Decomposition Methods in the Social Sciences

GESIS Training Course

January 29 – February 1, 2024, Cologne

Johannes Giesecke (Humboldt University Berlin)

Ben Jann (University of Bern)

4. Functional form

Some issues with the Oaxaca-Blinder decomposition

- The OB decomposition seems useful and easy to understand, but there are several complications we need to discuss.
  - The index problem
  - The transformation problem / base category problem
  - **Functional form**

Contents

Nonlinear effects and interactions

- The OB decomposition is based on linearity and additive separability.
- If important nonlinearities and interaction effects are ignored, the results may be misleading.
- Hence, care should be exercised when specifying the regression equation on which the decomposition is based.
- Detailed decomposition:
  - The detailed decomposition rests on the assumption of additive separability of the variable for which detailed results are to be obtained.
  - Thus, for example, if modeling polynomials, it does not make much sense to report results for the single terms. The sum of the contributions across all terms, however, has a clear interpretation.
  - Likewise, in case of interactions, it is not really clear how to separate the contributions of the individual variables.
- Reweighting (see later) may be a method to detect misspecification.

1. Nonlinear effects and interactions

2. Extension to nonlinear models
   - Aggregate decomposition
   - Detailed decomposition
   - Example analysis
   - Note on separation of direct and indirect effects

Extension to nonlinear models

- The dependent variable is not always continuous and unbounded.
- In many applications we are interested in other types of variables.
  - dichotomous variables (logit/probit)
  - polytomous variables (unordered: mlogit, ordered: ologit)
  - count data (poisson regression, nbreg, zero-inflated models)
  - censored data (tobit)
  - truncated data (truncreg)
- How can group differences in expected values (proportions in case of categorical variables) be decomposed for these types of variables?
- (There is also some literature on decompositions for survival analysis; see Powers and Yun 2009.)

Aggregate decomposition for nonlinear models

- The general setup is still the same, that is we are interested in a decomposition such as

$$\begin{aligned}\Delta^{\mu} &= \mu(F_{Y|G=0}) - \mu(F_{Y|G=1}) \\ &= \{\mu(F_{Y|G=0}) - \mu(F_{Y^0|G=1})\} + \{\mu(F_{Y^0|G=1}) - \mu(F_{Y|G=1})\} \\ &= \{E(Y|G=0) - E(Y^0|G=1)\} + \{E(Y^0|G=1) - E(Y|G=1)\} \\ &= \Delta^{\mu}_X + \Delta^{\mu}_S\end{aligned}$$

  where $E(Y)$ is the expected value or Y (the mean or a proportion).

- In linear regression we have $Y = m(X, \epsilon) = X\beta + \epsilon$ with $E(\epsilon|X) = 0$ such that

$$E(Y) = E(X\beta + \epsilon) = E(X)\beta$$

  and thus

$$\begin{aligned}\Delta^{\mu} &= \{E(Y|G=0) - E(Y^0|G=1)\} + \{E(Y^0|G=1) - E(Y|G=1)\} \\ &= (E(X|G=0) - E(X|G=1))\beta^0 + E(X|G=1)(\beta^0 - \beta^1) \\ &= \Delta^{\mu}_X + \Delta^{\mu}_S\end{aligned}$$

Aggregate decomposition for nonlinear models

- In general, we can write $\mathrm{E}(Y|X) = h(X;\beta)$.
- In linear regression we have $h(X;\beta) = X\beta$ (linear function).
- In nonlinear models, however, where $h()$ is a nonlinear function.
- For example, if Y is a binary outcome and we use logistic regression, we have
$$\mathrm{E}(Y|X) = h(X;\beta) = \frac{1}{1+e^{-X\beta}}$$
- If $h()$ is nonlinear, then
$$\mathrm{E}(Y) = \mathrm{E}(\mathrm{E}(Y|X)) = \mathrm{E}(h(X;\beta)) \neq h(\mathrm{E}(X);\beta)$$
- That is, we cannot just plug in $\mathrm{E}(X)$ into $h()$ to obtain $\mathrm{E}(Y)$, as is done in the linear OB decomposition.

Aggregate decomposition for nonlinear models

- Estimating expressions such as $\mathrm{E}(Y|G=g)$ is no problem because Y is observed; instead of computing $h(\bar{X}^g;\widehat{\beta}^g)$ as in the linear OB decomposition we can simply compute the mean of Y in the $G=g$ subsample.
- How can we estimate a counterfactual such as $\mathrm{E}(Y^0|G=1)$?
- Using $h(\bar{X}^1;\widehat{\beta}^0)$ as in the linear OB decomposition does not work because in the nonlinear case

$$\mathrm{E}(h(X;\beta^0)|G=1) \neq h(\mathrm{E}(X|G=1);\beta^0)$$

- Instead we have to estimate $\mathrm{E}(h(X;\beta^0)|G=1)$ directly.
- The general solution is to make out-of-sample predictions from the estimated models, and then average over these predictions, that is, compute $\widehat{Y}_i^0 = h(X_i;\widehat{\beta}^0)$ and then take the average $\frac{1}{N^1}\sum_{G_i=1}\widehat{Y}_i^0$ where N^1 is the number of observations in group 1.

Aggregate decomposition for nonlinear models

- The decomposition estimate then is

$$
\begin{aligned}
\widehat{\Delta}^{\mu} &= \left\{ \widehat{\mathrm{E}}(Y|G=0) - \widehat{\mathrm{E}}(\widehat{Y}^0|G=1) \right\} + \left\{ \widehat{\mathrm{E}}(\widehat{Y}^0|G=1) - \widehat{\mathrm{E}}(Y|G=1) \right\} \\
&= \widehat{\Delta}^{\mu}_{X} + \widehat{\Delta}^{\mu}_{S}
\end{aligned}
$$

- In practice, all we need to know is how to generate $\widehat{Y} = \widehat{E}(Y|X) = h(X; \widehat{\beta})$, that is, we need to know function $h()$.
- This illustrates that an aggregate decomposition is possible for just about any model and variable type.
- Bauer and Sinning (2008) provide an overview for various models and also provide a command called `nldecompose` that computes the aggregate decomposition (Sinning et al. 2008).
  - Supported models are `regress`, `logit`, `probit`, `ologit`, `oprobit`, `tobit`, `intreg`, `truncreg`, `poisson`, `nbreg`, `zip`, `zinb`, `ztp`, and `ztnb`.

1. Nonlinear effects and interactions

2. Extension to nonlinear models
   - Aggregate decomposition
   - Detailed decomposition
   - Example analysis
   - Note on separation of direct and indirect effects

Detailed decomposition for nonlinear models

- Decompositions for nonlinear models have the same general complications as the linear OB decomposition (index problem, transformation problem, base category problem for categorical predictors, correct model specification).
- In addition, obtaining a detailed decomposition is not as straightforward as in the linear decomposition.
  - Due to the nonlinearity Δ_X^μ and Δ_S^μ cannot be easily subdivided into additive components; the contribution of a particular X depends on the values of all other covariates.
  - There is no "best" way for dealing with this problem.
- Some solutions:
  - Use average marginal effects.
  - Use a series of counterfactuals switching covariates sequentially.
  - Linearization around $\mathrm{E}(X)\beta$.
  - For binary outcomes: apply the standard OB decomposition to a linear probability model (LPM).

Using marginal effects

- The idea is to use the standard formulas of the OB decomposition, but replace the coefficients by average marginal effects.
- That is, use

$$\widehat{\Delta}^{\mu} = \widehat{\Delta}^{\mu}_{X} + \widehat{\Delta}^{\mu}_{S} = (\bar{X}^0 - \bar{X}^1)\widehat{\delta}^0 + \bar{X}^1(\widehat{\delta}^0 - \widehat{\delta}^1)$$

  where $\widehat{\delta}$ are average marginal effects of the covariates on $\mathrm{E}(Y|X)$.
- The contributions of a single covariate X_k then are

$$\widehat{\Delta}^{\mu}_{X,X_k} = \widehat{\delta}^0_k(\bar{X}^0_k - \bar{X}^1_k) \quad \text{and} \quad \widehat{\Delta}^{\mu}_{S,X_k} = (\widehat{\delta}^0_k - \widehat{\delta}^1_k)\bar{X}^1_k$$

- One problem is that the individual contributions do not add up to the total.
- See Bartus (2006), who provides command `gdecomp`.

Using sequential counterfactuals

- For computing the contributions to Δ_X^μ, Fairlie (2005) proposes to sequentially adjust the X variables from one group to the other (similar approach: Gomulka and Stern 1990).
- Let
$$\widehat{\Delta}_X^\mu = \frac{1}{N^0} \sum_{G_i=0} h(X_i \widehat{\beta}^0) - \frac{1}{N^1} \sum_{G_i=1} h(X_i \widehat{\beta}^0)$$
- Let the two groups be of equal size: $N = N^0 = N^1$.
- We can then rearrange the data such that the variables of the two groups are placed side by side (one-to-one matching of observations between groups); let X^0 and X^1 denote the variables of group 0 and group 1, respectively.

Using sequential counterfactuals

- The decomposition term can then be written as

$$
\begin{aligned}
\widehat{\Delta}^{\mu}_{X} &= \frac{1}{N}\sum_{i=1}^{N}\Big\{h(X_i^0\widehat{\beta}^0) - h(X_i^1\widehat{\beta}^0)\Big\} \\
&= \frac{1}{N}\sum_{i=1}^{N}\Big\{h(\widehat{\beta}_0^0 + \widehat{\beta}_1^0 X_{1i}^0 + \widehat{\beta}_2^0 X_{2i}^0 + \cdots + \widehat{\beta}_K^0 X_{Ki}^0) \\
&\qquad\qquad - h(\widehat{\beta}_0^0 + \widehat{\beta}_1^0 X_{1i}^1 + \widehat{\beta}_2^0 X_{2i}^1 + \cdots + \widehat{\beta}_K^0 X_{Ki}^1)\Big\}
\end{aligned}
$$

- This idea now is to start with X_k^0 in both terms and then sequentially replace X_k^0 by X_k^1 moving from left to right:

$$
\begin{aligned}
\widehat{\Delta}^{\mu}_{X,X_1} &= \frac{1}{N}\sum_i\Big\{h(\widehat{\beta}_0^0 + \widehat{\beta}_1^0 X_{1i}^0 + \widehat{\beta}_2^0 X_{2i}^0 + \cdots + \widehat{\beta}_K^0 X_{Ki}^0) - h(\widehat{\beta}_0^0 + \widehat{\beta}_1^0 X_{1i}^1 + \widehat{\beta}_2^0 X_{2i}^0 + \cdots + \widehat{\beta}_K^0 X_{Ki}^0)\Big\} \\
\widehat{\Delta}^{\mu}_{X,X_2} &= \frac{1}{N}\sum_i\Big\{h(\widehat{\beta}_0^0 + \widehat{\beta}_1^0 X_{1i}^1 + \widehat{\beta}_2^0 X_{2i}^0 + \cdots + \widehat{\beta}_K^0 X_{Ki}^0) - h(\widehat{\beta}_0^0 + \widehat{\beta}_1^0 X_{1i}^1 + \widehat{\beta}_2^0 X_{2i}^1 + \cdots + \widehat{\beta}_K^0 X_{Ki}^0)\Big\} \\
&\vdots \\
\widehat{\Delta}^{\mu}_{X,X_K} &= \frac{1}{N}\sum_i\Big\{h(\widehat{\beta}_0^0 + \widehat{\beta}_1^0 X_{1i}^1 + \widehat{\beta}_2^0 X_{2i}^1 + \cdots + \widehat{\beta}_K^0 X_{Ki}^0) - h(\widehat{\beta}_0^0 + \widehat{\beta}_1^0 X_{1i}^1 + \widehat{\beta}_2^0 X_{2i}^1 + \cdots + \widehat{\beta}_K^0 X_{Ki}^1)\Big\}
\end{aligned}
$$

Using sequential counterfactuals

- If the sample sizes differ, the suggestion is to use a random sample of observations from the larger group (and repeat the decomposition R times and report the average).
  - In case of sampling weights, the one-to-one matching is problematic. A solution here is to draw samples form both groups with sampling probabilities proportional to the weights (and average over R repetitions).
- The sequential approach leads to results that are path dependent. The suggestion is to randomize the order of the covariates (and average over R repetitions).
- A question also is how to match the observations. In practice the observations are matched by their ranks in the (group-specific) distribution of predicted outcomes. (Fairlie (2005) claims, that the exact procedure should not have a large effect on the results.)

Using linearization

- Yun (2004) suggest determining the individual contributions of the covariates to Δ_X^μ and Δ_S^μ in relation to their relative contributions in a decomposition at the level of the linear predictor.
- Let $\widehat{\mathrm{E}}(X|G=g) = \bar{X}^g$ and $\widehat{\mathrm{E}}(h(X\beta)|G=g) = \overline{h(X\beta)}^g$. The aggregate decomposition can then be written as

$$\widehat{\Delta}^\mu = \left\{\overline{h(X\widehat{\beta}^0)}^0 - \overline{h(X\widehat{\beta}^0)}^1\right\} + \left\{\overline{h(X\widehat{\beta}^0)}^1 - \overline{h(X\widehat{\beta}^1)}^1\right\} = \widehat{\Delta}_X^\mu + \widehat{\Delta}_S^\mu$$

- The proposal now is to determine the individual contributions as

$$\widehat{\Delta}_{X,X_k}^\mu = \frac{(\bar{X}_k^0 - \bar{X}_k^1)\widehat{\beta}_k^0}{(\bar{X}^0 - \bar{X}^1)\widehat{\beta}^0}\widehat{\Delta}_X^\mu \quad \text{and} \quad \widehat{\Delta}_{S,\beta_k}^\mu = \frac{\bar{X}_k^1(\widehat{\beta}_k^0 - \widehat{\beta}_k^1)}{\bar{X}^1(\widehat{\beta}^0 - \widehat{\beta}^1)}\widehat{\Delta}_S^\mu$$

such that $\sum_{i=1}^K \widehat{\Delta}_{X,X_k}^\mu = \widehat{\Delta}_X^\mu$ and $\sum_{i=1}^K \widehat{\Delta}_{S,X_k}^\mu = \widehat{\Delta}_S^\mu$.

Using linearization

- Yun (2004) derives this solution by approximating $\widehat{\Delta}^{\mu}$ by evaluating the functions at the means of the covariates, that is,

$$\widehat{\Delta}^{\mu} \approx [h(\bar{X}^0\widehat{\beta}^0) - h(\bar{X}^1\widehat{\beta}^0)] + [h(\bar{X}^1\widehat{\beta}^0) - h(\bar{X}^1\widehat{\beta}^1)]$$

  and then further linearizing the differences around $\bar{X}^0\widehat{\beta}^0$ and $\bar{X}^1\widehat{\beta}^1$ using a first order Taylor expansion:

$$\widehat{\Delta}^{\mu} \approx ((\bar{X}^0 - \bar{X}^1)\widehat{\beta}^0) \cdot d^0 + (\bar{X}^1(\widehat{\beta}^0 - \widehat{\beta}^1)) \cdot d^1$$

  where d^g denotes the derivative of $h(\bar{X}^g\widehat{\beta}^g)$.

- The relative contributions to this approximate decomposition are

$$\frac{((\bar{X}^0_k - \bar{X}^1_k)\widehat{\beta}^0_k)d^0}{((\bar{X}^0 - \bar{X}^1)\widehat{\beta}^0)d^0} = \frac{(\bar{X}^0_k - \bar{X}^1_k)\widehat{\beta}^0_k}{(\bar{X}^0 - \bar{X}^1)\widehat{\beta}^0} \quad \text{and} \quad \frac{(\bar{X}^1_k(\widehat{\beta}^0_k - \widehat{\beta}^1_k))d^1}{(\bar{X}^1(\widehat{\beta}^0 - \widehat{\beta}^1))d^1} = \frac{\bar{X}^1_k(\widehat{\beta}^0_k - \widehat{\beta}^1_k)}{\bar{X}^1(\widehat{\beta}^0 - \widehat{\beta}^1)}$$

  which are then multiplied by $\widehat{\Delta}^{\mu}_X$ and $\widehat{\Delta}^{\mu}_S$ to ensure that the individual contributions sum up to the correct total.

Using linearization

- A problem of this approach is that it is not clear how good the approximation is.
- If the bulk of the data is in highly nonlinear regions of $h()$, if differences in coefficients are large, or differences in the means of the covariates are large, the approximation may be poor.

Using LPM

- Finally, for binary outcomes, why not simply apply a standard OB decomposition using a linear probability model (LPM)? (i.e. just apply `oaxaca` with default options)
- After all, the LPM also models conditional probabilities (albeit making crudely simplifying functional form assumptions).
- It is not apriori clear why an approximate approach such as the Yun decomposition should be better than an approximate approach such as the LPM decomposition.
- Both approaches will run into similar problems if linearization approximation is poor.

1. Nonlinear effects and interactions

2. Extension to nonlinear models
   - Aggregate decomposition
   - Detailed decomposition
   - Example analysis
   - Note on separation of direct and indirect effects

Stata implementations

`nldecompose`	aggregate decomposition for various nonlinear models; no detailed decomposition (Bauer and Sinning 2008)
`gdecomp`	detailed decomposition based on marginal effects for several nonlinear models (requires `margeff`) (Bartus 2006)
`fairlie`	Fairlie decomposition for logit and probit (Jann 2006)
`mvdcmp`	Yun decomposition for several nonlinear models (Powers et al. 2011)
`oaxaca`	LPM decomposition; Yun decomposition for logit and probit (requires the version of `oaxaca` from the SSC Archive; the version archived at the Stata Journal site is an outdated version that does not support the Yun decomposition)

Example: Leadership position and gender

```
. use gsoep-extract, clear
(Example data based on the German Socio-Economic Panel)
. keep if wave==2015
(29,970 observations deleted)
. keep if inrange(age, 25, 55)
(5,671 observations deleted)
. // Y: supervising others/leadership position
. fre supvis
supvis — supervision
```

		Freq.	Percent	Valid	Cum.
Valid	0 no	4174	57.11	72.50	72.50
	1 yes	1583	21.66	27.50	100.00
	Total	5757	78.77	100.00	
Missing	.	1552	21.23		
Total		7309	100.00		

```
. // covariates
. generate byte male = sex==1
. generate byte female = 1 - male
. summarize yeduc expft exppt male
```

Variable	Obs	Mean	Std. dev.	Min	Max
yeduc	7,121	12.28823	2.783974	7	18
expft	7,274	11.63359	9.556508	0	39.5
exppt	7,274	3.271481	5.052598	0	35.25
male	7,309	.4338487	.4956386	0	1

Gender gap in supervision

```
. svyset psu [pw=weight], strata(strata)
Sampling weights: weight
             VCE: linearized
     Single unit: missing
        Strata 1: strata
 Sampling unit 1: psu
           FPC 1: <zero>
. svy: mean supvis if !missing(yeduc, expft, exppt), over(male)
(running mean on estimation sample)
Survey: Mean estimation
Number of strata =    15           Number of obs   =      5,604
Number of PSUs   = 2,064           Population size = 12,551,189
                                   Design df       =      2,049

              |             Linearized
              |       Mean   std. err.     [95% conf. interval]
c.supvis@male |
            0 |   .2208335   .0131402      .1950639    .2466031
            1 |   .3676081    .015579      .3370558    .3981604
```

Gender differences in characteristics

```
. svy: mean yeduc expft exppt if !missing(supvis), over(male)
(running mean on estimation sample)
Survey: Mean estimation
Number of strata =     15          Number of obs   =      5,604
Number of PSUs   =  2,064          Population size = 12,551,189
                                   Design df       =      2,049
```

	Mean	Linearized std. err.	[95% conf.	interval]
c.yeduc@male				
0	12.89307	.0919486	12.71275	13.07339
1	12.68223	.0976222	12.49078	12.87368
c.expft@male				
0	10.69754	.2706955	10.16668	11.22841
1	17.02503	.3402843	16.35769	17.69237
c.exppt@male				
0	5.46444	.1884403	5.094886	5.833995
1	1.255	.0998899	1.059104	1.450896

Outcome model by gender

```
. bysort male: logit supvis yeduc expft exppt [pw=weight], cluster(psu) nolog

-> male = 0

Logistic regression                                     Number of obs =  2,910
                                                        Wald chi2(3)  =  25.29
                                                        Prob > chi2   = 0.0000
Log pseudolikelihood = -3119723.2                       Pseudo R2     = 0.0246

                                (Std. err. adjusted for 1,697 clusters in psu)

                           Robust
      supvis  Coefficient  std. err.      z    P>|z|     [95% conf. interval]

       yeduc    .1233739   .0277228     4.45   0.000     .0690382    .1777095
       expft    .0234178   .0086989     2.69   0.007     .0063683    .0404673
       exppt   -.0045416   .0143519    -0.32   0.752    -.0326708    .0235877
       _cons   -3.119188    .429148    -7.27   0.000    -3.960302   -2.278073

-> male = 1

Logistic regression                                     Number of obs =  2,694
                                                        Wald chi2(3)  =  30.38
                                                        Prob > chi2   = 0.0000
Log pseudolikelihood = -4156367.2                       Pseudo R2     = 0.0267

                                (Std. err. adjusted for 1,560 clusters in psu)

                           Robust
      supvis  Coefficient  std. err.      z    P>|z|     [95% conf. interval]

       yeduc    .1359181    .025392     5.35   0.000     .0861508    .1856855
       expft    .0135046   .0073652     1.83   0.067     -.000931    .0279402
       exppt   -.0455562   .0275738    -1.65   0.099    -.0995998    .0084875
       _cons   -2.459027   .3795256    -6.48   0.000    -3.202883    -1.71517
```

Aggregate decomposition using nldecompose

```
. nldecompose, by(male): svy: logit supvis yeduc expft exppt
                                                   Number of obs (A) =    2694
                                                   Number of obs (B) =    2910

------------------------------------------------------------------------------
      Results |      Coef.  Percentage
--------------+---------------------------------------------------------------
 Omega = 1    |
         Char |   .0494556   33.69495%
         Coef |    .097319   66.30505%
--------------+---------------------------------------------------------------
 Omega = 0    |
         Char |   .0240527   16.38752%
         Coef |   .1227219   83.61248%
--------------+---------------------------------------------------------------
          Raw |   .1467746        100%
------------------------------------------------------------------------------
```

Fairlie decomposition of explained part

```
. /* pweights are allowed in fairlie, but clustering is not possible (does not
>    really matter much because the standard errors are unreliable anyhow).*/
. fairlie supvis yeduc expft exppt [pw=weight], by(female)
(sum of wgt is 6,493,402.3937788)
Iteration 0:  Log pseudolikelihood =  -1771.764
Iteration 1:  Log pseudolikelihood =   -1724.64
Iteration 2:  Log pseudolikelihood = -1724.4048
Iteration 3:  Log pseudolikelihood = -1724.4047

Logistic regression                                     Number of obs     =       2694
                                                        Wald chi2(3)      =      31.01
                                                        Prob > chi2       =     0.0000
Log pseudolikelihood = -1724.4047                       Pseudo R2         =     0.0267

------------------------------------------------------------------------------
             |               Robust
      supvis | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
       yeduc |   .1359182   .0252053     5.39   0.000     .0865168    .1853196
       expft |   .0135046   .0073175     1.85   0.065    -.0008374    .0278466
       exppt |  -.0455563   .0279737    -1.63   0.103    -.1003836    .0092711
       _cons |  -2.459027   .3759087    -6.54   0.000    -3.195794   -1.722259
------------------------------------------------------------------------------

Decomposition replications (100)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..................................................    50
..................................................   100

Non-linear decomposition by female (G)                     Number of obs = 5,604
                                                        N of obs G=0      =       2694
                                                        N of obs G=1      =       2910
                                                        Pr(Y!=0|G=0)      =   .3676081
                                                        Pr(Y!=0|G=1)      =  .22083351
                                                        Difference        =  .14677458
                                                        Total explained   =  .04945562

------------------------------------------------------------------------------
      supvis | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
       yeduc |  -.0079051   .0016821    -4.70   0.000    -.0112018   -.0046083
       expft |   .0198582    .010628     1.87   0.062    -.0009722    .0406887
       exppt |   .0378633   .0207113     1.83   0.068    -.0027302    .0784568
------------------------------------------------------------------------------
```

Fairlie results depend on the order of the variables!

```
. fairlie supvis yeduc expft exppt [pw=weight], by(female) noest
Decomposition replications (100)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..................................................    50
..................................................   100
Non-linear decomposition by female (G)             Number of obs = 5,604
                                                   N of obs G=0    =      2694
                                                   N of obs G=1    =      2910
                                                   Pr(Y!=0|G=0)    =  .3676081
                                                   Pr(Y!=0|G=1)    = .22083351
                                                   Difference      = .14677458
                                                   Total explained = .04945562

------------------------------------------------------------------------------
      supvis | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
       yeduc |  -.0079359   .0016799    -4.72   0.000    -.0112284   -.0046433
       expft |   .0198785   .0106373     1.87   0.062    -.0009703    .0407273
       exppt |   .0375057   .0205228     1.83   0.068    -.0027183    .0777296
------------------------------------------------------------------------------

. fairlie supvis exppt expft yeduc [pw=weight], by(female) noest
Decomposition replications (100)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..................................................    50
..................................................   100
Non-linear decomposition by female (G)             Number of obs = 5,604
                                                   N of obs G=0    =      2694
                                                   N of obs G=1    =      2910
                                                   Pr(Y!=0|G=0)    =  .3676081
                                                   Pr(Y!=0|G=1)    = .22083351
                                                   Difference      = .14677458
                                                   Total explained = .04945562

------------------------------------------------------------------------------
      supvis | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
       exppt |   .0346824   .0185298     1.87   0.061    -.0016353        .071
       expft |   .0168858   .0091316     1.85   0.064    -.0010118    .0347834
       yeduc |   -.002137   .0021662    -0.99   0.324    -.0063826    .0021086
------------------------------------------------------------------------------
```

Use option "ro" to average over randomized order

```
. fairlie supvis yeduc expft exppt [pw=weight], ///
>     by(female) ro noest nodots reps(1000)
Non-linear decomposition by female (G)                 Number of obs = 5,604
                                                  N of obs G=0     =      2694
                                                  N of obs G=1     =      2910
                                                  Pr(Y!=0|G=0)     =    .3676081
                                                  Pr(Y!=0|G=1)     =  .22083351
                                                  Difference       =  .14677458
                                                  Total explained =   .04945562

------------------------------------------------------------------------------
      supvis | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
       yeduc |  -.0050829   .0019619   -2.59   0.010    -.0089281   -.0012377
       expft |   .0184994   .0099459    1.86   0.063    -.0009943    .0379931
       exppt |   .0360213    .019539    1.84   0.065    -.0022744     .074317
------------------------------------------------------------------------------

. fairlie supvis exppt expft yeduc [pw=weight], ///
>     by(female) ro noest nodots reps(1000)
Non-linear decomposition by female (G)                 Number of obs = 5,604
                                                  N of obs G=0     =      2694
                                                  N of obs G=1     =      2910
                                                  Pr(Y!=0|G=0)     =    .3676081
                                                  Pr(Y!=0|G=1)     =  .22083351
                                                  Difference       =  .14677458
                                                  Total explained =   .04945562

------------------------------------------------------------------------------
      supvis | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
       exppt |   .0362014   .0196477    1.84   0.065    -.0023073    .0747102
       expft |   .0184674   .0099307    1.86   0.063    -.0009963    .0379312
       yeduc |   -.005155   .0019414   -2.66   0.008      -.00896     -.00135
------------------------------------------------------------------------------
```

Detailed decomposition using mvdcmp

```
. /* missings are an issue with mvdcmp: we must make sure to exclude these
>    observations from the computations; however, mvdcmp does not support the
>    if qualifier, so we have to remove the observations from the data; we can
>    do this temporarily using -preserve- and -restore- */
. preserve
. keep if !missing(supvis, yeduc, expft, exppt)
(1,705 observations deleted)
. mvdcmp male: logit supvis yeduc expft exppt [pw=weight], cluster(psu)
Decomposition Results                                    Number of obs =   5,604
>
------------------------------------------------------------------------------------------------
         Reference group  (A):male==1             Mean = 0.3676
         Comparison group (B):male==0             Mean = 0.2208
------------------------------------------------------------------------------------------------
                         supvis |     Coef.  Std. Err.        z    P>|z|    [95% Conf. Interval]      Pct.
--------------------------------+---------------------------------------------------------------
                              E |   0.04946    0.02023     2.44    0.015    0.00980    0.08911     33.69
                              C |   0.09732    0.02877     3.38    0.001    0.04093    0.15371     66.31
--------------------------------+---------------------------------------------------------------
                              R |   0.14677    0.02017     7.28    0.000    0.10725    0.18630
                         Due to Difference in Characteristics (E)
------------------------------------------------------------------------------------------------
                         supvis |     Coef.  Std. Err.        z    P>|z|    [95% Conf. Interval]      Pct.
--------------------------------+---------------------------------------------------------------
                          yeduc |  -0.00570    0.00104    -5.51    0.000   -0.00773   -0.00367     -3.88
                          expft |   0.01700    0.00942     1.80    0.071   -0.00147    0.03547     11.58
                          exppt |   0.03816    0.02038     1.87    0.061   -0.00178    0.07809     26.00
------------------------------------------------------------------------------------------------

                         Due to Difference in Coefficients (C)
------------------------------------------------------------------------------------------------
                         supvis |     Coef.  Std. Err.        z    P>|z|    [95% Conf. Interval]      Pct.
--------------------------------+---------------------------------------------------------------
                          yeduc |   0.03201    0.09713     0.33    0.742   -0.15837    0.22239     21.81
                          expft |  -0.02099    0.02381    -0.88    0.378   -0.06765    0.02567    -14.30
                          exppt |  -0.04436    0.03539    -1.25    0.210   -0.11373    0.02501    -30.22
                          _cons |   0.13065    0.10885     1.20    0.230   -0.08269    0.34399     89.02
------------------------------------------------------------------------------------------------

. restore
```

Replication of results from mvdcmp using oaxaca

```
. oaxaca supvis yeduc expft exppt, by(female) svy weight(1) logit fixed
Blinder-Oaxaca decomposition

Number of strata =     15                     Number of obs     =      5,604
Number of PSUs   =  2,064                     Population size   = 12,551,189
                                              Design df         =      2,049
                                              Model             =      logit
Group 1: female = 0                           N of obs 1        =      2,694
Group 2: female = 1                           N of obs 2        =      2,910

    explained: (X1 - X2) * b1
  unexplained: X2 * (b1 - b2)
```

supvis	Coefficient	Linearized std. err.	t	P>|t|	[95% conf.	interval]
overall						
group_1	.3676081	.0153498	23.95	0.000	.3375052	.3977109
group_2	.2208335	.0130637	16.90	0.000	.1952141	.2464529
difference	.1467746	.0202957	7.23	0.000	.1069723	.1865769
explained	.0494556	.020231	2.44	0.015	.0097801	.0891312
unexplained	.097319	.0286585	3.40	0.001	.0411161	.1535219
explained						
yeduc	-.0057019	.0010292	-5.54	0.000	-.0077202	-.0036836
expft	.0170019	.0094286	1.80	0.071	-.0014887	.0354926
exppt	.0381555	.0203491	1.88	0.061	-.0017515	.0780626
unexplained						
yeduc	.0320093	.0979024	0.33	0.744	-.1599893	.2240079
expft	-.020988	.0240864	-0.87	0.384	-.0682244	.0262483
exppt	-.0443567	.0355656	-1.25	0.212	-.1141052	.0253918
_cons	.1306544	.1094276	1.19	0.233	-.0839464	.3452552

. . . with consistent standard errors

```
. oaxaca supvis yeduc expft exppt, by(female) svy weight(1) logit
Blinder-Oaxaca decomposition
Number of strata =     15                    Number of obs      =      5,604
Number of PSUs   =  2,064                    Population size    = 12,551,189
                                             Design df          =      2,049
                                             Model              =      logit
Group 1: female = 0                          N of obs 1         =      2,694
Group 2: female = 1                          N of obs 2         =      2,910
    explained: (X1 - X2) * b1
  unexplained: X2 * (b1 - b2)
```

supvis	Coefficient	Linearized std. err.	t	P>|t|	[95% conf.	interval]
overall						
group_1	.3676081	.0156415	23.50	0.000	.3369332	.398283
group_2	.2208335	.0132286	16.69	0.000	.1948906	.2467764
difference	.1467746	.0205677	7.14	0.000	.1064388	.1871104
explained	.0494556	.0206577	2.39	0.017	.0089434	.0899678
unexplained	.097319	.0287181	3.39	0.001	.0409992	.1536387
explained						
yeduc	-.0057019	.0032543	-1.75	0.080	-.0120839	.0006801
expft	.0170019	.0094945	1.79	0.073	-.0016179	.0356218
exppt	.0381555	.0204462	1.87	0.062	-.0019419	.078253
unexplained						
yeduc	.0320093	.0979031	0.33	0.744	-.1599907	.2240093
expft	-.020988	.0240933	-0.87	0.384	-.0682379	.0262619
exppt	-.0443567	.0355949	-1.25	0.213	-.1141627	.0254493
_cons	.1306544	.1094296	1.19	0.233	-.0839504	.3452592

Detailed decomposition based on LPM

```
. oaxaca supvis yeduc expft exppt, by(female) svy weight(1)
Blinder-Oaxaca decomposition
Number of strata =    15                    Number of obs      =      5,604
Number of PSUs   = 2,064                    Population size    = 12,551,189
                                            Design df          =      2,049
                                            Model              =     linear
Group 1: female = 0                         N of obs 1         =      2,694
Group 2: female = 1                         N of obs 2         =      2,910
    explained: (X1 - X2) * b1
  unexplained: X2 * (b1 - b2)
```

supvis	Coefficient	Linearized std. err.	t	P>|t|	[95% conf.	interval]
overall						
group_1	.3676081	.0156271	23.52	0.000	.3369614	.3982548
group_2	.2208335	.0132361	16.68	0.000	.1948759	.2467912
difference	.1467746	.0205591	7.14	0.000	.1064557	.1870934
explained	.0503704	.021876	2.30	0.021	.0074689	.093272
unexplained	.0964042	.0296137	3.26	0.001	.0383281	.1544802
explained						
yeduc	-.0065329	.0037626	-1.74	0.083	-.0139119	.000846
expft	.0189387	.0103364	1.83	0.067	-.0013322	.0392096
exppt	.0379646	.020897	1.82	0.069	-.0030169	.0789461
unexplained						
yeduc	.1269097	.0971622	1.31	0.192	-.0636372	.3174566
expft	-.0095913	.0238722	-0.40	0.688	-.0564076	.0372249
exppt	-.0447754	.0297591	-1.50	0.133	-.1031367	.0135858
_cons	.0238612	.1051331	0.23	0.820	-.1823177	.2300402

1. Nonlinear effects and interactions

2. Extension to nonlinear models
   - Aggregate decomposition
   - Detailed decomposition
   - Example analysis
   - Note on separation of direct and indirect effects

Note on separation of direct and indirect effects

- The Fairlie decomposition is sometimes used in social mobility research to separate direct and indirect effects of parental status.
- Example: dependent variable is college graduation, predictors are ability (e.g., measured by standardized tests at end of secondary school) and parental socio-economic status (SES).
- If parental SES has only two values (high, low) one could use the Fairlie decomposition to evaluate how much of the difference in graduation rates between the low SES class and the high SES class is explained by ability (this is the indirect effect; the unexplained part is the direct effect).
- However, different methods are usually employed in this research field (see, e.g., Karlson et al. 2012 and Breen et al. 2013).

Exercise 4

References

- Bartus, Tamás (2006). Marginal effects and extending the Blinder-Oaxaca decomposition to nonlinear models. Presentation at the 12th UK Stata Users Group meeting, available from https://ideas.repec.org/p/boc/usug06/05.html.
- Bauer, Thomas K., Mathias Sinning (2008). An extension of the Blinder–Oaxaca decomposition to nonlinear models. Advances in Statistical Analysis 92:197–206.
- Breen, Richard, Kristian B. Karlson, Anders Holm (2013). Total, Direct, and Indirect Effects in Logit and Probit Models. Sociological Methods & Research 42(2): 164–191.
- Fairlie, Robert W. (2005). An extension of the Blinder-Oaxaca decomposition technique to logit and probit models. Journal of Economic and Social Measurement 30:305–316.
- Gomulka, Joanna, Nicholas Stern (1990). The Employment of Married Women in the United Kingdom 1970-83. Economica 57:171—199.
- Jann, Ben (2006). fairlie: Stata module to generate nonlinear decomposition of binary outcome differentials. Available from http://ideas.repec.org/c/boc/bocode/s456727.html.

References

- Karlson, Kristian B., Anders Holm, Richard Breen (2012). Comparing Regression Coefficients Between Same-Sample Nested Models using Logit and Probit: A New Method. Sociological Methodology 42:286-313.
- Powers, Daniel A., Myeong-Su Yun (2009). Multivariate Decomposition for Hazard Rate Models. Sociological Methodology 39(1):233–263.
- Powers, Daniel A., Hirotoshi Yoshioka, Myeong-Su Yun (2011). mvdcmp: Multivariate decomposition for nonlinear response models. The Stata Journal 11(4): 556–576.
- Sinning, Mathias, Markus Hahn, Thomas K. Bauer (2008). The Blinder-Oaxaca decomposition for nonlinear regression models. The Stata Journal 8(4):480–492.
- Yun, Myeong-Su (2004). Decomposing differences in the first moment. Economics Letters 82(2):275–280.