Lecture 8: Decomposition Method

*Introduction to Econometrics, Fall 2019*

Zhaopeng Qu

Nanjing University

11/12/2020
1. Review Previous Lectures
2. Oaxaca-Blinder Decomposition
3. Reference group problem
4. Detailed Decomposition
5. Standard Errors
6. Introduction to bootstrap (Maybe Skipped)
7. Representative Applications
Section 1

Review Previous Lectures
Topics covered

Main Content

- Build a framework of Causal Inference
- Review Basic Probability and Statistics
- Simple OLS: Estimation and Inference
- Multiple OLS: Estimation and Inference
- Function forms: Nonlinear in independent variables
- Nonlinear Regression model: Dummy dependent variable
- Comprehensive Evaluations to OLS regression Model
Two explicite Assumptions

- So far, all models we learned have to be satisfied two strong hypotheses:

  1. **No Heterogeneity**: If the sample could be divided by \( m \) heterogenous groups, then we assume that the estimate coefficient \( \beta_j \) for the \( j \)th independent variable, \( X_j \) are the same among all groups of the sample. Thus

     \[
     \beta_{j,1} = \beta_{j,2} = \ldots = \beta_{j,M}
     \]

     for any group \( G_m : m = 1, 2, \ldots, M \)

  2. **No Endogeneity (Internal Valid)**: there is no endogeneity in these estimating models. Essentially, the 1st Assumption of identification in OLS model is satisfied. Thus

     \[
     E(u_i | X_1, X_2, \ldots, X_k) = 0
     \]
An Simple Extension: Decomposition Method

1. Heterogeneity: Gap between two groups (more than interactions)

2. Exogenous conditional on controlling variables

   - Ignorable or Conditional Independence Assumption (CIA), thus assume that

   \[ E(u_i | X_1, X_2, ..., X_k) = 0 \]

   , the first assumption of OLS regression.
Wage Decomposition: Introduction

- Wage Decomposition methods are used to analyze **distributional differences** in an outcome variable **between groups or time points**.

- In particular, the methods decompose the **observed difference** between two groups (or across two time points) into a component that is due to **compositional differences** between the groups, and a component that is due to **differential mechanisms**.
A Classical Case: **Gender Wage Gap**

- How can the difference in average wages between men and women be explained?

- Is the gap due to

  1. group differences in wage determinants (i.e. in characteristics that are relevant for wages, such as education)? (*compositional differences*)
   2. differential compensation for these determinants (e.g. different returns to education for men and women, or wage discrimination against women)? (*differential mechanisms*)

- The typical question is needed to answer is “what the pay (or other outcomes) would be *if women had* the same characteristics as men?”

- It will help us construct a *counterfactual* state by using decomposition method to recovery the causal effect (sort of causal) of a certain factor or a groups of determinants.
Decomposition Methods: Introduction

- The method can be tracked back from the seminal work by Solow (1957) for “growth accounting”.

- Seminal Work: Oaxaca (1973) and Blinder (1973), who analyzed mean wage differences between groups (males vs. females, whites vs. blacks).

- Once widely used in the area of labor economics, especially in the topics of earnings inequality since 1990s. Now it is popular used in many topics in many fields of economics.

- These more recent developments focus on topics such as
  - distributional measures other than the mean
  - non-linear models for categorical variables
  - taking into selection bias and other type endogenous.
  - combine with other quasi-experimental methods
  - extending into spatial econometric model
Decomposition Methods to Gaps: Two Categories

1. In Mean
   - Oaxaca-Blinder (1974): OB
   - Fairlie (1999): Fairlie

2. In Distribution (Skipped)
   - Juhn, Murphy and Pierce (1993): JMP
   - Machado and Mata (2005): MM
   - DiNardo, Fortin and Lemieux (1996): DFL
   - Firpo, Fortin and Lemieux (2007, 2010): FFL
Decomposition Methods: Pros and Cons

- **Pros**
  - It is a naturally way to distangle cause and effect based on OLS or other Linear regressions.

- **Cons**
  - In particular, decomposition methods inherently follow a partial equilibrium.
  - The results cannot be fully explained as causal inference.
Decomposition Methods to Gaps

- Although some of methods listed above is quite sophisticated and frontier in the filed, the OB is so fundamental that all other methods can explained by it.

- Therefore, in our lecture, we will only cover OB and its extension versions.
Section 2

Oaxaca-Blinder Decomposition
A naive way to identification gender gap

- Use a dummy variable in a regression function

\[ Y = \beta_0 + \beta_1 D + \Gamma X' + u \]

- \( D = 1 \) denotes that the gender of the sample is male, and \( D = 0 \) denotes female.
- \( X' \) denotes a series control variables, thus personal characteristics such as education, working experience, etc.
- So if \( \hat{\beta}_1 \) is large enough and significant statistically,
- then the result can only answer to that question: “is there a wage gap between men and women in the labor market when other things equal(\( X \))?"
Assume that a multiple OLS regression equation is

\[ Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + u_i \]

where \( Y_i \) is dependent variable, \( X_i \)s are a series independent(controlling) variables which affect \( Y_i \). And \( u_i \) are error terms which satisfied by \( E(u_i | X_1, \ldots, X_k) = 0 \)

The means of \( Y_i \)

\[ E(Y) = \beta_0 + \beta_1 E(X_1) + \ldots + \beta_k E(X_k) + E(u_i) \]

Using the sample estimator to replace the population parameters and considering the definition of error term, thus \( \sum u_i = 0 \), then

\[ \bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X}_1 + \ldots + \hat{\beta}_k \bar{X}_k \]
Oaxaca-Blinder Decomposition: Two groups

- If we assume that whole sample can be divided into 2 groups: A and B, then we could regress the similar regression using A and B subsamples, respectively. Thus,

\[
Y_{Ai} = \beta_{A0} + \beta_{A1}X_{1i} + \ldots + \beta_{Ak}X_{ki} + u_{Ai}
\]
\[
Y_{Bi} = \beta_{B0} + \beta_{B1}X_{1i} + \ldots + \beta_{Bk}X_{ki} + u_{Bi}
\]

- Accordingly, we can obtain the means of outcome \( Y \) for group A and group B are

\[
\bar{Y}_A = \hat{\beta}_0^A + \hat{\beta}_1^A \bar{X}_{A1} + \ldots + \hat{\beta}_k^A \bar{X}_{Ak}
\]
\[
\bar{Y}_B = \hat{\beta}_0^B + \hat{\beta}_1^B \bar{X}_{B1} + \ldots + \hat{\beta}_k^B \bar{X}_{Bk}
\]
Oaxaca-Blinder Decomposition: Two groups

- Denote
  \[ \bar{X}_A = (1, \bar{X}_{A1}, \bar{X}_{A2}, ..., \bar{X}_{Ak}) \]

- And
  \[ \hat{\beta}_A = (\hat{\beta}_0^A, \hat{\beta}_1^A, \hat{\beta}_2^A, ..., \hat{\beta}_k^A) \]

- Then
  \[ \bar{Y}^A = \hat{\beta}_A \bar{X}'_A \]

- Denote as the same way
  \[ \bar{Y}^B = \hat{\beta}_B \bar{X}'_B \]
Oaxaca-Blinder Decomposition: difference in mean

- The difference in mean of $Y_i$ of group A and B is
  \[ \bar{Y}_A - \bar{Y}_B = \hat{\beta}_A \bar{X}_A' - \hat{\beta}_B \bar{X}_B' \]

- A small trick: plus and minus a term $\hat{\beta}_B \bar{X}_A'$, then
  \[
  \begin{align*}
  \bar{Y}_A - \bar{Y}_B &= \hat{\beta}_A \bar{X}_A' - \hat{\beta}_B \bar{X}_B' \\
  &= \hat{\beta}_A \bar{X}_A' - \hat{\beta}_B \bar{X}_B' + \hat{\beta}_B \bar{X}_A' - \hat{\beta}_B \bar{X}_B' \\
  &= (\hat{\beta}_A - \hat{\beta}_B) \bar{X}_A' + \hat{\beta}_B (\bar{X}_A' - \bar{X}_B')
  \end{align*}
  \]

- Then the second term is **characteristics effect** which describes how much the difference of outcome, $Y$, in mean is due to differences in the levels of explanatory variables (characteristics).

- the first term is **coefficients effect** which describes how much the difference of outcome, $Y$, in mean is due to differences in the magnitude of regression coefficients.
Male-female average wage gap can be attributed into two parts:

1. **Explained Part:** due to differences in the levels of explanatory variables: such as schooling years, experience, tenure, industry, occupation, etc.

   - **characteristics effect**
   - **endowment effect**
   - **composition effect**

In the literature of labor economics, we think that the wage gap due to this part is reasonable...
Male-female average wage gap can be attributed into two parts:

- **Unexplained Part:** due to differences in the coefficients to explanatory variables: such as *returns* to schooling years, experience and tenure and *premium* in industry and occupation, etc
  - coefficients effect
  - returns effect
  - structure effect

- In the literature of labor economics, we think that the wage gap due to this part is unreasonable, often it is called *discrimination* part...
Gustafsson and Li (2000): Gender gaps in China

Table 7. Results of decomposition of gender difference of earnings in urban China

<table>
<thead>
<tr>
<th>1988</th>
<th>$\beta mX_m - \beta mX_f$</th>
<th>Percent of total</th>
<th>$\beta mX_f - \beta fX_f$</th>
<th>Percent of total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0</td>
<td>0</td>
<td>0.3628</td>
<td>203.12</td>
</tr>
<tr>
<td>Age group</td>
<td>0.0340</td>
<td>19.02</td>
<td>0.0110</td>
<td>6.14</td>
</tr>
<tr>
<td>Minority status</td>
<td>0.00005</td>
<td>0.03</td>
<td>0.0011</td>
<td>0.59</td>
</tr>
<tr>
<td>Party membership</td>
<td>0.0124</td>
<td>6.92</td>
<td>-0.0057</td>
<td>-3.19</td>
</tr>
<tr>
<td>Education</td>
<td>0.0056</td>
<td>3.14</td>
<td>0.0059</td>
<td>3.33</td>
</tr>
<tr>
<td>Ownership</td>
<td>0.0184</td>
<td>10.32</td>
<td>-0.0354</td>
<td>-19.83</td>
</tr>
<tr>
<td>Occupation</td>
<td>0.0122</td>
<td>6.85</td>
<td>-0.1476</td>
<td>-82.64</td>
</tr>
<tr>
<td>Economic sector</td>
<td>-0.0003</td>
<td>-0.16</td>
<td>-0.1240</td>
<td>-69.41</td>
</tr>
<tr>
<td>Type of job</td>
<td>0.0039</td>
<td>2.17</td>
<td>0.0067</td>
<td>3.76</td>
</tr>
<tr>
<td>Province</td>
<td>-0.0014</td>
<td>-0.78</td>
<td>0.0190</td>
<td>10.62</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>0.0849</strong></td>
<td><strong>47.51</strong></td>
<td><strong>0.0937</strong></td>
<td><strong>52.49</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>1995</th>
<th>$\beta mX_m - \beta mX_f$</th>
<th>Percent of total</th>
<th>$\beta mX_f - \beta fX_f$</th>
<th>Percent of total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0</td>
<td>0</td>
<td>0.0462</td>
<td>19.87</td>
</tr>
<tr>
<td>Age group</td>
<td>0.0169</td>
<td>7.28</td>
<td>0.0645</td>
<td>27.74</td>
</tr>
<tr>
<td>Minority status</td>
<td>0.0001</td>
<td>0.02</td>
<td>0.0014</td>
<td>0.59</td>
</tr>
<tr>
<td>Party membership</td>
<td>0.0142</td>
<td>6.12</td>
<td>-0.0037</td>
<td>-1.60</td>
</tr>
<tr>
<td>Education</td>
<td>0.0172</td>
<td>7.40</td>
<td>0.0001</td>
<td>0.02</td>
</tr>
<tr>
<td>Ownership</td>
<td>0.0208</td>
<td>8.96</td>
<td>-0.0163</td>
<td>-7.03</td>
</tr>
<tr>
<td>Occupation</td>
<td>0.0114</td>
<td>4.92</td>
<td>-0.0199</td>
<td>-8.58</td>
</tr>
<tr>
<td>Economic sector</td>
<td>0.0003</td>
<td>0.14</td>
<td>0.0087</td>
<td>3.76</td>
</tr>
<tr>
<td>Type of job</td>
<td>0.0026</td>
<td>1.12</td>
<td>0.0060</td>
<td>2.59</td>
</tr>
<tr>
<td>Province</td>
<td>0.0020</td>
<td>0.84</td>
<td>0.0601</td>
<td>25.86</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>0.0855</strong></td>
<td><strong>36.80</strong></td>
<td><strong>0.1469</strong></td>
<td><strong>63.20</strong></td>
</tr>
</tbody>
</table>

*Source:* Urban household income surveys 1989 and 1996.
OB Decomposition is a tool for separating the influences of quantities and prices on an observed mean difference.

The aim of the OB decomposition is to explain how much of the difference in mean outcomes across two groups is due to group differences in the levels of explanatory variables, and how much is due to differences in the magnitude of regression coefficients (Oaxaca 1973; Blinder 1973).

Although most applications of the technique can be found in the labor market and discrimination literature, it can also be useful in other fields.

In general, the technique can be employed to study group differences in any (continuous or categorical) outcome variable.
Section 3

Reference group problem
A different reference group

- what if we use a different reference group: plus and minus a term $\hat{\beta}_A \bar{X}'_B$, then

$$\bar{Y}_A - \bar{Y}_B = \hat{\beta}_A \bar{X}'_A - \hat{\beta}_B \bar{X}'_B$$

$$= \hat{\beta}_A \bar{X}'_A - \hat{\beta}_A \bar{X}'_B + \hat{\beta}_A \bar{X}'_B - \hat{\beta}_B \bar{X}'_B$$

$$= (\hat{\beta}_A - \hat{\beta}_B) \bar{X}'_A + \hat{\beta}_B (\bar{X}'_A - \bar{X}'_B)$$

- Then again the first term is characteristics effect or endowment effect as the amount of $X_j$ can be seen as an endowment for group A or B.

- The second term is coefficients effect or price(returns) effect as the estimate coefficients $\hat{\beta}_j$ can be seen as the market price of or the returns to a certain $X_j$.

- Question: is the result as same as the first decomposition?
What is the **true** coefficient or characteristics effect?

**Oaxaca-Blinder Decomposition**

- \( \overline{Y}_A \) vs. \( \overline{Y}_B \)
- \( \overline{X}_B \) vs. \( \overline{X}_A \)
- \( \frac{1}{2} (\overline{X}_A - \overline{X}_B) \)
Let $Y^*$ be a **nondiscriminatory potential** outcome, so $\beta^*$ is such a **nondiscriminatory** coefficient **vector**, and $X$ is still a vector of many $x$(characteristics). Then they satisfy as a following equation

\[
Y^* = X\beta^* + \epsilon
\]

where $\epsilon$ is the error term and satisfies $E(\epsilon|X) = 0$
Oaxaca-Blinder Decomposition: a general framework

Then the difference of the potential outcomes between two groups can then be decomposed as follows:

$$\bar{Y}_A - \bar{Y}_B = \bar{X}_A' \hat{\beta}_A - \bar{X}_B' \hat{\beta}_B$$

$$= \bar{X}_A' \hat{\beta}_A - \bar{X}_A' \hat{\beta}^* + \bar{X}_A' \hat{\beta}^* - \bar{X}_B' \hat{\beta}^* + \bar{X}_B' \hat{\beta}^* - \bar{X}_B' \hat{\beta}_B$$

$$= (\bar{X}_A' - \bar{X}_B') \hat{\beta}^* + [\bar{X}_A' (\hat{\beta}_A - \hat{\beta}^*) + \bar{X}_B' (\hat{\beta}^* - \hat{\beta}_B)]$$
Oaxaca-Blinder Decomposition: a general framework

- The first term, \((\bar{X}_A' - \bar{X}_B')\hat{\beta}^*\) is the explained part as usual
  - characteristics effect
  - endowment effect
  - composition effect
The second term, the **unexplained part** can further be subdivided into:

1. “discrimination” in favor of group A (such as Men)
   \[ \bar{X}_A' (\hat{\beta}_A - \hat{\beta}^*) \]

2. “discrimination” against group B (such as Women)
   \[ \bar{X}_B' (\hat{\beta}^* - \hat{\beta}_B) \]

All variables are known but the nondiscriminatory coefficients $\beta^*$. So how to determine it?
Oaxaca-Blinder Decomposition: One reference group

- Assume that discrimination is directed toward only one group.

- Recall

\[ \bar{Y}_A - \bar{Y}_B = (\bar{X}'_A - \bar{X}'_B)\hat{\beta}^* + [\bar{X}'_A(\hat{\beta}_A - \hat{\beta}^*) + \bar{X}'_B(\hat{\beta}^* - \hat{\beta}_B)] \]

- Assume that wage discrimination is directed only against women (denoted as group B) and there is no (positive) discrimination (favor) of men (denoted as group A).

- Then \( \beta^* = \beta_A \) and the wage gap can be decomposed into

\[ \bar{Y}_A - \bar{Y}_B = (\bar{X}'_A - \bar{X}'_B)\hat{\beta}_A + \bar{X}'_B(\hat{\beta}_A - \hat{\beta}_B) \]
Similarly, if there is only (positive) discrimination (favor) of men but no discrimination of women, then $\beta^* = \beta_B$, and the decomposition is

$$\bar{Y}_A - \bar{Y}_B = (\bar{X}'_A - \bar{X}'_B)\hat{\beta}_B + \bar{X}'_A(\hat{\beta}_A - \hat{\beta}_B)$$
Oaxaca-Blinder Decomposition: Weighted reference group

- However, there is no specific reason to assume that the coefficients of one or the other group are nondiscriminating.

- So the value of $\beta^*$ should be a math combination of $\hat{\beta}_A$ and $\hat{\beta}_B$,

  - Reimers (1983) therefore proposes using the average coefficients over both groups as an estimate for the nondiscriminatory parameter vector; that is,

    $$\hat{\beta}^* = 0.5 \hat{\beta}_A + 0.5 \hat{\beta}_B$$

- Cotton (1988) suggests to weight the coefficients by the group sizes, $n_A$ and $n_B$,

  $$\hat{\beta}^* = \frac{n_A}{n_A + n_B} \hat{\beta}_A + \frac{n_B}{n_A + n_B} \hat{\beta}_B$$
Oaxaca-Blinder Decomposition: Weighted Matrix

More general, let $W$ be a $k + 1$ diagonal matrix of weights, such that

$$
\beta^* = W \hat{\beta}_A + (1 - W) \hat{\beta}_B
$$

Note here

$$
\hat{\beta}_{A,B} = (\hat{\beta}_{0}^{A,B}, \hat{\beta}_{1}^{A,B}, \hat{\beta}_{2}^{A,B}, \ldots, \hat{\beta}_{k}^{A,B})
$$
Oaxaca-Blinder Decomposition: Weighted Matrix

- Then the difference between two groups can be expressed as

\[\bar{Y}_A - \bar{Y}_B = (\bar{X}_A' - \bar{X}_B')[W\hat{\beta}_A + (I - W)\hat{\beta}_B] \]
\[[(I - W)'\bar{X}_A + W\bar{X}_B](\hat{\beta}_A - \hat{\beta}_B)\]

- \(W\) is a matrix of relative weights given to the coefficients of group \(A\), and \(I\) is the identity matrix.

  - e.g. If we choose \(W = I\), then it is equivalent to setting

    \[\beta^* = \beta_A\]

  - e.g. If we choose \(W = 0.5I\), then it is equivalent to setting

    \[\beta^* = 0.5\beta_A + 0.5\beta_B\]
Oaxaca-Blinder Decomposition: Weighted Matrix

- Oaxaca and Ransom (1994) show that $\hat{W}$ can be used following the equation to estimate:

$$\hat{W} = \Omega = (X'X)^{-1}(X_A'X_A)$$

- Neumark (1988) also use the coefficients from a pooled model over both groups as the reference coefficients, thus:

We pool all data and run a regression:

$$Y = X\beta$$

Then $\beta^*$ can be obtained by

$$\beta^* = (X'X)^{-1}(X'Y)$$
However, Oaxaca and Ransom (1994) and Neumark (1998) can inappropriately transfer some of the unexplained parts of the differential into the explained component. 

Assume a simple OLS equation: \( Y_i \) on a single regressor \( X_i \) and a group specific intercepts \( \alpha_A \) and \( \alpha_B \)

\[
Y_{Ai} = \alpha_A + \gamma_A X_{Ai} + u_{Ai} \\
Y_{Bi} = \alpha_B + \gamma_B X_{Ai} + u_{Bi}
\]

Let \( \alpha_A = \alpha \) and \( \alpha_B = \alpha + \delta \), where \( \delta \) is the discrimination parameter. Then the model can also be expressed as

\[
Y = \alpha + \gamma X + \delta D + u
\]

where \( D \) as an indicator for group B, such as “female” in gender wage gap case.
Assume that $\gamma > 0$ (positive relation between $X$ and $Y$) and $\delta < 0$ (discrimination against women).

The true model is

$$Y = \alpha + \gamma X + \delta D + u$$

But if as Oaxaca and Ransom (1994) suggested, we only estimate

$$Y = \alpha + \gamma X + e$$

Then following the *Omitted Variable Bias* formula, we can obtain

$$\gamma^* = \gamma + \delta \frac{Cov(X, D)}{Var(X)}$$
Oaxaca-Blinder Decomposition: OVB and Weighted

Then the **explained part** of the differential is

\[
(\bar{X}_A - \bar{X}_B)\gamma^* = (\bar{X}_A - \bar{X}_B)[\gamma + \delta \frac{Cov(X, D)}{Var(X)}]
\]

Note: \(\delta\), the discrimination parameter, which belongs to the **unexplained** parts of the gap, now attributes to the **explained** part of the gap.
To address the OVB problem in decomposition, Jann (2008) suggested estimate a pooled regression over both groups but controlling group membership (a dummy variable $D$), that is

$$Y = \beta^* X + \delta D + \varepsilon$$

In this case,

$$\hat{\beta}^* = ((X, D)'(X, D))^{-1}(X, D)'Y$$

And the coefficient effect or unexplained part of the difference is $\hat{\delta}$, which is the coefficient of D in the pooled regression now.

The most widely used weighted method for OB decomposition right now.
Reference group: Wrap up

- On the different circumstances, the value could be quite different.
  - One reference group: A or B
  - Weighted reference group
    - simple weight:
    - weighted in matrix: Omega and Pooled

- In practice,
  1. We could use one reference group method but both A and B. If the two result are similar, then it is OK.
  2. If the simple method does not work (not similar), then we have to adjusted the weight.
Section 4

Detailed Decomposition
The detailed contributions of the single predictors or sets of predictors are subject to investigation.

For example, one might want to evaluate how much of the gender wage gap is due to differences in education and how much is due to differences in work experience.

Similarly, it might be informative to determine how much of the unexplained gap is related to differing returns to education and how much is related to differing returns to work experience.
Identifying the contributions of the individual predictors to the explained part of the differential is relative easy.

Because the total component is a simple sum over the individual contributions. Thus

\[(\bar{X}_A - \bar{X}_B)'\hat{\beta}_A = (\bar{X}_{1A} - \bar{X}_{1B})\hat{\beta}_{1A} + (\bar{X}_{2A} - \bar{X}_{2B})\hat{\beta}_{2A} + \ldots\]

The first summand reflects the contribution of the group differences in $X_1$; the second, of differences in $X_2$; and so on.
Detailed Decomposition: the unexplained part

- the individual contributions to the unexplained part are the summands in

\[
\tilde{X}_B'(\hat{\beta}_A - \hat{\beta}_B) = (\hat{\beta}_{0A} - \hat{\beta}_{0B}) + (\hat{\beta}_{1A} - \hat{\beta}_{1B})\tilde{X}_{1B} + (\hat{\beta}_{2A} - \hat{\beta}_{2B})\tilde{X}_{2B} \ldots
\]
Detailed Decomposition: sets of covariates

- Furthermore, it is easy to subsume the detailed decomposition by sets of covariates

- The **explained part** of every set

\[
(\bar{X}_A - \bar{X}_B)' \hat{\beta}_A = \sum_{k=1}^{a} \hat{\beta}_k A(\bar{X}_{kA} - \bar{X}_{kB}) + \sum_{j=a+1}^{b} \hat{\beta}_j A(\bar{X}_{jA} - \bar{X}_{jB}) + \ldots
\]

- The **unexplained part** of every set

\[
\bar{X}_B' (\hat{\beta}_A - \hat{\beta}_B) = (\hat{\beta}_{0A} - \hat{\beta}_{0B}) + \sum_{k=1}^{a} (\hat{\beta}_{kA} - \hat{\beta}_{kB}) \bar{X}_{kB} + \sum_{j=a+1}^{b} (\hat{\beta}_{jA} - \hat{\beta}_{jB}) \bar{X}_{j}\]
Section 5

Standard Errors
The computation of the decomposition components is straightforward:

- Estimate OLS models and insert the coefficients and the means of the regressors into the formulas.

For a long time, results from OB decompositions were reported **without** information on statistical inference (standard errors, confidence intervals).

Without reporting s.e. or C.I is problematic

- because it is hard to evaluate the significance of reported decomposition results without knowing anything about their sampling distribution.
Think of a term such as $\bar{X}\hat{\beta}$, where $\bar{X}$ is a row vector of sample means and $\hat{\beta}$ is a column vector of regression coefficients (the result is a scalar).

- How can its sampling variance, $V(\bar{X}\hat{\beta})$ be estimated?

- Following Jann(2005), the sampling variance is

$$Var(\bar{X}\hat{\beta}) = \bar{X}Var(\hat{\beta})\bar{X}' + \hat{\beta}'Var(\bar{X})\hat{\beta} + trace[Var(\bar{X})Var(\hat{\beta})]$$
The last term, $\text{trace}[\text{Var}(\bar{X})\text{Var}(\hat{\beta})]$, will be asymptotically vanishing and can be ignored when $n$ is enough large.

To estimate $\text{Var}(\bar{X}\hat{\beta})$, plug in estimates for $\text{Var}(\hat{\beta})$ (the variance-covariance matrix of the regression coefficients) and $\text{Var}(\bar{X})$ (the variance-covariance matrix of the means), which are readily available.
Recall OB decomposition:

\[ \bar{Y}_A - \bar{Y}_B = \bar{X}_A(\hat{\beta}_A - \hat{\beta}_B) + (\bar{X}_A - \bar{X}_B)\hat{\beta}_B \]

So corresponding the first term’s variance is as follows

\[
Var[\bar{X}_A(\hat{\beta}_A - \hat{\beta}_B)] = Var[\bar{X}_A\hat{\beta}_A - \bar{X}_A\hat{\beta}_B] \\
= Var[\bar{X}_A\hat{\beta}_A] - Var[\bar{X}_A\hat{\beta}_B] \\
\approx \bar{X}_A[V(\hat{\beta}_A) + V(\hat{\beta}_B)]\bar{X}'_A + (\hat{\beta}_A - \hat{\beta}_B)'V(\bar{X}_A)(\hat{\beta}_A - \hat{\beta}_B)
\]

Similarly,

\[
Var[(\bar{X}_A - \bar{X}_B)\hat{\beta}_B] \approx (\bar{X}_A - \bar{X}_B)V(\hat{\beta}_B)(\bar{X}_A - \bar{X}_B)' + \hat{\beta}_B'V(\bar{X}_A + \bar{X}_B)\hat{\beta}_B
\]

Equations for other variants of the decomposition, for elements of the...
Section 6

Introduction to bootstrap (Maybe Skipped)
Introduction

- In many cases where formulas for standard errors are hard to obtain mathematically, or where they are thought not to be very good approximations to the true sampling variation of an estimator, we can rely on a **resampling method**.

- The general idea is to treat the observed data as a population that we can draw samples from. The most common resampling method is the **bootstrap**.
Introduction

- In short, the bootstrap takes the sample (the values of the independent and dependent variables) as the population and the estimates of the sample as true values.

- Instead of drawing from a specified distribution (such as the normal) by a random number generator, the bootstrap draws with replacement from the sample.

- It therefore takes the empirical distribution function as the true distribution function.

- The great advantage is that we neither make assumption about the distributions nor about the true values of the parameters.
The Method: Nonparametric Bootstrap

- actually there are several bootstrap methods.

- A very simple approach is to use the quantiles of the bootstrap sampling distribution of the estimator to establish the end points of a confidence interval nonparametrically.
Bootstrap Standard Errors

- The empirical standard deviation of a series of bootstrap replications of $\hat{\beta}$ can be used to approximate the standard error $se(\hat{\beta})$

1. Draw $B$ independent bootstrap samples $(Y_i^*, X_i^*)$ of size $N$ from original sample $(Y_i, X_i)$. Usually $B = 100$ replications are sufficient.

2. Estimate the parameter $\beta$ of interest for each bootstrap sample:

$$\hat{\beta}_b^* \text{ for } b = 1, 2, ..., B$$
Bootstrap Standard Errors

3. Estimate $se(\hat{\beta})$ by

$$se_{Boot}(\hat{\beta}) = \sqrt{\frac{1}{B - 1} \sum_{b=1}^{B} (\hat{\beta}_b^* - \bar{\hat{\beta}}^*)^2}$$

where $\bar{\hat{\beta}}^* = \frac{1}{B} \sum_{b=1}^{B} \hat{\beta}_b^*$, thus the average of the B bootstrap estimates.

When the estimator $\hat{\beta}$ is consistent and asymptotically normally distributed, bootstrap standard errors can be used to construct approximate confidence intervals and to perform asymptotic tests based on the normal distribution.

4. Then we can construct 95% confidence interval for $\beta$

$$[\hat{\beta} - 1.96 \times se_{Boot}(\hat{\beta}), \hat{\beta} + 1.96 \times se_{Boot}(\hat{\beta})]$$
Bootstrap: Alternative way to construct Confidence Intervals

- We can construct a two-sided equal-tailed $1 - \alpha$ confidence interval for an estimate $\hat{\beta}$ from the empirical distribution function of a series of bootstrap replications.
- The $\frac{\alpha}{2}$ and the $1 - \frac{\alpha}{2}$ empirical percentiles of the bootstrap replications are used as lower and upper confidence bounds. This procedure is called percentile bootstrap.

1. Draw $B$ independent bootstrap samples $(Y^*_i, X^*_i)$ of size $N$ from original sample $(Y_i, X_i)$. Usually $B = 1000$ replications are sufficient.

2. Estimate the parameter $\beta$ of interest for each bootstrap sample:

$$\hat{\beta}^*_b \text{ for } b = 1, 2, \ldots, B$$
Order the bootstrap replications of $\hat{\beta}$ such that $\hat{\beta}_1^* \leq \ldots \leq \hat{\beta}_B^*$.

- The lower and upper confidence bounds are the $B \times \frac{\alpha}{2}$th and $B \times (1 - \frac{\alpha}{2}) - th$ ordered elements, respectively.

- For example, $B = 1000$ and $\alpha = 0.05$, then these are the $25th$ and $975th$ ordered elements.

- The estimated $1 - \alpha$ confidence interval of $\hat{\beta}$ is

$$[\hat{\beta}_{B \frac{\alpha}{2}}^*, \hat{\beta}_{B(1-\frac{\alpha}{2})}^*]$$
Bootstrap: t-statistic

- Review: Assume that we have consistent estimates of $\hat{\beta}$ and $\hat{se}(\hat{\beta})$ at hand and that the asymptotic distribution of the t-statistic is the standard normal, thus

$$t = \frac{\hat{\beta} - \beta_0}{\hat{se}(\hat{\beta})} \xrightarrow{d} N(0, 1)$$

- Then we can calculate approximate critical values from percentiles of the empirical distribution of a series of bootstrap replications for the t-statistic.

1. Consistently estimate $\beta$ and $se(\beta)$ using the originally observed sample:

$$\hat{\beta}, \hat{se}(\hat{\beta})$$
**Bootstrap: t-statistic**

1. Draw $B$ independent bootstrap samples $(Y^*_i, X^*_i)$ of size $N$ from original sample $(Y_i, X_i)$. Usually $B = 1000$ replications are sufficient.

2. Estimate the $t$-value assuming $\beta_0 = \hat{\beta}$ for each bootstrap sample:

   $$t_b^* = \frac{\hat{\beta}_b^* - \hat{\beta}}{\hat{se}_b^*(\hat{\beta})} \text{ for } b = 1, 2, ..., B$$

3. Order the bootstrap replications of $t$ such that $t_1^* \leq ... \leq t_B^*$.

   - The lower and upper confidence bounds are the $B \times \frac{\alpha}{2} - th$ and $B \times (1 - \frac{\alpha}{2}) - th$ ordered elements, respectively.

   - For example, $B = 1000$ and $\alpha = 0.05$, then these are the $25th$ and $975th$ ordered elements.

   - So the critical values are

     $$t_{\frac{\alpha}{2}} = t_B^* \frac{\alpha}{2}, \quad t_{1-\frac{\alpha}{2}} = t_B^*(1- \frac{\alpha}{2})$$
What are the estimators in decomposition?

the unexplained part:

$$\bar{X}_B' (\hat{\beta}_A - \hat{\beta}_B)$$

the explained part:

$$(\bar{X}_A - \bar{X}_B)' \hat{\beta}_A$$
If the bootstrap is so simple and of such broad application, why isn’t it used more in the social sciences?

Because the bootstrap is computationally intensive. This barrier to bootstrapping is more apparent than real.

When the outcome of one of many small steps immediately affects the next, rapid results are important.
Section 7

Representative Applications
Examples

1. Labor Economics: Wage or Income Gaps
   - Gender: Male-Female
   - Urban-Rural (or Urban-Migrant)
   - Minority-Majority (Racial Gaps)
   - Poor-Nonpoor
   - Public-Private Sectors gaps
   - Union-NonUnion gaps
### 代表应用

章莉等 (2014)：中国劳动力市场上工资收入的户籍歧视

### 表 4 工资户籍差异 Oaxaca-Blinder 分解结果 (CHIPs2007)

<table>
<thead>
<tr>
<th></th>
<th>(1)标准分解</th>
<th>(2)反向分解</th>
<th>(3)Omega分解</th>
<th>(4)全样本分解</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>E[ln(w_a)] - E[ln(w_m)]</strong></td>
<td>0.6456 (0.0130)</td>
<td>0.6456 (0.0130)</td>
<td>0.6456 (0.0130)</td>
<td>0.6456 (0.0130)</td>
</tr>
</tbody>
</table>

可解释部分

| A | 年龄 | -0.0458 (0.0094) | -0.0444 (0.0084) | -0.0291 (0.0065) | -0.0490 (0.0067) |
| B | 教育 | 0.1652 (0.0105) | 0.1598 (0.0112) | 0.2071 (0.0088) | 0.1710 (0.0089) |
| C | 工作经验 | 0.1246 (0.0093) | 0.0376 (0.0210) | 0.1354 (0.0074) | 0.1209 (0.0073) |
| D | 性别 | -0.0079 (0.0021) | -0.0050 (0.0014) | -0.0057 (0.0015) | -0.0065 (0.0017) |
| E | 民族 | -0.0001 (0.0005) | 0.0005 (0.0004) | 0.0005 (0.0003) | 0.0004 (0.0003) |
| F | 婚姻状况 | 0.0240 (0.0056) | 0.0177 (0.0044) | 0.0232 (0.0037) | 0.0231 (0.0036) |
### 不可解释部分

<table>
<thead>
<tr>
<th></th>
<th>年龄</th>
<th></th>
<th>工作经验</th>
<th></th>
<th>性别</th>
<th></th>
<th>民族</th>
<th></th>
<th>婚姻状况</th>
<th></th>
<th>地区</th>
<th></th>
<th>行业</th>
<th></th>
<th>职业</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>-0.0059</td>
<td></td>
<td>-0.0073</td>
<td></td>
<td>-0.0226</td>
<td></td>
<td>-0.0027</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0526)</td>
<td></td>
<td>(0.0651)</td>
<td></td>
<td>(0.0605)</td>
<td></td>
<td>(0.0605)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>b</td>
<td>0.0139</td>
<td></td>
<td>0.0192</td>
<td></td>
<td>-0.0281</td>
<td></td>
<td>0.0080</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0384)</td>
<td></td>
<td>(0.0532)</td>
<td></td>
<td>(0.0497)</td>
<td></td>
<td>(0.0495)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>c</td>
<td>0.0107</td>
<td></td>
<td>0.0977</td>
<td></td>
<td>-0.0001</td>
<td></td>
<td>-0.0144</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0129)</td>
<td></td>
<td>(0.0287)</td>
<td></td>
<td>(0.0174)</td>
<td></td>
<td>(0.0173)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>d</td>
<td>0.0481</td>
<td></td>
<td>0.0453</td>
<td></td>
<td>0.0460</td>
<td></td>
<td>0.0467</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0147)</td>
<td></td>
<td>(0.0138)</td>
<td></td>
<td>(0.0144)</td>
<td></td>
<td>(0.0144)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>e</td>
<td>-0.0973</td>
<td></td>
<td>-0.0979</td>
<td></td>
<td>-0.0978</td>
<td></td>
<td>-0.0978</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.1017)</td>
<td></td>
<td>(0.1023)</td>
<td></td>
<td>(0.0925)</td>
<td></td>
<td>(0.0925)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>f</td>
<td>0.0201</td>
<td></td>
<td>0.0265</td>
<td></td>
<td>0.0210</td>
<td></td>
<td>0.0210</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0224)</td>
<td></td>
<td>(0.0294)</td>
<td></td>
<td>(0.0268)</td>
<td></td>
<td>(0.0268)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>g</td>
<td>0.0738</td>
<td></td>
<td>0.0675</td>
<td></td>
<td>0.0704</td>
<td></td>
<td>0.0707</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0247)</td>
<td></td>
<td>(0.0238)</td>
<td></td>
<td>(0.0238)</td>
<td></td>
<td>(0.0238)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>h</td>
<td>-0.1421</td>
<td></td>
<td>-0.0711</td>
<td></td>
<td>-0.1163</td>
<td></td>
<td>-0.1049</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0423)</td>
<td></td>
<td>(0.0340)</td>
<td></td>
<td>(0.0409)</td>
<td></td>
<td>(0.0409)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>i</td>
<td>-0.2503</td>
<td></td>
<td>-0.1197</td>
<td></td>
<td>-0.2511</td>
<td></td>
<td>-0.2311</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Fortin and Lemiux (2011)

<table>
<thead>
<tr>
<th>Reference Group</th>
<th>(1) Using Male Coef. from col. 2, Table 2</th>
<th>(2) Using Male Coef. from col. 4, Table 2</th>
<th>(3) Using Female Coef.</th>
<th>(4) Using Weighted Sum</th>
<th>(5) Using Pooled from col. 5, Table 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unadjusted mean log wage gap : ( E[\ln(w_m)] - E[\ln(w_f)] )</td>
<td>0.233 (0.015)</td>
<td>0.233 (0.015)</td>
<td>0.233 (0.015)</td>
<td>0.233 (0.015)</td>
<td>0.233 (0.015)</td>
</tr>
<tr>
<td>Composition effects attributable to</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age, race, region, etc.</td>
<td>0.012 (0.003)</td>
<td>0.012 (0.003)</td>
<td>0.009 (0.003)</td>
<td>0.011 (0.003)</td>
<td>0.010 (0.003)</td>
</tr>
<tr>
<td>Education</td>
<td>-0.012 (0.006)</td>
<td>-0.012 (0.006)</td>
<td>-0.008 (0.004)</td>
<td>-0.010 (0.005)</td>
<td>-0.010 (0.005)</td>
</tr>
<tr>
<td>AFQT</td>
<td>0.011 (0.003)</td>
<td>0.011 (0.003)</td>
<td>0.011 (0.003)</td>
<td>0.011 (0.003)</td>
<td>0.011 (0.003)</td>
</tr>
<tr>
<td>L.T. withdrawal due to family</td>
<td>0.033 (0.011)</td>
<td>0.033 (0.011)</td>
<td>0.035 (0.008)</td>
<td>0.034 (0.007)</td>
<td>0.028 (0.007)</td>
</tr>
<tr>
<td>Life-time work experience</td>
<td>0.137 (0.011)</td>
<td>0.137 (0.011)</td>
<td>0.087 (0.01)</td>
<td>0.112 (0.008)</td>
<td>0.092 (0.007)</td>
</tr>
<tr>
<td>Industrial sectors</td>
<td>0.017 (0.006)</td>
<td>0.017 (0.006)</td>
<td>0.003 (0.005)</td>
<td>0.010 (0.004)</td>
<td>0.009 (0.004)</td>
</tr>
<tr>
<td>Total explained by model</td>
<td>0.197 (0.018)</td>
<td>0.197 (0.018)</td>
<td>0.136 (0.014)</td>
<td>0.167 (0.013)</td>
<td>0.142 (0.012)</td>
</tr>
<tr>
<td>Wage structure effects attributable to</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age, race, region, etc.</td>
<td>-0.098 (0.234)</td>
<td>-0.098 (0.234)</td>
<td>-0.096 (0.232)</td>
<td>-0.097 (0.233)</td>
<td>-0.097 (0.24)</td>
</tr>
<tr>
<td>Education</td>
<td>0.045 (0.034)</td>
<td>0.045 (0.034)</td>
<td>0.041 (0.033)</td>
<td>0.043 (0.034)</td>
<td>0.043 (0.031)</td>
</tr>
<tr>
<td>AFQT</td>
<td>0.003 (0.023)</td>
<td>0.003 (0.023)</td>
<td>0.003 (0.025)</td>
<td>0.003 (0.024)</td>
<td>0.002 (0.025)</td>
</tr>
<tr>
<td>L.T. withdrawal due to family</td>
<td>0.003 (0.017)</td>
<td>0.003 (0.017)</td>
<td>0.001 (0.004)</td>
<td>0.002 (0.011)</td>
<td>0.007 (0.01)</td>
</tr>
<tr>
<td>Life-time work experience</td>
<td>0.048 (0.062)</td>
<td>0.048 (0.062)</td>
<td>0.098 (0.067)</td>
<td>0.073 (0.064)</td>
<td>0.092 (0.065)</td>
</tr>
<tr>
<td>Industrial sectors</td>
<td>-0.092 (0.033)</td>
<td>0.014 (0.028)</td>
<td>-0.077 (0.029)</td>
<td>-0.085 (0.031)</td>
<td>-0.084 (0.032)</td>
</tr>
<tr>
<td>Constant</td>
<td>0.128 (0.213)</td>
<td>0.022 (0.212)</td>
<td>0.193 (0.211)</td>
<td>0.128 (0.213)</td>
<td>0.128 (0.216)</td>
</tr>
<tr>
<td>Total wage structure -</td>
<td>0.036 (0.019)</td>
<td>0.036 (0.019)</td>
<td>0.097 (0.016)</td>
<td>0.066 (0.015)</td>
<td>0.092 (0.014)</td>
</tr>
<tr>
<td>Unexplained log wage gap</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Examples

2 Other Fields:

- Educational Performance: Fortin, Oreopoulos and Phipps (2017)
- Family Origins: Li, Ling and Qu (2018)
- House Price:
- Health Status:
Concluding Remarks and Discussions:

- OB decomposition can be easily extended in some nonlinear regression models.
- But OB method decompose the gap only on the mean.
- The result may depend on the choice of counterfactual fact if you neglect the reference group problem.
- Intrinsically, a partial equilibrium approach to analyze a general equilibrium question.
- Question: how is extent to trust that the result have a causal explanation in the decomposition?
王美艳（2005）,“城市劳动力市场上的就业机会与工资差异——外来劳动力就业与报酬研究”,《中国社会科学》第5期，第36-46页。

章莉，李实，William Darity, Rhonda Sharpe（2014）,“中国劳动力市场上工资收入的户籍歧视”《管理世界》第11期，第35-46页。


Li, Shi, Xiaoguang Ling and Zhaopeng Qu（2018）,“The Red and The Black in the 21st Century: The Long-Term Effects of Parental Socioeconomic Status on Children’s Well-Beings in China”, Working paper