Matching, weighting, or regression? Evidence from a comprehensive simulation study of Stata treatment-effect estimators

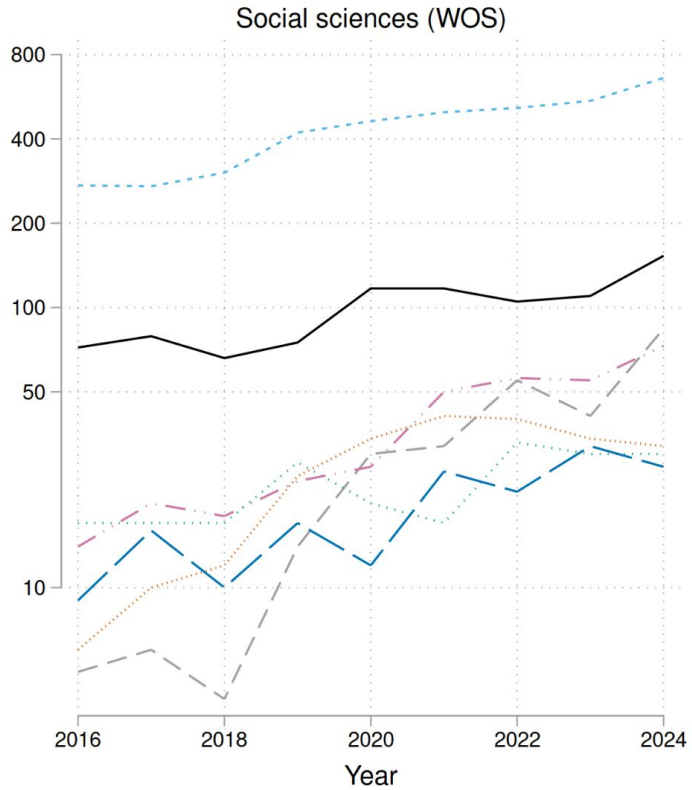
Felix Bittmann
Leibniz-Institute for Educational Trajectories

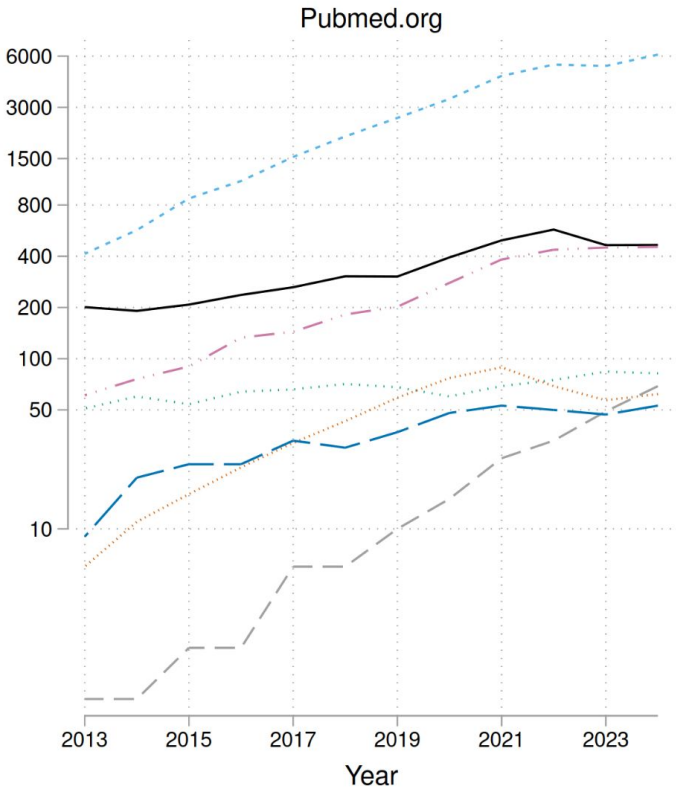
Stata User Meeting Germany 2026

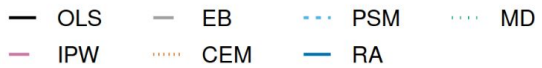
How can we approximate causal effects?

- Conducting experiments (RCTs)
- Longitudinal analyses
- Exploiting certain data constellations (e.g. IV-approaches)
- “Control” approaches using cross-sectional data
  - Stata offers a wide range of approaches:
  - Regression
  - Matching
  - Weighting

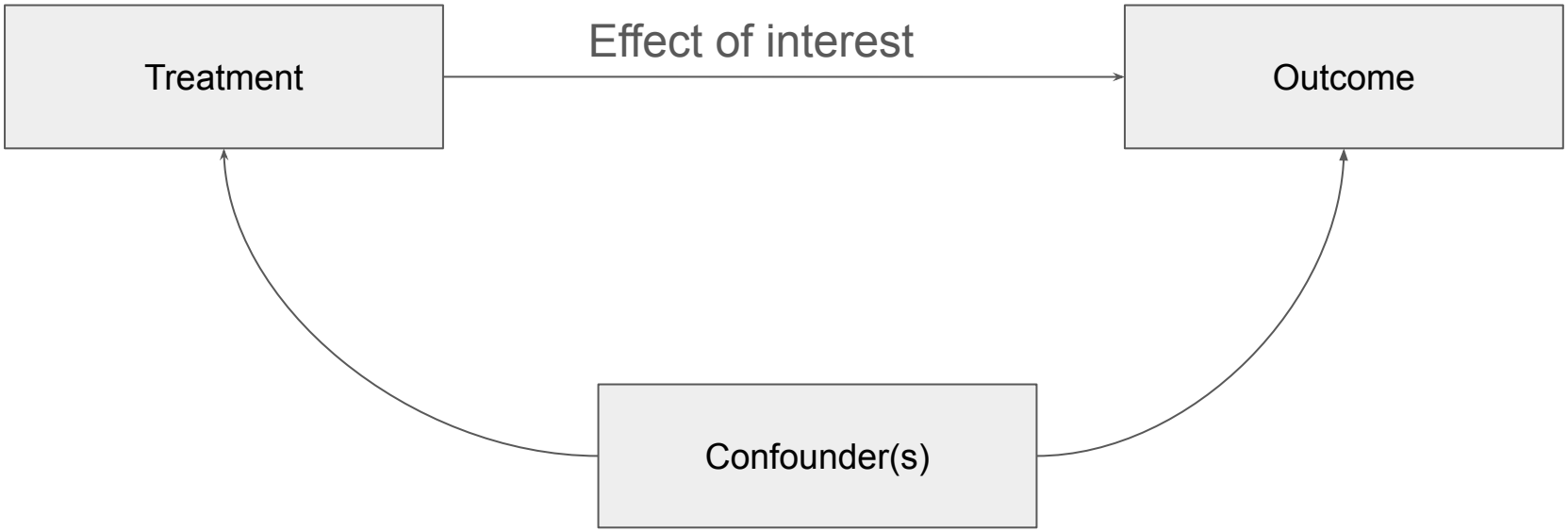
Which one to pick?


Social sciences (WOS)
800
400
200
100
50
10
2016
2018
2020
2022
2024
Year


Pubmed.org
6000
3000
1500
800
400
200
100
50
10
2013
2015
2017
2019
2021
2023
Year


OLS
EB
PSM
MD
IPW
CEM
RA

General framework



When controlling approaches fail

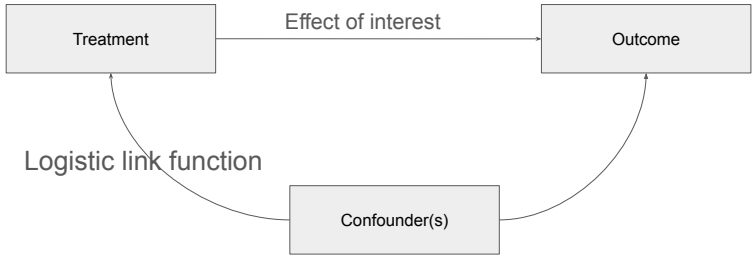
- Unmeasured confounder(s) present
- Functional forms are incorrectly specified (e.g. nonlinearities, interactions)
- Treatment effects are heterogeneous

Analytical approach

- Simulate a dataset with known parameters
- Apply various Stata commands to the dataset
- Analyze which command recovers the known parameters the best
- Various scenarios for coverage of real-world data constellations

Simulation setup

- Treatment is a binary variable
- Outcome is a continuous variable
- Confounders are continuous variables



- Sample sizes: 500 / 1500 / 5000
- Relative size of treated group: 15% / 25% / 50%
- Number of confounders: 1 / 3 / 7
- Correlation among confounders: 0 / 0.10 / 0.20
- True effect size of treatment: 0 / 0.15 / 0.50 / 1
- **Full factorial design: 324 conditions**
- **300 replications each**

Specific conditions

A.) **Unbiased**: all confounders are included in the estimation commands (best-case scenario)

B.) **Omitted-variable bias**: one confounder is randomly omitted (at least one measured confounder is still present in any setup)

C.) **Nonlinear bias**: the squared term of a randomly selected confounder is included in the data generation but not in the estimation command

D.) **Heterogeneous treatment bias**: an interaction term is included between the treatment variable and a randomly selected confounder (ATT != ATC)

Official Stata commands

- Linear (OLS) regression: *regress*
- Treatment effects suite (*teffects*)
  - Augmented inverse probability weighting: *aipw*
  - Inverse probability weighting: *ipw*
  - IPW regression adjustment: *ipwra*
  - Nearest neighbour matching: *nnmatch*
  - Propensity score matching: *psmatch*
  - Regression adjustment: *ra*

User-written commands

- Kmatch (Jann, 2017)
  - Propensity score matching (kernel): *ps*
  - Multivariate distance matching: *md*
  - Entropy balancing: *eb*
  - Coarsened exact matching: *em*
  - Regression adjustment: *ra*
  - Inverse probability weighting: *ipw*
- Teffects2 (Słoczyński, Uysal, and Wooldridge, 2025)
  - *IPW* (AIPW, IPWRA)

Specifications: mostly defaults

- teffects: logistic model for treatment
- teffects nnmatch / psmatch: 1 NN, bias adjustment
- kmatch eb: matching as many moments as possible
- kmatch ps / eb / ipw / md / ra: common support sample selection (*teffects* enforces this)

Evaluation of results

- Either regression coefficient (*regress*) or ATE (all other methods)
- Main statistic of interest: **absolute difference** between target parameter and empirical point estimate
- Empirical standard errors (as is)
- Coverage (Prob. that a CI contains the true population parameter)
- Power (Prob. to detect a statistically significant effect)
- Approach: linear regression models
- Trimming results above percentile 99 (absolute difference; entire simulated dataset is discarded)
- Alternative: *simsum* (White, 2010), see appendix

Table 1: Convergence statistics by method and simulation specification

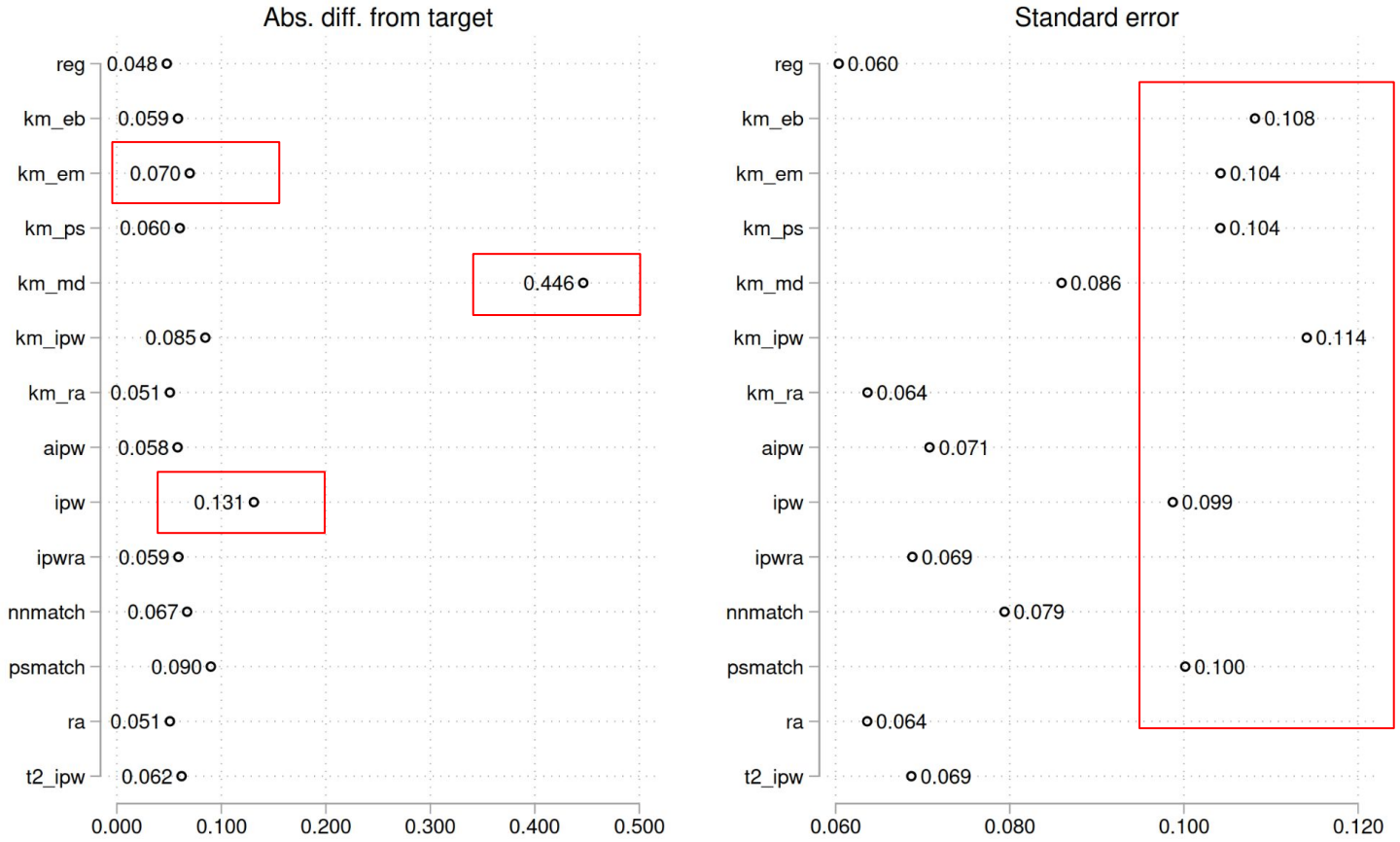
Method	Unbiased	Biased	Incorr. funct. form	Heterog.	Total
reg	96,228	96,228	95,930	96,228	384,614
km_eb	96,228	96,228	95,930	96,228	384,614
km_em	64,793	64,959	64,794	64,798	259,342
km_ps	96,228	96,228	95,930	96,228	384,614
km_md	96,228	96,228	95,930	96,228	384,614
km_ipw	96,228	96,228	95,930	96,228	384,614
km_ra	96,228	96,228	95,930	96,228	384,614
aipw	96,228	96,228	95,930	96,228	384,614
ipw	96,228	96,228	95,930	96,228	384,614
ipwra	96,228	96,228	95,930	96,228	384,614
nnmatch	96,228	96,228	95,930	96,228	384,614
psmatch	96,228	96,228	95,930	96,228	384,614
ra	96,228	96,228	95,930	96,228	384,614
t2_ipw	96,167	96,219	94,497	96,162	383,045
Total	1,315,696	1,315,914	1,310,449	1,315,696	5,257,755

Note: reported are the number of valid results. The only command that displays serious issues is `kmatch em`, as it has problems handling multiple control variables.

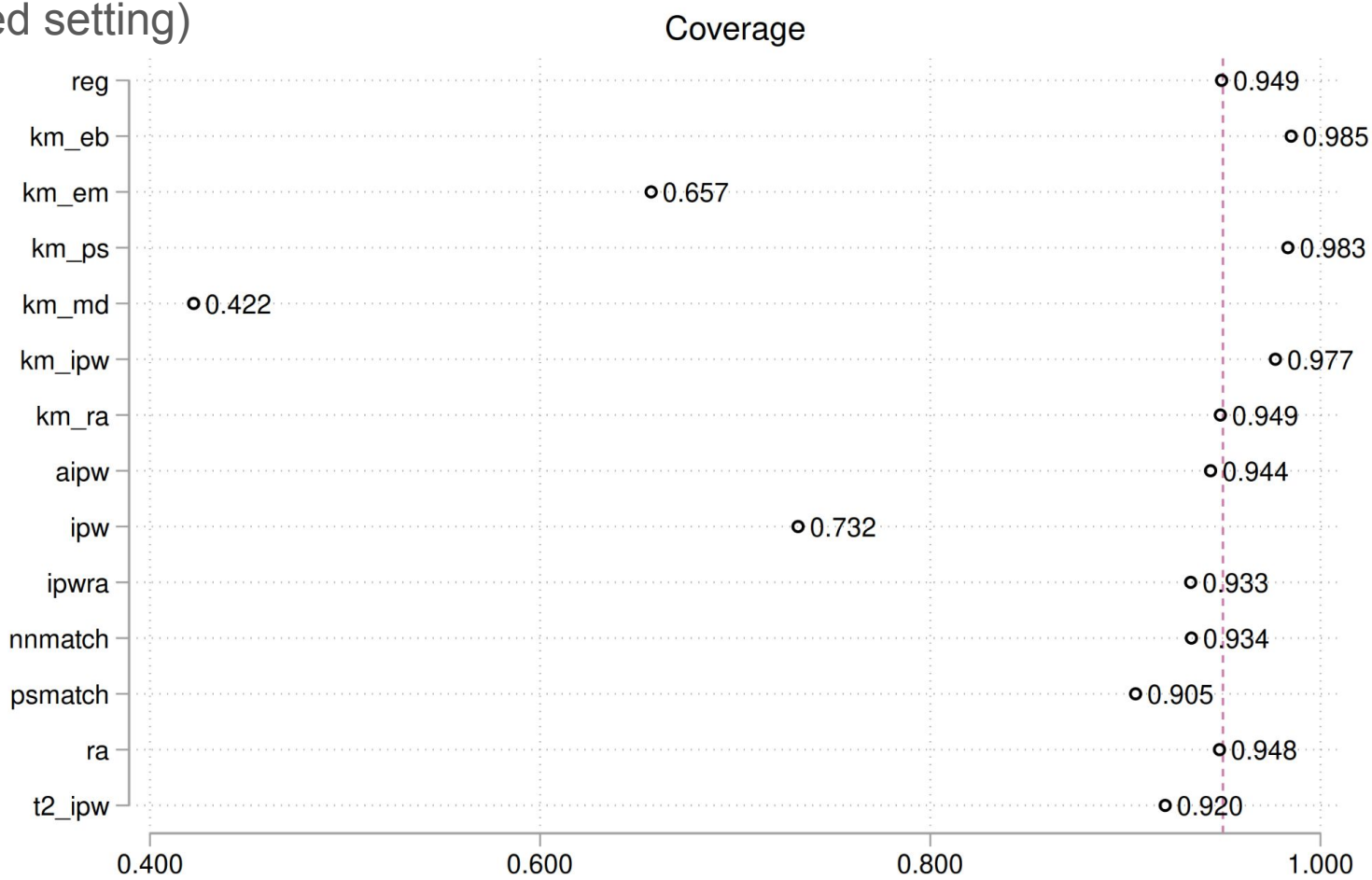
CEM

teffects2

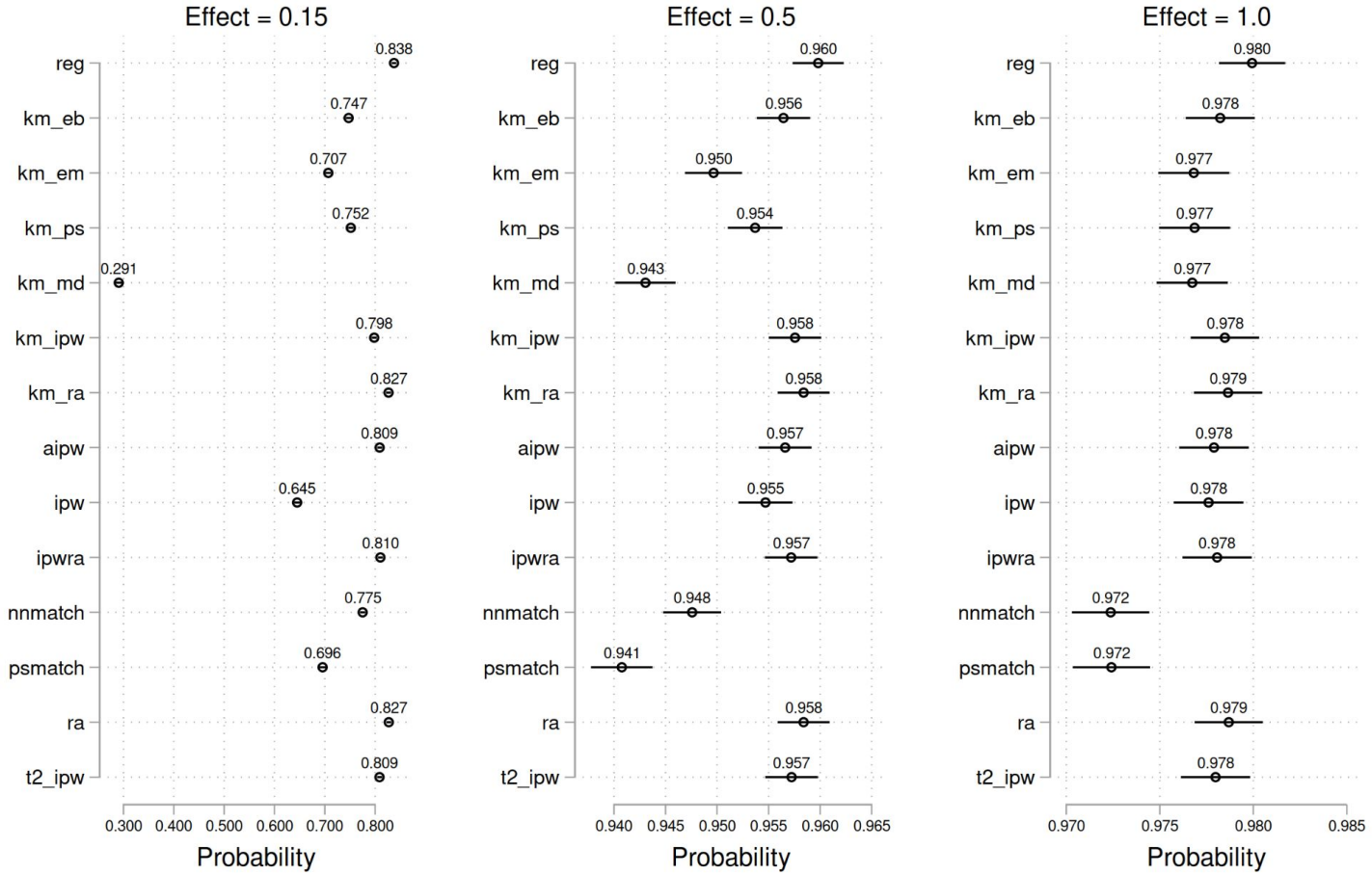
Unbiased setting



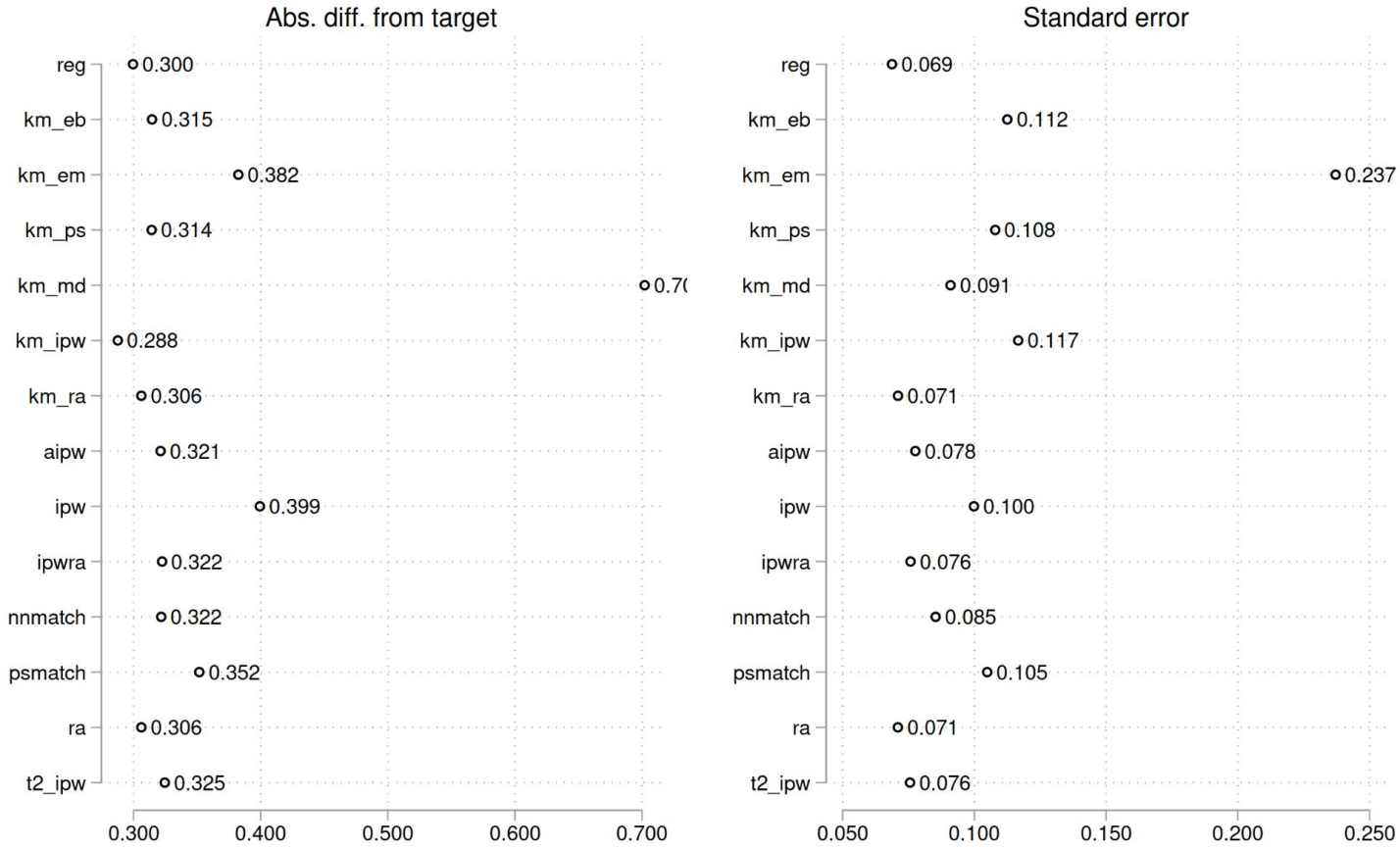
(Unbiased setting)



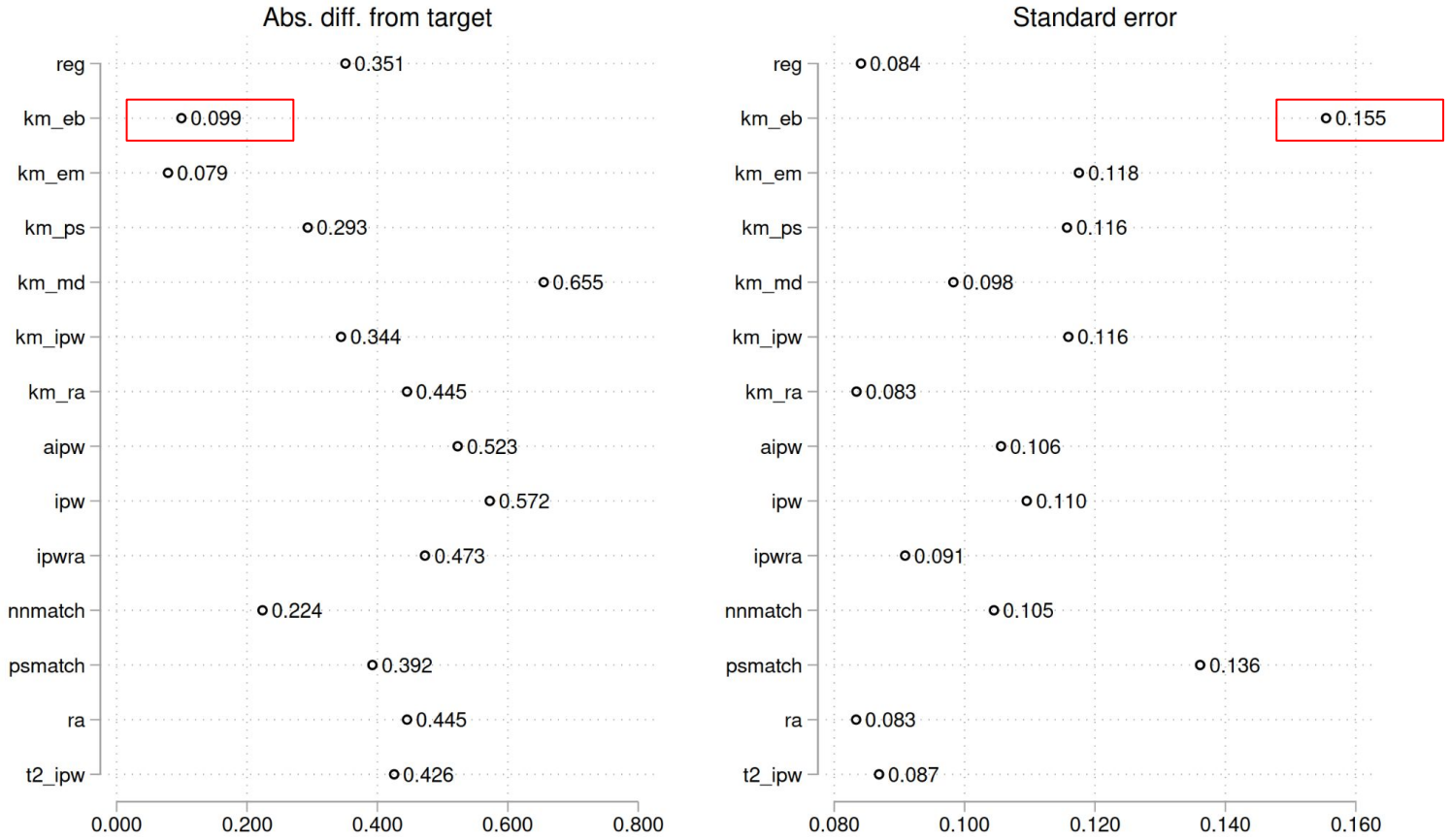
Statistical power (unbiased setting)



Omitted variable bias

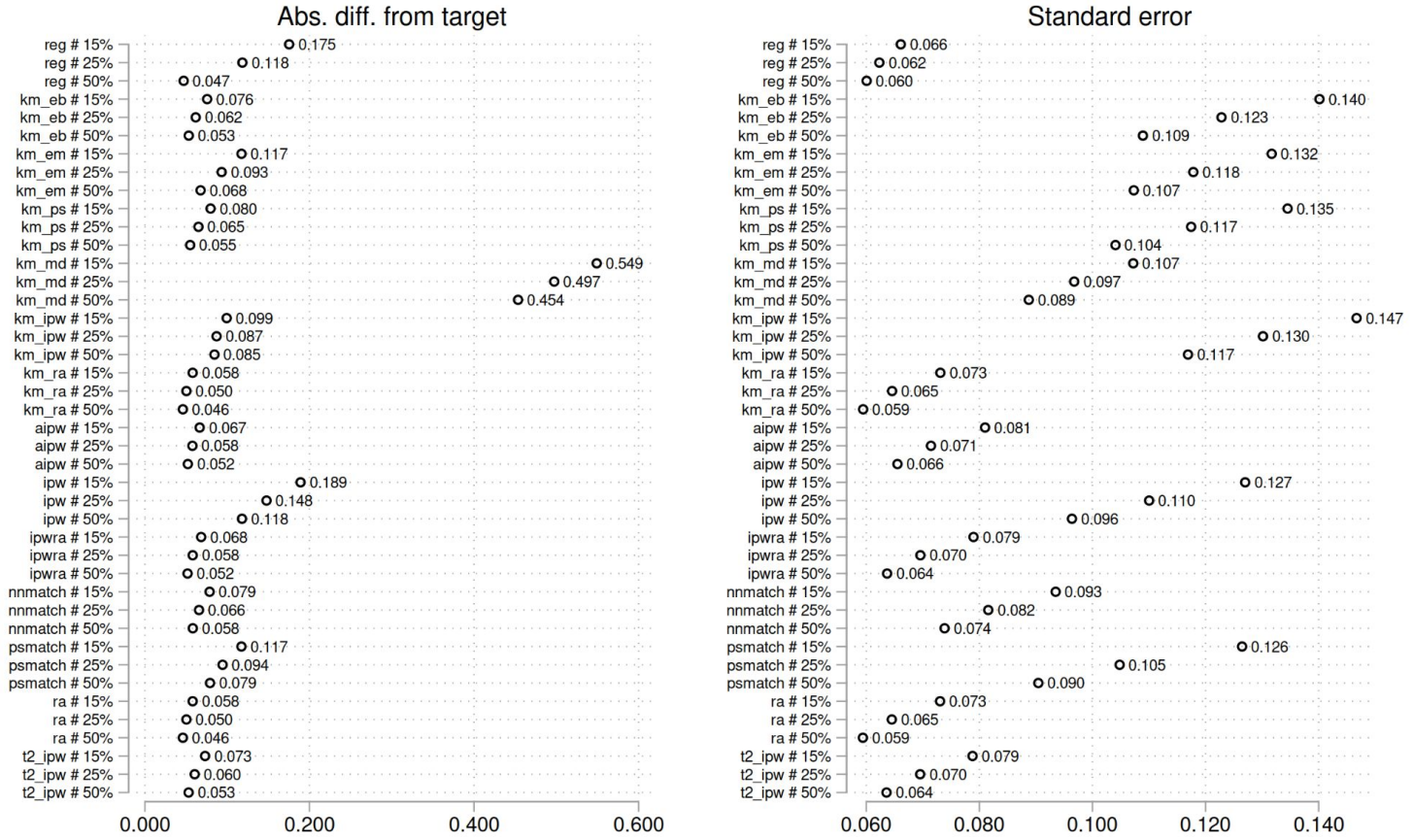


Incorrect functional form



By rel. group size

Heterogeneous treatment effects



Final conclusions

- Not all methods are equally good; some are better avoided
- *regress* is usually not a bad choice
  - Functional form test: Ramsey RESET test (*estat ovtest*)
  - Heterogeneous treatment effects: *hettreatreg* (Słoczyński, 2019) / regression adjustment
- When functional forms are problematic: entropy balancing
  - Matching multiple moments
  - Accounts for heterogeneous treatment effects
- Testing various methods for robustness

Limitations

- No simulation can cover all possible data constellations
- Adjusting estimation command options is usually required and beneficial

WP available on Zenodo.org

https://zenodo.org/records/18267321

Matching, weighting, or regression? Evidence from a comprehensive simulation study of Stata treatment effect estimators

Felix Bittmann
Leibniz Institute for Educational Trajectories
Bamberg, Germany
felix.bittmann@lifbi.de
0000-0003-0802-5854

March 2026 (v1.3)

Abstract. Estimating treatment effects with cross-sectional data is one of the most widespread approaches in empirical research. Provided that researchers are able to measure all relevant control variables, it is possible to approximate unbiased (causal) treatment effects. To this end, Stata offers a wide range of standard and user-written commands. Naturally, the question remains which of these methods is most robust for producing unbiased point estimates and valid inference. We address this question by evaluating 14 different commands in a comprehensive simulation study. Using four different settings (unbiased, biased, and incorrect functional form, heterogeneous treatment effects), we analyze a variety of empirically relevant scenarios. Our results indicate that linear (OLS) regression exhibits the lowest bias, the smallest standard errors, and the most accurate coverage in almost all simulation specifications. Entropy balancing and some matching approaches offer advantages when nonlinearities are incorrectly specified. When heterogeneous treatment effects are present, regression adjustment or AIPW approaches deliver the best results. Surprisingly, several methods deviate substantially from the target estimands, even in unbiased "best-case" scenarios.

Keywords: treatment effect, regression, matching, balancing, simulation, causal inference, Stata

1 Introduction

Estimating treatment effects is a common task in almost all fields of empirical research. Accordingly, a large part of statistics is concerned with developing and evaluating methods to achieve this goal. Over the past hundred years, a variety of approaches and techniques have been developed to quantify effects of interest. An overview of the most common approaches is provided by Cunningham (2021).

Appendix

simsum results

Table 2: Standardized simulation analysis

Method	reg	km_eb	km_em	km_ps	km_md	km_ipw	km_ra	aipw	ipw	ipwra	nnmatch	psmatch	ra	t2_ipw
Unbiased setting														
Non-missing point estimates	96228	96228	64793	96228	96228	96217	96228	96224	96222	96228	96228	96228	96228	96166
Non-missing SEs	96228	96228	64793	96228	96228	96217	96228	96224	96222	96228	96228	96228	96228	96166
Bias in point estimate	0.001	0.001	-0.006	0.001	-0.430	0.050	0.001	0.001	-0.094	0.001	0.001	-0.033	0.001	0.001
Mean of point estimate	0.001	0.001	-0.006	0.001	-0.430	0.050	0.001	0.001	-0.094	0.001	0.001	-0.033	0.001	0.001
Empirical SE	0.066	0.082	0.100	0.083	0.495	0.116	0.070	0.082	0.155	0.082	0.094	0.124	0.070	0.087
% precision gain rel. to reg	0.000	-35.166	-56.735	-37.418	-98.239	-68.117	-12.283	-35.295	-82.067	-35.568	-50.710	-71.959	-12.283	-43.311
Mean squared error	0.004	0.007	0.010	0.007	0.429	0.016	0.005	0.007	0.033	0.007	0.009	0.016	0.005	0.008
Root mean squared error	0.066	0.082	0.100	0.083	0.655	0.127	0.070	0.082	0.181	0.082	0.094	0.128	0.070	0.087
RMS model-based SE	0.066	0.123	0.120	0.117	0.095	0.134	0.070	0.079	0.120	0.075	0.088	0.117	0.070	0.075
Mean CI width	0.237	0.424	0.408	0.408	0.337	0.447	0.250	0.277	0.387	0.270	0.311	0.393	0.249	0.269
Relative % error in SE	0.893	50.933	20.612	41.339	-80.838	15.288	0.172	-3.054	-22.846	-7.844	-6.313	-5.999	0.094	-13.791
% cov. of nom. 95% CI	94.914	98.488	97.537	98.330	42.220	97.675	94.833	94.369	73.223	93.348	93.388	90.524	94.827	92.100
% power of 5% level test	5.086	1.512	2.463	1.670	57.780	2.325	5.167	5.631	26.777	6.652	6.612	9.476	5.173	7.900
Biased setting														
Non-missing point estimates	96228	96228	64959	96228	96228	96228	96228	96228	96228	96228	96228	96228	96228	96219
Non-missing SEs	96228	96228	64959	96228	96228	96228	96228	96228	96228	96228	96228	96228	96228	96219
Bias in point estimate	-0.296	-0.309	-0.332	-0.308	-0.701	-0.261	-0.302	-0.316	-0.393	-0.317	-0.315	-0.345	-0.302	-0.319
Mean of point estimate	-0.296	-0.309	-0.332	-0.308	-0.701	-0.261	-0.302	-0.316	-0.393	-0.317	-0.315	-0.345	-0.302	-0.319
Empirical SE	0.204	0.217	0.326	0.217	0.377	0.239	0.210	0.220	0.233	0.220	0.225	0.234	0.210	0.223
% precision gain rel. to reg	0.000	-11.423	-60.682	-11.467	-70.608	-27.095	-5.408	-13.491	-23.086	-14.052	-17.109	-23.522	-5.408	-15.890
Mean squared error	0.129	0.142	0.216	0.142	0.633	0.126	0.135	0.148	0.209	0.149	0.149	0.174	0.135	0.151
Root mean squared error	0.359	0.377	0.465	0.377	0.796	0.354	0.368	0.385	0.457	0.386	0.387	0.417	0.368	0.389
RMS model-based SE	0.076	0.127	0.311	0.120	0.100	0.134	0.079	0.086	0.117	0.083	0.094	0.120	0.079	0.083
Mean CI width	0.269	0.441	0.930	0.423	0.357	0.457	0.278	0.304	0.391	0.297	0.334	0.411	0.278	0.296
Relative % error in SE	-62.828	-41.463	-4.481	-44.567	-73.429	-43.848	-62.603	-60.665	-49.898	-62.232	-58.040	-48.611	-62.632	-62.805
% cov. of nom. 95% CI	23.379	38.342	52.238	36.842	6.923	44.303	23.732	24.490	19.885	23.777	27.421	28.924	23.728	23.437
% power of 5% level test	76.621	61.658	47.762	63.158	93.077	55.697	76.268	75.510	80.115	76.223	72.579	71.076	76.272	76.563

	reg	km_eb	km_em	km_ps	km_md	km_ipw	km_ra	aipw	ipw	ipwra	nnmatch	psmatch	ra	t2_ipw
Incorrect functional form														
Non-missing point estimates	95930	95930	64792	95930	95930	95930	95930	95930	95930	95930	95930	95930	95930	94497
Non-missing SEs	95930	95930	64792	95930	95930	95930	95930	95930	95930	95930	95930	95930	95930	94497
Bias in point estimate	-0.347	-0.016	-0.012	-0.267	-0.645	-0.324	-0.442	-0.519	-0.567	-0.468	-0.202	-0.375	-0.442	-0.419
Mean of point estimate	-0.347	-0.016	-0.012	-0.267	-0.645	-0.324	-0.442	-0.519	-0.567	-0.468	-0.202	-0.375	-0.442	-0.419
Empirical SE	0.227	0.154	0.110	0.298	0.618	0.291	0.273	0.333	0.333	0.295	0.200	0.338	0.273	0.271
% precision gain rel. to reg	0.000	117.957	323.138	-42.263	-86.571	-39.367	-31.115	-53.805	-53.703	-41.156	28.327	-54.921	-31.115	-30.259
Mean squared error	0.172	0.024	0.012	0.160	0.799	0.190	0.270	0.380	0.432	0.306	0.081	0.255	0.270	0.249
Root mean squared error	0.414	0.154	0.111	0.401	0.894	0.435	0.519	0.617	0.657	0.554	0.285	0.505	0.519	0.499
RMS model-based SE	0.094	0.176	0.132	0.130	0.107	0.137	0.093	0.125	0.131	0.100	0.116	0.160	0.093	0.096
Mean CI width	0.330	0.609	0.461	0.454	0.385	0.454	0.327	0.414	0.429	0.356	0.410	0.534	0.327	0.341
Relative % error in SE	-58.371	14.845	20.123	-56.416	-82.660	-52.849	-66.058	-62.374	-60.594	-66.049	-42.028	-52.733	-66.084	-64.589
% cov. of nom. 95% CI	22.389	95.806	97.202	54.585	31.084	38.123	16.778	16.281	14.281	16.979	50.753	41.767	16.766	18.730
% power of 5% level test	77.611	4.194	2.798	45.415	68.916	61.877	83.222	83.719	85.719	83.021	49.247	58.233	83.234	81.270
Heterog. treatment effects														
Non-missing point estimates	96228	96228	64798	96228	96228	96217	96228	96225	96224	96228	96228	96227	96228	96162
Non-missing SEs	96228	96228	64798	96228	96228	96217	96228	96225	96224	96228	96228	96227	96228	96162
Bias in point estimate	-0.091	-0.016	-0.055	-0.015	-0.488	0.042	0.000	0.000	-0.112	0.000	0.000	-0.039	0.000	0.000
Mean of point estimate	-0.091	-0.016	-0.055	-0.015	-0.488	0.042	0.000	0.000	-0.112	0.000	0.000	-0.039	0.000	0.000
Empirical SE	0.126	0.087	0.122	0.091	0.528	0.128	0.071	0.082	0.174	0.082	0.094	0.135	0.071	0.087
% precision gain rel. to reg	0.000	108.836	6.663	90.259	-94.343	-3.517	210.081	132.915	-48.050	132.631	79.944	-13.632	210.081	106.881
Mean squared error	0.024	0.008	0.018	0.009	0.517	0.018	0.005	0.007	0.043	0.007	0.009	0.020	0.005	0.008
Root mean squared error	0.155	0.088	0.133	0.092	0.719	0.134	0.071	0.082	0.207	0.082	0.094	0.141	0.071	0.087
RMS model-based SE	0.069	0.141	0.136	0.134	0.108	0.154	0.072	0.081	0.135	0.077	0.092	0.126	0.072	0.077
Mean CI width	0.246	0.486	0.466	0.465	0.382	0.514	0.257	0.285	0.435	0.277	0.325	0.420	0.257	0.277
Relative % error in SE	-45.054	61.938	12.191	46.896	-79.520	20.205	1.569	-1.485	-22.356	-5.958	-1.963	-6.435	1.490	-11.568
% cov. of nom. 95% CI	63.729	98.772	94.018	98.470	39.601	98.090	95.341	94.791	71.115	93.803	94.387	90.232	95.337	92.619
% power of 5% level test	36.271	1.228	5.982	1.530	60.399	1.910	4.659	5.209	28.885	6.197	5.613	9.768	4.663	7.381