsymbol{X})\right)\left(l(\boldsymbol{X})-l_0(\boldsymbol{X})\right)\right\}-E\left\{\left[Y-l_0(\boldsymbol{X})-\theta_0\left(D-m_0(\boldsymbol{X})\right)\right]\left(r(\boldsymbol{X})-r_0(\boldsymbol{X})\right)\right\}=0,$$

for any function $\eta=(l,m,r)$. This straightforwardly follows from the definitions of functions $l_0(\boldsymbol{X})=E(Y\mid\boldsymbol{X}),m_0(\boldsymbol{X})=E(D\mid\boldsymbol{X}),r_0(\boldsymbol{X})=E(Z\mid\boldsymbol{X})$.

Remark A3. In the context of a special linear IV model with a single endogenous treatment D and a single instrument Z, Imbens (2014) state that the target parameter θ_0 of the two stage least squares (2SLS) procedure can be written as the ratio of two parameters, i.e., $\theta_0=\zeta_1/\beta_1$, where ζ_1,β_1 are respectively the coefficients of the instrument Z in the following two linear regression models:

$$Y=\zeta_0+\zeta_1 Z+\zeta_2^\top\boldsymbol{X}+\varepsilon_1,\quad E(\varepsilon_1\mid Z,\boldsymbol{X})=0$$

$$D=\beta_0+\beta_1 Z+\beta_2^\top\boldsymbol{X}+\varepsilon_2,\quad E(\varepsilon_2\mid Z,\boldsymbol{X})=0$$

This conclusion carries also applies to the PLIV model. Specifically, we can consider the following two partially linear regressions:

$$Y=\zeta_1' Z+u_0(\boldsymbol{X})+\varepsilon_1,\quad E(\varepsilon_1\mid Z,\boldsymbol{X})=0,$$

$$D=\beta_1' Z+v_0(\boldsymbol{X})+\varepsilon_2,\quad E(\varepsilon_2\mid Z,\boldsymbol{X})=0,$$

where ζ_1',β_1' indicate the coefficients of the instrumental variable Z. These two coefficients can be written as

$$\zeta_1'=\frac{E\left[(Y-E(Y\mid\boldsymbol{X}))(Z-E(Z\mid\boldsymbol{X}))\right]}{E\left[(Z-E(Z\mid\boldsymbol{X}))^2\right]}=\frac{E\left[(Y-l_0(\boldsymbol{X}))(Z-r_0(\boldsymbol{X}))\right]}{E\left[(Z-r_0(\boldsymbol{X}))^2\right]},$$

$$\beta_1'=\frac{E\left[(D-E(D\mid\boldsymbol{X}))(Z-E(Z\mid\boldsymbol{X}))\right]}{E\left[(Z-E(Z\mid\boldsymbol{X}))^2\right]}=\frac{E\left[(D-m_0(\boldsymbol{X}))(Z-r_0(\boldsymbol{X}))\right]}{E\left[(Z-r_0(\boldsymbol{X}))^2\right]}.$$

The ratio of the these two coefficients is hence

$$\theta_0=\frac{E\left[(Y-l_0(\boldsymbol{X}))(Z-r_0(\boldsymbol{X}))\right]}{E\left[(D-m_0(\boldsymbol{X}))(Z-r_0(\boldsymbol{X}))\right]},$$

which exactly solves the estimating equation in Equation (15). Therefore, the DML estimator for the PLIV model based on eq. (15) can be interpreted as a nonlinear version of 2SLS.

F.3. DID Model

In the context of the DID model under the repeated cross section setting, recall the estimating function proposed by Chang (2020):

$$E\left[\psi_{\mathrm{RCS}}\left(\boldsymbol{W};\theta_0,p_0,\lambda_0,h_0(\boldsymbol{X}),m_0(\boldsymbol{X})\right)\right]=0,$$
$$\text{where } \psi_{\mathrm{RCS}}(\boldsymbol{W};\theta,p,\lambda,h(\boldsymbol{X}),m(\boldsymbol{X}))=\frac{(T-\lambda)Y-h(\boldsymbol{X})}{\lambda(1-\lambda)p}\frac{D-m(\boldsymbol{X})}{1-m(\boldsymbol{X})}-\theta, \tag{16}$$
$$\text{and } p_0=P(D=1),\ \lambda_0=P(T=1),\ m_0(\boldsymbol{X})=E\left[D\mid\boldsymbol{X}\right],\ h_0=E\left[(T-\lambda_0)Y\mid\boldsymbol{X},D=0\right].$$

The pathwise derivative of $E\left[\psi_{\mathrm{RCS}}\left(\boldsymbol{W};\theta_0,p,\lambda,h(\boldsymbol{X}),m(\boldsymbol{X})\right)\right]$ at the point $\eta_0=(p_0,\lambda_0,h_0,m_0)$ is given by the following: for any $\eta=(p,\lambda,h,m)$,

$$\begin{aligned}
&\partial_\eta E\left[\psi_{\mathrm{RCS}}\left(\boldsymbol{W};\theta_0,p_0,\lambda_0,h_0(\boldsymbol{X}),m_0(\boldsymbol{X})\right)\right][\eta-\eta_0]\\
=&E\left[\frac{(D-1)\left((T-\lambda_0)Y-h_0(\boldsymbol{X})\right)}{p_0\lambda_0(1-\lambda_0)\left(1-m_0(\boldsymbol{X})\right)^2}\left(m(\boldsymbol{X})-m_0(\boldsymbol{X})\right)\right]-E\left[\frac{D-m_0(\boldsymbol{X})}{p_0\lambda_0(1-\lambda_0)\left(1-m_0(\boldsymbol{X})\right)}\left(h(\boldsymbol{X})-h_0(\boldsymbol{X})\right)\right]\\
=&-E\left[\frac{m(\boldsymbol{X})-m_0(\boldsymbol{X})}{p_0\lambda_0(1-\lambda_0)\left(1-m_0(\boldsymbol{X})\right)}\left(h_0(\boldsymbol{X})-h_0(\boldsymbol{X})\right)\right]-E\left[\frac{h(\boldsymbol{X})-h_0(\boldsymbol{X})}{p_0\lambda_0(1-\lambda_0)(1-m_0(\boldsymbol{X}))}(m_0(\boldsymbol{X})-m_0(\boldsymbol{X}))\right]=0.
\end{aligned}$$

Chang (2020) also consider the DMLDID estimating function for the scenario of panel data:

$$E\left[\psi_{\mathrm{Panel}}\left(\boldsymbol{W};\theta_0,p_0,s_0(\boldsymbol{X}),m_0(\boldsymbol{X})\right)\right]=0,$$
$$\text{where } \psi_{\mathrm{Panel}}\left(\boldsymbol{W};\theta,p,s(\boldsymbol{X}),m(\boldsymbol{X})\right)=\frac{Y_1-Y_0-s(\boldsymbol{X})}{p}\frac{D-m(\boldsymbol{X})}{1-m(\boldsymbol{X})}-\theta, \tag{17}$$
$$\text{and } p_0=P(D=1),\ m_0(\boldsymbol{X})=E(D\mid\boldsymbol{X}),\ s_0(\boldsymbol{X})=E[Y_1-Y_0\mid\boldsymbol{X},D=0].$$

The pathwise derivative of $E\left[\psi_{\mathrm{Panel}}\left(\boldsymbol{W};\theta_0,p,s(\boldsymbol{X}),m(\boldsymbol{X})\right)\right]$ at the point $\eta_0=(p_0,s_0,m_0)$ also vanishes: for any $\eta=(p,s,m)$,

$$\begin{aligned}
&\partial_\eta E\left[\psi_{\mathrm{Panel}}\left(\boldsymbol{W};\theta_0,p_0,s_0(\boldsymbol{X}),m_0(\boldsymbol{X})\right)\right][\eta-\eta_0]\\
=&E\left[\frac{(D-1)\left(Y_1-Y_0-s_0(\boldsymbol{X})\right)}{p_0\left(1-m_0(\boldsymbol{X})\right)^2}\left(m(\boldsymbol{X})-m_0(\boldsymbol{X})\right)\right]-E\left[\frac{D-m_0(\boldsymbol{X})}{p_0\left(1-m_0(\boldsymbol{X})\right)}\left(s(\boldsymbol{X})-s_0(\boldsymbol{X})\right)\right]\\
=&-E\left[\frac{s_0(\boldsymbol{X})-s_0(\boldsymbol{X})}{p_0\left(1-m_0(\boldsymbol{X})\right)}(m(\boldsymbol{X})-m_0(\boldsymbol{X}))\right]-E\left[\frac{m_0(\boldsymbol{X})-m_0(\boldsymbol{X})}{p_0\left(1-m_0(\boldsymbol{X})\right)}(s(\boldsymbol{X})-s_0(\boldsymbol{X}))\right]=0.
\end{aligned}$$

F.4. Linear regression model with ML-generated covariates

Recall the doubly robust estimating function:

$$\begin{aligned}
&\psi_{\mathrm{DR}}(\boldsymbol{W};\boldsymbol{\theta}_0,\boldsymbol{\gamma}_0,\eta_0(Y,\boldsymbol{Z},\boldsymbol{F}))=\frac{R}{\rho_0(Y,\boldsymbol{Z},\boldsymbol{F})}\psi_{\mathrm{OLS}}(\boldsymbol{W};\boldsymbol{\theta}_0,\boldsymbol{\gamma}_0)-\frac{R-\rho_0(Y,\boldsymbol{Z},\boldsymbol{F})}{\rho_0(Y,\boldsymbol{Z},\boldsymbol{F})}E\left[\psi_{\mathrm{OLS}}(\boldsymbol{W};\boldsymbol{\theta}_0,\boldsymbol{\gamma}_0)\mid Y,\boldsymbol{Z},\boldsymbol{F},R=1\right]\\
=&\begin{pmatrix}\left(\frac{R}{\rho_0(Y,\boldsymbol{Z},\boldsymbol{F})}X-\frac{R-\rho_0(Y,\boldsymbol{Z},\boldsymbol{F})}{\rho_0(Y,\boldsymbol{Z},\boldsymbol{F})}\mu_{10}(Y,\boldsymbol{Z},\boldsymbol{F})\right)\left(Y-\boldsymbol{Z}^\top\boldsymbol{\gamma}_0\right)-\left(\frac{R}{\rho_0(Y,\boldsymbol{Z},\boldsymbol{F})}\boldsymbol{X}\boldsymbol{X}^\top-\frac{R-\rho_0(Y,\boldsymbol{Z},\boldsymbol{F})}{\rho_0(Y,\boldsymbol{Z},\boldsymbol{F})}\mu_{20}(Y,\boldsymbol{Z},\boldsymbol{F})\right)\boldsymbol{\theta}_0\\ \boldsymbol{Z}\left[Y-\left(\frac{R}{\rho_0(Y,\boldsymbol{Z},\boldsymbol{F})}\boldsymbol{X}-\frac{R-\rho_0(Y,\boldsymbol{Z},\boldsymbol{F})}{\rho_0(Y,\boldsymbol{Z},\boldsymbol{F})}\mu_{10}(Y,\boldsymbol{Z},\boldsymbol{F})\right)^\top\boldsymbol{\theta}_0-\boldsymbol{Z}^\top\boldsymbol{\gamma}_0\right]\end{pmatrix},
\end{aligned}$$

where the nuisance functions $\eta_0 = (\rho_0, \mu_{10}, \mu_{20})$ are given by

$$\rho_0(Y, \boldsymbol{Z}, \boldsymbol{F}) = P(R=1 \mid Y, \boldsymbol{Z}, \boldsymbol{F}),\ \mu_{10}(Y, \boldsymbol{Z}, \boldsymbol{F}) = E[\boldsymbol{X} \mid Y, \boldsymbol{Z}, \boldsymbol{F}, R=1],\ \mu_{20}(Y, \boldsymbol{Z}, \boldsymbol{F}) = E[\boldsymbol{X}\boldsymbol{X}^\top \mid Y, \boldsymbol{Z}, \boldsymbol{F}, R=1].$$

The pathwise derivative of $E\left[\psi_{\mathrm{DR}}\left(\boldsymbol{W}; \theta_0, \boldsymbol{\gamma}_0, \eta_0(Y, \boldsymbol{Z}, \boldsymbol{F})\right)\right]$ at the point η_0 vanishes: for any $\eta = (\rho, \mu_1, \mu_2)$,

$$\begin{aligned}
&\partial_\eta E\left[\psi_{\mathrm{DR}}\left(\boldsymbol{W}; \boldsymbol{\theta}_0, \boldsymbol{\gamma}_0, \eta_0\right)\right][\eta - \eta_0] \\
=&E\left[\begin{pmatrix} R(\boldsymbol{X} - \mu_{10}(Y, \boldsymbol{Z}, \boldsymbol{F}))(Y - \boldsymbol{Z}^\top \boldsymbol{\gamma}_0) - R(\boldsymbol{X}\boldsymbol{X}^\top - \mu_{20}(Y, \boldsymbol{Z}, \boldsymbol{F}))\boldsymbol{\theta}_0 \\ -\boldsymbol{Z}R(\boldsymbol{X} - \mu_{10}(Y, \boldsymbol{Z}, \boldsymbol{F}))^\top \boldsymbol{\theta}_0 \end{pmatrix} \left(\frac{1}{\rho(Y, \boldsymbol{Z}, \boldsymbol{F})} - \frac{1}{\rho_0(Y, \boldsymbol{Z}, \boldsymbol{F})}\right)\right] \\
+&E\left[\begin{pmatrix} -\frac{R - \rho_0(Y, \boldsymbol{Z}, \boldsymbol{F})}{\rho_0(Y, \boldsymbol{Z}, \boldsymbol{F})}(Y - \boldsymbol{Z}^\top \boldsymbol{\gamma}_0) I_{d_\theta \times d_\theta} \\ \frac{R - \rho_0(Y, \boldsymbol{Z}, \boldsymbol{F})}{\rho_0(Y, \boldsymbol{Z}, \boldsymbol{F})} \boldsymbol{Z}\boldsymbol{\theta}_0^\top \end{pmatrix} (\mu_1(Y, \boldsymbol{Z}, \boldsymbol{F}) - \mu_{10}(Y, \boldsymbol{Z}, \boldsymbol{F}))\right] \\
+&E\left[\begin{pmatrix} \frac{R - \rho_0(Y, \boldsymbol{Z}, \boldsymbol{F})}{\rho_0(Y, \boldsymbol{Z}, \boldsymbol{F})} \boldsymbol{\theta}_0^\top \otimes I_{d_\theta \times d_\theta} \\ 0 \end{pmatrix} (\mathrm{vec}(\mu_2(Y, \boldsymbol{Z}, \boldsymbol{F})) - \mathrm{vec}(\mu_{20}(Y, \boldsymbol{Z}, \boldsymbol{F})))\right] = 0.
\end{aligned}$$

where $\mathrm{vec}(\cdot)$ is the vectorization operator. In the equation above, the first term is equal to zero because $E[\boldsymbol{X} - \mu_{10}(Y, \boldsymbol{Z}, \boldsymbol{F}) \mid Y, \boldsymbol{Z}, \boldsymbol{F}, R=1] = 0$ and $E[\boldsymbol{X}\boldsymbol{X}^\top - \mu_{20}(Y, \boldsymbol{Z}, \boldsymbol{F}) \mid Y, \boldsymbol{Z}, \boldsymbol{F}, R=1]$, the second and third terms are equal to zero because $E\left[R - \rho_0(Y, \boldsymbol{Z}, \boldsymbol{F}) \mid Y, \boldsymbol{Z}, \boldsymbol{F}\right] = 0$.

Therefore, the doubly robust estimating equation is Neyman Orthogonal.

Appendix G: Asymptotic properties of DML estimators

As stated in Section 2, the DML estimators, based on Neyman Orthogonal estimating equations and cross-fitting, have strong theoretical guarantees. Chernozhukov et al. (2018) provide a general theory for the statistical property of DML estimators. They show that as long as the estimating equation satisfies some mild smoothness conditions and the nuisance estimators satisfy certain convergence rate conditions, then the DML estimators are $\sqrt{N}$-consistent and asymptotically normal.

Below we provide an an informal summary of the main theoretical conclusions in Chernozhukov et al. (2018). Specifically, we consider a Neyman Orthogonal estimating equation $E[\psi(\boldsymbol{W}; \theta_0, \eta_0(\boldsymbol{W}))] = 0$ for the target parameter $\theta_0 \in \mathbb{R}^d$ with nuisance functions η_0, where ψ is vector-valued mapping[6] that takes values

[6] Here the dimension of the estimating function ψ is identical to the dimension of the unknown parameter θ_0, so that d equations are used to determine the d-dimensional parameter θ_0. This corresponds to the *just identification* setting. The other two settings are the *under-identification* setting where the dimension of the estimating function is smaller than the dimension of the target parameter so the target parameter cannot be uniquely determined by the equation, or the *over-identification* setting where the dimension of the estimating function is larger than the dimension of the target parameter. Chernozhukov et al. (2018) focuses on the just identification setting. All examples in our paper also belong to the just identification setting. Chernozhukov et al. (2022b) discuss DML in possibly over-identification setting under the generalized method of moments (GMM) framework.

in $\mathbb{R}^d$. Let $\hat{\theta}$ be an DML estimator for the parameter θ_0 based on K-fold cross-fitted nuisance estimators $\hat{\eta}_1, \ldots, \hat{\eta}_K$ for η_0 and independent and identically distributed sample $(\boldsymbol{W}_i)_{i=1}^N$ according to Algorithm 1.

THEOREM A1 **(Asymptotic Property of DML)**. *Under mild regularity conditions on the estimating equation, and assume that the nuisance estimators $\hat{\eta}_1, \ldots, \hat{\eta}_K$ satisfy certain convergence rate conditions, then as the sample size $N \to \infty$,*

$$\sqrt{N}\left(\hat{\theta} - \theta_0\right) \rightsquigarrow \mathcal{N}(0, \Sigma),$$

where Σ is the asymptotic variance-covariance matrix given by

$$\Sigma = J_0^{-1} E\left[\psi\left(\boldsymbol{W}; \theta_0, \eta_0\right) \psi\left(\boldsymbol{W}; \theta_0, \eta_0\right)^\top\right] \left(J_0^{-1}\right)^\top, \quad \textit{where } J_0 = \frac{\partial}{\partial \theta} E\left[\psi\left(\boldsymbol{W}; \theta, \eta_0\right)\right]\big|_{\theta=\theta_0}.$$

The specific form of the convergence rate conditions on the nuisance estimators $\hat{\eta}_1, \ldots, \hat{\eta}_K$ vary across different problems. A common sufficient (but not necessary) condition is to require that the root mean squared errors (RMSE) of all estimators vanish to 0 at a rate faster than $N^{-1/4}$. In the rest of this section, we will show concrete forms of the convergence rate conditions in our major examples, which are weaker than the $N^{-1/4}$ requirement. These rate conditions are generic, without restricting the types of nuisance estimators. Notably, the RMSE convergence rates could be slower than the canonical $N^{-1/2}$ convergence rate. It is known that the $N^{-1/2}$ convergence rate can be only achieved by highly parametric methods, which would exclude most machine learning methods. By imposing only generic and slow convergence rate conditions, the DML theory can accommodate a wide range of powerful machine learning techniques for nuisance estimation. Even though the nuisance estimators converge at slow rates, Theorem A1 shows that the resulting DML estimator still has a $N^{-1/2}$ convergence rate and an asymptotic normal distribution.

Although the asymptotic variance-covariance matrix Σ is unknown, it can be consistently estimated, again by the cross-fitted nuisance estimators:

$$\hat{\Sigma} = \hat{J}_0^{-1}\left[\frac{1}{N}\sum_{k=1}^{K}\sum_{i\in I_k}\psi\left(\boldsymbol{W}_i; \hat{\theta}, \hat{\eta}_k(\boldsymbol{W}_i)\right)\psi\left(\boldsymbol{W}_i; \hat{\theta}, \hat{\eta}_k(\boldsymbol{W}_i)\right)^\top\right]\left(\hat{J}_0^{-1}\right)^\top, \text{where } \hat{J}_0 = \frac{1}{N}\sum_{k=1}^{K}\sum_{i\in I_k}\frac{\partial}{\partial\theta}\psi\left(\boldsymbol{W}_i; \hat{\theta}, \hat{\eta}_k(\boldsymbol{W}_i)\right).$$

Chernozhukov et al. (2018) shows that under mild conditions this matrix estimator converges to the true asymptotic variance-covariance matrix Σ as $N \to \infty$.

The asymptotic analysis in Theorem A1 and the consistent variance estimator above provide a solid foundation for statistical inference on the parameter θ_0. In practice, we are often interested in a linear combination of the components of θ_0, i.e., $c^\top\theta_0$ for a vector $c \in \mathbb{R}^d$. For example, if c is a canonical basis

vector with value 1 in one coordinate and value 0 in all other coordinates, then $c^\top\theta_0$ gives the component of θ_0 corresponding to the nonzero coordinate of c. This linear combination can be easily estimated by $c^\top\hat{\theta}$, where $\hat{\theta}$ is the DML estimator of θ_0. According to Theorem A1, we have

$$\sqrt{N}(c^\top\hat{\theta} - c^\top\theta_0) \rightsquigarrow \mathcal{N}(0, c^\top\Sigma c), \text{ when } N \to \infty.$$

The asymptotic variance $c^\top\Sigma c$ can be consistently estimated by $c^\top\hat{\Sigma}c$. It then follows that

$$\frac{\sqrt{N}(c^\top\hat{\theta} - c^\top\theta_0)}{\sqrt{c^\top\hat{\Sigma}c}} \rightsquigarrow \mathcal{N}(0,1). \tag{18}$$

Given the asymptotic variance estimator and the asymptotic normal distribution in Equation (18), we can easily construct a $1-\alpha$ confidence interval for $c^\top\theta_0$:

$$\text{CI} = \left[c^\top\hat{\theta} - \Phi(1-\alpha/2)\sqrt{c^\top\hat{\Sigma}c/N}, c^\top\hat{\theta} + \Phi(1-\alpha/2)\sqrt{c^\top\hat{\Sigma}c/N}\right],$$

where $\Phi^{-1}(1-\alpha/2)$ is the $1-\alpha/2$ quantile of the standard normal distribution. This confidence interval is asymptotically valid, in the sense that the probability of the confidence interval covering the true value of $c^\top\theta_0$ approaches the nominal confidence level $1-\alpha$ when the sample size N grows to infinity. Thus for a large enough sample size N, the coverage probability of the confidence interval would be close to the confidence level $1-\alpha$.

Moreover, we can also use the left hand side of Equation (18) to construct hypothesis tests. For example, consider testing the null hypothesis $H_0 : c^\top\theta_0 = b$ versus the alternative hypothesis $H_a : c^\top\theta_0 \neq b$, for a constant b. Then we could calculate the value of the test statistic $z = \sqrt{N}(c^\top\hat{\theta} - b)/\sqrt{c^\top\hat{\Sigma}c}$. Then we can use the asymptotic normal distribution in Equation (18) to approximate the associated P-value by $2(1-\Phi(|z|))$, where Φ is the cumulative distribution function (CDF) of the standard normal distribution. The null hypothesis H_0 is rejected at a pre-specified significance level α (e.g., $\alpha = 0.05$) if the P-value is smaller than α. This provides a hypothesis test whose type I error rate is close to the significance level α when the sample size N is large enough. We can similarly construct tests for other forms of hypotheses based on the asymptotic normal distribution in Equation (18).

The discussions above outline the general asymptotic theory for the DML estimators and provide provably valid confidence intervals and hypothesis tests. In the following subsections, we further leverage the general theory to establish the asymptotic properties of DML estimators for PLR model, PLIV model, DID model, and the linear regression model with ML-generated covariates, respectively.

Notation. Below we will need to characterize the convergence rates of nuisance estimators. Here we define the notations that will be used in the characterization. Specifically, for a generic function $g(\boldsymbol{W})$ and its estimator $\hat{g}(\boldsymbol{W})$, the mean squared error of the estimator $\hat{g}$ is given by $\|\hat{g}-g\|_2 = (E_{\boldsymbol{W}}[(\hat{g}(\boldsymbol{W})-g(\boldsymbol{W}))^2])^{1/2}$. Here $E_{\boldsymbol{W}}[\cdot]$ means that the expectation is taken only with respect to the distribution of $\boldsymbol{W}$ but not the distribution of data used to construct the estimator $\hat{g}$. Thus $\|\hat{g}-g\|_2$ is still a random variable since the estimator $\hat{g}$ is constructed from the random sample data drawn from the population distribution. For a sequence of random variables Z_N and a deterministic sequence a_N, we denote $Z_N = O_P(a_N)$ if for any positive constant ε, there exists a finite positive constant M such that $P(|Z_N/a_N| > M) < \varepsilon$, and we denote $Z_N = o_P(a_N)$ if for any positive constant ε, $P(|Z_N/a_N| > \varepsilon) \to 0$ as $N \to \infty$.

G.1. Asymptotic properties of DML estimator for PLR model

Recall the estimating equation of PLR model used in DML method:

$$\begin{aligned} &E\left[\psi_{\mathrm{PLR}}^{\mathrm{DML}}(\boldsymbol{W};\theta_0,l_0(\boldsymbol{X}),m_0(\boldsymbol{X}))\right]=0 \\ &\text{where } \psi_{\mathrm{PLR}}^{\mathrm{DML}}(\boldsymbol{W};\theta,l(\boldsymbol{X}),m(\boldsymbol{X}))=[Y-l(\boldsymbol{X})-\theta\,(D-m(\boldsymbol{X}))]\,(D-m(\boldsymbol{X})) \\ &\text{and } l_0(\boldsymbol{X})=E(Y\mid\boldsymbol{X}),\ m_0(\boldsymbol{X})=E(D\mid\boldsymbol{X}). \end{aligned} \tag{19}$$

To construct the DML estimator, we need to first estimate the nuisance functions $\eta_0 = (l_0, m_0)$. Let $\hat{\eta}_k = (\hat{l}_k, \hat{m}_k)$ for $k = 1, \ldots, K$ be the nuisance estimators resulted from the cross-fitting procedure in Algorithm 1. We need these nuisance estimators to satisfy certain convergence rate conditions in order to achieve the asymptotic normality in Theorem A1.

Assumption A1 **(Nuisance Estimation Convergence, PLR Model)**. *Suppose that for $k = 1, \ldots, K$, the nuisance estimators $\hat{\eta}_k = (\hat{l}_k, \hat{m}_k)$ are consistent estimators for $\eta_0 = (l_0, m_0)$, and they satisfy the following conditions:*

$$\|\hat{m}_k - m_0\|_2^2 = o_P(N^{-1/2}),\ \ \|\hat{m}_k - m_0\|_2 \times \|\hat{l}_k - l_0\|_2 = o_P(N^{-1/2}).$$

This assumption is satisfied if $\|\hat{m}_k - m_0\|_2 = o_P(N^{-1/4})$ and $\|\hat{l}_k - l_0\|_2 = o_P(N^{-1/4})$ for all $k = 1, \ldots, K$. Alternatively, if the nuisance estimators $\hat{m}_k$ can converge at rates much faster than $N^{-1/4}$, then the convergence rates of $\hat{l}_k$ are allowed to be accordingly slower. With this assumption, we can obtain the following proposition that characterizes the asymptotic distribution of the corresponding DML estimator.

THEOREM A2 **(DML Asymptotic Normality, PLR model)**. *Let $\hat{\theta}$ be the DML estimator based on the estimating equation in Equation (19) and $\hat{\eta}_k = (\hat{l}_k, \hat{m}_k)$, $k = 1, \ldots, K$ be cross-fitted nuisance estimators that satisfy Assumption A1. Under some mild regularity conditions, the DML estimator $\hat{\theta}$ is asymptotically normal: as $N \to \infty$,*

$$\sqrt{N}\left(\hat{\theta} - \theta_0\right) \rightsquigarrow N(0, \sigma^2),$$

where

$$\sigma^2 = \frac{E\left[(D - m_0(\boldsymbol{X}))^2 (Y - l_0(\boldsymbol{X}) - \theta_0(D - m_0(\boldsymbol{X})))^2\right]}{\left\{E\left[(D - m_0(\boldsymbol{X}))^2\right]\right\}^2}.$$

Theorem A2 establishes the asymptotic normality of the DML estimator and also provides the concrete form of its asymptotic variance. Furthermore, this asymptotic variance σ^2 can be consistently estimated by the following estimator:

$$\hat{\sigma}^2 = \frac{\frac{1}{N}\sum_{k=1}^{K}\sum_{i \in I_k}\left\{\left[Y_i - \hat{l}_k(\boldsymbol{X}_i) - \hat{\theta}(D_i - \hat{m}_k(\boldsymbol{X}_i))\right](D_i - \hat{m}_k(\boldsymbol{X}_i))\right\}^2}{\left\{\frac{1}{N}\sum_{k=1}^{K}\sum_{i \in I_k}(D_i - \hat{m}_k(\boldsymbol{X}_i))^2\right\}^2}.$$

G.2. Asymptotic properties of DML estimator for PLIV model

Recall the estimating equation of PLIV model used in DML method:

$$\begin{aligned} &E\left[\psi_{\text{PLIV}}(\boldsymbol{W}; \theta_0, l_0(\boldsymbol{X}), m_0(\boldsymbol{X}), r_0(\boldsymbol{X}))\right] = 0, \\ &\text{where } \psi_{\text{PLIV}}(\boldsymbol{W}; \theta, l(\boldsymbol{X}), m(\boldsymbol{X}), r(\boldsymbol{X})) = [Y - l(\boldsymbol{X}) - \theta(D - m(\boldsymbol{X}))](Z - r(\boldsymbol{X})), \\ &\text{and } l_0(\boldsymbol{X}) = E(Y \mid \boldsymbol{X}),\ m_0(\boldsymbol{X}) = E(D \mid \boldsymbol{X}),\ r_0(\boldsymbol{X}) = E(Z \mid \boldsymbol{X}). \end{aligned} \tag{20}$$

To construct the DML estimator, we need to first estimate the nuisance functions $\eta_0 = (l_0, m_0, r_0)$. Let $\hat{\eta}_k = (\hat{l}_k, \hat{m}_k, \hat{r}_k)$ for $k = 1, \ldots, K$ be the cross-fitted nuisance estimators. We assume that these nuisance estimators satisfy the following convergence rate conditions.

ASSUMPTION A2 **(Nuisance Estimation Convergence, PLIV Model)**. *Suppose that for $k = 1, \ldots, K$, the nuisance estimators $\hat{\eta}_k = (\hat{l}_k, \hat{m}_k, \hat{r}_k)$ are consistent estimators for $\eta_0 = (l_0, m_0, r_0)$, and they satisfy the following conditions:*

$$\|\hat{m}_k - m_0\|_2 \times \|\hat{r}_k - r_0\|_2 = o_P(N^{-1/2}),\ \ \left\|\hat{l}_k - l_0\right\|_2 \times \|\hat{r}_k - r_0\|_2 = o_P(N^{-1/2}).$$

With the convergence rate condition above, the following lemma shows that the DML estimator for the PLIV model is also asymptotically normal.

THEOREM A3 **(DML Asymptotic Normality, PLIV model)**. *Let $\hat{\theta}$ be the DML estimator based on the estimating equation in Equation (20) and $\hat{\eta}_k = (\hat{l}_k, \hat{m}_k, \hat{r}_k)$, $k = 1, \ldots, K$ be cross-fitted nuisance estimators that satisfy Assumption A2. Under some mild regularity conditions, the DML estimator $\hat{\theta}$ is asymptotically normal: as $N \to \infty$,*

$$\sqrt{N}\left(\hat{\theta} - \theta_0\right) \rightsquigarrow N(0, \sigma^2),$$

where

$$\sigma^2 = \frac{E\left[(Z - r_0(\boldsymbol{X}))^2 (Y - l_0(\boldsymbol{X}) - \theta_0(D - m_0(\boldsymbol{X})))^2\right]}{\left\{E\left[(Z - r_0(\boldsymbol{X}))(D - m_0(\boldsymbol{X}))\right]\right\}^2}.$$

Again, the asymptotic variance σ^2 in Theorem A3 can be consistently estimated by

$$\hat{\sigma}^2 = \frac{\frac{1}{N}\sum_{k=1}^{K}\sum_{i \in I_k}\left\{\left(Z_i - \hat{r}_k\left(\boldsymbol{X}_i\right)\right)\left[Y_i - \hat{l}_k\left(\boldsymbol{X}_i\right) - \hat{\theta}\left(D_i - \hat{m}_k\left(\boldsymbol{X}_i\right)\right)\right]\right\}^2}{\left\{\frac{1}{N}\sum_{k=1}^{K}\sum_{i \in I_k}\left(Z_i - \hat{r}_k\left(\boldsymbol{X}_i\right)\right)\left(D_i - \hat{m}_k(\boldsymbol{X}_i)\right)\right\}^2}.$$

G.3. Asymptotic properties of DML estimator for DID model

For the repeated cross section setting, the estimating function is

$$\begin{aligned} &E\left[\psi_{\mathrm{RCS}}\left(\boldsymbol{W}; \theta_0, p_0, \lambda_0, h_0(\boldsymbol{X}), m_0(\boldsymbol{X})\right)\right] = 0, \\ &\text{where } \psi_{\mathrm{RCS}}(\boldsymbol{W}; \theta, p, \lambda, h(\boldsymbol{X}), m(\boldsymbol{X})) = \frac{(T - \lambda)Y - h(\boldsymbol{X})}{\lambda(1 - \lambda)\, p}\frac{D - m(\boldsymbol{X})}{1 - m(\boldsymbol{X})} - \theta, \\ &\text{and } p_0 = P(D = 1),\ \lambda_0 = P(T = 1),\ m_0(\boldsymbol{X}) = E[D \mid \boldsymbol{X}],\ h_0 = E\left[(T - \lambda_0)Y \mid \boldsymbol{X}, D = 0\right]. \end{aligned} \tag{21}$$

In the DML estimator, we need to first estimate the nuisance parameters p_0, λ_0 and the nuisance functions m_0, h_0. The nuisance parameters p_0, λ_0 can be straightforwardly estimated by $\hat{p} = \frac{1}{N}\sum_{i=1}^{N} D_i$ and $\hat{\lambda} = \frac{1}{N}\sum_{i=1}^{N} T_i$. The nuisance functions m_0, h_0 can be estimated by cross-fitted ML estimators, denoted by $\hat{\eta}_k = (\hat{m}_k, \hat{h}_k)$ for $k = 1, \ldots, K$. Below we impose some convergence rate conditions on them.

ASSUMPTION A3 **(Nuisance Estimation Convergence, DID Repeated Cross Section Model)**. *Suppose that for $k = 1, \ldots, K$, the nuisance estimators $\hat{m}_k, \hat{h}_k$ are consistent estimators for m_0, h_0, and they satisfy the following conditions:*

$$\|\hat{m}_k - m_0\|_2^2 = o_P(N^{-1/2}),\ \ \|\hat{m}_k - m_0\|_2 \times \|\hat{h}_k - h_0\|_2 = o_P(N^{-1/2}).$$

The convergence rate assumption above ensures that the resulting DML estimator has an asymptotic normal distribution.

THEOREM A4 **(DML Asymptotic Normality, DID Repeated Cross Section Model)**. *Let* $\hat{\theta}$ *be the DML estimator based on the estimating equation in Equation* (21) *and* $\hat{\eta}_k = (\hat{p}, \hat{\lambda}, \hat{h}_k, \hat{m}_k)$, $k = 1, \ldots, K$ *be cross-fitted nuisance estimators that satisfy Assumption A3. Under some mild regularity conditions, the DML estimator* $\hat{\theta}$ *is asymptotically normal: as* $N \to \infty$,

$$\sqrt{N}\left(\hat{\theta} - \theta_0\right) \rightsquigarrow N(0, \sigma^2),$$

where

$$\sigma^2 = E\Big\{\big[\psi_{\text{RCS}}\left(\boldsymbol{W}; \theta_0, p_0, \lambda_0, h_0(\boldsymbol{X}), m_0(\boldsymbol{X})\right) + \frac{\partial}{\partial p}\psi_{\text{RCS}}\left(\boldsymbol{W}; \theta_0, p_0, \lambda_0, h_0(\boldsymbol{X}), m_0(\boldsymbol{X})\right)(D - p_0) + \frac{\partial}{\partial \lambda}\psi_{\text{RCS}}\left(\boldsymbol{W}; \theta_0, p_0, \lambda_0, h_0(\boldsymbol{X}), m_0(\boldsymbol{X})\right)(T - \lambda_0)\big]^2\Big\}.$$

The asymptotic variance σ^2 in Theorem A4 can be consistently estmated by

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{k=1}^{K}\sum_{i \in I_k}\Big\{\big[\psi_{\text{RCS}}\left(\boldsymbol{W}_i; \hat{\theta}, \hat{p}, \hat{\lambda}, \hat{h}_k(\boldsymbol{X}_i), \hat{m}_k(\boldsymbol{X}_i)\right) + \frac{\partial}{\partial p}\psi_{\text{RCS}}\left(\boldsymbol{W}_i; \hat{\theta}, \hat{p}, \hat{\lambda}, \hat{h}_k(\boldsymbol{X}_i), \hat{m}_k(\boldsymbol{X}_i)\right)(D_i - \hat{p}) + \frac{\partial}{\partial \lambda}\psi_{\text{RCS}}\left(\boldsymbol{W}_i; \hat{\theta}, \hat{p}, \hat{\lambda}, \hat{h}_k(\boldsymbol{X}_i), \hat{m}_k(\boldsymbol{X}_i)\right)(T_i - \hat{\lambda})\big]^2\Big\}.$$

For the panel data setting with repeated outcome measurements, the Neyman Orthogonal estimating function is

$$\begin{aligned} &E\left[\psi_{\text{Panel}}\left(\boldsymbol{W}; \theta_0, p_0, s_0(\boldsymbol{X}), m_0(\boldsymbol{X})\right)\right] = 0, \\ &\text{where } \psi_{\text{Panel}}\left(\boldsymbol{W}; \theta, p, s(\boldsymbol{X}), m(\boldsymbol{X})\right) = \frac{Y_1 - Y_0 - s(\boldsymbol{X})}{p}\frac{D - m(\boldsymbol{X})}{1 - m(\boldsymbol{X})} - \theta, \\ &\text{and } p_0 = P(D = 1),\ m_0(\boldsymbol{X}) = E(D \mid \boldsymbol{X}),\ s_0(\boldsymbol{X}) = E[Y_1 - Y_0 \mid \boldsymbol{X}, D = 0]. \end{aligned} \tag{22}$$

Again, the nuisance parameter p_0 can be estimated by $\hat{p} = \frac{1}{N}\sum_{i=1}^{N} D_i$ and the nuisance functions m_0, s_0 can be estimated by cross-fitted ML estimators $\hat{m}_k, \hat{s}_k$ for $k = 1, \ldots, K$. We also need to impose some convergence rate conditions on the cross-fitted nuisance estimators.

ASSUMPTION A4 **(Nuisance Estimation Convergence, DID Panel Data Model)**. *Suppose that for* $k = 1, \ldots, K$, *the nuisance estimators* $\hat{m}_k, \hat{s}_k$ *are consistent estimators for* m_0, s_0, *and they satisfy the following conditions:*

$$\|\hat{m}_k - m_0\|_2^2 = o_P(N^{-1/2}),\ \ \|\hat{m}_k - m_0\|_2 \times \|\hat{s}_k - s_0\|_2 = o_P(N^{-1/2}).$$

We can again prove that the resulting DML estimator has an asymptotic normal distribution.

THEOREM A5 **(DML Asymptotic Normality, DID Panel Data Model)**. *Let $\hat{\theta}$ be the DML estimator based on the estimating equation in Equation (22) and $\hat{\eta}_k = (\hat{p}, \hat{m}_k, \hat{s}_k)$, $k = 1, \ldots, K$ be the cross-fitted nuisance estimators that satisfy Assumption A4. Under some mild regularity conditions, the DML estimator $\hat{\theta}$ is asymptotically normal: as $N \to \infty$,*

$$\sqrt{N}\left(\hat{\theta} - \theta_0\right) \rightsquigarrow N(0, \sigma^2),$$

where

$$\sigma^2 = E\left\{\left[\psi_{\text{panel}}\left(\boldsymbol{W}; \theta_0, p_0, s_0(\boldsymbol{X}), m_0(\boldsymbol{X})\right) + \frac{\partial}{\partial p}\psi_{\text{panel}}\left(\boldsymbol{W}; \theta_0, p_0, s_0(\boldsymbol{X}), m_0(\boldsymbol{X})\right)(D - p_0)\right]^2\right\}.$$

The asymptotic variance σ^2 in Theorem A5 can be consistently estimated by

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{k=1}^{K}\sum_{i \in I_k}\left\{\left[\psi_{\text{panel}}\left(\boldsymbol{W}_i; \hat{\theta}, \hat{p}, \hat{s}_k(\boldsymbol{X}_i), \hat{m}_k(\boldsymbol{X}_i)\right) + \frac{\partial}{\partial p}\psi_{\text{panel}}\left(\boldsymbol{W}_i; \hat{\theta}, \hat{p}, \hat{s}_k(\boldsymbol{X}_i), \hat{m}_k(\boldsymbol{X}_i)\right)(D_i - \hat{p})\right]^2\right\}.$$

G.4. Asymptotic properties of DML estimator for the linear regression model with ML-generated covariates

Here we provide the theoretical underpinnings of the DML estimator $\hat{\theta}$ for the linear regression model with ML-generated covariates. Recall that the DML estimator is based on the following doubly robust estimating function:

$$\begin{aligned}
&\psi_{\text{DR}}(\boldsymbol{W}; \boldsymbol{\theta}_0, \boldsymbol{\gamma}_0, \eta_0(Y, \boldsymbol{Z}, \boldsymbol{F})) = \frac{R}{\rho_0(Y, \boldsymbol{Z}, \boldsymbol{F})}\psi_{\text{OLS}}(\boldsymbol{W}; \boldsymbol{\theta}_0, \boldsymbol{\gamma}_0) - \frac{R - \rho_0(Y, \boldsymbol{Z}, \boldsymbol{F})}{\rho_0(Y, \boldsymbol{Z}, \boldsymbol{F})}E\left[\psi_{\text{OLS}}(\boldsymbol{W}; \boldsymbol{\theta}_0, \boldsymbol{\gamma}_0) \mid Y, \boldsymbol{Z}, \boldsymbol{F}, R = 1\right] \\
&= \begin{pmatrix} \left(\frac{R}{\rho_0(Y,\boldsymbol{Z},\boldsymbol{F})}X - \frac{R-\rho_0(Y,\boldsymbol{Z},\boldsymbol{F})}{\rho_0(Y,\boldsymbol{Z},\boldsymbol{F})}\mu_{10}(Y, \boldsymbol{Z}, \boldsymbol{F})\right)\left(Y - \boldsymbol{Z}^\top\boldsymbol{\gamma}_0\right) - \left(\frac{R}{\rho_0(Y,\boldsymbol{Z},\boldsymbol{F})}\boldsymbol{X}\boldsymbol{X}^\top - \frac{R-\rho_0(Y,\boldsymbol{Z},\boldsymbol{F})}{\rho_0(Y,\boldsymbol{Z},\boldsymbol{F})}\mu_{20}(Y, \boldsymbol{Z}, \boldsymbol{F})\right)\boldsymbol{\theta}_0 \\ \boldsymbol{Z}\left[Y - \left(\frac{R}{\rho_0(Y,\boldsymbol{Z},\boldsymbol{F})}\boldsymbol{X} - \frac{R-\rho_0(Y,\boldsymbol{Z},\boldsymbol{F})}{\rho_0(Y,\boldsymbol{Z},\boldsymbol{F})}\mu_{10}(Y, \boldsymbol{Z}, \boldsymbol{F})\right)^\top\boldsymbol{\theta}_0 - \boldsymbol{Z}^\top\boldsymbol{\gamma}_0\right] \end{pmatrix}.
\end{aligned} \tag{23}$$

The nuisance functions $\eta_0 = (\rho_0, \mu_{10}, \mu_{20})$ can be estimated according to the procedure outlined in Section 5. We denote the cross-fitted nuisance estimators as $\hat{\eta}_k = (\hat{\rho}_k, \hat{\mu}_{1k}, \hat{\mu}_{2k})$ for $k = 1, \ldots, K$. Again, we need to impose some convergence rate conditions on these nuisance estimators.

ASSUMPTION A5. *Suppose that for $k = 1, \ldots, K$, the nuisance estimators $\hat{\eta}_k = (\hat{\rho}_k, \hat{\mu}_{1k}, \hat{\mu}_{2k})$ are consistent estimators for $\eta_0 = (\rho_0, \mu_{10}, \mu_{20})$, and they satisfy the following conditions:*

$$\|\hat{\rho}_k^{-1} - \rho_0^{-1}\|_2 \times \|\hat{\mu}_{1k} - \mu_{10}\|_2 = o_P(N^{-1/2}), \ \|\hat{\rho}_k^{-1} - \rho_0^{-1}\|_2 \times \|\hat{\mu}_{2k} - \mu_{20}\|_2 = o_P(N^{-1/2}).$$

The theorem below shows the asymptotic normality of the resulting DML estimator.

THEOREM A6 **(DML Asymptotic Normality, Linear Regression with ML-generated Covariates)**. *Let $(\hat{\theta},\hat{\gamma})$ be the DML estimator based on the doubly robust estimating function in Equation (23) and nuisance estimators $\hat{\eta}_k=(\hat{\rho}_k,\hat{\mu}_{1k},\hat{\mu}_{2k})$, $k=1,\ldots,K$ that satisfy Assumption A5. Under some mild regularity conditions described in Appendix H Assumption A6, the DML estimator $\hat{\theta}$ is asymptotically normal: as $N\to\infty$,*

$$\sqrt{N}\left[\begin{pmatrix}\hat{\theta}\\ \hat{\gamma}\end{pmatrix}-\begin{pmatrix}\theta_0\\ \gamma_0\end{pmatrix}\right]\rightsquigarrow N(0,\Sigma)$$

where

$$\Sigma=J_0^{-1}E\left[\psi_{DR}(\boldsymbol{W};\theta_0,\gamma_0,\eta_0(Y,\boldsymbol{Z},\boldsymbol{F}))\psi_{DR}(\boldsymbol{W};\theta_0,\gamma_0,\eta_0(Y,\boldsymbol{Z},\boldsymbol{F}))^\top\right](J_0^{-1})^\top,$$

$$J_0=-E\left[\begin{pmatrix}\frac{R}{\rho_0}\boldsymbol{X}\boldsymbol{X}^\top-\frac{R-\rho_0}{\rho_0}\hat{\mu}_{20} & \left(\frac{R}{\rho_0}\boldsymbol{X}-\frac{R-\rho_0}{\rho_0}\mu_{10}\right)\boldsymbol{Z}^\top\\ \boldsymbol{Z}\left(\frac{R}{\rho_0}\boldsymbol{X}^\top-\frac{R-\rho_0}{\rho_0}\hat{\mu}_{10}^\top\right) & \boldsymbol{Z}\boldsymbol{Z}^\top\end{pmatrix}\right].$$

The asymptotic variance-covariance matrix Σ in Theorem A6 can be consistently estimated by

$$\hat{\Sigma}=\frac{1}{N}\sum_{k=1}^{K}\sum_{i\in I_k}\hat{J}_0^{-1}\left[\psi_{DR}\left(W_i;\hat{\theta},\hat{\gamma},\hat{\eta}_k\right)\psi_{DR}\left(W_i;\hat{\theta},\hat{\gamma},\hat{\eta}_k\right)^\top\right]\left(\hat{J}_0^{-1}\right)^\top,$$

where $\hat{J}_0$ is the estimator for J_0 given as follows:

$$\hat{J}_0=\frac{1}{N}\sum_{k=1}^{K}\sum_{i\in I_k}\left[-\begin{pmatrix}\frac{R_i}{\hat{\rho}_{k,i}}\boldsymbol{X}_i\boldsymbol{X}_i^\top-\frac{R_i-\hat{\rho}_{k,i}}{\hat{\rho}_{k,i}}\hat{\mu}_{2k,i} & \left(\frac{R_i}{\hat{\rho}_{k,i}}\boldsymbol{X}_i-\frac{R_i-\hat{\rho}_{k,i}}{\hat{\rho}_{k,i}}\hat{\mu}_{1k,i}\right)\boldsymbol{Z}_i^\top\\ \boldsymbol{Z}_i\left(\frac{R_i}{\hat{\rho}_{k,i}}\boldsymbol{X}_i^\top-\frac{R_i-\hat{\rho}_{k,i}}{\hat{\rho}_{k,i}}\hat{\mu}_{1k,i}^\top\right) & \boldsymbol{Z}_i\boldsymbol{Z}_i^\top\end{pmatrix}\right],$$

where

$$\hat{\rho}_{k,i}=\hat{\rho}_k(Y_i,\boldsymbol{Z}_i,\boldsymbol{F}_i),\quad \hat{\mu}_{1k,i}=\hat{\mu}_{1k}(Y_i,\boldsymbol{Z}_i,\boldsymbol{F}_i),\quad \hat{\mu}_{2k,i}=\hat{\mu}_{2k}(Y_i,\boldsymbol{Z}_i,\boldsymbol{F}_i).$$

The proof of Theorem A6 is presented in Appendix H.

Appendix H: Proof for Theorem A6

Let c, ε, $q>2$ be positive constants greater than 0, $K\geq 2$ be some fixed integer, and let $(\delta_N)_{n=1}^{\infty}$ be a sequence of positive constants converging to 0 such that $\delta_N\geq N^{-1/2}$. Moreover, for any generic function $g(W)$ and positive constant $q\geq 2$, the q-norm of $g(W)$ is given by $\|g(W)\|_q=\left(E_{\boldsymbol{W}}\left[(g(\boldsymbol{W}))^q\right]\right)^{1/q}$.

ASSUMPTION A6 **(Regularity conditions)**. *For the doubly robust estimating function $\psi_{DR}(W;\theta,\gamma,\eta)$, the following mild conditions hold:*

(a) $E[\psi_{DR}(W;\theta_0,\gamma_0,\eta_0)]=0$;

(b) $\|X\|_q+\|Z\|_q+\|Y\|_q\leq c$;

(c) $P(\varepsilon \le \rho_0 \le 1-\varepsilon) = 1$;

(d) The matrix J_0 *and* $J_1 = E\left(Cov\left[\begin{pmatrix} X(Y - Z^\top\gamma_0 - X^\top\theta_0) \\ ZX^\top\theta_0 \end{pmatrix} \mid Y, Z, F, R=1 \right] \right)$ *are non-singular;*

(e) For $k = 1, \ldots, K$, *the nuisance estimators* $\hat{\eta}_k = (\hat{\rho}_k, \hat{\mu}_{1k}, \hat{\mu}_{2k})$ *satisfies the following conditions:*

$$\|\hat{\eta}_k - \eta_0\|_q \le c, \ \ \|\hat{\eta}_k - \eta_0\|_2 \le \delta_N,$$
$$\left\|\hat{\rho}_k^{-1} - \rho_0^{-1}\right\|_2 \times \|\hat{\mu}_{1k} - \mu_{10}\|_2 \le \delta_N N^{-1/2},$$
$$\left\|\hat{\rho}_k^{-1} - \rho_0^{-1}\right\|_2 \times \|\hat{\mu}_{2k} - \mu_{20}\|_2 \le \delta_N N^{-1/2}$$

Proof of Theorem A6: Observe that the doubly robust estimating function $\psi_{DR}(W;\theta,\gamma,\eta)$ is linear in (θ,γ) that $\psi_{DR}(W;\theta,\gamma,\eta) = \psi^a(W;\eta)(\theta^\top,\gamma^\top)^\top + \psi^b(W;\eta)$, where

$$\psi^a(W;\eta) = -\begin{pmatrix} \frac{R}{\rho}XX^\top - \frac{R-\rho}{\rho}\mu_2 & \left(\frac{R}{\rho}X - \frac{R-\rho}{\rho}\mu_1\right)Z^\top \\ Z\left(\frac{R}{\rho}X^\top - \frac{R-\rho}{\rho}\mu_1^\top\right) & ZZ^\top \end{pmatrix}, \ \psi^b(W;\eta) = \begin{pmatrix} \left(\frac{R}{\lambda}X - \frac{R-\rho}{\rho}\mu_1\right)Y \\ ZY \end{pmatrix}$$

Therefore, according to the work of Chernozhukov et al. (2018), all we need is to verify Assumption 3.1 and 3.2 in Chernozhukov et al. (2018). We proceed our proof in four steps[7].

Step 1 We first verify Assumption 3.1(d) and (e). According to the calculation in Appendix F.4, the doubly robust estimating equation $E[\psi_{DR}(W;\theta_0,\gamma_0,\eta_0)] = 0$ is Neyman orthogonal. Furthermore, in accordance with Assumption A6(d), the matrix J_0 is non-singular.

Step 2 Then, we verify Assumption 3.2(d). $\forall(\theta^\top,\gamma^\top) \in \Theta$, we have

$$\begin{aligned}(\theta^\top,\gamma^\top)\psi_{DR}(W;\theta_0,\gamma_0,\eta_0) = {} & \theta^\top\left[\frac{R}{\rho_0}(X-\mu_{10})+\mu_{10}\right](Y - Z^\top\gamma_0) - \theta^\top\left[\frac{R}{\rho_0}(XX^\top - \mu_{20})+\mu_{20}\right]\theta_0 \\ & + Z^\top\gamma Y - (Z^\top\gamma)(Z^\top\gamma_0) - (Z^\top\gamma)\theta_0^\top\left[\frac{R}{\rho_0}(X-\mu_{10})+\mu_{10}\right] \\ = {} & \left(\frac{R}{\rho_0}\theta^\top\left[(X-\mu_{10})(Y - Z^\top\gamma_0) - (XX^\top - \mu_{20})\theta_0\right] - \frac{R}{\rho_0}\gamma^\top Z\theta_0^\top(X-\mu_{10})\right) \\ & + \left(\theta^\top\mu_{10}(Y - Z^\top\gamma_0) - \theta^\top\mu_{20}\theta_0 + Z^\top\gamma Y - (Z^\top\gamma)(Z^\top\gamma_0) - (Z^\top\gamma)\theta_0^\top\mu_{10}\right)\end{aligned}$$

We denote that

$$\mathcal{I}_1 := \frac{R}{\rho_0}\theta^\top\left[(X-\mu_{10})(Y - Z^\top\gamma_0) - (XX^\top - \mu_{20})\theta_0\right] - \frac{R}{\rho_0}\gamma^\top Z\theta_0^\top(X-\mu_{10})$$
$$\mathcal{I}_2 := \theta^\top\mu_{10}(Y - Z^\top\gamma_0) - \theta^\top\mu_{20}\theta_0 + Z^\top\gamma Y - (Z^\top\gamma)(Z^\top\gamma_0) - (Z^\top\gamma)\theta_0^\top\mu_{10}$$

[7] Indeed, Assumption 3.1(a)-(c) and Assumption 3.2(a) hold trivially under our construction, so we simplify the discussion of these assumptions.

then because $E[\mathcal{I}_1\mathcal{I}_2] = E[\mathcal{I}_2\rho_0 E(\mathcal{I}_1 \mid Y,Z,R=1)] = 0$, we have that

$$(\theta^\top,\gamma^\top)E\left[\psi_{DR}(W;\theta_0,\gamma_0,\eta_0)\psi_{DR}(W;\theta_0,\gamma_0,\eta_0)^\top\right](\theta^\top,\gamma^\top)^\top = E\left[(\mathcal{I}_1+\mathcal{I}_2)^2\right] = E\left[\mathcal{I}_1^2+\mathcal{I}_2^2\right] \geq E\left[\mathcal{I}_1^2\right]$$

According to the definition of $\mathcal{I}_1$ and Assumption A6(d),

$$\begin{aligned}
E\left[\mathcal{I}_1^2\right] &= E\left\{\left(\frac{R}{\rho_0}\right)^2\left[(\theta^\top,\gamma^\top)\begin{pmatrix}(X-\mu_{10})(Y-Z^\top\gamma_0)-(XX^\top-\mu_{20})\theta_0\\ Z\theta_0^\top(X-\mu_{10})\end{pmatrix}\right]^2\right\}\\
&= E\left\{\left(\frac{R}{\rho_0}\right)^2\left[(\theta^\top,\gamma^\top)\begin{pmatrix}X(Y-Z^\top\gamma_0-X^\top\theta_0)-E[X(Y-Z^\top\gamma_0-X^\top\theta_0)\mid Y,Z,F,R=1]\\ Z\theta_0^\top X-E(Z\theta_0^\top X\mid Y,Z,F,R=1)\end{pmatrix}\right]^2\right\}\\
&= \frac{1}{\rho_0}(\theta^\top,\gamma^\top)E\left\{\mathrm{Cov}\left[\begin{pmatrix}X(Y-Z^\top\gamma_0-X^\top\theta_0)\\ ZX^\top\theta_0\end{pmatrix}\mid Y,Z,F,R=1\right]\right\}(\theta^\top,\gamma^\top)^\top > 0
\end{aligned}$$

This implies the matrix $E\left[\psi_{DR}(W;\theta_0,\gamma_0,\eta_0)\psi_{DR}(W;\theta_0,\gamma_0,\eta_0)^\top\right]$ is non-singular.

Step 3 Here we verify Assumption 3.2(b). In accordance with the triangle inequality and Jensen inequality, We have

$$\begin{aligned}
&\|\mu_{10}\|_q = \|E[X\mid Y,Z,F,R=1]\|_q \leq \|X\|_q\\
&\|\mu_{20}\|_q = \left\|E\left[XX^\top\mid Y,Z,F,R=1\right]\right\|_q \leq \left(\|X\|_q\right)^2\\
&\|\mu_1\|_q \leq \|\mu_{10}\|_q + \|\mu_1-\mu_{10}\|_q \leq \|X\|_q + c\\
&\|\mu_2\|_q \leq \|\mu_{20}\|_q + \|\mu_2-\mu_{20}\|_q \leq \left(\|X\|_q\right)^2 + c\\
&\left\|\frac{1}{\rho}\right\|_q \leq \left\|\frac{1}{\rho_0}\right\|_q + \left\|\frac{1}{\rho}-\frac{1}{\rho_0}\right\|_q \leq \frac{1}{\varepsilon}+c
\end{aligned}$$

Thus, for any $\eta=(\rho,\mu_1,\mu_2)$, we have

$$\begin{aligned}
\left(E\left||\psi^a(W;\eta)\right||^q\right])^{1/q} = \|\psi^a(W;\eta)\|_q &= \left\|\begin{pmatrix}\frac{R}{\rho}XX^\top-\frac{R-\rho}{\rho}\mu_2 & \left(\frac{R}{\rho}X-\frac{R-\rho}{\rho}\mu_1\right)Z^\top\\ Z\left(\frac{R}{\rho}X^\top-\frac{R-\rho}{\rho}\mu_1^\top\right) & ZZ^\top\end{pmatrix}\right\|_q\\
&\leq \left\|\frac{R}{\rho}XX^\top-\frac{R-\rho}{\rho}\mu_2\right\|_q + 2\left\|\left(\frac{R}{\rho}X-\frac{R-\rho}{\rho}\mu_1\right)Z^\top\right\|_q + \left\|ZZ^\top\right\|_q\\
&\leq O\left(c^3+c^2/\varepsilon\right)
\end{aligned}$$

by Assumption A6, which gives the upper bound on $\left(E\left||\psi^a(W;\eta)\right||^q\right])^{1/q}$. Also, since the finite-dimensional matrix J_0 is non-singular, without loss of generality, we denote the eigenvalues of its inverse matrix

$$\left(E\left[\begin{pmatrix}\mu_{20} & \mu_{10}Z^\top\\ Z\mu_{10}^\top & ZZ^\top\end{pmatrix}\right]\right)^{-1}$$

are upper bounded by c_0. It follows that (θ_0, γ_0) satisfies

$$\begin{aligned}\|(\theta_0^\top, \gamma_0^\top)\| &= \left\| \left(E\left[\begin{pmatrix} \mu_{20} & \mu_{10} Z^\top \\ Z\mu_{10}^\top & ZZ^\top \end{pmatrix} \right] \right)^{-1} E\left[\begin{pmatrix} \mu_{10} \\ Z \end{pmatrix} Y \right] \right\| \\ &\le \left\| \left(E\left[\begin{pmatrix} \mu_{20} & \mu_{10} Z^\top \\ Z\mu_{10}^\top & ZZ^\top \end{pmatrix} \right] \right)^{-1} \right\| \left\| E\left[\begin{pmatrix} \mu_{10} \\ Z \end{pmatrix} Y \right] \right\| \\ &\le c_0 \left(\|X\|_q + \|Z\|_q \right) \|Y\|_q \le c_0 c^2 \end{aligned}$$

Hence, we have

$$\begin{aligned}&\left(E\left[||\psi_{DR}(W;\theta_0,\gamma_0,\eta)||^q\right]\right)^{1/q} = \|\psi_{DR}(W;\theta_0,\gamma_0,\eta)\|_q \\ &\le \left\| \begin{pmatrix} \frac{R}{\rho} XX^\top - \frac{R-\rho}{\rho}\mu_2 & \left(\frac{R}{\rho}X - \frac{R-\rho}{\rho}\mu_1\right) Z^\top \\ Z\left(\frac{R}{\rho}X^\top - \frac{R-\rho}{\rho}\mu_1^\top\right) & ZZ^\top \end{pmatrix} \right\|_q \|(\theta_0^\top, \gamma_0^\top)\| + \left\| \begin{pmatrix} \left(\frac{R}{\rho}X - \frac{R-\rho}{\rho}\mu_1\right) Y \\ ZY \end{pmatrix} \right\|_q \\ &\le O\left(c_0 c^3/\varepsilon\right)\end{aligned}$$

This gives the upper bound on $\left(E\left[||\psi_{DR}(W;\theta_0,\gamma_0,\eta)||^q\right]\right)^{1/q}$.

Step 4 Finally, we verify Assumption 3.2(c), i.e, three conditions on the convergence rate. For any $\eta = (\rho, \mu_1, \mu_2)$, we have

$$\begin{aligned}\|E\left[\psi^a(W;\eta)\right] - E\left[\psi^a(W;\eta_0)\right]\| &= \left\| E\left[\begin{pmatrix} \rho_0\left(\frac{1}{\rho} - \frac{1}{\rho_0}\right)(\mu_2 - \mu_{20}) & \rho_0\left(\frac{1}{\rho} - \frac{1}{\rho_0}\right)(\mu_1 - \mu_{10}) Z^\top \\ \rho_0\left(\frac{1}{\rho} - \frac{1}{\rho_0}\right) Z(\mu_1 - \mu_{10})^\top & 0 \end{pmatrix} \right] \right\| \\ &\le (1-\varepsilon) \left\| \frac{1}{\rho} - \frac{1}{\rho_0} \right\|_2 \left(\|\mu_2 - \mu_{20}\|_2 + 2\|\mu_1 - \mu_{10}\|_2 \|Z\|_2 \right) \\ &\le O_P\left(\delta_N N^{-1/2}\right)\end{aligned}$$

which gives the upper bound on $\|E\left[\psi^a(W;\eta)\right] - E\left[\psi^a(W;\eta_0)\right]\|$. Further, by the triangle inequality,

$$\begin{aligned}\left(E\left[||\psi_{DR}(W;\theta_0,\gamma_0,\eta) - \psi_{DR}(W;\theta_0,\gamma_0,\eta_0)||^2\right]\right)^{1/2} &= \|\psi_{DR}(W;\theta_0,\gamma_0,\eta) - \psi_{DR}(W;\theta_0,\gamma_0,\eta_0)\|_2 \\ &\le \left\| \begin{pmatrix} \mathcal{I}_1 & \mathcal{I}_2 \\ \mathcal{I}_2^\top & 0 \end{pmatrix} \right\|_2 \|(\theta_0^\top, \gamma_0^\top)\| + \|\mathcal{I}_3\|_2 \\ &\le \left(\|\mathcal{I}_1\|_2 + 2\|\mathcal{I}_2\|_2 \right) \|(\theta_0^\top, \gamma_0^\top)\| + \|\mathcal{I}_3\|_2\end{aligned}$$

where

$$\mathcal{I}_1 := RXX^\top \left(\frac{1}{\rho} - \frac{1}{\rho_0} \right) - R\left(\frac{\mu_2}{\rho} - \frac{\mu_{20}}{\rho_0} \right) + (\mu_2 - \mu_{20})$$

$$\mathcal{I}_2 := \left[RX\left(\frac{1}{\rho}-\frac{1}{\rho_0}\right) - R\left(\frac{\mu_1}{\rho}-\frac{\mu_{10}}{\rho_0}\right)+(\mu_1-\mu_{10})\right]Z^\top$$
$$\mathcal{I}_3 := \left[RX\left(\frac{1}{\rho}-\frac{1}{\rho_0}\right) - R\left(\frac{\mu_1}{\rho}-\frac{\mu_{10}}{\rho_0}\right)+(\mu_1-\mu_{10})\right]Y$$

To bound $\|\mathcal{I}_1\|_2$, $\|\mathcal{I}_2\|_2$ and $\|\mathcal{I}_3\|_2$, we have

$$\|\mathcal{I}_1\|_2 \le \left\|\frac{1}{\rho}-\frac{1}{\rho_0}\right\|_2 \left\|XX^\top\right\|_2 + \left\|\frac{\mu_2}{\rho}-\frac{\mu_{20}}{\rho_0}\right\|_2 + \|\mu_2-\mu_{20}\|_2 \le O_P(\delta_N)$$
$$\|\mathcal{I}_2\|_2 \le \left(\left\|X\left(\frac{1}{\rho}-\frac{1}{\rho_0}\right)\right\|_2 + \left\|\frac{\mu_1}{\rho}-\frac{\mu_{10}}{\rho_0}\right\|_2 + \|\mu_1-\mu_{10}\|_2\right)\|Z\|_2 \le O_P(\delta_N)$$
$$\|\mathcal{I}_3\|_2 \le \left(\left\|X\left(\frac{1}{\rho}-\frac{1}{\rho_0}\right)\right\|_2 + \left\|\frac{\mu_1}{\rho}-\frac{\mu_{10}}{\rho_0}\right\|_2 + \|\mu_1-\mu_{10}\|_2\right)\|Y\|_2 \le O_P(\delta_N)$$

This gives the upper bound on $\left(E\left[||\psi_{DR}(W;\theta_0,\gamma_0,\eta)-\psi_{DR}(W;\theta_0,\gamma_0,\eta_0)||^2\right]\right)^{1/2}$. Finally, let

$$f(r) := E\left[\psi_{DR}(W;\theta_0,\gamma_0,\eta_0+r(\eta-\eta_0))\right],\ r\in(0,1)$$

Then, for any $r\in(0,1)$,

$$\begin{aligned}\left\|\frac{\partial^2}{\partial r^2}f(r)\right\| &= \left\|E\left\{2\rho_0\left(\frac{1}{\rho}-\frac{1}{\rho_0}\right)\left[\begin{pmatrix}\mu_2-\mu_{20} & (\mu_1-\mu_{10})Z^\top\\ Z(\mu_1-\mu_{10})^\top & 0\end{pmatrix}(\theta_0^\top,\gamma_0^\top)^\top - \begin{pmatrix}(\mu_1-\mu_{10})Y\\ 0\end{pmatrix}\right]\right\}\right\| \\ &\le 2(1-\varepsilon)\left\|\frac{1}{\rho}-\frac{1}{\rho_0}\right\|_2\left[\left(\|\mu_2-\mu_{20}\|_2+2\left\|(\mu_1-\mu_{10})Z^\top\right\|_2\right)\|(\theta_0^\top,\gamma_0^\top)\|+\|(\mu_1-\mu_{10})Y\|_2\right] \\ &\le O_P\left(\delta_N N^{-1/2}\right)\end{aligned}$$

This gives the upper bound on $\left\|\frac{\partial^2}{\partial r^2}E\left[\psi_{DR}(W;\theta_0,\gamma_0,\eta_0+r(\eta-\eta_0))\right]\right\|$. Thus, all conditions of Assumption 3.1 and 3.2 in Chernozhukov et al. (2018) are verified. This completes the proof.

□

Appendix I: Supplementary Experiments for the simulation in Section 2

In this section, we present supplementary estimation results that were not reported in Section 2, but are based on similar simulation settings. Below, we provide a brief overview of the details of these additional experiments.

I.1. Supplementary Experiments for the simulation in Section 2.3

As a variation of the simulation in Section 2.3, we reduce the variances of X_1 and X_2 from 2^2 to 1^2. Using the same estimation methods, we re-estimate the target parameter θ. The results are reported in Table A16.

Compared to the estimation results under the original setting, all estimation methods exhibit a clear reduction in bias when the variances of the covariates are decreased. This is primarily because with smaller

Table A16 Estimation results for $\theta_0 = 1$ from different estimators across 100 experiment replications.

	$N = 100$	$N = 500$	$N = 1000$	$N = 5000$	$N = 10000$
Linear regression	0.2486	0.3678	0.3609	0.3495	0.3497
	(0.5661)	(0.2305)	(0.1626)	(0.0814)	(0.0583)
Bias-Corrected PSM					
Logit, 1:1	1.3244	1.1360	1.0981	1.0241	1.0090
	(0.6678)	(0.2492)	(0.1497)	(0.0739)	(0.0427)
Probit, 1:1	1.3029	1.1290	1.0915	1.0231	1.0080
	(0.6814)	(0.2565)	(0.1506)	(0.0737)	(0.0440)
Logit, 1:5	1.4614	1.2101	1.1546	1.0473	1.0215
	(0.6450)	(0.2224)	(0.1371)	(0.0593)	(0.0433)
Probit, 1:5	1.4598	1.2061	1.1497	1.0445	1.0199
	(0.6484)	(0.2204)	(0.1366)	(0.0597)	(0.0430)
Weighted Regression					
Logit	1.2301	1.0947	1.0644	1.0018	0.9908
	(0.8392)	(0.5027)	(0.3618)	(0.1751)	(0.1393)
Probit	1.1766	0.9646	0.9006	0.7597	0.7133
	(0.8832)	(0.6124)	(0.4877)	(0.2887)	(0.2831)
DML					
Random forest	**1.0021**	**1.0042**	**1.0045**	**0.9967**	**0.9948**
	(0.3927)	**(0.1242)**	**(0.0951)**	**(0.0381)**	**(0.0272)**
XGBoost	**0.9141**	**0.9574**	**0.9577**	**0.9799**	**0.9808**
	(0.4416)	**(0.1206)**	**(0.0921)**	**(0.0374)**	**(0.0252)**

variances, the covariates are more concentrated around zero, diminishing the influence of quadratic terms and making the model approximately linear. Nevertheless, the DML estimator retains a clear advantage in bias reduction, consistent with our earlier findings in Section 2.3.

I.2. Supplementary Experiments for the simulation in Section 2.4

The supplementary experiments for the simulation in Section 2.4 consist of two parts. In the first part, we replace all propensity score estimation methods with logistic regression to examine the estimation performance when the propensity score model is correctly specified. The results are reported in Table A17, Table A18 and Table A19. We observe that the bias of all estimators, except for the Speckman method, is substantially reduced to nearly zero.

This is not surprising as the bias of the DML estimator is impacted by the second-order terms $\|\hat{m} - m_0\|\|\hat{l} - l_0\|$ and $\|\hat{m} - m_0\|^2$. When the propensity score is estimated by correctly specified logistic regression, the error term $\|\hat{m} - m_0\|$ should vanish to zero at the canonical parametric $O(N^{-1/2})$ rate under regularity conditions, and the second-order terms are rapidly vanishing. This is why we observe nearly zero bias when the propensity score is estimated by correctly specified logistic regression. Moreover, we note that the exception of the Speckman method is due to bias arises from overfitting in the absence of cross-fitting. When we

increase the number of cross-fitting folds for the Speckman method to $K = 5$, it also yields an unbiased estimate.

In the second part, as a variation of the simulation in Section I.1, we reduce the variance of the noise term ε from 2^2 to 1^2 in both simulation settings of Section 2.4, and report the corresponding estimation results under the low-variance case. Compared to the original results, this variance reduction only leads to a general decrease in the estimated variances, without affecting our conclusions regarding the performance sensitivity of DML.

I.3. Supplementary discussion of the non-orthogonal estimating equation for the simulation in Section 2.4

In this section, we clarify that non-orthogonality has impact on multiple factors: bias of the non-orthogonal estimator, non-normality of the orthogonal estimator, and bias in the standard error estimation. These factors together contribute to the low coverage rate when using the non-orthogonal estimators.

Let us first explain these via the PLR example. The PLR orthogonal estimating equation considered in our paper is $\mathbb{E}\left[(Y - l_0(\boldsymbol{X}) - \theta_0(D - m_0(\boldsymbol{X})))(D - m_0(\boldsymbol{X}))\right] = 0$ (See equations (3)-(4) in our paper). The DML estimator $\hat{\theta}^{\mathrm{DML}}$ that empirically solves this estimating equation with nuisance estimators $\hat{m}_k, \hat{l}_k$ $(k = 1, \ldots, K)$ have the following decomposition:

$$\sqrt{N}\left(\hat{\theta}^{\mathrm{DML}} - \theta_0\right) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{(Y - l_0(\boldsymbol{X}) - \theta_0(D - m_0(\boldsymbol{X})))(D - m_0(\boldsymbol{X}))}{\mathbb{E}\left[(D - m_0(X))^2\right]} + R,$$

where the remainder error term R can be upper bounded by $\sum_{k=1}^{K} \|\hat{m}_k - m_0\| \|\hat{l}_k - l_0\| + \|\hat{m}_k - m_0\|^2$ (up to proportional constants) with high probability. The first term starting with $\frac{1}{\sqrt{N}}$ converges to a mean-zero Normal distribution at the $1/\sqrt{N}$ rate, under mild regularity conditions according to the central limit theorem. When the nuisance estimators are accurate enough so that the remainder error term R vanishes to 0 faster than the $1/\sqrt{N}$ rate (e.g., when $\|\hat{m}_k - m_0\|$ and $\|\hat{l}_k - l_0\|$ both vanish at faster than $N^{-1/4}$ rate), the remainder error term R is negligible. As a result, the DML estimator $\hat{\theta}^{\mathrm{DML}}$ is asymptotically Normal, where the distribution of the DML estimator is determined by the distribution of the first term (that is free from the nuisance estimation error). This is exactly how DML estimator is robust to the nuisance estimation error. One can then leverage the form of the first term to estimate the asymptotic variance of the DML estimator and accordingly construct the confidence interval using the asymptotic normal distribution.

Now let us consider the non-orthogonal estimating equation $\mathbb{E}\left[(Y-\theta_0 D-g_0(X))D\right]=0$. Correspondingly, we can empirically solve this estimating equation to form a non-orthogonal estimator (which we denote by $\hat{\theta}^{\text{no}}$). Even if we use cross-fitting, e.g., construct estimators $\hat{g}_k$ $(k=1,\ldots,K)$ for the nuisance function g_0, the resulting estimator $\hat{\theta}^{\text{no}}$ may still not have a tractable asymptotic distribution. Indeed, one can show that the non-orthogonal estimator $\hat{\theta}^{\text{no}}$ has the following decomposition:

$$\sqrt{N}\left(\hat{\theta}^{\text{no}}-\theta_0\right)=\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{(Y-\theta_0 D-g_0(X))D}{\mathbb{E}\left[D^2\right]}+R^{\text{no}},$$

where the first term is again asymptotically normal according to the central limit theorem, but the second remainder error term R^{no} is much more challenging. One can show that R^{no} is upper bounded by $\sum_{k=1}^{K}\|\hat{g}_k-g_0\|$ (up to proportional constants) with high probability. However, this term generally cannot converge faster than the $N^{-1/2}$ rate to be negligible. This is exactly the problem: the remainder error R^{no} is not negligible, and it can contribute to significant bias, variance and it may also distort the asymptotic normal distribution. Worse yet, the non-negligible impact of the remainder error R^{no} is very hard to characterize, especially when the nuisance function is estimated via complex black-box models[8]. As a result, the non-orthogonal estimator $\hat{\theta}^{\text{no}}$ could be severely biased, asymptotically non-normal and its asymptotic standard error cannot be precisely estimated using the form of the first term in the decomposition above. Any of these issues or their combination could cause the coverage rate of confidence intervals to be lower than the nominal level.

Our results in Tables A2 and A3 already confirm that some non-orthogonal estimators have significant bias, which could explain the low coverage rate of confidence intervals. In fact, even when the bias is not large, the coverage could be low because the non-orthogonal estimator's distribution deviates from a normal distribution or its asymptotic variance is not accurately estimated. We illustrate these in the new Figures A4 and A5. In Figure A4, we compare the distribution of the orthogonal DML estimator and its non-orthogonal counterpart across the 200 experiment replications. We benchmark both against normal distributions via QQ-plot and also report the results from normality tests. We can observe that the distribution of the non-orthogonal estimator sometimes significantly deviates from a normal distribution due to outliers (i.e., extreme

[8] In fact, the remainder term can be explicitly analyzed for classical nonparametric estimators such as kernel and series regressions because they are much more tractable nonparametric estimators. This is why kernel and series regressions are very commonly applied in classic semi-parametric literature. For these simple nonparametric estimators, one can prove when hyperparameters in the kernel and series regressions satisfy certain undersmoothing conditions, the final estimator is still asymptotically normal, but the distribution of the estimator is determined by both the first term and the remainder term. However, these may not be the case when the nuisance function is estimated via complex black-box models. We hope this comment further clarify our responses to your last comment.

estimates). In contrast, the DML estimator' empirical distribution tends to be closer to normal distribution. Moreover, in Figure A5, we plot the empirical distribution of the estimated standard errors of the orthogonal DML estimator and its non-orthogonal counterpart across 200 experiment replications. We also plot the standard deviations of these two estimators in the 200 replications, as Monte Carlo approximation for the true standard errors of the orthogonal and non-orthogonal estimators. We can observe that the estimated standard error of the non-orthogonal estimate is significantly biased[9] while that of the DML estimator is much closer to the approximated truth despite finite-sample errors.

In summary, generally, the bias of the estimator, the consistency of variance estimation, and the validity of asymptotic normality can all serve as channels through which non-orthogonal estimating equations affect the coverage rate of confidence intervals.

[9] This example suffers from underestimated standard errors that result in under-coverage of the confidence intervals. However, the error remainder term could have very complicated impact. In principle, overestimated standard error is also possible, leading to over-coverage and excessive conservativeness of the confidence intervals.

Table A17 Estimation results for $\theta_0 = 1$ using different estimators over 200 replications, with propensity scores estimated via logistic regression

Scenario 1	$p=2$		$p=5$		$p=10$	
	$N=100$	$N=1000$	$N=100$	$N=1000$	$N=100$	$N=1000$
Linear regression	0.9954	0.9587	0.7635	0.9834	0.3042	0.8947
	(0.5187)	(0.1623)	(0.6666)	(0.2079)	(1.1236)	(0.2707)
	96.0%	95.5%	97.5%	99.5%	98.5%	100.0%
Traditional Semiparametrics						
Kernel (Speckman)	0.9210	0.9305	0.1437	0.1569	0.0007	0.0001
	(0.4613)	(0.1490)	(0.1347)	(0.0426)	(0.0046)	(0.0002)
	92.0%	93.5%	0.5%	0.0%	0.0%	0.0%
DML						
Kernel	1.0653	0.9743	1.0317	1.0053	0.7568	0.9806
	(0.5281)	(0.1558)	(0.7415)	(0.2031)	(1.2280)	(0.3447)
	93.0%	95.0%	92.0%	95.5%	94.5%	95.5%
Decision tree	1.0394	0.9687	0.9508	1.0003	1.2245	0.9685
	(0.6092)	(0.1609)	(1.0624)	(0.2647)	(1.9238)	(0.4538)
	93.5%	96.0%	96.5%	96.0%	97.0%	98.0%
Random forest	1.0337	0.9731	0.9443	1.0055	0.9739	0.9539
	(0.5557)	(0.1566)	(0.7271)	(0.1929)	(1.3323)	(0.3396)
	92.0%	95.0%	96.5%	98.0%	98.0%	98.5%
XGBoost	1.0199	0.9714	0.9185	1.0124	0.9103	0.9477
	(0.5515)	(0.1570)	(0.7376)	(0.2031)	(1.4275)	(0.3452)
	92.0%	96.0%	97.0%	97.0%	95.5%	97.5%

Scenario 2	$p=2$		$p=5$		$p=10$	
	$N=100$	$N=1000$	$N=100$	$N=1000$	$N=100$	$N=1000$
Linear regression	0.9543	0.9245	0.9750	1.0460	0.7397	1.0045
	(0.8387)	(0.2678)	(1.5546)	(0.4910)	(2.7256)	(0.7665)
	96.5%	95.0%	93.5%	95.5%	89.0%	93.5%
Traditional Semiparametrics						
Kernel (Speckman)	0.8947	0.9272	0.1460	0.1576	0.0010	0.0001
	(0.4809)	(0.1502)	(0.1424)	(0.0432)	(0.0060)	(0.0003)
	91.0%	93.0%	1.0%	0.0%	0.0%	0.0%
DML						
Kernel	1.0274	0.9656	0.9947	1.0477	0.6630	1.0150
	(0.6234)	(0.1641)	(1.3261)	(0.3242)	(2.4606)	(0.7019)
	94.5%	96.5%	94.5%	95.5%	95.0%	95.0%
Decision tree	1.0186	0.9701	0.9610	1.0392	0.8840	0.9903
	(0.7092)	(0.1762)	(1.4354)	(0.3112)	(2.4532)	(0.6386)
	93.0%	96.0%	94.5%	96.0%	93.5%	95.0%
Random forest	0.9838	0.9726	0.9468	1.0417	0.8968	1.0005
	(0.6299)	(0.1638)	(1.1674)	(0.2621)	(2.1562)	(0.5019)
	94.5%	95.5%	94.0%	94.5%	91.0%	93.0%
XGBoost	1.0404	0.9774	0.9719	1.0466	0.8882	1.0100
	(0.6835)	(0.1661)	(1.0906)	(0.2438)	(2.0976)	(0.3710)
	93.0%	96.5%	97.0%	94.0%	92.0%	95.0%

Table A18 Estimation results of DML estimators with varying numbers of cross-fitting folds in Scenario 1, using logistic regression for propensity score estimation (200 replications)

$p=2$	$K=1$	$K=2$	$K=5$	$K=10$
DML				
Kernel	0.9305	0.9738	0.9732	0.9743
	(0.1490)	(0.1587)	(0.1567)	(0.1559)
	93.5%	94.5%	95.0%	95.0%
Decision tree	0.9025	0.9691	0.9662	0.9687
	(0.1523)	(0.1712)	(0.1623)	(0.1609)
	92.0%	94.5%	94.5%	96.0%
Random forest	0.8987	0.9736	0.9740	0.9731
	(0.1467)	(0.1628)	(0.1559)	(0.1566)
	90.0%	95.0%	95.0%	95.0%
XGBoost	0.9117	0.9700	0.9682	0.9714
	(0.1483)	(0.1634)	(0.1613)	(0.1570)
	90.5%	94.0%	95.5%	96.0%

$p=5$	$K=1$	$K=2$	$K=5$	$K=10$
DML				
Kernel	0.1569	0.9984	1.0029	1.0053
	(0.0426)	(0.2240)	(0.2104)	(0.2031)
	0.0%	96.5%	95.5%	95.5%
Decision tree	0.7592	0.9927	0.9949	1.0003
	(0.1919)	(0.2747)	(0.2607)	(0.2647)
	71.5%	98.5%	97.0%	96.0%
Random forest	0.4832	0.9840	1.0034	1.0055
	(0.0966)	(0.1992)	(0.1900)	(0.1929)
	0.5%	98.5%	98.0%	98.0%
XGBoost	0.5434	0.9918	1.0071	1.0124
	(0.1077)	(0.2221)	(0.1937)	(0.2031)
	2.0%	96.5%	97.0%	97.0%

$p=10$	$K=1$	$K=2$	$K=5$	$K=10$
DML				
Kernel	0.0001	0.9656	0.9801	0.9806
	(0.0002)	(0.3734)	(0.3614)	(0.3447)
	0.0%	96.0%	95.0%	95.5%
Decision tree	0.9562	0.9476	0.9645	0.9685
	(0.3071)	(0.5082)	(0.4525)	(0.4538)
	99.5%	98.0%	98.5%	98.0%
Random forest	0.5576	0.9474	0.9674	0.9539
	(0.1621)	(0.3652)	(0.3361)	(0.3396)
	40.0%	97.0%	97.5%	98.5%
XGBoost	0.1349	0.9340	0.9356	0.9477
	(0.0818)	(0.4492)	(0.3666)	(0.3452)
	0.0%	94.0%	97.0%	97.5%

Table A19 Estimation results of DML estimators with varying numbers of cross-fitting folds in Scenario 2, using logistic regression for propensity score estimation (200 replications)

$p=2$	$K=1$	$K=2$	$K=5$	$K=10$
DML				
Kernel	0.9272	0.9656	0.9667	0.9656
	(0.1502)	(0.1730)	(0.1649)	(0.1641)
	93.0%	95.0%	96.5%	96.5%
Decision tree	0.8305	0.9662	0.9689	0.9701
	(0.1588)	(0.1865)	(0.1840)	(0.1762)
	80.0%	95.5%	94.5%	96.0%
Random forest	0.8350	0.9715	0.9715	0.9726
	(0.1457)	(0.1673)	(0.1661)	(0.1638)
	75.5%	96.0%	96.0%	95.5%
XGBoost	0.9018	0.9681	0.9739	0.9774
	(0.1511)	(0.1808)	(0.1687)	(0.1661)
	89.0%	95.5%	94.5%	96.5%

$p=5$	$K=1$	$K=2$	$K=5$	$K=10$
DML				
Kernel	0.1576	1.0336	1.0489	1.0477
	(0.0432)	(0.3526)	(0.3334)	(0.3242)
	0.0%	95.5%	95.5%	95.5%
Decision tree	0.6622	1.0446	1.0420	1.0392
	(0.1993)	(0.3378)	(0.3308)	(0.3112)
	57.5%	97.0%	94.0%	96.0%
Random forest	0.6114	1.0255	1.0425	1.0417
	(0.1376)	(0.2931)	(0.2722)	(0.2621)
	21.5%	94.0%	94.5%	94.5%
XGBoost	0.8133	1.0480	1.0427	1.0466
	(0.1539)	(0.2595)	(0.2347)	(0.2438)
	78.5%	94.0%	95.5%	94.0%

$p=10$	$K=1$	$K=2$	$K=5$	$K=10$
DML				
Kernel	0.0001	0.9953	1.0234	1.0150
	(0.0003)	(0.7297)	(0.6993)	(0.7019)
	0.0%	96.0%	95.5%	95.0%
Decision tree	0.6278	0.9781	1.0094	0.9903
	(0.4089)	(0.7040)	(0.6591)	(0.6386)
	81.0%	94.5%	93.5%	95.0%
Random forest	0.6516	0.9806	1.0128	1.0005
	(0.2863)	(0.5426)	(0.5137)	(0.5019)
	77.0%	94.5%	93.5%	93.0%
XGBoost	0.5005	1.0072	0.9928	1.0100
	(0.1427)	(0.4107)	(0.3521)	(0.3710)
	8.0%	96.0%	95.0%	95.0%

Table A20 Estimation results for $\theta_0 = 1$ from different estimators across 200 experiment replications

Scenario 1	$p=2$		$p=5$		$p=10$	
	$N=100$	$N=1000$	$N=100$	$N=1000$	$N=100$	$N=1000$
Linear regression	0.5577	0.5442	−0.5699	−0.5057	−2.2496	−2.3252
	(0.3812)	(0.1238)	(0.8889)	(0.2979)	(2.2641)	(0.7543)
	82.5%	4.0%	64.0%	0.0%	56.5%	0.5%
Traditional Semiparametrics						
Kernel (Backfitting)	1.5119	1.2679	1.3872	1.1825	-	-
	(0.2839)	(0.0790)	(0.4943)	(0.1379)	-	-
	-	-	-	-	-	-
Kernel (Speckman)	1.0397	0.9905	1.0703	1.0159	0.4241	0.9769
	(0.2704)	(0.0780)	(0.7607)	(0.1959)	(23.0645)	(2.3772)
	92.5%	95.0%	78.0%	89.0%	41.5%	33.5%
DML						
Kernel	1.0404	0.9894	1.1464	1.0383	1.5006	1.2597
	(0.2767)	(0.0783)	(0.5034)	(0.1239)	(1.1702)	(0.2604)
	93.0%	95.0%	86.5%	88.5%	81.0%	75.0%
Decision tree	0.8327	0.9130	0.6663	0.6881	1.1496	0.8214
	(0.3392)	(0.0911)	(0.6510)	(0.1909)	(1.3864)	(0.4142)
	85.5%	82.5%	93.3%	62.5%	95.9%	91.0%
Random forest	0.9914	0.9827	0.9191	0.9112	1.4559	0.8354
	(0.3192)	(0.0795)	(0.5583)	(0.1170)	(1.2585)	(0.2968)
	92.0%	94.5%	97.5%	96.0%	96.0%	95.0%
XGBoost	0.9685	0.9724	0.8658	0.9557	1.0043	0.8941
	(0.3003)	(0.0786)	(0.6290)	(0.1444)	(1.2164)	(0.3122)
	93.0%	94.5%	95.0%	92.5%	95.0%	94.5%

Scenario 2	$p=2$		$p=5$		$p=10$	
	$N=100$	$N=1000$	$N=100$	$N=1000$	$N=100$	$N=1000$
Linear regression	0.2300	0.1998	−0.0357	−0.0089	−0.3378	−0.1230
	(0.9121)	(0.2731)	(1.5904)	(0.4455)	(2.6538)	(0.7256)
	85.0%	16.5%	89.5%	45.5%	86.0%	69.0%
Traditional Semiparametrics						
Kernel (Backfitting)	1.8740	1.5754	1.2832	1.1632	-	-
	(0.3476)	(0.0894)	(0.5726)	(0.1518)	-	-
	-	-	-	-	-	-
Kernel (Speckman)	0.9775	0.9775	0.9785	0.9758	−3.8155	0.7934
	(0.2841)	(0.0786)	(0.8090)	(0.2077)	(37.0323)	(2.3699)
	94.0%	95.5%	82.0%	90.0%	35.0%	33.5%
DML						
Kernel	0.9147	0.9664	0.6739	0.8609	0.1363	0.4681
	(0.4770)	(0.1025)	(1.0982)	(0.2693)	(2.0523)	(0.5264)
	94.5%	97.0%	90.0%	89.5%	86.0%	81.5%
Decision tree	0.8875	0.9191	0.7090	0.7286	0.7197	0.5169
	(0.4668)	(0.1149)	(0.9923)	(0.2605)	(1.5374)	(0.4614)
	92.6%	88.0%	94.8%	78.5%	95.7%	75.8%
Random forest	0.9635	0.9854	1.0275	0.9922	0.9892	0.9555
	(0.4882)	(0.0918)	(0.9176)	(0.2314)	(1.5604)	(0.3577)
	97.5%	95.0%	97.5%	97.0%	96.0%	97.0%
XGBoost	0.9856	0.9636	0.8762	0.9601	0.5861	0.8849
	(0.4733)	(0.1014)	(0.8326)	(0.2349)	(1.5712)	(0.3142)
	95.0%	95.0%	97.5%	93.0%	92.5%	93.5%

Table A21 Estimation results of the DML estimators under orthogonal and their counterparts using non-orthogonal estimating equations, with different cross-fitting folds in Scenario 1 (across 200 experiment replications)

$p=2$	Orthogonal				Non-orthogonal			
	$K=1$	$K=2$	$K=5$	$K=10$	$K=1$	$K=2$	$K=5$	$K=10$
DML								
Kernel	0.9905	0.9886	0.9892	0.9894	1.2679	1.3556	1.3073	1.2956
	(0.0780)	(0.0826)	(0.0801)	(0.0783)	(0.0790)	(0.0838)	(0.0803)	(0.0800)
	95.0%	94.0%	94.0%	95.0%	-	-	-	-
Decision tree	0.8660	0.8878	0.9012	0.9130	0.8969	0.9608	0.9653	0.9671
	(0.0868)	(0.1033)	(0.0925)	(0.0911)	(0.0863)	(0.0974)	(0.0917)	(0.0914)
	59.5%	72.0%	80.0%	82.5%	34.0%	68.5%	65.0%	65.0%
Random forest	0.9359	0.9895	0.9836	0.9827	0.9816	1.0433	1.0364	1.0315
	(0.1378)	(0.0828)	(0.0787)	(0.0795)	(0.0884)	(0.0904)	(0.0862)	(0.0873)
	77.5%	95.5%	94.5%	94.5%	53.5%	67.5%	65.0%	66.0%
XGBoost	0.9257	0.9677	0.9687	0.9724	0.6268	0.9699	0.9834	0.9861
	(0.0861)	(0.0850)	(0.0816)	(0.0786)	(0.0641)	(0.0940)	(0.0943)	(0.0922)
	79.0%	90.5%	94.0%	94.5%	0.0%	60.0%	58.5%	64.0%

$p=5$	Orthogonal				Non-orthogonal			
	$K=1$	$K=2$	$K=5$	$K=10$	$K=1$	$K=2$	$K=5$	$K=10$
DML								
Kernel	1.0159	1.0558	1.0438	1.0383	1.1825	1.6424	1.5554	1.5389
	(0.1959)	(0.1385)	(0.1285)	(0.1239)	(0.1379)	(0.1480)	(0.1400)	(0.1401)
	89.0%	86.5%	88.0%	88.5%	-	-	-	-
Decision tree	0.5455	0.6370	0.6541	0.6881	0.6161	0.9564	0.9542	0.9631
	(0.1968)	(0.2442)	(0.2028)	(0.1909)	(0.1900)	(0.2604)	(0.2347)	(0.2221)
	19.2%	54.0%	53.5%	62.5%	11.1%	80.8%	78.9%	82.0%
Random forest	0.6384	0.8895	0.8999	0.9112	0.9671	1.0974	1.1005	1.1107
	(0.1896)	(0.1533)	(0.1278)	(0.1170)	(0.1156)	(0.1691)	(0.1267)	(0.1232)
	82.5%	88.0%	91.0%	96.0%	43.5%	75.5%	78.5%	76.5%
XGBoost	0.7351	0.9317	0.9551	0.9557	0.6255	0.9157	0.9361	0.9278
	(0.7723)	(0.1643)	(0.1322)	(0.1444)	(0.1014)	(0.1588)	(0.1241)	(0.1274)
	8.5%	90.0%	94.5%	92.5%	0.0%	66.0%	74.0%	68.5%

$p=10$	Orthogonal				Non-orthogonal			
	$K=1$	$K=2$	$K=5$	$K=10$	$K=1$	$K=2$	$K=5$	$K=10$
DML								
Kernel	0.9769	1.3194	1.2715	1.2597	-	-	-	-
	(2.3772)	(0.2618)	(0.2730)	(0.2605)	-	-	-	-
	33.5%	72.0%	71.5%	75.0%	-	-	-	-
Decision tree	1.5321	0.7518	0.7941	0.8214	1.7057	1.3554	1.4766	1.4892
	(0.8160)	(0.4499)	(0.4341)	(0.4142)	(0.6872)	(0.5254)	(0.4454)	(0.4296)
	80.9%	86.3%	89.9%	91.0%	28.4%	74.0%	69.0%	65.2%
Random forest	0.7362	0.7730	0.8496	0.8354	0.8813	1.0873	1.0765	1.0686
	(0.2861)	(0.3511)	(0.2848)	(0.2950)	(0.2732)	(0.3887)	(0.2965)	(0.2820)
	89.0%	92.5%	95.5%	95.0%	45.5%	84.0%	91.5%	92.5%
XGBoost	0.4690	0.8549	0.8794	0.8941	0.5984	0.9493	0.9752	0.9621
	(0.9081)	(0.4058)	(0.3075)	(0.3122)	(0.1867)	(0.3762)	(0.2817)	(0.2605)
	8.5%	90.5%	95.0%	94.5%	2.5%	82.5%	88.5%	90.5%

Table A22 Estimation results of the DML estimators under orthogonal and their counterparts using non-orthogonal estimating equations, with different cross-fitting folds in Scenario 2 (across 200 experiment replications)

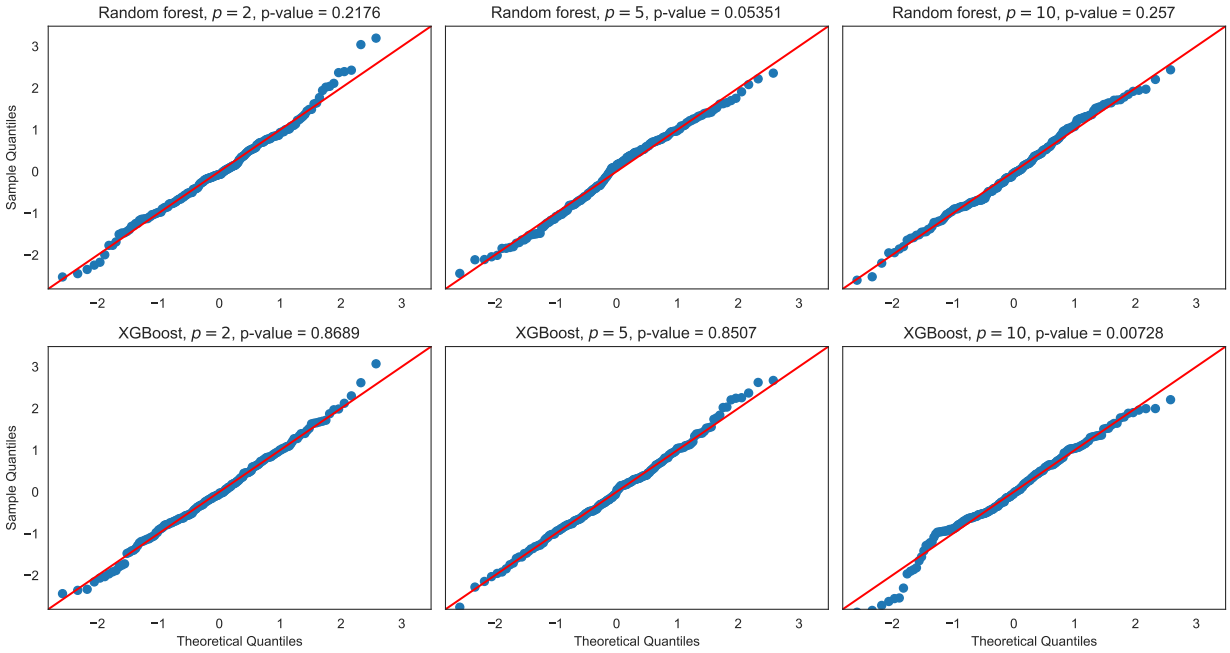
$p=2$	Orthogonal				Non-orthogonal			
	$K=1$	$K=2$	$K=5$	$K=10$	$K=1$	$K=2$	$K=5$	$K=10$
DML								
Kernel	0.9775	0.9593	0.9663	0.9664	1.5754	1.7993	1.7030	1.6789
	(0.0786)	(0.1272)	(0.1070)	(0.1025)	(0.0894)	(0.1096)	(0.1007)	(0.0988)
	95.5%	92.5%	96.0%	97.0%	-	-	-	-
Decision tree	0.7625	0.8899	0.9069	0.9191	0.7931	0.9704	0.9757	0.9763
	(0.0962)	(0.1354)	(0.1215)	(0.1149)	(0.1051)	(0.1420)	(0.1182)	(0.1167)
	22.0%	83.0%	86.0%	89.0%	8.0%	63.5%	71.5%	76.0%
Random forest	0.8382	0.9843	0.9809	0.9854	0.9808	1.0625	1.0542	1.0501
	(0.1401)	(0.0905)	(0.0927)	(0.0918)	(0.0954)	(0.1201)	(0.1020)	(0.1041)
	38.5%	98.0%	95.5%	95.0%	51.5%	66.0%	72.0%	69.0%
XGBoost	0.9262	0.9559	0.9607	0.9636	0.6677	0.9758	0.9727	0.9822
	(0.0897)	(0.1251)	(0.1072)	(0.1014)	(0.0751)	(0.1238)	(0.1222)	(0.1148)
	85.0%	93.0%	96.0%	95.0%	0.5%	65.0%	69.5%	67.5%

$p=5$	Orthogonal				Non-orthogonal			
	$K=1$	$K=2$	$K=5$	$K=10$	$K=1$	$K=2$	$K=5$	$K=10$
DML								
Kernel	0.9758	0.8194	0.8540	0.8609	1.1632	2.5498	2.3868	2.3487
	(0.2077)	(0.2855)	(0.2702)	(0.2693)	(0.1518)	(0.1873)	(0.1815)	(0.1845)
	90.0%	91.0%	91.0%	89.5%	-	-	-	-
Decision tree	0.5553	0.7325	0.7449	0.7286	0.5973	0.9394	0.9370	0.9066
	(0.2165)	(0.3103)	(0.2677)	(0.2605)	(0.1944)	(0.3253)	(0.2620)	(0.2861)
	36.7%	80.3%	82.4%	78.5%	9.6%	63.6%	71.2%	68.3%
Random forest	0.7861	0.9854	0.9959	0.9922	1.0418	1.2853	1.2467	1.2416
	(0.1838)	(0.2805)	(0.2232)	(0.2314)	(0.2202)	(0.2765)	(0.2420)	(0.2341)
	57.5%	97.0%	98.0%	97.0%	33.0%	40.0%	42.5%	42.0%
XGBoost	1.0516	0.9582	0.9531	0.9601	0.7320	0.9750	0.9367	0.9256
	(0.9500)	(0.2229)	(0.2027)	(0.2349)	(0.1166)	(0.2338)	(0.2095)	(0.2343)
	67.5%	94.5%	95.5%	93.0%	5.0%	59.0%	59.5%	61.0%

$p=10$	Orthogonal				Non-orthogonal			
	$K=1$	$K=2$	$K=5$	$K=10$	$K=1$	$K=2$	$K=5$	$K=10$
DML								
Kernel	0.7934	0.4076	0.4807	0.4681	-	-	-	-
	(2.3699)	(0.5702)	(0.5244)	(0.5264)	-	-	-	-
	33.5%	76.5%	82.0%	81.5%	-	-	-	-
Decision tree	0.7432	0.5456	0.5454	0.5169	0.6434	0.8979	0.8362	0.8182
	(1.2219)	(0.4841)	(0.4286)	(0.4614)	(0.7615)	(0.5650)	(0.4917)	(0.5024)
	78.5%	81.1%	79.9%	75.8%	29.5%	69.0%	76.1%	68.5%
Random forest	1.1815	0.9712	0.9602	0.9555	1.0464	1.3650	1.3794	1.3510
	(0.4762)	(0.3998)	(0.3469)	(0.3577)	(0.3447)	(0.4515)	(0.3784)	(0.3625)
	95.0%	96.5%	98.0%	97.0%	47.0%	53.0%	48.0%	55.5%
XGBoost	1.5755	0.8295	0.8707	0.8849	0.6981	0.8408	0.8069	0.8107
	(2.4058)	(0.3524)	(0.3138)	(0.3142)	(0.2255)	(0.4468)	(0.3672)	(0.3820)
	49.0%	95.0%	95.5%	93.5%	8.0%	46.5%	55.0%	52.0%

Figure A4 **Q–Q plots of the DML estimator obtained from (a) a non-orthogonal estimating equation and (b) an orthogonal estimating equation. The associated p-value is computed using the Shapiro–Wilk normality test.**

(a) Non-orthogonal Estimating Equation



(b) Orthogonal Estimating Equation

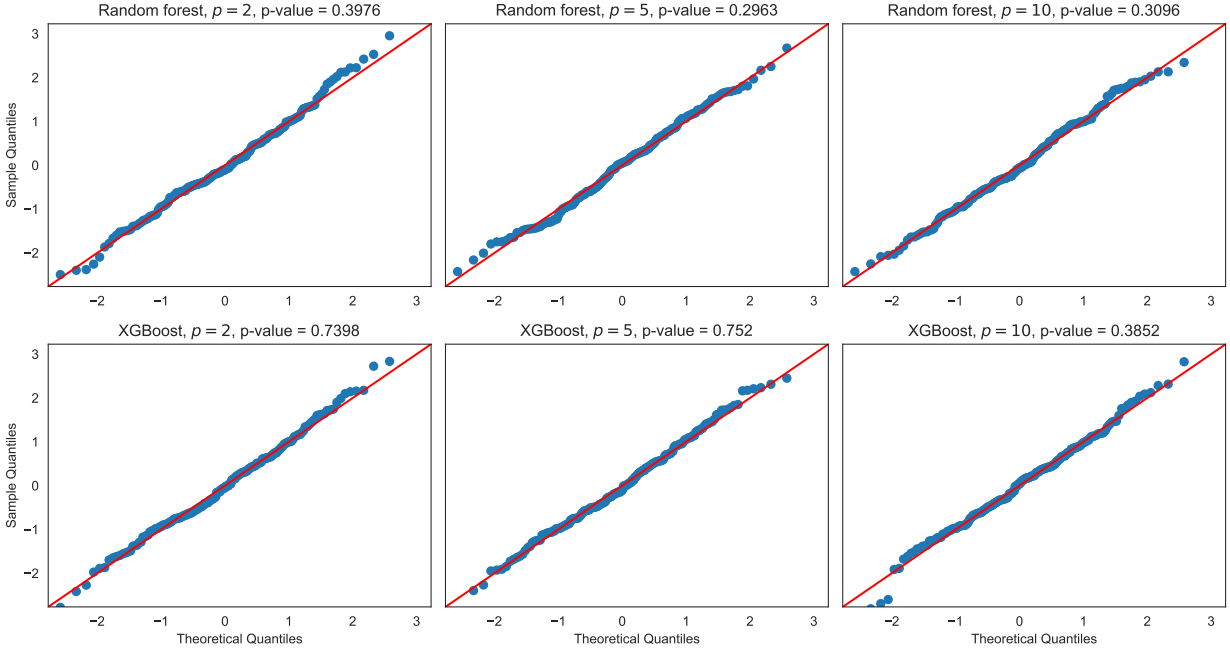
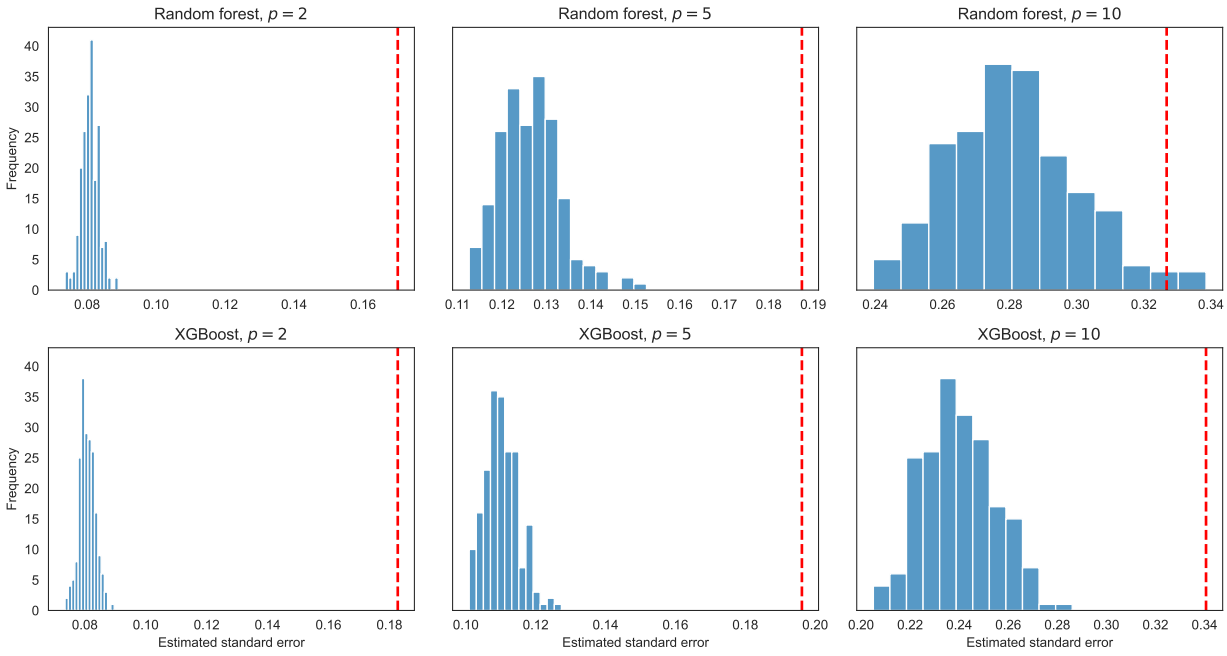
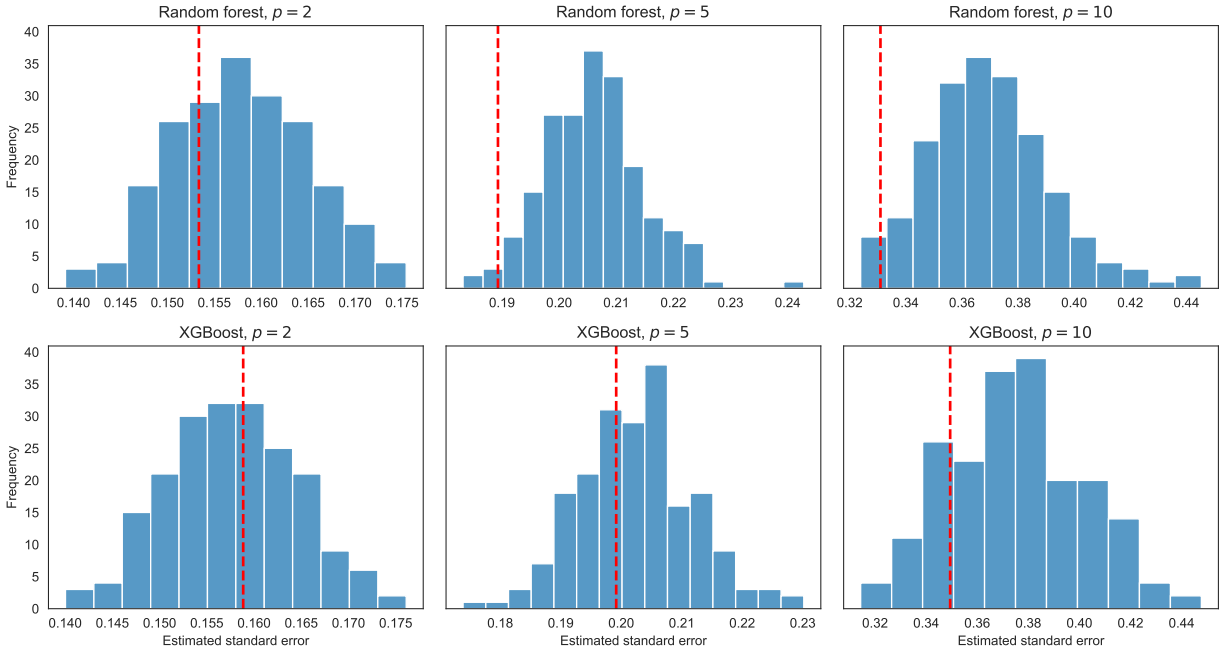


Figure A5 **The empirical distribution of the estimated standard errors (blue histogram) is compared with the Monte Carlo approximation of the true standard error across simulations (red dashed line), for (a) the estimator based on a non-orthogonal estimating equation and (b) the counterpart DML estimator based on an orthogonal estimating equation.**

(a) Non-orthogonal Estimating Equation



(b) Orthogonal Estimating Equation



References

Abadie A (2005) Semiparametric difference-in-differences estimators. *The review of economic studies* 72(1):1–19.

Ahrens A, Hansen CB, Schaffer ME, Wiemann T (2024a) ddml: Double/debiased machine learning in stata. *The Stata Journal* 24(1):3–45.

Ahrens A, Hansen CB, Schaffer ME, Wiemann T (2024b) Model averaging and double machine learning. *arXiv preprint arXiv:2401.01645* .

Angrist JD, Pischke JS (2009) *Mostly harmless econometrics: An empiricist's companion* (Princeton university press).

Arkhangelsky D, Athey S, Hirshberg DA, Imbens GW, Wager S (2021) Synthetic difference-in-differences. *American Economic Review* 111(12):4088–4118.

Arkhangelsky D, Imbens GW (2022) Doubly robust identification for causal panel data models. *The Econometrics Journal* 25(3):649–674.

Bach P, Chernozhukov V, Kurz MS, Spindler M (2021) Doubleml–an object-oriented implementation of double machine learning in r. *arXiv preprint arXiv:2103.09603* .

Bach P, Chernozhukov V, Kurz MS, Spindler M (2022) Doubleml: an object-oriented implementation of double machine learning in python. *The Journal of Machine Learning Research* 23(1):2469–2474.

Bach P, Schacht O, Chernozhukov V, Klaassen S, Spindler M (2024) Hyperparameter tuning for causal inference with double machine learning: A simulation study. *Causal Learning and Reasoning*, 1065–1117 (PMLR).

Bang H, Robins JM (2005) Doubly robust estimation in missing data and causal inference models. *Biometrics* 61(4):962–973.

Battocchi K, Dillon E, Hei M, Lewis G, Oprescu M, Syrgkanis V (2021) Estimating the long-term effects of novel treatments. *Advances in Neural Information Processing Systems* 34:2925–2935.

Belloni A, Chernozhukov V, Fernandez-Val I, Hansen C (2017) Program evaluation and causal inference with high-dimensional data. *Econometrica* 85(1):233–298.

Belloni A, Chernozhukov V, Hansen C (2014) Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies* 81(2):608–650.

Ben-Michael E, Feller A, Rothstein J (2021) The augmented synthetic control method. *Journal of the American Statistical Association* 116(536):1789–1803.

Benkeser D, Ran J (2021) Nonparametric inference for interventional effects with multiple mediators. *Journal of Causal Inference* 9(1):172–189.

Bennett A, Kallus N, Mao X, Newey W, Syrgkanis V, Uehara M (2023a) Inference on strongly identified functionals of weakly identified functions.

Bennett A, Kallus N, Mao X, Newey W, Syrgkanis V, Uehara M (2023b) Source condition double robust inference on functionals of inverse problems.

Beraja M, Kao A, Yang DY, Yuchtman N (2023) Ai-tocracy. *The Quarterly Journal of Economics* 138(3):1349–1402.

Bia M, Huber M, Lafférs L (2024) Double machine learning for sample selection models. *Journal of Business & Economic Statistics* 42(3):958–969.

Bickel PJ (1982) On adaptive estimation. *The Annals of Statistics* 647–671.

Bickel PJ, Klaassen CA, Bickel PJ, Ritov Y, Klaassen J, Wellner JA, Ritov Y (1993) *Efficient and adaptive estimation for semiparametric models*, volume 4 (Springer).

Bodory H, Huber M, Lafférs L (2022) Evaluating (weighted) dynamic treatment effects by double machine learning. *The Econometrics Journal* 25(3):628–648.

Bonaccolto-Töpfer M, Satlukal S (2024) Gender differences in reservation wages: New evidence for germany. *Labour Economics* 91:102649.

Bradic J, Ji W, Zhang Y (2024) High-dimensional inference for dynamic treatment effects. *The Annals of Statistics* 52(2):415–440.

Caetano C, Callaway B (2024) Difference-in-differences when parallel trends holds conditional on covariates. *arXiv preprint arXiv:2406.15288* .

Caetano C, Callaway B, Payne S, Rodrigues HS (2022) Difference in differences with time-varying covariates. *arXiv preprint arXiv:2202.02903* .

Callaway B, Sant'Anna PH (2021) Difference-in-differences with multiple time periods. *Journal of econometrics* 225(2):200–230.

Card D, Krueger AB (1993) Minimum wages and employment: A case study of the fast food industry in new jersey and pennsylvania.

Chang NC (2020) Double/debiased machine learning for difference-in-differences models. *The Econometrics Journal* 23(2):177–191.

Chen J, Ritzwoller DM (2023) Semiparametric estimation of long-term treatment effects. *Journal of Econometrics* 237(2):105545.

Chen Q, Syrgkanis V, Austern M (2022) Debiased machine learning without sample-splitting for stable estimators. *Advances in Neural Information Processing Systems* 35:3096–3109.

Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, Newey W, Robins J (2018) Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal* 21(1).

Chernozhukov V, Cinelli C, Newey W, Sharma A, Syrgkanis V (2022a) Long story short: Omitted variable bias in causal machine learning. Technical report, National Bureau of Economic Research.

Chernozhukov V, Escanciano JC, Ichimura H, Newey WK, Robins JM (2022b) Locally robust semiparametric estimation. *Econometrica* 90(4):1501–1535.

Chernozhukov V, Hansen C, Spindler M (2015a) Post-selection and post-regularization inference in linear models with many controls and instruments. *American Economic Review* 105(5):486–490.

Chernozhukov V, Hansen C, Spindler M (2015b) Valid post-selection and post-regularization inference: An elementary, general approach. *Annu. Rev. Econ.* 7(1):649–688.

Chernozhukov V, Kasahara H, Schrimpf P (2021a) Causal impact of masks, policies, behavior on early covid-19 pandemic in the us. *Journal of econometrics* 220(1):23–62.

Chernozhukov V, Newey M, Newey WK, Singh R, Srygkanis V (2023) Automatic debiased machine learning for covariate shifts. *arXiv preprint arXiv:2307.04527* .

Chernozhukov V, Newey M, Newey WK, Singh R, Srygkanis V (2025) Automatic debiased machine learning for covariate shifts. URL https://arxiv.org/abs/2307.04527.

Chernozhukov V, Newey W, Quintas-Martınez VM, Syrgkanis V (2022c) Riesznet and forestriesz: Automatic debiased machine learning with neural nets and random forests. *International Conference on Machine Learning*, 3901–3914 (PMLR).

Chernozhukov V, Newey W, Singh R, Syrgkanis V (2022d) Automatic debiased machine learning for dynamic treatment effects and general nested functionals. *arXiv preprint arXiv:2203.13887* .

Chernozhukov V, Newey WK, Quintas-Martinez V, Syrgkanis V (2021b) Automatic debiased machine learning via neural nets for generalized linear regression. *arXiv preprint arXiv:2104.14737* .

Chernozhukov V, Newey WK, Singh R (2022e) Debiased machine learning of global and local parameters using regularized riesz representers. *The Econometrics Journal* 25(3):576–601.

Chiang HD, Kato K, Ma Y, Sasaki Y (2021) Multiway cluster robust double/debiased machine learning. *Journal of Business & Economic Statistics* 39(4):1080–1080.

Clarke P, Polselli A (2023) Double machine learning for static panel models with fixed effects. *arXiv preprint arXiv:2312.08174* .

Colangelo K, Lee YY (2020) Double debiased machine learning nonparametric inference with continuous treatments. *arXiv preprint arXiv:2004.03036* .

Cui Y, Pu H, Shi X, Miao W, Tchetgen Tchetgen E (2024) Semiparametric proximal causal inference. *Journal of the American Statistical Association* 119(546):1348–1359.

Cui Y, Tchetgen Tchetgen E (2024) Selective machine learning of doubly robust functionals. *Biometrika* 111(2):517–535.

De Chaisemartin C, d'Haultfoeuille X (2023) Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: A survey. *The Econometrics Journal* 26(3):C1–C30.

Devlin J, Chang MW, Lee K, Toutanova K (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* .

Díaz I (2019) Statistical inference for data-adaptive doubly robust estimators with survival outcomes. *Statistics in medicine* 38(15):2735–2748.

Díaz I (2020) Machine learning in the estimation of causal effects: targeted minimum loss-based estimation and double/debiased machine learning. *Biostatistics* 21(2):353–358.

Díaz I, Williams N, Hoffman KL, Schenck EJ (2023) Nonparametric causal effects based on longitudinal modified treatment policies. *Journal of the American Statistical Association* 118(542):846–857.

Dorn J, Guo K, Kallus N (2021) Doubly-valid/doubly-sharp sensitivity analysis for causal inference with unmeasured confounding. *arXiv preprint arXiv:2112.11449* .

Ellickson PB, Kar W, Reeder III JC (2023) Estimating marketing component effects: Double machine learning from targeted digital promotions. *Marketing Science* 42(4):704–728.

Ellul S, Carlin JB, Vansteelandt S, Moreno-Betancur M (2024) Causal machine learning methods and use of sample splitting in settings with high-dimensional confounding. *arXiv preprint arXiv:2405.15242* .

Farbmacher H, Huber M, Lafférs L, Langen H, Spindler M (2022) Causal mediation analysis with double machine learning. *The Econometrics Journal* 25(2):277–300.

Farrell MH (2015) Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics* 189(1):1–23.

Farrell MH, Liang T, Misra S (2021a) Deep learning for individual heterogeneity: An automatic inference framework. URL https://arxiv.org/abs/2010.14694.

Farrell MH, Liang T, Misra S (2021b) Deep neural networks for estimation and inference. *Econometrica* 89(1):181–213.

Feng X, Ma X, Lu J, Tang Q, Chen Z (2025) Assessing the impact of the digital economy on sustainable development in the underdeveloped regions of western china. *Cities* 156:105552.

Feyzollahi M, Rafizadeh N (2024) Double/Debiased Machine Learning for Economists: Practical Guidelines, Best Practices, and Common Pitfalls. URL http://dx.doi.org/10.2139/ssrn.4703243.

Fingerhut N, Sesia M, Romano Y (2022) Coordinated double machine learning. *International Conference on Machine Learning*, 6499–6513 (PMLR).

Fisher A, Kennedy EH (2021) Visually communicating and teaching intuition for influence functions. *The American Statistician* 75(2):162–172.

Fong C, Tyler M (2021) Machine learning predictions as regression covariates. *Political Analysis* 29(4):467–484.

Fuhr J, Berens P, Papies D (2024) Estimating causal effects with double machine learning–a method evaluation. *arXiv preprint arXiv:2403.14385* .

Fuhr J, Papies D (2024) Double machine learning meets panel data–promises, pitfalls, and potential solutions. *arXiv preprint arXiv:2409.01266* .

Ghassami A, Ying A, Shpitser I, Tchetgen ET (2022) Minimax kernel machine learning for a class of doubly robust functionals with application to proximal causal inference. *International conference on artificial intelligence and statistics*, 7210–7239 (PMLR).

Giaquinta M, Hildebrandt S (2004) *Calculus of Variations I*. Grundlehren der mathematischen Wissenschaften (Springer Berlin Heidelberg).

Giraldo C, Giraldo I, Gomez-Gonzalez JE, Uribe JM (2024) Financial integration and banking stability: A post-global crisis assessment. *Economic Modelling* 139:106835.

Goh KY, Heng CS, Lin Z (2013) Social media brand community and consumer behavior: Quantifying the relative impact of user-and marketer-generated content. *Information systems research* 24(1):88–107.

Gordon BR, Moakler R, Zettelmeyer F (2023) Close enough? a large-scale exploration of non-experimental approaches to advertising measurement. *Marketing Science* 42(4):768–793.

Haddad MF, Huber M, Zhang LZ (2024) Difference-in-differences with time-varying continuous treatments using double/debiased machine learning. *arXiv preprint arXiv:2410.21105* .

Hines CL, Hines OJ (2024) Automatic debiasing of neural networks via moment-constrained learning. *arXiv preprint arXiv:2409.19777* .

Hines O, Diaz-Ordaz K, Vansteelandt S (2023) Optimally weighted average derivative effects. *arXiv preprint arXiv:2308.05456* .

Hines O, Dukes O, Diaz-Ordaz K, Vansteelandt S (2022) Demystifying statistical learning based on efficient influence functions. *The American Statistician* 76(3):292–304.

Hoffmann NI (2024) Double robust, flexible adjustment methods for causal inference: an overview and an evaluation. *UCLA Thesis* .

Holland PW (1986) Statistics and causal inference. *Journal of the American statistical Association* 81(396):945–960.

Hsu YC, Huber M, Yen YM (2023) Doubly robust estimation of direct and indirect quantile treatment effects with machine learning. URL https://arxiv.org/abs/2307.01049.

Ibragimov I, Has' minskii R, et al. (1981) Statistical estimation: Asymptotic theory. *Springer Book Archive-Mathematics* .

Imai K, Khanna K (2016) Improving ecological inference by predicting individual ethnicity from voter registration records. *Political Analysis* 24(2):263–272.

Imbens G (2014) Instrumental variables: an econometrician's perspective. Technical report, National Bureau of Economic Research.

Imbens G, Kallus N, Mao X, Wang Y (2024) Long-term causal inference under persistent confounding via data combination. *Journal of the Royal Statistical Society Series B: Statistical Methodology* qkae095.

Imbens GW, Rubin DB (2015) *Causal inference in statistics, social, and biomedical sciences* (Cambridge University Press).

Jung Y, Díaz I, Tian J, Bareinboim E (2024) Estimating causal effects identifiable from a combination of observations and experiments. *Advances in Neural Information Processing Systems* 36.

Jung Y, Tian J, Bareinboim E (2021) Estimating identifiable causal effects through double machine learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 12113–12122.

Kallus N, Mao X (2024) On the role of surrogates in the efficient estimation of treatment effects with limited outcome data. *Journal of the Royal Statistical Society Series B: Statistical Methodology* qkae099.

Kallus N, Mao X, Uehara M (2021) Causal inference under unmeasured confounding with negative controls: A minimax learning approach. *arXiv preprint arXiv:2103.14029* .

Kallus N, Mao X, Uehara M (2024) Localized debiased machine learning: Efficient inference on quantile treatment effects and beyond. *Journal of Machine Learning Research* 25(16):1–59.

Kato M (2023) Covariate shift adaptation robust to density-ratio estimation. *arXiv preprint arXiv:2310.16638* .

Kennedy EH (2019) Nonparametric causal effects based on incremental propensity score interventions. *Journal of the American Statistical Association* 114(526):645–656.

Kennedy EH (2022) Semiparametric doubly robust targeted double machine learning: a review. *arXiv preprint arXiv:2203.06469* .

Kennedy EH, Ma Z, McHugh MD, Small DS (2017) Non-parametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 79(4):1229–1245.

Klaassen S, Teichert-Kluge J, Bach P, Chernozhukov V, Spindler M, Vijaykumar S (2024) Doublemldeep: Estimation of causal effects with multimodal data. *arXiv preprint arXiv:2402.01785* .

Klosin S (2021) Automatic double machine learning for continuous treatment effects. *arXiv preprint arXiv:2104.10334* .

Klyne H, Shah RD (2023) Average partial effect estimation using double machine learning. *arXiv preprint arXiv:2308.09207* .

Knaus MC (2022) Double machine learning-based programme evaluation under unconfoundedness. *The Econometrics Journal* 25(3):602–627.

Lee Y, Kennedy EH, Mitra N (2023) Doubly robust nonparametric instrumental variable estimators for survival outcomes. *Biostatistics* 24(2):518–537.

Levit BY (1976) On the efficiency of a class of non-parametric estimates. *Theory of Probability & Its Applications* 20(4):723–740.

Lewis G, Syrgkanis V (2020) Double/debiased machine learning for dynamic treatment effects via g-estimation. *arXiv preprint arXiv:2002.07285* .

Li HH, Owen AB (2024) Double machine learning and design in batch adaptive experiments. *Journal of Causal Inference* 12(1):20230068.

Lin L, Khamaru K, Wainwright MJ (2023) Semi-parametric inference based on adaptively collected data. URL https://arxiv.org/abs/2303.02534.

Ling S, Jin S, Wang H, Zhang Z, Feng Y (2024) Transportation infrastructure upgrading and green development efficiency: Empirical analysis with double machine learning method. *Journal of Environmental Management* 358:120922.

Little RJ, Rubin DB (2019) *Statistical analysis with missing data*, volume 793 (John Wiley & Sons).

Liu F, Liu G, Wang X, Feng Y (2024a) Whether the construction of digital government alleviate resource curse? empirical evidence from chinese cities. *Resources Policy* 90:104811.

Liu L, Mukherjee R, Newey WK, Robins JM (2017) Semiparametric efficient empirical higher order influence function estimators. *arXiv preprint arXiv:1705.07577* .

Liu L, Mukherjee R, Robins JM, Tchetgen ET (2021a) Adaptive estimation of nonparametric functionals. *Journal of Machine Learning Research* 22(99):1–66.

Liu M, Zhang Y, Zhou D (2021b) Double/debiased machine learning for logistic partially linear model. *The Econometrics Journal* 24(3):559–588.

Liu R, Williams NT, Rudolph KE, Díaz I (2024b) General targeted machine learning for modern causal mediation analysis. *arXiv preprint arXiv:2408.14620* .

Liu X, Huang KW (2024) Controlling homophily in social network regression analysis by machine learning. *INFORMS Journal on Computing* .

Mackey L, Syrgkanis V, Zadik I (2018) Orthogonal machine learning: Power and limitations. *International Conference on Machine Learning*, 3375–3383 (PMLR).

Manzoor E, Chen GH, Lee D, Smith MD (2023) Influence via ethos: On the persuasive power of reputation in deliberation online. *Management Science* .

Meza I, Singh R (2021) Nested nonparametric instrumental variable regression: Long term, mediated, and time varying treatment effects. *arXiv preprint arXiv:2112.14249* .

Moccia C, Moirano G, Popovic M, Pizzi C, Fariselli P, Richiardi L, Ekstrøm CT, Maule M (2024) Machine learning in causal inference for epidemiology. *European Journal of Epidemiology* 1–12.

Molinari F (2020) Microeconometrics with partial identification. *Handbook of econometrics* 7:355–486.

Morenz ER, Wolock CJ, Carone M (2024) Debiased machine learning for counterfactual survival functionals based on left-truncated right-censored data. *arXiv preprint arXiv:2411.09017* .

Naimi AI, Mishler AE, Kennedy EH (2023) Challenges in obtaining valid causal effect estimates with machine learning algorithms. *American Journal of Epidemiology* 192(9):1536–1544.

Newey WK (1990) Semiparametric efficiency bounds. *Journal of Applied Econometrics* 5(2):99–135.

Newey WK (1994) The asymptotic variance of semiparametric estimators. *Econometrica: Journal of the Econometric Society* 1349–1382.

Newey WK, Hsieh F, Robins JM (2004) Twicing kernels and a small bias property of semiparametric estimators. *Econometrica* 72(3):947–962.

Newey WK, Robins JR (2018) Cross-fitting and fast remainder rates for semiparametric estimation. *arXiv preprint arXiv:1801.09138* .

Pang S, Hua G (2024) How does digital tax administration affect r&d manipulation? evidence from dual machine learning. *Technological Forecasting and Social Change* 208:123691.

Parikh H, Morucci M, Orlandi V, Roy S, Rudin C, Volfovsky A (2023) A double machine learning approach to combining experimental and observational data. *arXiv preprint arXiv:2307.01449* .

Pearl J (2009) *Causality* (Cambridge university press).

Qiao M, Huang KW (2021) Correcting misclassification bias in regression models with variables generated via data mining. *Information Systems Research* 32(2):462–480.

Robins JM, Rotnitzky A (1995) Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association* 90(429):122–129.

Robinson PM (1988) Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society* 931–954.

Roth J, Sant'Anna PH, Bilinski A, Poe J (2023) What's trending in difference-in-differences? a synthesis of the recent econometrics literature. *Journal of Econometrics* .

Rudolph KE, Williams N, Díaz I (2024a) Using instrumental variables to address unmeasured confounding in causal mediation analysis. *Biometrics* 80(1):ujad037.

Rudolph KE, Williams NT, Diaz I (2024b) Practical causal mediation analysis: extending nonparametric estimators to accommodate multiple mediators and multiple intermediate confounders. *Biostatistics* kxae012.

Sant'Anna PH, Zhao J (2020) Doubly robust difference-in-differences estimators. *Journal of Econometrics* 219(1):101–122.

Schindl K, Shen S, Kennedy EH (2024) Incremental effects for continuous exposures. *arXiv preprint arXiv:2409.11967* .

Scott DW (2015) *Multivariate density estimation: theory, practice, and visualization* (John Wiley & Sons).

Semenova V (2023a) Debiased machine learning of set-identified linear models. *Journal of Econometrics* 235(2):1725–1746.

Semenova V (2023b) Generalized lee bounds. URL https://arxiv.org/abs/2008.12720.

Semenova V (2024) Aggregated intersection bounds and aggregated minimax values. URL https://arxiv.org/abs/2303.00982.

Shan J, Li W, Ai C (2024) Efficient nonparametric inference of causal mediation effects with nonignorable missing confounders. *arXiv preprint arXiv:2402.05384* .

Singh R, Sun L (2024) Double robustness for complier parameters and a semi-parametric test for complier characteristics. *The Econometrics Journal* 27(1):1–20.

Sun B, Cui Y, Tchetgen ET (2022) Selective machine learning of the average treatment effect with an invalid instrumental variable. *Journal of Machine Learning Research* 23(204):1–40.

Tamer E (2010) Partial identification in econometrics. *Annu. Rev. Econ.* 2(1):167–195.

Tchetgen EJT, Ying A, Cui Y, Shi X, Miao W (2024) An Introduction to Proximal Causal Inference. *Statistical Science* 39(3):375 – 390, URL http://dx.doi.org/10.1214/23-STS911.

Tsiatis AA (2006) *Semiparametric theory and missing data* (Springer).

Van der Laan MJ, Rose S (2018) *Targeted learning in data science* (Springer).

Van der Laan MJ, Rose S, et al. (2011) *Targeted learning: causal inference for observational and experimental data*, volume 4 (Springer).

Van Der Laan MJ, Rubin D (2006) Targeted maximum likelihood learning. *The international journal of biostatistics* 2(1).

Van der Vaart A (1991) On differentiable functionals. *The Annals of Statistics* 178–204.

Van der Vaart AW (2000) *Asymptotic statistics*, volume 3 (Cambridge university press).

Vansteelandt S, Dukes O (2022) Assumption-lean inference for generalised linear model parameters. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 84(3):657–685.

Vansteelandt S, Dukes O, Van Lancker K, Martinussen T (2024) Assumption-lean cox regression. *Journal of the American Statistical Association* 119(545):475–484.

Wang C, Wu Q (2019) Flo: Fast and lightweight hyperparameter optimization for automl. *arXiv preprint arXiv:1911.04706* .

Wang L, Cheng Z (2024) Impact of the belt and road initiative on enterprise green transformation. *Journal of Cleaner Production* 468:143043.

Wang L, Lyu J, Zhang J (2024) Explicating the role of agricultural socialized services on chemical fertilizer use reduction: Evidence from china using a double machine learning model. *Agriculture* 14(12):2148.

Wooldridge JM (2010) *Econometric analysis of cross section and panel data* (MIT press).

Xie T, Ge Y, Kannan K (2025) Crypto airdrop success blueprint: A high-dimensional causal study using double machine learning. *Available at SSRN 5215263* .

Xu S, Liu L, Liu Z (2022a) Deepmed: Semiparametric causal mediation analysis with debiased deep learning. *Advances in Neural Information Processing Systems* 35:28238–28251.

Xu Y, Ghose A, Xiao B (2024) Mobile payment adoption: An empirical investigation of alipay. *Information Systems Research* 35(2):807–828.

Xu Y, Shi C, Luo S, Wang L, Song R (2022b) Quantile off-policy evaluation via deep conditional generative learning. *arXiv preprint arXiv:2212.14466* .

Yang J, Shao Y, Liu J, Wang L (2025) Double machine learning for partially linear mediation models with high-dimensional confounders. *Neurocomputing* 614:128766.

Yang M, Adomavicius G, Burtch G, Ren Y (2018) Mind the gap: Accounting for measurement error and misclassification in variables generated via data mining. *Information Systems Research* 29(1):4–24.

Yang M, McFowland III E, Burtch G, Adomavicius G (2022) Achieving reliable causal inference with data-mined variables: A random forest approach to the measurement error problem. *INFORMS Journal on Data Science* .

Yang M, Ren Y, Adomavicius G (2019) Understanding user-generated content and customer engagement on facebook business pages. *Information Systems Research* 30(3):839–855.

Yang Y, Lian D, Zhang Y, Wang D, Wang J (2024) Analysis of the impact of resource misallocation and socialized services on low-carbon agricultural production with dml based on random forest. *International Review of Economics & Finance* 95:103452.

Yin X, Li J, Wu J, Cao R, Xin S, Liu J (2024) Impacts of geographical indications on agricultural growth and farmers' income in rural china. *Agriculture* 14(1):113.

Zhang L (2024) Continuous difference-in-differences with double/debiased machine learning. *arXiv preprint arXiv:2408.10509* .

Zhang S, Lee D, Singh PV, Srinivasan K (2022) What makes a good image? airbnb demand analytics leveraging interpretable image features. *Management Science* 68(8):5644–5666.

Zhao L, Liu G, Jiao H, Hu S, Feng Y (2024) China's endeavor to reduce energy intensity: Does the green financial reform and innovation pilot zones policy matter? *Journal of environmental management* 370:122631.

Zhou M, Abhishek V, Kennedy EH, Srinivasan K, Sinha R (2024) Linking clicks to bricks: Understanding the effects of email advertising on multichannel sales. *Information Systems Research* .

Zimmert M (2020) Efficient difference-in-differences estimation with high-dimensional common trend confounding. *arXiv preprint arXiv:1809.01643* .

Zivich PN, Breskin A (2021) Machine learning for causal inference: on the use of cross-fit estimators. *Epidemiology* 32(3):393–401.

Zou T, Li F, Guo P (2024) Advancing effective energy transition: The effects and mechanisms of china's dual-pilot energy policies. *Energy* 307:132538.