Decomposition methods in the social sciences

Bamberg Graduate School of Social Sciences, June 7–8, 2018

Ben Jann

University of Bern, Institut of Sociology

Difference-in-difference decompositions

Contents

Difference-in-difference decompositions

- Up to now we were concerned with a single outcome differential (e.g. a gender wage gap) at a specific point in time and in a specific region or population.
- Often, however, comparisons over time or between countries or regions are of interest.
  - How did the gender wage gap change over time and how much of this change is due to changes with respect to covariates?
  - How would the gender wage gap in country A look like if it had the wage structure of country B?
- One way of analyzing changes over time or between populations is to compare separate decomposition results. Some questions, however, require a "double" or "difference-in-difference" decomposition.
- A very famous application of such methodology, for example, is the "Swimming upstream" paper by Blau and Kahn (1997).

Smith-Welch decomposition

(Smith and Welch 1987, also see e.g. Heckman et al. 2000)

- Given is a linear model

$$Y^{gt} = X^{gt}\beta^{gt} + \epsilon^{gt}, \quad \mathsf{E}(e^{gt}|X^{gt}) = 0$$

  for two groups, $g = m, f$ (males and females), at two time points, $t = 0, 1$.

- Using the male coefficients, β^{mt}, as reference, the decomposition of the group difference in average Y at time t can be written as

$$\Delta^{\mu t} = (\bar{X}^{mt} - \bar{X}^{ft})\beta^{mt} + \bar{X}^{ft}(\beta^{mt} - \beta^{ft}) = \Delta_X^{\mu t} + \Delta_S^{\mu t}$$

- We are now interested in decomposing the change in the wage gap over time.

Smith-Welch decomposition

- Let $\Delta\bar{X}^t = \bar{X}^{mt} - \bar{X}^{ft}$ and $\Delta\beta^t = \beta^{mt} - \beta^{ft}$. Using the male coefficients from the first time point, β^{m0}, as reference, this double decomposition can be written as

$$\begin{aligned} d\Delta^\mu = \Delta^{\mu 1} - \Delta^{\mu 0} &= \left\{(\Delta\bar{X}^1 - \Delta\bar{X}^0)\beta^{m0} + \Delta\bar{X}^1(\beta^{m1} - \beta^{m0})\right\} \\ &\quad + \left\{\bar{X}^{f1}(\Delta\beta^1 - \Delta\beta^0) + (\bar{X}^{f1} - \bar{X}^{f0})\Delta\beta^0\right\} \\ &= d\Delta^\mu_X + d\Delta^\mu_S \end{aligned}$$

- Interpretation:

$(\Delta\bar{X}^1 - \Delta\bar{X}^0)\beta^{m0}$	main endowments effect: shows how the wage gap changed because men and women became more similar or dissimilar in X (negative, if they became more similar; positive, if they became more dissimilar)
$\Delta\bar{X}^1(\beta^{m1} - \beta^{m0})$	secondary endowments effect due to change in reference wage structure over time

Smith-Welch decomposition

$\bar{X}^{f1}(\Delta\beta^1 - \Delta\beta^0)$	primary coefficients effect: effect of change in wage structure difference between men and women (negative, if coefficients became more similar; positive, if coefficients became more dissimilar)
$(\bar{X}^{f1} - \bar{X}^{f0})\Delta\beta^0$	secondary coefficients effect due to change in reference endowments over time

- Of course, various other types of decompositions are possible depending on the choice of the reference group and the reference year. The index problem of the standard OB decomposition is now a double index problem, which can make it hard to keep an overview.
- See `help smithwelch` for a systematic discussion. It starts with the threefold decomposition and then shows how the formulas change if a reference group and/or a reference year is introduced. Of course, reference groups/years can also be results from pooled or averaged models.

Example analysis (using `smithwelch` by Jann 2005b)

```
. use gsoep29, clear
(BCPGEN: Nov 12, 2013 17:15:52-251 DBV29)
. // selection
. generate age = 2012 - bcgeburt
. keep if inrange(age, 25, 55)
(10,780 observations deleted)
. // compute gross wages and ln(wage)
. generate wage = labgro12 / (bctatzeit * 4.3) if labgro12>0 & bctatzeit>0
(1,936 missing values generated)
. generate lnwage = ln(wage)
(1,936 missing values generated)
. // X variables
. generate schooling = bcbilzeit if bcbilzeit>0
(318 missing values generated)
. generate ft_experience = expft12 if expft12>=0
(15 missing values generated)
. generate ft_experience2 = expft12^2 if expft12>=0
(15 missing values generated)
. generate public = oeffd12==1 if oeffd12>0
(2,274 missing values generated)
. generate female = bcsex==2
. // summarize
. summarize lnwage schooling ft_experience ft_experience2 public female
```

Variable	Obs	Mean	Std. Dev.	Min	Max
lnwage	8,090	2.615219	.5944705	-1.014929	6.818627
schooling	9,708	12.76118	2.73677	7	18
ft_experie~e	10,011	13.41052	10.03473	0	39
ft_experie~2	10,011	280.5277	324.8873	0	1521
public	7,752	.2582559	.4377037	0	1
female	10,026	.5398963	.4984306	0	1

Example analysis

```
. oaxaca lnwage schooling (experience: ft_experience ft_experience2) ///
>     if public==0, by(female) weight(1) nodetail
Blinder-Oaxaca decomposition                    Number of obs     =      5,476
                                                  Model           =     linear
Group 1: female = 0                               N of obs 1      =       2896
Group 2: female = 1                               N of obs 2      =       2580
```

lnwage	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
overall						
group_1	2.735162	.0109447	249.91	0.000	2.71371	2.756613
group_2	2.423025	.0115499	209.79	0.000	2.400388	2.445663
difference	.3121364	.0159118	19.62	0.000	.2809498	.3433231
explained	.1689735	.0116228	14.54	0.000	.1461933	.1917538
unexplained	.1431629	.0160575	8.92	0.000	.1116907	.1746351

```
. oaxaca lnwage schooling (experience: ft_experience ft_experience2) ///
>     if public==1, by(female) weight(1) nodetail
Blinder-Oaxaca decomposition                    Number of obs     =      1,912
                                                  Model           =     linear
Group 1: female = 0                               N of obs 1      =        753
Group 2: female = 1                               N of obs 2      =       1159
```

lnwage	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
overall						
group_1	2.853219	.0152189	187.48	0.000	2.823391	2.883048
group_2	2.682151	.0135559	197.86	0.000	2.655582	2.70872
difference	.1710683	.0203808	8.39	0.000	.1311226	.211014
explained	.1426937	.0155707	9.16	0.000	.1121756	.1732118
unexplained	.0283746	.0199081	1.43	0.154	-.0106445	.0673938

Example analysis

```
. regress lnwage schooling ft_experience ft_experience2 if female==0 & public==0
  (output omitted)
. estimates store male_private
. regress lnwage schooling ft_experience ft_experience2 if female==0 & public==1
  (output omitted)
. estimates store male_public
. regress lnwage schooling ft_experience ft_experience2 if female==1 & public==0
  (output omitted)
. estimates store female_private
. regress lnwage schooling ft_experience ft_experience2 if female==1 & public==1
  (output omitted)
. estimates store female_public
```

Example analysis

```
. smithwelch male_public female_public male_private female_private, ///
>     reference(1) benchmark(1)
```

Decompositions of individual differentials:

	D	E	C
Sample 1	.1710683	.1426937	.0283746
Sample 2	.3121364	.1689735	.1431629

Difference in (components of) differentials:

	dD	dE	dC
	.1410681	.0262799	.1147882

Decomposition of difference in differentials:

	D	E	C
dE	.0262799	.0162967	.0099831
dC	.1147882	-.0140906	.1288789

D = differential / difference in component of differential
E = part of D due to differences in endowments
C = part of D due to differences in coefficients

```
. smithwelch male_public female_public male_private female_private, ///
>     reference(1) benchmark(1) detail(schooling=schooling, experience=ft_exp*)
```

Decompositions of individual differentials:

Sample 1	D	E	C
schooling	.0941913	.0030393	.091152
experience	.2770018	.1396544	.1373475
_cons	-.2001249	0	-.2001249
Total	.1710683	.1426937	.0283746
Sample 2	**D**	**E**	**C**
schooling	.169705	.003751	.165954
experience	.3003422	.1652225	.1351197
_cons	-.1579108	0	-.1579108
Total	.3121364	.1689735	.1431629

Difference in (components of) differentials:

	dD	dE	dC
schooling	.0755137	.0007117	.074802
experience	.0233404	.0255682	-.0022278
_cons	.0422141	0	.0422141
Total	.1410681	.0262799	.1147882

Decomposition of difference in differentials:

dE	D	E	C
schooling	.0007117	-.0003119	.0010236
experience	.0255682	.0166087	.0089595
_cons	0	0	0
Total	.0262799	.0162967	.0099831
dC	**D**	**E**	**C**
schooling	.074802	-.0075696	.0823716
experience	-.0022278	-.006521	.0042932
_cons	.0422141	0	.0422141
Total	.1147882	-.0140906	.1288789

D = differential / difference in component of differential
E = part of D due to differences in endowments
C = part of D due to differences in coefficients

Juhn-Murphy-Pierce 1991 decomposition

(Juhn et al. 1991)

- Juhn et al. (1991) use an alternative setup by considering changes in the residual variance.
- The argument is that the gender wage gap will be large if the residual variance, that is, the variance of wages once controlling for observables such as education or work experience, is large.
- Conceptually, the residual variance can be viewed as the price of unobservables. The idea is, that there may be differences between men and women in such unobservables. If the prices increase, the gender wage gap will increase as well.
- For example, Blau and Kahn (1997) use the method in an analysis in which they argue that the gender wage gap would have declined more than it actually did, hadn't there been a strong increase in general wage inequality that had nothing to do with gender.

Juhn-Murphy-Pierce 1991 decomposition

- Consider a model

$$Y^{gt} = X^{gt}\beta^t + \epsilon^{gt}$$

for two groups, $g = m, f$ (males and females), at two time points, $t = 0, 1$, where β^t are some reference regression parameters (non-discriminatory prices of observables).

- The model can also be expressed as

$$Y^{gt} = X^{gt}\beta^t + r^{gt}\sigma^t$$

where σ is a reference residual standard deviation (non-discriminatory prices of unobservables) and $r = (Y - X\beta^t)/\sigma^t$ is a standardized residual.

- Thus, the equation now has a two-component error term. The residuals are expressed as a function of the general residual inequality at time t and the positions of the residuals in the residual distribution.

Juhn-Murphy-Pierce 1991 decomposition

- The mean outcome differential between men an women at time t can then be decomposed as follows:

$$\begin{aligned}\Delta^{\mu,t} = \bar{Y}^{mt} - \bar{Y}^{ft} &= (\bar{X}^{mt} - \bar{X}^{ft})\beta^t + (\bar{r}^{mt} - \bar{r}^{ft})\sigma^t \\ &= \Delta^{\mu}_{X} + \Delta^{\mu}_{S}\end{aligned}$$

- The first term is the "predicted gap" and the second term is the "residual gap". They are equal to the "explained" part and the "unexplained" part in a standard OB decomposition using β^t as reference coefficients.
- Let $\Delta\bar{X}^t = (\bar{X}^{mt} - \bar{X}^{ft})$ and $\Delta\bar{r}^t = (\bar{r}^{mt} - \bar{r}^{ft})$. Given two time points $t = 0$ and $t = 1$ the change in the outcome differential can then be written as

$$\begin{aligned}\Delta^{\mu,1} - \Delta^{\mu,0} &= (\Delta\bar{X}^1\beta^1 - \Delta\bar{X}^0\beta^0) + (\Delta\bar{r}^1\sigma^1 - \Delta\bar{r}^0\sigma^0) \\ &= d\Delta^{\mu}_{X} + d\Delta^{\mu}_{S}\end{aligned}$$

Juhn-Murphy-Pierce 1991 decomposition

- The change in the "predicted gap" can be further decomposed as

$$d\Delta_X^{\mu} = (\Delta\bar{X}^1 - \Delta\bar{X}^0)\beta^0 + \Delta\bar{X}^0(\beta^1 - \beta^0) + (\Delta\bar{X}^1 - \Delta\bar{X}^0)(\beta^1 - \beta^0)$$

  where the first term is the main "observed quantities" effect due to a change in gender differences in X, the second term is a secondary effect due a change in "prices" for observed quantities, and the third term is an interaction term.

- Likewise, the change in the "residual gap" can be decomposed as

$$d\Delta_S^{\mu} = (\Delta\bar{r}^1 - \Delta\bar{r}^0)\sigma^0 + \Delta\bar{r}^0(\sigma^1 - \sigma^0) + (\Delta\bar{r}^1 - \Delta\bar{r}^0)(\sigma^1 - \sigma^0)$$

  where the first term is the so-called "gap effect" due to changes in the group differences in residual positions (i.e. changes in the group differences in "unobserved quantities" and changes in discrimination), the second term is the part due to changes in residual inequality (i.e. changes in "prices" for unobserved quantities), and the third term is again an interaction term.

Juhn-Murphy-Pierce 1991 decomposition

- Similar to other decompositions, it is common practice to use the "prices" of one of the years as the reference prices (or use some average or pooled results). For example, if we use $t = 0$ as the reference, the decomposition simplifies to:

$$d\Delta_X^\mu = (\Delta\bar{X}^1 - \Delta\bar{X}^0)\beta^0 + \Delta\bar{X}^1(\beta^1 - \beta^0)$$

$$d\Delta_S^\mu = (\Delta\bar{r}^1 - \Delta\bar{r}^0)\sigma^0 + \Delta\bar{r}^1(\sigma^1 - \sigma^0)$$

- Furthermore, a detailed decomposition can be obtained for the components of $d\Delta_X^\mu$ in the usual way (but obviously not for $d\Delta_S^\mu$).
- Note that results for $d\Delta_X^\mu$ are the same as for the Smith-Welch decomposition (if using the same setup). For $d\Delta_S^\mu$ only the total is the same; that is, Smith-Welch and Juhn-Murphy-Pierce lead to a different breakup of $d\Delta_S^\mu$.

Juhn-Murphy-Pierce 1991 decomposition: estimation

- Estimation of the components of $d\Delta_X^{\mu}$ is straightforward.
- Estimation of the components of $d\Delta_S^{\mu}$ is more involved and requires some discussion. Two approaches are used in the literature, a **parametric** approach and a **nonparametric** approach.
- Parametric
  - Since, by definition, $\epsilon^t = r^t\sigma^t$, expression $\Delta\bar{r}^t\sigma^t$ can simply be estimated as the mean difference in residuals ϵ between men an women at time t.
  - But what about expressions such as $\Delta\bar{r}^1\sigma^0$?
  - An obvious solution is to estimate the residual standard deviation at time $t = 0$ and then multiply it by the mean difference in standardized residuals of $t = 1$.
- Nonparametric
  - The parametric approach is simple, but neglects changes in the distributional shape (apart from the variance).
  - The following nonparametric procedure has therefore been proposed.

Juhn-Murphy-Pierce 1991 decomposition: estimation

- Let $F^t()$ be the distribution function of the residuals at time t. Furthermore, let p^t represent the relative positions of the residuals in the residual distribution at time t, that is

$$p^{gt} = F^t(\epsilon^{gt}) \quad \text{and thus} \quad \epsilon^{gt} = Q^t(p^{gt})$$

  where $Q() = F^{-1}()$ is the quantile function (inverse of $F()$).
- The solution now is to apply the quantile function of one time point to the residual ranks of the other time point.
- For example, $\Delta\bar{r}^1\sigma^0$ is estimated by assigning each individual at $t = 1$ a percentile number corresponding to its position in the residual distribution of $t = 1$ (i.e., compute p^1), then using these relative ranks to derive hypothetical residuals given the $t = 0$ residual distribution (i.e. compute $Q^0(p^1)$), and finally taking the mean difference in these hypothetical residuals between men and women.

- The JMP 1991 procedure relies on some strong assumptions and is not free of critique (e.g. Yun 2009).

Example analysis (using `jmpierce2` by Jann 2005b)

```
. jmpierce2 male_public female_public male_private female_private, ///
>     reference(1) benchmark(1)
```

Decomposition of individual differentials:

	raw differential	quantity effect	residual gap
Sample 1	.1710683	.1426937	.0283746
Sample 2	.3121364	.1689735	.1431629

Difference in (components of) differentials:

	D	E	U
Total	.1410681	.0262799	.1147882

Decomposition of difference in predicted gap:

	E	Q	P
Total	.0262799	.0162967	.0099831

Decomposition of diffence in residual gap:

	U	Q	P
Total	.1147882	.0681732	.0466151

D = difference in differential
E = difference in predicted gap
U = difference in residual gap
Q = quantity effect
P = price effect

Example analysis

```
. jmpierce2 male_public female_public male_private female_private, ///
>     reference(1) benchmark(1) parametric
```

Decomposition of individual differentials:

	raw differential	quantity effect	residual gap
Sample 1	.1710683	.1426937	.0283746
Sample 2	.3121364	.1689735	.1431629

Difference in (components of) differentials:

	D	E	U
Total	.1410681	.0262799	.1147882

Decomposition of difference in predicted gap:

	E	Q	P
Total	.0262799	.0162967	.0099831

Decomposition of diffence in residual gap:

	U	Q	P
Total	.1147882	.0683891	.0463991

```
D  = difference in differential
E  = difference in predicted gap
U  = difference in residual gap
Q  = quantity effect
P  = price effect
```

```
. jmpierce2 male_public female_public male_private female_private, ///
>     reference(1) benchmark(1) detail(schooling=schooling, experience=ft_exp*)
```

Decomposition of individual differentials:

	raw differential	quantity effect	residual gap
Sample 1	.1710683	.1426937	.0283746
Sample 2	.3121364	.1689735	.1431629

Difference in (components of) differentials:

	D	E	U
Total	.1410681	.0262799	.1147882

Decomposition of difference in predicted gap:

	E	Q	P
Total	.0262799	.0162967	.0099831
schooling	.0007117	-.0003119	.0010236
experience	.0255682	.0166087	.0089595

Decomposition of diffence in residual gap:

	U	Q	P
Total	.1147882	.0681732	.0466151

```
D  = difference in differential
E  = difference in predicted gap
U  = difference in residual gap
Q  = quantity effect
P  = price effect
```

Exercise 6

- Repeat the example analysis, but evaluate how changing the reference and "benchmark" estimates (as they are called in the help file) affects the results.
- Also try to compute results based on pooled models.

References

- Blau, Francine D., Lawrence M. Kahn (1997). Swimming Upstream: Trends in the Gender Wage Differential in the 1980s. Journal of Labor Economics 15(1):1–42.
- Heckman, James J., Thomas M. Lyons, Petra E. Todd (2000). Understanding Black-White Wage Differentials. American Economic Review 90(2): 344-349.
- Ben Jann (2005a). jmpierce2: Stata module to compute trend decomposition of outcome differentials. Available from https://ideas.repec.org/c/boc/bocode/s448804.html.
- Ben Jann (2005b). smithwelch: Stata module to compute trend decomposition of outcome differentials. Available from http://ideas.repec.org/c/boc/bocode/s448805.html.
- Juhn, Chinhui, Kevin M. Murphy, Brooks Pierce (1991). Accounting for the Slowdown in Black-White Wage Convergence. In Marvin Kosters (Ed.), Workers and Their Wages (pp. 107–143). Washington, DC: AEI Press.
- Smith, James P., Finis R. Welch (1989). Black Economic Progress After Myrdal. Journal of Economic Literature 27(2):519–564.

References

- Yun, Myeong-Su (2009). Wage Differentials, Discrimination and Inequality: A Cautionary Note on the Juhn, Murphy and Pierce Decomposition Method. Scottish Journal of Political Economy 56(1):123–137.