Using Decomposition methods to Solve Economic and Social Problems

Hugo Reis (Banco de Portugal)
Pedro Raposo (Catolica Lisbon School Business and Economics)
Paulo Rodrigues (Banco de Portugal)

13th LABOUR ECONOMICS MEETING - XIII JORNADAS DE ECONOMA LABORAL

June 2019

Roadmap

- **Motivation**
- Refresher on Oaxaca-Blinder decomposition
- Going beyond the mean
- Quantile regression based decomposition (Machado and Mata)
- RIF-regression (Firpo, Fortin, and Lemieux)
- When do covariates matter? Gelbach decomposition

Motivation: what if ... I hadn't taken that train?

- Recently there has been a renewed interest for changes in wage inequality.
- These changes have been characterized as the "polarization" of the U.S. labor market into high-wage and low-wage jobs at the expense of middle-skill jobs (Autor, Katz and Kearney, 2006).
- We would like to characterize the changes in the wage distribution (compositional changes and structural changes)
- What would have happened to the wage distribution today if we had the covariates of 30 years ago (female, occupation, education distribution of 30 years ago)?
- What would have happened to the wage distribution today if we had the coefficients (prices) of 30 years ago (female, occupation, education returns of 30 years ago)?

Motivation: what if ... I hadn't taken that train?

- Using statistical decomposition techniques to identify the main causes of distributional differences in wages started with the methods proposed by Oaxaca(1973) and Blinder (1973).
- Considerable research has been devoted to go beyond the analysis of the mean differences, e.g. Juhn et al. (1993), DiNardo et al.(1996), Machado and Mata (2005), or Firpo et al (2009).
- The use of the entire distribution is important to understand the differences on the bottom or top part of the distribution of study.

Roadmap

- Motivation
- **Refresher on Oaxaca-Blinder decomposition**
- Going beyond the mean
- Quantile regression based decomposition (Machado and Mata)
- RIF-regression (Firpo, Fortin, and Lemieux)
- When do covariates matter? Gelbach decomposition

Refresher on Oaxaca-Blinder decomposition

- Oaxaca (1973) (along with Blinder, 1973) first study aimed at understanding the sources of the large gender gap in wage: (difference in mean wages) of 43 % in the 1967
- Question: how much of the gender gap can be "explained" by male/female differences in human capital (education and labour market experience), occupational choices, etc?
  - The "unexplained" part of the gender gap is often interpreted as representing labour market discrimination, though other interpretations (unobserved skills) are possible too
- Estimate OLS regressions of (log) wages on covariates/characteristics
- Construct a counterfactual wage such as "what would be the average wage of women if they had the same characteristics as men?"
- This forms the basis of the decomposition

Refresher on Oaxaca-Blinder decomposition

- We want to decompose the difference in the mean of an "outcome variable Y between two groups A and B"
- Groups could also be periods, regions, etc.
- If we model the conditional expectation of the variable of interest in state j as $E[W(j)|X] = X\hat{\beta}(j)$ $(j = 0, 1)$, the decomposition reads

$$E[W(1)] - E[W(0)] = \underbrace{\{E[X(1)] - E[X(0)]\}\hat{\beta}(0)}_{\text{covariates}} + \underbrace{E[X(1)][\hat{\beta}(1) - \hat{\beta}(0)]}_{\text{coefficients}}$$

Refresher on Oaxaca-Blinder decomposition

- The covariates part will be called, "explained" or "composition effect" (X) since it reflects differences in the distribution of the X's between the two groups.
- The coefficients part will be called, "unexplained", or "prices" or "structural effect" (β) since it reflects differences in the β between the two groups, i.e. in the way the X's are "priced" (valued) by the market.
- In the "aggregate" decomposition, we only divide into two components (wage structure effect) and (composition effect).
- In the "detailed" decomposition we also look at the contribution of each individual covariate (or β)

Refresher on Oaxaca-Blinder decomposition

- The Oaxaca technique is useful for decomposing mean differences.
- It is not applicable to decomposing densities or differences between densities as formulated above.
- It is a sequential decomposition, meaning that the order of the decomposition (or more specifically, the choice of weights) may affect the conclusions.
- These questions implicitly assume that changing the quantity available of something has no effect on its price (that is, the βs are independent of the Xs).

Data

- Longitudinal linked employer-employee data: *Quadros de Pessoal*
- Years: 1986 to 2017 (1990 and 2001 not available)
- Variables:
  - worker's gender, birth date, schooling, occupation, date of hire, earnings (several components), hours of work (normal and overtime), collective bargaining agreement, worker category in the agreement
  - firm's industry, location, etc.
  - unique identifiers for workers, firms and job titles
- Final dataset (after constraints full-timer, base wage at least 80% min wage, non-missing relevant data, etc; connected set):
  - 28 million observations (.1% sample)
- Hourly wage = (overall monthly earnings, incl, overtime)/(sum of normal and overtime hours)

stata code and results - Oaxaca-Blinder

```
. oaxaca $y $x, by(male) weight(0)

Blinder-Oaxaca decomposition                    Number of obs     =    458,203
                                                  Model           =     linear
Group 1: male = 0                                 N of obs 1      =     191800
Group 2: male = 1                                 N of obs 2      =     266403
```

lhw	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
overall						
group_1	.1823541	.0011366	160.45	0.000	.1801265	.1845817
group_2	.3844937	.0010775	356.83	0.000	.3823818	.3866056
difference	-.2021396	.0015661	-129.07	0.000	-.2052092	-.19907
explained	.0210986	.0011334	18.62	0.000	.0188772	.02332
unexplained	-.2232382	.0011522	-193.75	0.000	-.2254965	-.22098
explained						
age	-.0927789	.0020147	-46.05	0.000	-.0967277	-.0888301
age2	.0670021	.0015122	44.31	0.000	.0640383	.0699659
lfirm	-.0155763	.0005458	-28.54	0.000	-.016646	-.0145065
educ	.0624517	.0008462	73.81	0.000	.0607932	.0641101
unexplained						
age	-.6014306	.0262211	-22.94	0.000	-.652823	-.5500383
age2	.2638998	.0132422	19.93	0.000	.2379457	.289854
lfirm	-.0558149	.0020775	-26.87	0.000	-.0598868	-.051743
educ	.046216	.0027132	17.03	0.000	.0408982	.0515337
_cons	.1238915	.0135584	9.14	0.000	.0973176	.1504654

stata code and results - Oaxaca-Blinder

```
. reg $y $x if male==0
```

Source	SS	df	MS	Number of obs	=	191,800
				F(4, 191795)	=	47639.72
Model	23682.7839	4	5920.69598	Prob > F	=	0.0000
Residual	23836.4113	191,795	.124280671	R-squared	=	0.4984
				Adj R-squared	=	0.4984
Total	47519.1953	191,799	.247755177	Root MSE	=	.35253

lhw	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	.0408534	.0005176	78.93	0.000	.039839	.0418679
age2	-.0003148	6.63e-06	-47.50	0.000	-.0003278	-.0003018
lfirm	.0653484	.0003535	184.84	0.000	.0646554	.0660413
educ	.0819369	.0002289	357.90	0.000	.0814882	.0823856
_cons	-1.813302	.0097762	-185.48	0.000	-1.832463	-1.79414

```
. reg $y $x if male==1
```

Source	SS	df	MS	Number of obs	=	266,403
				F(4, 266398)	=	58197.15
Model	38425.4485	4	9606.36213	Prob > F	=	0.0000
Residual	43973.2139	266,398	.165065856	R-squared	=	0.4663
				Adj R-squared	=	0.4663
Total	82398.6624	266,402	.309301966	Root MSE	=	.40628

lhw	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	.0571479	.0004865	117.47	0.000	.0561944	.0581014
age2	-.0004938	6.06e-06	-81.49	0.000	-.0005057	-.0004819
lfirm	.0786261	.0003449	227.94	0.000	.07795	.0793021
educ	.0763441	.0002353	324.47	0.000	.0758829	.0768053
_cons	-1.937193	.0093944	-206.21	0.000	-1.955606	-1.91878

```
. table male, c(mean age mean age2 mean lfirm mean educ mean yhat)
```

male	mean(age)	mean(age2)	mean(lfirm)	mean(educ)	mean(yhat)
0	36.91005	1474.125	4.203663	8.263494	.1823541
1	38.53354	1609.809	4.401769	7.445464	.3844937

Interpretation of the results - Oaxaca-Blinder

- The gender gap mean-difference is 20.2%
- Composition effect (explained): +2.1% - differences in covariates
- Structural effect (unexplained): -22.3%
- Totally driven by structural effect: potential role for discrimination
- Individual components:
- Composition effect: +6% - if education structure was the same the gap would be smaller by 6%
- Structural effect: mainly age - returns to experience

Roadmap

- Motivation
- Refresher on Oaxaca-Blinder decomposition
- **Going beyond the mean**
- Quantile regression based decomposition (Machado and Mata)
- RIF-regression (Firpo, Fortin, and Lemieux)
- When do covariates matter? Gelbach decomposition

Going beyond the mean

- The discussed Oaxaca-Blinder procedures and their extensions to non-linear models focus on the decomposition of differences in the expected value (mean) of an outcome variable.
- In many cases, however, one is interested in other distributional statistics, say the Gini coefficient or a the D9/D1 quantile ratio, or even in whole distributions (density curves, Lorenz curves).
- The basic setup is the same; an estimate of $F_{Y^g|G\neq g}$ is needed to be able to compute a decomposition such as

$$\Delta^{\nu} = \nu(F_{Y|G=0}) - \nu(F_{Y|G=1})$$
$$\Delta^{\nu} = \nu(F_{Y|G=0}) - \nu(F_{Y^0|G=1}) + \nu(F_{Y^0|G=1}) - \nu(F_{Y|G=1})$$
$$\Delta^{m} = \Delta^{\nu}_{X} + \Delta^{\nu}_{S}$$

where $F_{Y^g|G\neq g}(y) = \int F_{Y|X,G=g}(y|x) f_{X|G\neq g}(x) dx$

Going beyond the mean

- Several approaches have been proposed in the literature:
- Estimating $F_{Y^g|G\neq g}$ by reweighting (DiNardo et al. 1996).
- Imputing values for Y^g in group $G \neq g$
  - based on regression residuals (Juhn et al. 1993)
  - based on quantile regression (Machado and Mata 2005)
- Estimating $F_{Y^g|G\neq g}$ by distribution regression (Chernozhukov et al.2013)
- Estimating $m(F_{Y^g|G\neq g})$ via recentered influence function regression (Firpo et al. 2007, 2009)
- Today, we will only look at quantile and RIF regression.

Roadmap

- Motivation
- Refresher on Oaxaca-Blinder decomposition
- Going beyond the mean
- **Quantile regression based decomposition (Machado and Mata)**
- RIF-regression (Firpo, Fortin, and Lemieux)
- When do covariates matter? Gelbach decomposition

Quantile regression based decomposition (Machado and Mata)

- Their method can be viewed as a generalization of the Oaxaca-Blinder decomposition
- Consider two groups, A and B, with characteristics given by the stochastic vectors X_A for group A and X_B for group B.
- We denote realizations of these stochastic vectors by x_A and x_B.
- Assume that X_A and X_B both have dimension k and have distribution functions G_{X_A} and G_{X_B}, respectively.
- The endogenous variable is Y_A for group A and Y_B for group B with unconditional distribution functions F_{Y_A} and F_{Y_B}, respectively.

Quantile regression based decomposition (Machado and Mata)

- They assume that the regression quantiles are $\beta^A(\tau)$ for group A and $\beta^B(\tau)$ for group B for each $\tau \in [0,1]$:

$$Q_\tau(Y_A|X_A = x_A) = x_A\beta^A(\tau)$$

$$Q_\tau(Y_B|X_B = x_B) = x_B\beta^B(\tau)$$

- The distribution of Y_A conditional on $X_A = x_A$ is completely characterized by the collection of regression quantiles $\{\beta^A(\tau); \tau \in [0,1]\}$, and likewise for the distribution of Y_B conditional on $X_B = x_B$.
- Consider a counterfactual random variable Y_{AB} with the property that its quantiles conditional on x_A are given by

$$Q_\tau(Y_{AB}|X_A = x_A) = x_A\beta^B(\tau)$$

Quantile regression based decomposition (Machado and Mata)

- The MM method generates a sample from the unconditional distribution of Y_{AB} as follows:
  1. Sample τ from a standard uniform distribution.
  2. Compute $\hat{\beta}^B(\tau)$, i.e., estimate the τth regression quantile of Y_B on x_B.
  3. Sample x_A from the empirical distribution $\hat{G}_{X_A}$.
  4. Compute $\hat{y_{AB}} = x_A\hat{\beta}^B(\tau)$.
  5. Repeat steps 1 to 4 R times.

Quantile regression based decomposition (Machado and Mata)

- The method is based on the estimation of marginal wage distributions consistent with a conditional distribution estimated by quantile regression as well as with any hypothesized distribution for the covariates.
- Comparing the marginal distributions implied by different distributions for the covariates, one is then able to perform counterfactual exercises.
- The basic idea of the procedure is quite simple.
- In their paper, they were looking after the distribution of wages that would have prevailed in 1995 if gender had been distributed as in 1986
- That is, the 1995 marginal distribution of wages is a mixture of the wage distribution for men and of the wage distribution for women with weights that equal proportion of men and women in the 1995 workforce.

Quantile regression based decomposition (Machado and Mata)

- It is clear, however, that looking just at means is overly restrictive as a method for analyzing cases such as inequality, where the critical indicators relate to spread and tail weight.
  Machado and Mata (2005) proposes a method to decompose the changes in a given distribution(W) in two sub-populations (indexed by 0 and 1) in several factors contributing to those changes: that is, an Oaxaca-Blinder type decomposition for the **entire** distribution,

$$\text{distrib.}W(0) \rightarrow \text{distrib.}W(1) = \begin{cases} \text{distrib.}X(0) \rightarrow \text{distrib.}X(1) \\ \\ \text{cond. distrib.}W(0)|X \rightarrow \text{cond. distril} \end{cases}$$

stata code and results - Machado and Mata

```
. rqdeco $y $x, by(male) nquantreg(5)

Decomposition of differences in distribution using quantile regression
         Total number of observations                 458203
           Number of observations in group 0          191800
           Number of observations in group 1          266403

         Number of quantile regressions estimated          5

The variance has not been computed.
Use the option vce if you want to compute it.
```

Component	Effects	Std. Err.	t	P>|t|	[95% Conf.	Interval]
Quantile .1						
Raw difference	.118845	.	.	.	.	.
Characteristics	-.022631	.	.	.	.	.
Coefficients	.141475	.	.	.	.	.
Quantile .2						
Raw difference	.146429	.	.	.	.	.
Characteristics	-.024819	.	.	.	.	.
Coefficients	.171248	.	.	.	.	.
Quantile .3						
Raw difference	.168236	.	.	.	.	.
Characteristics	-.026961	.	.	.	.	.
Coefficients	.195197	.	.	.	.	.
Quantile .4						
Raw difference	.190217	.	.	.	.	.
Characteristics	-.028607	.	.	.	.	.
Coefficients	.218823	.	.	.	.	.
Quantile .5						
Raw difference	.208143	.	.	.	.	.
Characteristics	-.030366	.	.	.	.	.
Coefficients	.23851	.	.	.	.	.
Quantile .6						
Raw difference	.225976	.	.	.	.	.
Characteristics	-.032751	.	.	.	.	.
Coefficients	.258727	.	.	.	.	.
Quantile .7						
Raw difference	.24499	.	.	.	.	.
Characteristics	-.033962	.	.	.	.	.
Coefficients	.278952	.	.	.	.	.
Quantile .8						
Raw difference	.256764	.	.	.	.	.
Characteristics	-.035452	.	.	.	.	.
Coefficients	.292216	.	.	.	.	.
Quantile .9						
Raw difference	.259963	.	.	.	.	.
Characteristics	-.036574	.	.	.	.	.
Coefficients	.296537	.	.	.	.	.

stata code and results - Machado and Mata

```
. qreg $y $x if male==0, q(50) nolog

Median regression                                   Number of obs =    191,800
  Raw sum of deviations 34522.27 (about .05319354)
  Min sum of deviations 25114.49                    Pseudo R2     =     0.2725
```

lhw	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	.0357038	.0006157	57.99	0.000	.0344971	.0369104
age2	-.0002755	7.88e-06	-34.94	0.000	-.0002909	-.00026
lfirm	.0580385	.0004205	138.01	0.000	.0572143	.0588627
educ	.0743593	.0002723	273.06	0.000	.0738256	.0748931
_cons	-1.633221	.0116286	-140.45	0.000	-1.656013	-1.61043

```
. qreg $y $x if male==1, q(50) nolog

Median regression                                   Number of obs =    266,403
  Raw sum of deviations 55507.19 (about .26097685)
  Min sum of deviations 40874.85                    Pseudo R2     =     0.2636
```

lhw	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	.0555373	.0005564	99.81	0.000	.0544467	.0566278
age2	-.0004958	6.93e-06	-71.54	0.000	-.0005094	-.0004822
lfirm	.0798557	.0003945	202.41	0.000	.0790824	.0806289
educ	.0703811	.0002691	261.54	0.000	.0698536	.0709085
_cons	-1.873013	.0107445	-174.32	0.000	-1.894072	-1.851954

```
. table male, c(mean age mean age2 mean lfirm mean educ mean yhat50)
```

male	mean(age)	mean(age2)	mean(lfirm)	mean(educ)	mean(yhat50)
0	36.91005	1474.125	4.203663	8.263494	.1369738
1	38.53354	1609.809	4.401769	7.445464	.3444039

Interpretation of the results - Machado and Mata

- The observed median gender gap is 20.9%.
- About -3% is explained by gender differences in the distribution of the covariates. If it was for the differences in the distributions of the covariates, the gap would be smaller by 3%.
- About 24% is due to differing coefficients between men and women and can be interpreted as discrimination.

Roadmap

- Motivation
- Refresher on Oaxaca-Blinder decomposition
- Going beyond the mean
- Quantile regression based decomposition (Machado and Mata)
- **RIF-regression (Firpo, Fortin, and Lemieux)**
- When do covariates matter? Gelbach decomposition

Influence functions

- A very nice approach to compute Oaxaca-Blinder type decompositions for almost any distributional statistic of interest is based on influence functions.
- An influence functions is a function that quantifies how a target statistic changes in response to small changes in the data. That is, for each value y, the influence function $IF(y;\nu,F_Y)$provides an approximation of how the functional $\nu(F_Y)$ changes if a small probability mass is added at point y.
- Influence functions are used in robust statistics to describe the robustness properties of various statistic (a robust statistic has a bounded influence function).
- There is also a close connection to the sampling variance of a statistic. The asymptotic sampling variance of a statistic is equal to the sampling variance of the mean of the influence function.
- Therefore, influence functions provide an easy way to estimate standard errors for many statistics (e.g. inequality measured).

RIF-regression (Firpo, Fortin, and Lemieux)

- For example, the influence function of quantile Q_τ is simply
$$IF(y; Q_\tau, F_Y) = \frac{\tau - I(y \leq Q_\tau)}{f_Y(Q_\tau)}$$
- Influence functions are centered around zero (that is, have an expected value of zero). To center an influence function around the statistic of interest, we can simply add the statistic to the influence function. This is called a **recentered influence function**
$$RIF(y; \nu, F_Y) = \nu(F_Y) + IF(y; \nu, F_Y)$$
- The idea now is to model the conditional expectation of $RIF(y; \nu, F_Y)$ using regression models, e.g. using a linear model
$$E[RIF(y; \nu, F_Y|X)] = X\lambda$$
- Coefficient λ provides an approximation of how $\nu(F_Y)$ reacts to changes in X.

RIF-regression decomposition (Firpo, Fortin, and Lemieux)

- In practice, taking the example of a quantile, we would first compute the sample quantile $\hat{Q}_\tau$ and then use kernel density estimation to get $\hat{f}(\hat{Q}_\tau)$, the density of Y at point $\hat{Q}_\tau$.
- $RIF(Y_i; Q_p, F_Y)$ is then computed for each observation by plugging these estimates in to the above formula.
- Finally, we regress $RIF(Y_i; Q_\tau, F_Y)$ on X to get an estimate of λ.
- Using the coefficients from **RIF regression** in two groups, we can perform an Oaxaca-Blinder type decomposition for Q_τ:

$$\hat{\Delta}^{Q_\tau} = \hat{\Delta}_X^{Q_\tau} + \hat{\Delta}_S^{Q_\tau}$$

$$\hat{\Delta}^{Q_\tau} = (\overline{X}^0 - \overline{X}^1)\hat{\lambda}^0 + \overline{X}^1(\hat{\lambda}^0 - \hat{\lambda}^1)$$

- A similar procedure can be followed for any other statistic $\nu(F_Y)$. All you have to know is the influence function.

RIF-regression decomposition - STATA aplication

- Command **rifreg** can be used to run RIF regressions for quantiles, GINI coefficient, and the variance.
- The RIF variables stored by **rifreg** can be used in **oaxaca**.

stata code and results: Step 1 - RIF

```
. rifreg $y $x if male==0, q(0.5) re(rif_50f)
(266,403 real changes made)
```

Source	SS	df	MS	Number of obs =	191800
				F(4,191795) =	44304.90
Model	13492.5961	4	3373.14903	Prob > F =	0.0000
Residual	26750.382	191795	.139473823	R-squared =	0.3353
				Adj R-squared =	0.3353
Total	40242.9781	191799	.209818498	Root MSE =	.37346

rif_50f	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	.0297937	.00055	54.17	0.000	.0287157	.0308717
age2	-.0002318	7.20e-06	-32.20	0.000	-.0002459	-.0002176
lfirm	.0494659	.000356	138.96	0.000	.0487682	.0501636
educ	.0624462	.0002177	286.84	0.000	.0620195	.0628729
_cons	-1.428819	.0096715	-147.74	0.000	-1.447775	-1.409864

```
. estimates store female50

. rifreg $y $x if male==1, q(0.5) re(rif_50m)
(191,800 real changes made)
```

Source	SS	df	MS	Number of obs =	266403
				F(4,266398) =	49814.43
Model	27079.9033	4	6769.97582	Prob > F =	0.0000
Residual	66142.1371	266398	.24828316	R-squared =	0.2905
				Adj R-squared =	0.2905
Total	93222.0404	266402	.349929957	Root MSE =	.49828

rif_50m	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	.0574839	.0005667	101.43	0.000	.0563731	.0585947
age2	-.0005423	7.22e-06	-75.13	0.000	-.0005565	-.0005282
lfirm	.0764788	.0004005	190.95	0.000	.0756938	.0772638
educ	.0555381	.000267	207.98	0.000	.0550147	.0560615
_cons	-1.83117	.0101754	-179.96	0.000	-1.851114	-1.811227

```
. estimates store male50
```

stata code and results: Step 2 - RIF

```
. oaxaca8 female50 male50, detail notf weight(0)
(high estimates: male50; low estimates: female50)

                                        Mean prediction 1 =  .2609791
                                        Mean prediction 2 =  .0531964
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
difference	.2077827	.0015516	133.91	0.000	.2047416	.2108238

Linear decomposition

W=0	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
explained						
age	.0483697	.0013158	36.76	0.000	.0457907	.0509487
age2	-.0314454	.0011441	-27.49	0.000	-.0336877	-.0292031
lfirm	.0097995	.0003479	28.17	0.000	.0091176	.0104814
educ	-.0510828	.0006971	-73.28	0.000	-.0524491	-.0497165
Total	-.0243591	.0008386	-29.05	0.000	-.0260027	-.0227154
unexplained						
age	1.067	.0304382	35.05	0.000	1.007342	1.126658
age2	-.4999776	.0164186	-30.45	0.000	-.5321574	-.4677977
lfirm	.1189046	.0023618	50.34	0.000	.1142755	.1235337
educ	-.051434	.0025656	-20.05	0.000	-.0564626	-.0464055
_cons	-.402351	.0140384	-28.66	0.000	-.4298658	-.3748363
Total	.2321418	.0013336	174.07	0.000	.2295279	.2347557

Interpretation of the results - RIF

- In this example and setup it provides very similar results
- The gender gap is mainly driven by the structural effect (23.2%) - again age the main component
- The composition effect is only sizable for the education (5.1%)

Roadmap

- Motivation
- Refresher on Oaxaca-Blinder decomposition
- Going beyond the mean
- Quantile regression based decomposition (Machado and Mata)
- RIF-regression (Firpo, Fortin, and Lemieux)
- **When do covariates matter? Gelbach decomposition**

Motivation - Gelbach decomposition

- Suppose we are interested in the male-female wage gap.
- And suppose we begin by estimating the following model:

$$logWages_{it} = \alpha + \beta_1^{base} female_i + \epsilon_{it}$$

where the coefficient β_1^{base} is the mean difference between female and male log wages.

- Naturally, this may be a biased measure of a wage gap
- Males and females may differ along dimensions relevant to wages

Motivation - Gelbach decomposition

- Now, suppose we estimate a new model controlling for other factors:

$$logWages_{it} = \alpha + \beta_1^{full} female_i + \beta_2^{full} age_{it} + \beta_3^{full} education_i + \epsilon_{it}$$

and after the estimation $\beta_1^{base} \neq \beta_1^{full}$

- How much of the movement in the coefficients is attributable to age, and how much is attributable to education?
- How much of the male-female wage gap is explained by gender differences in age? What about education?

Sequence/Order dependence

- Researchers may want to "solve" this problem by estimating one additional intermediate model:

$$logWages_{it} = \alpha + \beta_1^{age} female_i + \beta_2^{age} age_{it} + \epsilon_{it}$$

- and attribute the difference $\beta_1^{base} - \beta_1^{age}$ to age differences
- and attribute the difference $\beta_1^{age} - \beta_1^{full}$ to education differences

Sequence/Order dependence

- BUT if instead we estimated the following intermediate model:

$$logWages_{it} = \alpha + \beta_1^{educ} female_i + \beta_2^{educ} education_i + \epsilon_{it}$$

- and attribute the difference $\beta_1^{base} - \beta_1^{educ}$ to education differences
- and attribute the difference $\beta_1^{educ} - \beta_1^{full}$ to age differences

$$\text{Is } \beta_1^{base} - \beta_1^{age} = \beta_1^{educ} - \beta_1^{full} \text{ ?}$$
$$\text{Is } \beta_1^{base} - \beta_1^{educ} = \beta_1^{age} - \beta_1^{full} \text{ ?}$$

- No, as the result is **order/sequence dependent**
  - Only if female is orthogonal to age and education will the order be irrelevant.

Correct Approach - Gelbach Decomposition

- Gelbach (2016) provides a solution for this problem using the well-known **omitted variables bias** formula
- Implementing the solution requires only three steps:
  1. Estimate the full model and recover $(\beta_1^{full}, \beta_2^{full}, \beta_3^{full})$
  2. Run a regression of age_i on $female_i$ and of $education_i$ on $female_i$; recover the coefficients Γ_{age} and Γ_{educ}, respectively.
  3. The male-female gap attributable to age_i is $\beta_2^{full}\Gamma_{age}$ and to $education_i$ is $\beta_3^{full}\Gamma_{educ}$

$$\beta_2^{full}\Gamma_{age} + \beta_3^{full}\Gamma_{educ} = \beta_1^{base} - \beta_1^{full}$$

Thus the contribution of each variable is its base impact (β_2^{full} and β_3^{full}) multiplied by the mean difference across groups (Γ_{age} and Γ_{educ}), respectively

Gelbach Decomposition

- Consider that the population regression function can be correctly written as a linear function of X_1 and X_2:

$$Y = X_1\beta_1 + X_2\beta_2 + \epsilon$$

- The covariates $X = [X_1, X_2]$ satisfy $E[X\epsilon] = 0$
- Under the assumptions, $\hat{\beta} = [\beta_1^{full}, \beta_2^{full}] = (X'X)^{-1}X'Y$ is a consistent estimator of $\beta = [\beta_1, \beta_2]$

Gelbach Decomposition - Omitted variable bias

- Now, consider the X_1 coefficient from the base specification ignoring X_2

$$\beta_1^{base} = (X_1'X_1)^{-1}X_1'Y$$

- How this coefficient relates to β_1^{full} ?
- We know that: $plim\beta_1^{base} = plim(X_1'X_1)^{-1}X_1'(X_1\beta_1 + X_2\beta_2 + \epsilon)$

$$plim\beta_1^{base} = \beta_1 + (X_1'X_1)^{-1}X_1X_2\beta_2$$

$$plim\beta_1^{base} = \beta_1 + \Gamma\beta_2$$

Gelbach decomposition - Omitted variable bias

- $\delta = \Gamma\beta_2$ is the omitted variable bias from excluding X_2 when estimating β_1, and is nonzero if $E[X_1 X_2] \neq 0$
- It also suggests a natural decomposition of the difference between the **base** and the **full** model coefficients on X_1
- $\delta = \Gamma_1 beta_{2,1} + ... + \Gamma_k beta_{2,k}$
- $\delta = \beta_1^{base} - \beta_1$

stata code and results - Gelbach

```
. group3hdfe nss npc um, largest(big)

.
. reghdfe lhw male age age2 lfirm educ if big, a(nss npc year, savefe)
(dropped 70827 singleton observations)
note: male is probably collinear with the fixed effects (all values close to zero after partialling-out; tol = 1.0e-09)
note: age is probably collinear with the fixed effects (all values close to zero after partialling-out; tol = 1.0e-09)
note: educ is probably collinear with the fixed effects (all values close to zero after partialling-out; tol = 1.0e-09)
(MWFE estimator converged in 268 iterations)
note: male omitted because of collinearity
note: age omitted because of collinearity
note: educ omitted because of collinearity

HDFE Linear regression                            Number of obs   =      8,577
Absorbing 3 HDFE groups                           F(   2,   3720) =      59.22
                                                  Prob > F        =     0.0000
                                                  R-squared       =     0.9709
                                                  Adj R-squared   =     0.9329
                                                  Within R-sq.    =     0.0309
                                                  Root MSE        =     0.1618

------------------------------------------------------------------------------
         lhw |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |          0  (omitted)
         age |          0  (omitted)
        age2 |  -.0003286   .0000309   -10.64   0.000    -.0003892   -.0002681
       lfirm |   .0134696    .008475     1.59   0.112    -.0031465    .0300856
        educ |          0  (omitted)
       _cons |   1.221131   .0854139    14.30   0.000     1.053668    1.388593
------------------------------------------------------------------------------

Absorbed degrees of freedom:
-----------------------------------------------------+
 Absorbed FE | Categories  - Redundant  = Num. Coefs |
-------------+---------------------------------------|
         nss |       4175           0        4175   |
         npc |        652           1         651   |
        year |         30           1          29   ? |
-----------------------------------------------------+
? = number of redundant parameters may be higher
```

stata code and results - Gelbach

Table: Decomposition of the gender gap

gap	full	worker fe	firm fe
0.2412	-	0.1993	0.0419

Note: Gelbach (2016).

Interpretation of the results - Gelbach

- The wage penalty of around 24% can be decomposed into two components: worker and firm unobserved heterogeneity
- A non-negligible fraction of the gender gap is explained by the heterogeneity of the firms' compensation policies. The allocation of workers into firms is responsible for 4%.
- If workers were randomly allocated to firms, the gender gap would be reduced by more than 15%.
- The unobserved (permanent) attributes of the workers is responsible for the remaining 85%
- Unobserved skills or, simply, some form of gender discrimination.