`wdiscrim`
Earnings discrimination statistics

Philippe Van Kerm
CEPS/INSTEAD, Luxembourg[‡]
philippe.vankerm@ceps.lu

September 2009
(updated March 2014)

Abstract This note describes `wdiscrim`, a user-written Stata package for computing the 'distributionally-sensitive' earnings discrimination measures proposed in Jenkins (*Journal of Econometrics*, 1994).

Keywords `wdiscrim` ; Stata ; discrimination; wage differentials

JEL Classification: J3; J7

1 Introduction

This note describes `wdiscrim`, a user-written Stata package for computing the 'distributionally-sensitive' earnings discrimination measures proposed in Jenkins (1994). The command is available online for installation in net-aware Stata.[1] At the command prompt, type

```
ssc install wdiscrim
```

2 Earnings discrimination measures

Consider that people are of one of two types—male or female, white or black, etc.. Measures of earnings "discrimination" typically quantify the amount of wage differences between agents of these two types that can not be attributed to other (observable) productivity-related characteristics, but rather to differential (possibly discriminatory) treatment of the two types.

Let y_i denote the earnings of an individual of a given type that can be predicted given her observable characteristics (human capital endowments, job type, race, gender, ...). Let r_i denote the earnings of the same individual that would be predicted if she had the same set of observable characteristics except that she would be of the other type (the reference type). It is common to compare y_i and r_i to capture the prejudice that individual i experiences and to aggregate these individual-level experiences over the population of individuals of the type considered to compute an aggregate 'earnings discrimination' statistic.

Typically, y_i and r_i are estimated from log-linear regression models and are thus of the form

$$\begin{aligned} y_i &= \exp(X_i\beta + 0.5\sigma^2) \\ r_i &= \exp(X_i\beta_r + 0.5\sigma_r^2) \end{aligned}$$

where the subscript r indicates that the parameters (β coefficients and residual variance σ^2) have been estimated in a sample from the reference population type. One may consider refined definitions of y_i and r_i that avoid the log-linear parametric assumptions or that consider higher order moments (see, e.g., Van Kerm, 2013). However the exact definition of y_i and r_i is orthogonal to the discussion of

[‡] Centre d'Etudes de Populations, de Pauvreté et de Politiques Socio-Economiques/International Networks for Studies in Technology, Environment, Alternatives, Development. 3 Av. de la fonte, L-4364 Esch/Alzette, Luxembourg. http://www.ceps.lu

[1] The latest version of the `wdiscrim` package is 2.1.0 (of 2014-03-14). Stata 9.2 or later is required.

the aggregation of y_i and r_i into a global earnings discrimination measure that is treated here. Any definition of y_i and r_i can indeed be handled by wdiscrim.

Given a sample of N observations on y_i and r_i, an ubiquitous measure of wage differentials is, for example,

$$D = \exp\left(\frac{1}{N}\sum_{i=1}^{N}(\log(y_i) - \log(r_i))\right) \tag{1}$$

that can be interpreted as the cents a person makes for every dollar an observationally equivalent person of the reference type makes on average.

Jenkins (1994) argues that this kind of measure fails to capture fine details of the (joint) distribution of y_i and r_i and proposes alternative classes of aggregate indices. His J-index is given by

$$J_\alpha = \frac{1}{N}\sum_{i=1}^{N}\frac{y_i}{\bar{y}}\left(1 - d_i^{-\alpha}\right)$$

where $d_i = 1 + |r_i - y_i|/\bar{r}$ and $\bar{y}$ and $\bar{r}$ are the sample means of y_i and r_i, and $\alpha > 0$. From J_α, Jenkins also suggests a measure of the social opportunity cost of discrimination in terms of average wage levels given by

$$W = \bar{y}\left(1 - J_\alpha\right).$$

Jenkins also proposes an R-index (ordinally equivalent to J_α for $\upsilon < 0$) given by

$$R_\upsilon = \frac{1}{\upsilon}\frac{1}{N}\sum_{i=1}^{N}\frac{y_i}{\bar{y}}\left(d_i^\upsilon - 1\right)$$

for any $\upsilon \neq 0$ and

$$R_0 = \frac{1}{N}\sum_{i=1}^{N}\frac{y_i}{\bar{y}}\log(d_i)$$

for $\upsilon = 0$. The justification and properties of these indices are discussed at length in Jenkins (1994).

In a related paper, del Río *et al.* (2011) also discuss aggregation issues in the measurement of discrimination and propose alternative measures; in particular, indices similar to the 'FGT' index of poverty:

$$P_\alpha = \frac{1}{N}\sum_{i=1}^{N} g_i^\alpha$$

with $\alpha \geqslant 0$ and $g_i = \max(0, (1 - \frac{y_i}{z_i}))$ (the 'relative' index) or $g_i = \max(0, z_i - y_i)$ (the 'absolute' index).[2]

wdiscrim computes all these indices from vectors y_i and r_i for a sample of the 'discriminated' population. It also shows summary statistics of the distribution of various 'individual-level' differentials. Optionally, it provides coordinates of the generalized Lorenz and concentration curves of y_i, r_i and $|y_i - r_i|$ whose role in the measurement of discrimination is discussed in Jenkins (1994).

[2] wdiscrim also reports the transformations $EDE_\alpha = P_\alpha^{\frac{1}{\alpha}}$.

3 The wdiscrim command

3.1 Syntax

`wdiscrim` *actvar refvar* [*if*] [*in*] [*weight*] [, `rindex adgc rdgc generate(`*newvarname*`)` `coordinates(`*newvarlist*`) format(`*string*`) install`]

`aweight` and `fweight` are allowed; see [U] **11.1.6 weight – Weights**.

`wdiscrim` takes two variables as input. The first (*actvar*) contains an earnings prediction from a model for the observed 'discriminated' population; the y_i's. The second (*refvar*) contains a counterfactual prediction from a model for a reference population (the non-discriminatory benchmark); the r_i's. Based on these pairs of actual and counterfactual wage predictions for a sample of individuals, `wdiscrim` reports descriptive statistics about the distribution of individual-level earnings differentials and computes Jenkins J and R indices, as well as, optionally, del Rio et al.'s 'FGT' discrimination measures.

3.2 Options

`rindex` requests computation of the R_υ index (not computed by default).

`adgc` requests computation of the absolute FGT statistics of del Río *et al.* (2011) (not computed by default).

`rdgc` requests computation of the relative FGT statistics of del Río *et al.* (2011) (not computed by default).

`format(`*string*`)` specifies a format for the displayed results. The default is `%4.3f`.

`generate(`*newvarname*`)` fills the new variable *newvarname* with the relative differences between *actvar* and *refvar*, that is $\exp(\ln(r_i) - \ln(y_i)) - 1$. The sub-option `replace` can be used to replace any already existing variable named *newvarname*.

`coordinates(`*newvarlist*`)` creates four new variables filled with generalized Lorenz and concentration curves ordinates. The first variable is filled with the x-ordinates (the cumulative population share ordered in increasing value of y). The second variable contains the ordinates of the generalized Lorenz curve of y. The third variable contains the ordinates of the generalized concentration curve of r. The fourth variable contains the ordinates of the generalized concentration curve of $|r - y|$. Exactly four new variable names must be supplied in *newvarlist*. The sub-option `replace` can be used to replace any already existing variable in *newvarlist*.

`install` checks if required user-written packages `makematrix` and `glcurve` are installed, and prompts for installation if needed.

3.3 Saved results

Matrices	
r(desc)	Summary statistics of individual-level differentials
r(jindex)	Estimates of the J-index
r(rindex)	Estimates of the R-index (if requested)
r(adgcindex)	Estimates of the absolute FGT-discrimination index (if requested)
r(rdgcindex)	Estimates of the relative FGT-discrimination index (if requested)
Scalars	
r(prop)	Proportion of observations 'discriminated', that is with $r_i > y_i$
r(N)	Number of observations
Macros	
r(actvar)	*actvar*
r(refvar)	*refvar*
r(generate)	*newvarname* if generate(*newvarname*) specified
r(pvar)	First element in *newvarlist* if coordinates(*newvarlist*) specified
r(glyvar)	Second element in *newvarlist* if coordinates(*newvarlist*) specified
r(glrvar)	Third element in *newvarlist* if coordinates(*newvarlist*) specified
r(gldvar)	Fourth element in *newvarlist* if coordinates(*newvarlist*) specified

3.4 Dependencies on user-written packages

wdiscrim requires two user-written packages.

The first is the makematrix package by Nicholas J. Cox available from the SSC archive. The second is the glcurve package by Stephen P. Jenkins and Philippe Van Kerm available from the SSC archive or Stata Journal website.

Both packages can be installed easily with the install option.

4 Example

The following example illustrates wdiscrim using data from the National Longitudinal Survey of Youth, available from the Stata Press website.

In the first step we open the dataset and construct actual and counterfactual predictions for black women in the data where the reference is the group of white women.

```
. cap use http://www.stata-press.com/data/r9/nlswork , clear

. regress  ln_wage age msp collgrad not_smsa south if race==2

      Source |       SS       df       MS              Number of obs =    8030
-------------+------------------------------           F(  5,  8024) =  775.19
       Model |  565.340677     5  113.068135           Prob > F      =  0.0000
    Residual |  1170.37424  8024  .145859202           R-squared     =  0.3257
-------------+------------------------------           Adj R-squared =  0.3253
       Total |  1735.71491  8029  .216180709           Root MSE      =  .38192

------------------------------------------------------------------------------
     ln_wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0163813   .0006687    24.50   0.000     .0150706    .0176921
         msp |   .0234725   .0086328     2.72   0.007       .00655    .0403949
    collgrad |   .4836731   .0140065    34.53   0.000     .4562166    .5111295
    not_smsa |  -.2059166   .0109524   -18.80   0.000    -.2273862    -.184447
       south |  -.2544368   .0095847   -26.55   0.000    -.2732253   -.2356483
       _cons |   1.243714   .0200443    62.05   0.000     1.204422    1.283006
------------------------------------------------------------------------------

. predict wact if race==2
(option xb assumed; fitted values)
(20504 missing values generated)

. replace wact = exp(wact+ 0.5*e(rss)/(e(N)-1))
(8030 real changes made)

. regress  ln_wage  age msp collgrad not_smsa south if race==1
```

```
      Source |       SS       df       MS              Number of obs =   20153
-------------+------------------------------           F(  5, 20147) = 1024.74
       Model |  928.456208     5  185.691242           Prob > F      =  0.0000
    Residual |  3650.79926 20147  .181208083           R-squared     =  0.2028
-------------+------------------------------           Adj R-squared =  0.2026
       Total |  4579.25546 20152  .227235781           Root MSE      =  .42569

------------------------------------------------------------------------------
     ln_wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0175918    .000452    38.92   0.000     .0167058    .0184778
         msp |  -.0171251   .0064287    -2.66   0.008    -.0297258   -.0045243
    collgrad |   .3307761    .007767    42.59   0.000     .3155521    .3460002
   not_smsa  |  -.1767216   .0066223   -26.69   0.000    -.1897018   -.1637414
       south |  -.0770721   .0064658   -11.92   0.000    -.0897456   -.0643985
       _cons |   1.229273   .0135934    90.43   0.000     1.202629    1.255917
------------------------------------------------------------------------------

. predict wref if race==2
(option xb assumed; fitted values)
(20504 missing values generated)

. replace wref = exp(wref+ 0.5*e(rss)/(e(N)-1))
(8030 real changes made)
```

wdiscrim can then be called to compute various summary statistics about the distribution of differences between r_i and y_i and compute the various aggregate measures.

```
. wdiscrim wact wref
 Distribution of individual-level differentials:
                                   mean      p10      p25      p50      p75      p90
              Difference [r-y]    0.581   -0.028    0.182    0.801    1.003    1.172
   Diff of logs [log(r)-log(y)]   0.121   -0.005    0.030    0.171    0.210    0.232
Rel diff [exp(log(r)-log(y))-1]   0.134   -0.005    0.030    0.187    0.234    0.261
                   Max(r-y,0)     0.636    0.000    0.182    0.801    1.003    1.172
                 Max(1-y/r,0)     0.115    0.000    0.029    0.158    0.190    0.207

 Proportion discriminated: 0.88

 J(alpha) indices (Jenkins, 1994):
       J-index        W
  a(0)    0.000    5.376
a(1/4)    0.025    5.240
a(1/2)    0.049    5.110
  a(1)    0.095    4.863
  a(2)    0.178    4.421
  a(5)    0.364    3.417
 a(10)    0.548    2.429

. wdiscrim wact wref , adgc rdgc rindex
 Distribution of individual-level differentials:
                                   mean      p10      p25      p50      p75      p90
              Difference [r-y]    0.581   -0.028    0.182    0.801    1.003    1.172
   Diff of logs [log(r)-log(y)]   0.121   -0.005    0.030    0.171    0.210    0.232
Rel diff [exp(log(r)-log(y))-1]   0.134   -0.005    0.030    0.187    0.234    0.261
                   Max(r-y,0)     0.636    0.000    0.182    0.801    1.003    1.172
                 Max(1-y/r,0)     0.115    0.000    0.029    0.158    0.190    0.207

 Proportion discriminated: 0.88

 J(alpha) indices (Jenkins, 1994):
       J-index        W
  a(0)    0.000    5.376
a(1/4)    0.025    5.240
a(1/2)    0.049    5.110
  a(1)    0.095    4.863
  a(2)    0.178    4.421
  a(5)    0.364    3.417
 a(10)    0.548    2.429
 R(upsilon) indices (Jenkins, 1994):
         R-index
 u(-10)    0.055
  u(-5)    0.073
  u(-2)    0.089
  u(-1)    0.095
u(-1/2)    0.099
u(-1/4)    0.101
   u(0)    0.103
 u(1/4)    0.105
 u(1/2)    0.107
   u(1)    0.111
   u(2)    0.120
   u(5)    0.155
  u(10)    0.252

 Absolute ´FGT´ discrimination indices (del Rio et al., 2011):
           P    EDE
a(1/2)  0.704  0.495
  a(1)  0.636  0.636
a(3/2)  0.610  0.719
  a(2)  0.605  0.778

 Relative ´FGT´ discrimination indices (del Rio et al., 2011):
           P    EDE
a(1/2)  0.298  0.089
  a(1)  0.115  0.115
a(3/2)  0.047  0.131
  a(2)  0.020  0.142

. wdiscrim wact wref , coordinates(p gly glr gldiff) gen(gap, replace)

 Distribution of individual-level differentials:
                                    mean     p10     p25     p50     p75     p90
               Difference [r-y]    0.581  -0.028   0.182   0.801   1.003   1.172
   Diff of logs [log(r)-log(y)]    0.121  -0.005   0.030   0.171   0.210   0.232
Rel diff [exp(log(r)-log(y))-1]    0.134  -0.005   0.030   0.187   0.234   0.261
                   Max(r-y,0)      0.636   0.000   0.182   0.801   1.003   1.172
                 Max(1-y/r,0)      0.115   0.000   0.029   0.158   0.190   0.207

 Proportion discriminated: 0.88

 J(alpha) indices (Jenkins, 1994):
        J-index        W
  a(0)    0.000    5.376
a(1/4)    0.025    5.240
a(1/2)    0.049    5.110
  a(1)    0.095    4.863
  a(2)    0.178    4.421
  a(5)    0.364    3.417
 a(10)    0.548    2.429

. twoway line gly glr p , sort

. sumdist gap

Warning: gap has 967 values < 0. Used in calculations

Distributional summary statistics, 10 quantile groups

---------------------------------------------------------------------------
Quantile  |
group     |     Quantile  % of median     Share, %      L(p), %        GL(p)
----------+----------------------------------------------------------------
        1 |       -0.005       -2.484       -4.163       -4.163       -0.006
        2 |        0.027       14.220        0.908       -3.255       -0.004
        3 |        0.035       18.890        2.133       -1.121       -0.002
        4 |        0.056       29.713        3.268        2.147        0.003
        5 |        0.187      100.000       10.332       12.479        0.017
        6 |        0.210      112.409       14.567       27.046        0.036
        7 |        0.230      122.764       17.167       44.212        0.059
        8 |        0.239      127.892       17.013       61.225        0.082
```

```
           9 |       0.261      139.310       18.428       79.653        0.107
          10 |                                20.347      100.000        0.134
--------------------------------------------------------------------------------

Share = quantile group share of total gap;
L(p)=cumulative group share; GL(p)=L(p)*mean(gap)
```

References

del Río, C., Gradín, C. & Cantó, O. (2011), 'The measurement of gender wage discrimination: The distributional approach revisited', *Journal of Economic Inequality*, **9**(1):57–86.

Jenkins, S. P. (1994), 'Earnings discrimination measurement: A distributional approach', *Journal of Econometrics*, **61**:81–102.

Van Kerm, P. (2013), 'Generalized measures of wage differentials', *Empirical Economics*, **45**(1):465–482.

Citation, liability, conditions of use

The program should work as described, but it is freely offered 'as-is'. Use at your own risk! Of course, bug reports, as well as comments and suggestions are appreciated.

Please cite as:
Van Kerm, P. (2009), '`wdiscrim` – Earnings discrimination statistics', v2.1 (updated March 2014), CEPS/INSTEAD, Esch/Alzette, Luxembourg.

Acknowledgements

This work is part of the MeDIM project (*Advances in the Measurement of Discrimination, Inequality and Mobility*) supported by the Luxembourg 'Fonds National de la Recherche' (contract FNR/06/15/08) and by core funding for CEPS/INSTEAD by the Ministry of Higher Education and Research of Luxembourg.