



nopo

Implementation of a matching-based decomposition technique with postestimation commands

Maximilian Sprengholz¹ Maik Hamjediers¹

¹Humboldt-Universität zu Berlin

19th German Stata Conference, Berlin

16.6.2023



DECOMPOSITIONS IN THE SOCIAL SCIENCES

- Decomposition techniques are a common way to examine gaps in socio-economic outcomes between two groups (e.g., sex, race, nativity)
 - To what extent contribute observed differences in group characteristics to gaps?
⇒ **Explained component**
 - Gaps not accounted for by observed differences in group characteristic might indicate differential returns or unobservables
⇒ **Unexplained component**



DECOMPOSITIONS IN THE SOCIAL SCIENCES

- Decomposition techniques are a common way to examine gaps in socio-economic outcomes between two groups (e.g., sex, race, nativity)
 - To what extent contribute observed differences in group characteristics to gaps?
⇒ **Explained component**
 - Gaps not accounted for by observed differences in group characteristic might indicate differential returns or unobservables
⇒ **Unexplained component**
- Methods invoke different assumptions, can lead to different results, and provide different insights (Strittmatter and Wunsch 2021; Hamjediers and Sprengholz 2023)
 - Many applications rely on regression-based techniques (Blinder 1973; Oaxaca 1973), Nopo (2008) proposed a matching-based approach



DECOMPOSITION À LA ÑOPO

1. Each member of group B can be matched to all potential matches of group A along a set of characteristics \mathbf{X} (one-to-many-matching), providing two pieces of information:
 - Who can be matched (subscript m) and who cannot be matched (subscript u)
 - Weights to calculate counterfactual outcome $\bar{Y}_{AB,m}$ that reflects
 - outcome of group A if it had the same characteristics as group B
 - outcome of group B if it had the same returns to characteristics as group A



DECOMPOSITION À LA ÑOPO

2. If $D = \bar{Y}_B - \bar{Y}_A$,
gap can be decomposed into four components after matching:

$$\begin{aligned}
 D &= D_0 + D_X + D_A + D_B \\
 &= \underbrace{\bar{Y}_{B,m} - \bar{Y}_{AB,m}}_{\text{splitting difference}} + \underbrace{\bar{Y}_{AB,m} - \bar{Y}_{A,m}}_{\text{among matched by}} + D_A + D_B \\
 &\quad \text{reweighted group A} \\
 &= \underbrace{\text{unexplained component} + \text{explained component}}_{\text{pertains only to matched units}} + D_A + D_B
 \end{aligned}$$



DECOMPOSITION À LA ÑOPO

2. If $D = \bar{Y}_B - \bar{Y}_A$,
gap can be decomposed into four components after matching:

$$D = D_0 + D_X + \underbrace{D_A + D_B}_{\text{out of support}}$$

$$D_A = \underbrace{(\bar{Y}_{A,m} - \bar{Y}_{A,u})}_{\text{gap between matched and unmatched } A} \cdot \underbrace{(N_{A,u}/N_A)}_{\text{share of unmatched } A}$$

$$D_B = \underbrace{(\bar{Y}_{B,u} - \bar{Y}_{B,m})}_{\text{gap between unmatched and matched } B} \cdot \underbrace{(N_{B,u}/N_B)}_{\text{share of unmatched } B}$$



LINKS TO OTHER APPROACHES

- Generally, similar to two-fold (Kitagawa-)Blinder-Oaxaca-Decomposition

(Hamjediers and Sprengholz 2023)

- Advantages of matching-based decomposition:

- + Non-parametric estimation → no assumptions about functional form
- + D_0 & D_X apply only to matched units → no model-based extrapolation

- Disadvantages of matching-based decomposition:

- Suffers from curse of dimensionality → risk of attributing too much to D_A & D_B
- Does not allow to disentangle explained component across predictors

⇒ Similar arguments as for regression- vs. matching-based adjustment for confounders in estimating (local) treatment effects



LINKS TO OTHER APPROACHES

- Component D_0 is equal to the average treatment effect on the treated ATT after matching

$$\begin{aligned} ATT &= P_{O_{t=1}^{T=1}} - P_{O_{t=0}^{T=1}} \\ &= \bar{Y}_{B,m} - \bar{Y}_{AB,m} = D_0 \end{aligned}$$

⇒ All other components of Nopo's approach are seldom assessed in estimations of treatment effects via matching



IT'S ALL ABOUT THE MATCHING

- Originally, exact matching on (coarsened) predictors
(cf. ado-file `nopomatch` of [Atal et al. \(2013\)](#))
- We extend it to Propensity Score Matching ([Rosenbaum and Rubin 1983](#))
and Multivariate Distance Matching
- Trade-off between reaching balance on predictors between B, m and A^B, m
vs. curse of dimensionality and lack of common support ([Iacus et al. 2012](#))



NEW COMMAND: NOPO

- More flexible, inference for all components
- Allows matching by different measures
 - `nopo` calls `kmatch` (Jann 2017) inherently or can be used as postestimation-command after matching via `kmatch` (Jann 2017)
- Provides postestimation commands for descriptives after matching, contribution of groups to D_A and D_B , and components across distribution of Y
- Illustration based on example from Hamjediers and Sprengholz (2023):
 - Data: GSOEP, 2014-2019, one observation per individual
 - Groups: Native men (A) and immigrant women (B)
 - Outcome: hourly gross wages
 - Predictors: age, married, educational attainment, labor market experience, 2-digit ISCO-08 occupations, part-time indicator (*all coarsened*)



STANDALONE USAGE

```
nopo decomp depvar varlist [if] [in] [weights], by(varname)
```



STANDALONE USAGE

```
. nopo decomp wage age_c married edu lmexp parttime isco2d, by(grp)
```

```

Nopo decomposition              N              =      8954
Exact matching:                 N strata      =      1783
                                N matched strata =      488
                                (unique combinations of matching set)

```

Group	N / %			Mean
	Matched	Unmatched	Total	wage
A: Native Men	3110	2939	6049	20.5
grp == 1	51.4	48.6		
B: Immigrant Women	1387	1518	2905	12.1
grp == 4 (ref)	47.7	52.3		

wage	Coefficient	Std. err.	z	P> z	[95% conf. interval]	
D	-8.384027	.2067519	-40.55	0.000	-8.789253	-7.978801
DO	-2.406294	.5407104	-4.45	0.000	-3.466067	-1.346521
DX	-5.05875	.6061777	-8.35	0.000	-6.246837	-3.870664
DA	.7731673	.1257708	6.15	0.000	.5266609	1.019674
DB	-1.69215	.1285267	-13.17	0.000	-1.944057	-1.440242



STANDALONE USAGE

```
. nopo decomp wage age_c married edu lmexp parttime isco2d, by(grp)
```

```
Nopo decomposition          N          =      8954
Exact matching:            N strata    =      1783
                           N matched strata =       488
                           (unique combinations of matching set)
```

Group	N / %			Mean wage
	Matched	Unmatched	Total	
A: Native Men grp == 1	3110 51.4	2939 48.6	6049	20.5
B: Immigrant Women grp == 4 (ref)	1387 47.7	1518 52.3	2905	12.1

wage	Coefficient	Std. err.	z	P> z	[95% conf. interval]	
D	-8.384027	.2067519	-40.55	0.000	-8.789253	-7.978801
D0	-2.406294	.5407104	-4.45	0.000	-3.466067	-1.346521
DX	-5.05875	.6061777	-8.35	0.000	-6.246837	-3.870664
DA	.7731673	.1257708	6.15	0.000	.5266609	1.019674
DB	-1.69215	.1285267	-13.17	0.000	-1.944057	-1.440242

Interpretations:

- D0** Among the matched, 2.41 Euro lower wages for group *B* are unexplained
- DX** Compositional differences account for 5.06 Euro of the gap among matched units
- DA** Unmatched units of group *A* earn lower wages than matched units, which accounts for 0.77 Euro of the gap
- DB** Unmatched units of group *B* earn lower wages than matched units, which accounts for 1.69 Euro of the gap



STANDALONE USAGE

- General Options:
 - Swap groups: `swap`
 - Defining matching direction: `bref(varname == #)`
 - Normalize outcome to the reference group of the matching: `normalize`



STANDALONE USAGE

```
. nopo decomp wage ${pred}, by(grp) bref(grp == 1) swap normalize
Normalized outcome generated: _wage_norm
```

```
Nopo decomposition          N          =      8954
Exact matching:             N strata     =      1783
                             N matched strata =      488
                             (unique combinations of matching set)
```

Group	N / %			Mean
	Matched	Unmatched	Total	_wage_norm
A: Immigrant Women	1387	1518	2905	.591
grp == 4	47.7	52.3		
B: Native Men	3110	2939	6049	1
grp == 1 (ref)	51.4	48.6		

_wage_norm	Coefficient	Std. err.	z	P> z	[95% conf. interval]
D	.4090725	.0100878	40.55	0.000	.3893008 .4288443
DO	.1813969	.0299544	6.06	0.000	.1226873 .2401065
DX	.1828367	.0328027	5.57	0.000	.1185446 .2471288
DA	.0825632	.0062711	13.17	0.000	.0702721 .0948542
DB	-.0377243	.0061366	-6.15	0.000	-.0497518 -.0256968



STANDALONE USAGE

- General Options:
 - Swap groups: `swap`
 - Defining matching direction: `bref(varname == #)`
 - Normalize outcome to the reference group of the matching: `normalize`
- Options to adjust matching procedure correspondingly to `kmatch`:
 - `kmatch()` allows for exact matching (`em`) (*the default*), propensity score matching (`ps`), and multivariate distance matching (`md`)
 - Matching-specific options from `kmatch` can be implemented via `kmatchopt()`

```
. qui: nopo decomp wage ${pred}, by(grp)
. qui: est store em
. qui: nopo decomp wage ${pred}, by(grp) kmatch(ps)
. qui: est store ps
. qui: nopo decomp wage ${pred}, by(grp) kmatch(md)
. qui: est store md
. qui: nopo decomp wage ${pred}, by(grp) kmatch(ps) kmopt(pscmd(probit) bw(0.0001))
. qui: est store ps_probbw
```




STANDALONE USAGE

```
. esttab em ps md ps_probbw, se nonumbers nonotes ///
>     mtitles("exact" "prop. score" "multi. dist." "probit ps") ///
>     stats(nA mshareuA nB mshareuB bwidth, label("N(A)" "% matched A" "N(B)" "% matched B" "Bandwidth"))
```

	exact	prop. score	multi. dist.	probit ps
D	-8.384*** (0.207)	-8.384*** (0.207)	-8.384*** (0.207)	-8.384*** (0.207)
D0	-2.406*** (0.541)	-2.652*** (0.504)	-3.957*** (0.402)	-3.680*** (0.788)
DX	-5.059*** (0.606)	-5.650*** (0.546)	-4.427*** (0.452)	-3.673*** (0.829)
DA	0.773*** (0.126)	0 (.)	0 (.)	0.260* (0.101)
DB	-1.692*** (0.129)	-0.0827*** (0.0232)	0 (.)	-1.291*** (0.111)
N(A)	6049	6049	6049	6049
% matched A	51.41	100	100	62.08
N(B)	2905	2905	2905	2905
% matched B	47.75	96.73	100	55.42
Bandwidth		0.00265	2.325	0.000100



AS POSTESTIMATION AFTER `KMATCH`

- Can be used after `kmatch` by just prompting `nopo decomp`
- Needs that following options of `kmatch` are specified:
 - `tval(#)` to define reference group (if different from `tval(1)`)
 - `att` and/or `atc`; should be coherent to matching direction



AS POSTESTIMATION AFTER KMATCH

```
. qui: kmatch ps grp ${pred} (wage), ///
> tval(4) atc att bw(0.001) pscmd(probit) generate wgenerate replace
```

```
. nopo decomp
```

```
Nopo decomposition                N                =      8954
Propensity-score matching:        Kernel bandwidth: =      1.0e-03
```

Group	N / %			Mean
	Matched	Unmatched	Total	wage
A: Native Men	5900	149	6049	20.5
grp == 1	97.5	2.5		
B: Immigrant Women	2516	389	2905	12.1
grp == 4 (ref)	86.6	13.4		

	wage	Coefficient	Std. err.	z	P> z	[95% conf. interval]	
D		-8.384027	.2067519	-40.55	0.000	-8.789253	-7.978801
DO		-3.02594	.6471339	-4.68	0.000	-4.294299	-1.757581
DX		-5.066576	.6814688	-7.43	0.000	-6.40223	-3.730921
DA		-.0101861	.0206239	-0.49	0.621	-.0506082	.0302361
DB		-.2813254	.049454	-5.69	0.000	-.3782534	-.1843973



AS POSTESTIMATION AFTER KMATCH

- Invoked `kmatch`-command in standalone usage is returned and can be copied for case-specific adjustments

```
. qui: nopo decomp wage ${pred}, by(grp) kmatch(ps)
```

```
. display "`e(kmatch_cmdline)'"
```

```
kmatch ps grp age_c married edu lmexp parttime isco2d (wage) , tval(4) att generate wgenerate replace
```



DESCRIPTION BY MATCHING STATUS

```
. qui: nopo decomp wage ${pred}, by(grp)
```

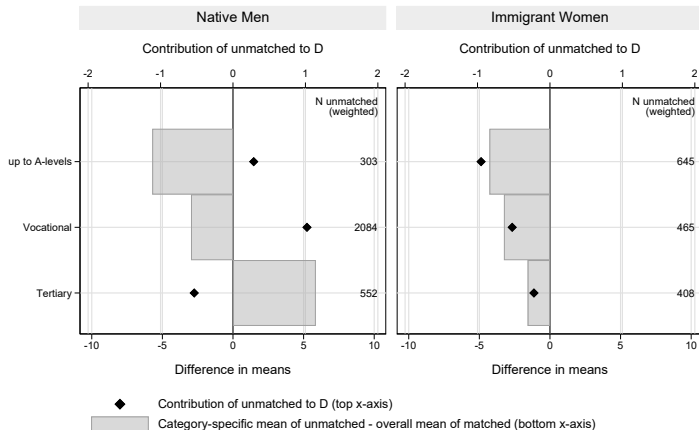
```
. nopo summarize wage age married edu_1 edu_2 edu_3, label
```

	unmatched	Native Men matched	matched ~d	Immigrant matched	Women unmatched
Hourly wag~)					
Mean	19.7	21.3	16.2	13.8	10.6
SD	9.71	10.4	9.74	7.65	5.52
Age					
Mean	44.8	43.9	38.9	38.9	41.6
SD	9.94	11	10.2	10	8.79
Married					
Mean	.612	.669	.607	.607	.759
SD	.487	.471	.488	.489	.428
edu==up to~s					
Mean	.103	.055	.199	.199	.425
SD	.304	.228	.399	.399	.494
edu==Vocat~l					
Mean	.709	.504	.49	.49	.306
SD	.454	.5	.5	.5	.461
edu==Terti~y					
Mean	.188	.441	.311	.311	.269
SD	.391	.497	.463	.463	.443



VARIABLE-SPECIFIC CONTRIBUTION TO D_A AND D_B

```
. nopo dadb edu
```



See application in [Sprengholz and Hamjediers \(2022\)](#), Figure 2



COMPONENTS ALONG OUTCOME-DISTRIBUTION

```
. nopo gapoverdist
```

```
Component distribution across 100 quantiles of wage requested
```

```
-----
```

	Estimate	Sum over q	Minimum among compared groups Unique q values	N
D	-8.38	-8.39	100	2905
DO	-2.41	-2.36	98	1387
DX	-5.06	-5.1	98	3110
DA	.773	.773	100	2939
DB	-1.69	-1.69	100	1387

```
-----
```

Note:

- The component sum across quantiles should correspond to the estimates with well populated quantiles.
- There are less unique quantile values than quantiles requested which means that across some quantiles, the value of wage does not change for (one of) the groups compared to estimate the component.
- Use the nquantiles(#) option to set the number of quantiles.



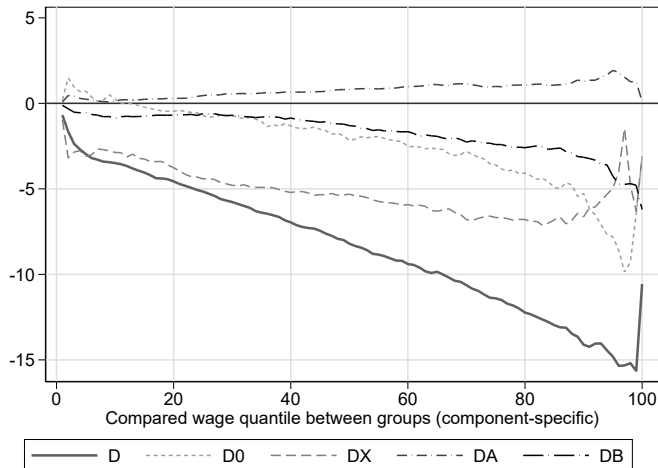
COMPONENTS ALONG OUTCOME-DISTRIBUTION

```
. nopo gapoverdist
Component distribution across 100 quant
```

	Estimate	Sum over q
D	-8.38	-8.39
D0	-2.41	-2.36
DX	-5.06	-5.1
DA	.773	.773
DB	-1.69	-1.69

Note:

- The component sum across quantiles sh well populated quantiles.
- There are less unique quantile values that across some quantiles, the value (one of) the groups compared to estim
- Use the nquantiles(#) option to set t



See application in [Nopo \(2008\)](#), Figure 2



OUTLOOK

- On our to-do-list:
 - Options for component-size relative to gap
 - Standard errors are still too large and need to be adjusted
 - bootstrap-prefix can be applied
 - Write a help-file
- Current version is available on git: github.com/mhamjediers/nopo_decomposition
- Any feedback is of course very welcome



REFERENCES I

- Atal, J.P., Hoyos, A., Nopo, H., 2013. NOPOMATCH: Stata Module to Implement Nopo's Decomposition. URL: <https://ideas.repec.org/c/boc/bocode/s457157.html>.
- Blinder, A.S., 1973. Wage Discrimination: Reduced Form and Structural Estimates. *The Journal of Human Resources* 8, 436–455.
- Hamjediers, M., Sprengholz, M., 2023. Comparing the Incomparable? Issues of Lacking Common Support, Functional-Form Misspecification, and Insufficient Sample Size in Decompositions. *Sociological Methodology* , 008117502311697doi:10.1177/00811750231169729.
- Iacus, S.M., King, G., Porro, G., 2012. Causal Inference without Balance Checking: Coarsened Exact Matching. *Political Analysis* 20, 1–24. doi:10.1093/pan/mpr013.
- Jann, B., 2017. KMATCH: Stata module module for multivariate-distance and propensity-score matching, including entropy balancing, inverse probability weighting, (coarsened) exact matching, and regression adjustment. URL: <https://ideas.repec.org/c/boc/bocode/s458346.html>.
- Kitagawa, E.M., . Components of a Difference Between Two Rates* 50, 1168–1194. doi:10/ggfgpj.
- Nopo, H., 2008. Matching as a Tool to Decompose Wage Gaps. *The Review of Economics and Statistics* 90, 290–299.
- Oaxaca, R., 1973. Male-Female Wage Differentials in Urban Labor Markets. *International Economic Review* 14, 693–709.
- Rosenbaum, P.R., Rubin, D.B., 1983. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika* 70, 41. doi:10.2307/2335942.
- Sprengholz, M., Hamjediers, M., 2022. Intersections and Commonalities: Using Matching to Decompose Wage Gaps by Gender and Nativity in Germany. *Work and Occupations* , 073088842211411doi:10.1177/07308884221141100.
- Strittmatter, A., Wunsch, C., 2021. The Gender Pay Gap Revisited with Big Data: Do Methodological Choices Matter? *SSRN Electronic Journal* doi:10.2139/ssrn.3794074.