ineqrbd:
Regression-based inequality decomposition, following Fields (2003)

Carlo V. Fiorio, University of Milan
&
Stephen P. Jenkins, ISER, University of Essex

UKSUG - 10 September 2007

Motivation of RB decomposition

- Decomposition analysis of inequality is important for understanding the main determinants of inequality and for policy analysis.
- The "traditional" approach to the subject was based purely on the analysis of the mathematical properties of inequality indices and is open to the criticism that the formal requirements for exact decomposition are perhaps too demanding for some practical applications.
  - It allows inequality accounting but not a causal analysis.
- Recent applied work has reawakened interest in inequality decomposition by focusing on the use of regression-based (RB) approaches to avoid some of the restrictions of the traditional methods.

The Fields' approach to RB decomposition of inequality

- Assuming that the income DGP is

$$\mathbf{y} = \mathbf{X}\beta + \epsilon \tag{1}$$

  where:
  - $\mathbf{y}$ is an $n \times 1$ vector of incomes;
  - $\mathbf{X}$ is an $n \times (K+1)$ matrix of individual and household characteristics (age, education, household size, residence, etc.) including the constant;
  - β is a $(K+1) \times 1$ vector of coefficients and ϵ is an $n \times 1$ vector of residuals.
- a sample of observations $\{y_i, \mathbf{x}_i, i = 1, 2, ...n\}$ can be used to estimate the model.

The Fields' approach to RB decomposition of inequality

- The linear model (1) can be rewritten as:

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots + \beta_K x_K + \epsilon && (2) \\ &= \beta_0 + z_1 + z_2 \ldots + z_K + \epsilon && (3) \end{aligned}$$

where:

- each z_k is a "composite" variable, equal to the product of a regression coefficient and its variable ($z_k = \beta_k x_k$), with $k = 0, 1, \ldots K$ and $x_0 = 1$.
- NB: For inequality decomposition calculations, the value of β_0 is irrelevant as it is constant for every observation.

The Fields' approach to RB decomposition of inequality

- Following Fields suggestion, the OLS estimate of (3) can be used for inequality decomposition:

$$y = b_0 + \widehat{z}_1 + \widehat{z}_2 ... + \widehat{z}_K + \widehat{\epsilon}_i \tag{4}$$

- Alternatively, one may look at the *predicted* income:

$$\widehat{y} = b_0 + \widehat{z}_1 + \widehat{z}_2 ... + \widehat{z}_K \tag{5}$$

  in which case there is no residual term.
  - $\widehat{z}_k = b_k x_k$ and b_k is the OLS estimate of β_k, $k = 0, 1, ...K$.

Our focus

- Neglecting the constant, equations (4) and (5) are of exactly the same form as the equation used by Shorrocks (1982) when deriving rules for inequality decomposition by factor components (e.g. total income is the sum of labour earnings, income from savings and other assets, private and public transfers. How much inequality in total income is attributable to each of these factors?)
- Shorrocks proved that a set of arguably persuasive axioms led to a unique additive and exact decomposition rule, with one term for each factor.
  - The decomposition rule did *not* depend on the choice of measure summarizing inequality in total income.

Our focus

- Fields (2003) exploited the parallel with the factor decomposition case, and applied the Shorrocks decomposition rule to relate inequality in $\widehat{y}$ to contributions from each of the RHS variables (x_k).
- There are two main issues to notice about Fields decomposition:
    1. One can only relate inequality in y to contributions from each of the **composite variables** z_k, not x_k.
    2. Decomposing $\widehat{y}$ instead of y does make a difference!

Our focus

- `ineqrbd` uses code from `ineqfac` by S.P. Jenkins, which performs Shorrocks' factor decomposition.
- `ineqrbd` provides a regression-based Shorrocks-type decomposition of a variable labelled *Total*, where *Total* is defined as *yvar* (y), unless the `fields` option is used, in which case *Total* refers to *yhat* $\widehat{y}$.
- In either case, the contribution to inequality in Total of each term is labelled x_k in the output.

An example: wage inequality

- Use LIS sample dataset (US, year 2000. Not a random sample of the original!): how relevant is the contribution of individual characteristics to explain the inequality of log-wages?
- We used gross individual wage of working-age people. A very simple model (i.e. no sample selection considered) is assumed and finally estimated with OLS.

$$lgrosswage = \beta_0 + \beta_1 age + \beta_2 age^2 + \beta_3 female + \sum_{k=4}^{7} \beta_k (educ_k) + \epsilon$$

(education=no title, high school, some college, college, postgrad)

An example: wage inequality

```
. use http://www.lisproject.org/dataccess/sample/us00samppp.dta
. keep if page>25 & page<65 & pgwage>0
. gen page2=page*page
. gen lpgwage=log(pgwage)
. recode peduc (-1/8=0)(9=1)(10=2)(11/13=3)(14/16=4)
. xi: ineqrbd lpgwage page page2 i.psex i.peduc
```

The ineqrbd output. OLS regression

```
Results
. xi: ineqrbd lpgwage page page2 i.psex i.peduc
i.psex            _Ipsex_1-2          (naturally coded; _Ipsex_1 omitted)
i.peduc           _Ipeduc_0-4         (naturally coded; _Ipeduc_0 omitted)

Regression of lpgwage on RHS variables

(analytic weights assumed)
(sum of wgt is   1.1800e+03)

      Source |       SS       df       MS              Number of obs =    1180
-------------+------------------------------           F(  7,  1172) =   40.16
       Model |  214.068896     7  30.5812709           Prob > F      =  0.0000
    Residual |  892.370392  1172  .761408184           R-squared     =  0.1935
-------------+------------------------------           Adj R-squared =  0.1887
       Total |  1106.43929  1179  .938455715           Root MSE      =  .87259

------------------------------------------------------------------------------
     lpgwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        page |   .0866839   .0229005     3.79   0.000     .0417534    .1316144
       page2 |   -.000963   .0002609    -3.69   0.000    -.0014747   -.0004512
    _Ipsex_2 |  -.5136189    .050975   -10.08   0.000    -.6136314   -.4136064
   _Ipeduc_1 |    .520429   .1311271     3.97   0.000     .2631589    .7776991
   _Ipeduc_2 |   .7096313    .133643     5.31   0.000     .4474251    .9718376
   _Ipeduc_3 |    .933268   .1306687     7.14   0.000     .6768973    1.189639
   _Ipeduc_4 |   1.423139   .1403917    10.14   0.000     1.147692    1.698586
       _cons |   7.897521   .4979311    15.86   0.000     6.920585    8.874456
------------------------------------------------------------------------------
```

The ineqrbd output. Default choice of LHS variable: y

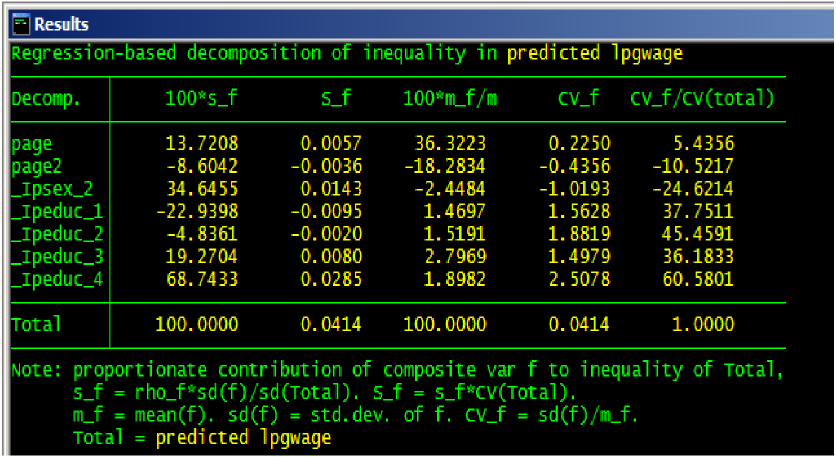
Results

Regression-based decomposition of inequality in lpgwage

Decomp.	100*s_f	S_f	100*m_f/m	CV_f	CV_f/CV(total)
residual	80.6524	0.0759	0.0000	6.64e+14	7.05e+15
page	2.6546	0.0025	36.3223	0.2250	2.3909
page2	-1.6647	-0.0016	-18.2834	-0.4356	-4.6281
_Ipsex_2	6.7031	0.0063	-2.4484	-1.0193	-10.8299
_Ipeduc_1	-4.4383	-0.0042	1.4697	1.5628	16.6051
_Ipeduc_2	-0.9357	-0.0009	1.5191	1.8819	19.9956
_Ipeduc_3	3.7284	0.0035	2.7969	1.4979	15.9155
_Ipeduc_4	13.3002	0.0125	1.8982	2.5078	26.6467
Total	100.0000	0.0941	100.0000	0.0941	1.0000

Note: proportionate contribution of composite var f to inequality of Total,
s_f = rho_f*sd(f)/sd(Total). S_f = s_f*CV(Total).
m_f = mean(f). sd(f) = std.dev. of f. CV_f = sd(f)/m_f.
Total = lpgwage

The ineqrbd output. fields option of LHS variable, $\widehat{y}$


Results

Regression-based decomposition of inequality in predicted lpgwage

Decomp.	100*s_f	s_f	100*m_f/m	CV_f	CV_f/CV(total)
page	13.7208	0.0057	36.3223	0.2250	5.4356
page2	-8.6042	-0.0036	-18.2834	-0.4356	-10.5217
_Ipsex_2	34.6455	0.0143	-2.4484	-1.0193	-24.6214
_Ipeduc_1	-22.9398	-0.0095	1.4697	1.5628	37.7511
_Ipeduc_2	-4.8361	-0.0020	1.5191	1.8819	45.4591
_Ipeduc_3	19.2704	0.0080	2.7969	1.4979	36.1833
_Ipeduc_4	68.7433	0.0285	1.8982	2.5078	60.5801
Total	100.0000	0.0414	100.0000	0.0414	1.0000

Note: proportionate contribution of composite var f to inequality of Total,
s_f = rho_f*sd(f)/sd(Total). S_f = s_f*CV(Total).
m_f = mean(f). sd(f) = std.dev. of f. CV_f = sd(f)/m_f.
Total = predicted lpgwage

Options

- `fields` implies decomposition of *predicted yvar* ($\widehat{y}$) rather than of *yvar* (y).
- `noregression` suppresses reporting of the OLS regression equation used to derive the composite variables and residual.
- `noconstant` excludes the intercept term from the regression.
- `stats` provides the means, sd, and ρ, of Total, the residual (unless the fields option is used), and the composite variables.
- `i2` summarises inequality using half the squared coefficient of variation (the Generalized Entropy measure I2), rather than the coefficient of variation (CV).

Saved results

- `r(total)` contains *predicted yvar* ($\widehat{y}$) if fields used; else contains *yvar* (y)
- `r(mean_tot)`, `r(sd_tot)`, `r(cv_tot)` mean, standard deviation, CV for Total
- `r(sf_Z0)`, `r(mean_Z0)`, proportionate inequality contribution, mean,
- `r(sd_Z0)`, `r(cv_Z0)` standard deviation, CV for the residual.
- `r(sf_Z0)` is not reported if fields option used.
- `r(sf_Zf)`, `r(mean_Zf)`, proportionate inequality contribution, mean,
- `r(sd_Zf)`, `r(cv_Zf)` standard deviation, CV for each of variables in rhsvars, where "f" is an integer `1,...,K`, indicating the order in which entered in rhsvars.