piecewise_ginireg[1]

Piecewise Gini Regressions in Stata

Jan Ditzen[1] Shlomo Yitzhaki[2]

[1]Heriot-Watt University, Edinburgh, UK
Center for Energy Economics Research and Policy (CEERP)

[2]The Hebrew University and Hadassah Academic College, Jerusalem, Israel

September 8, 2017

[1]Name subject to changes...

Note

This slide was added after the presentation at the Stata User Group Meeting in London. As of 11. September 2017 `picewise_ginireg` is **not** available on SSC or publicly otherwise.

For inquiries, questions or comments, please write me at

j.ditzen@hw.ac.uk

or see

www.jan.ditzen.net

Introduction

- OLS requires...
  1. ... linear relationship between conditional expectation of the dependent variable and explanatory variables and ...
  2. ... errors are iid and uncorrelated with the independent variables.
- Often monotonic transformations are applied to linearize the model, can lead to changes of the sign of the estimated coefficients.
- OLS sensitive to outliers.

Gini Regressions

Basics

- Idea: replace the (co-)variance in an OLS regression with the Gini notion of (co-)variance, i.e. the Gini's Mean Difference (GMD) as the measure of dispersion.
- Gini Mean Difference: $G_{YX} = E|Y - X|$ with gini covariance: $Gcov(Y, X) = cov\,(Y, F(X))$, where $F(X)$ is the cumulative population distribution function.
- Regressor $\beta^G = \frac{cov(Y,F(X))}{cov(X,F(X))}$.
- Can be interpreted as an IV regression, with $F(X)$ as an instrument for X.

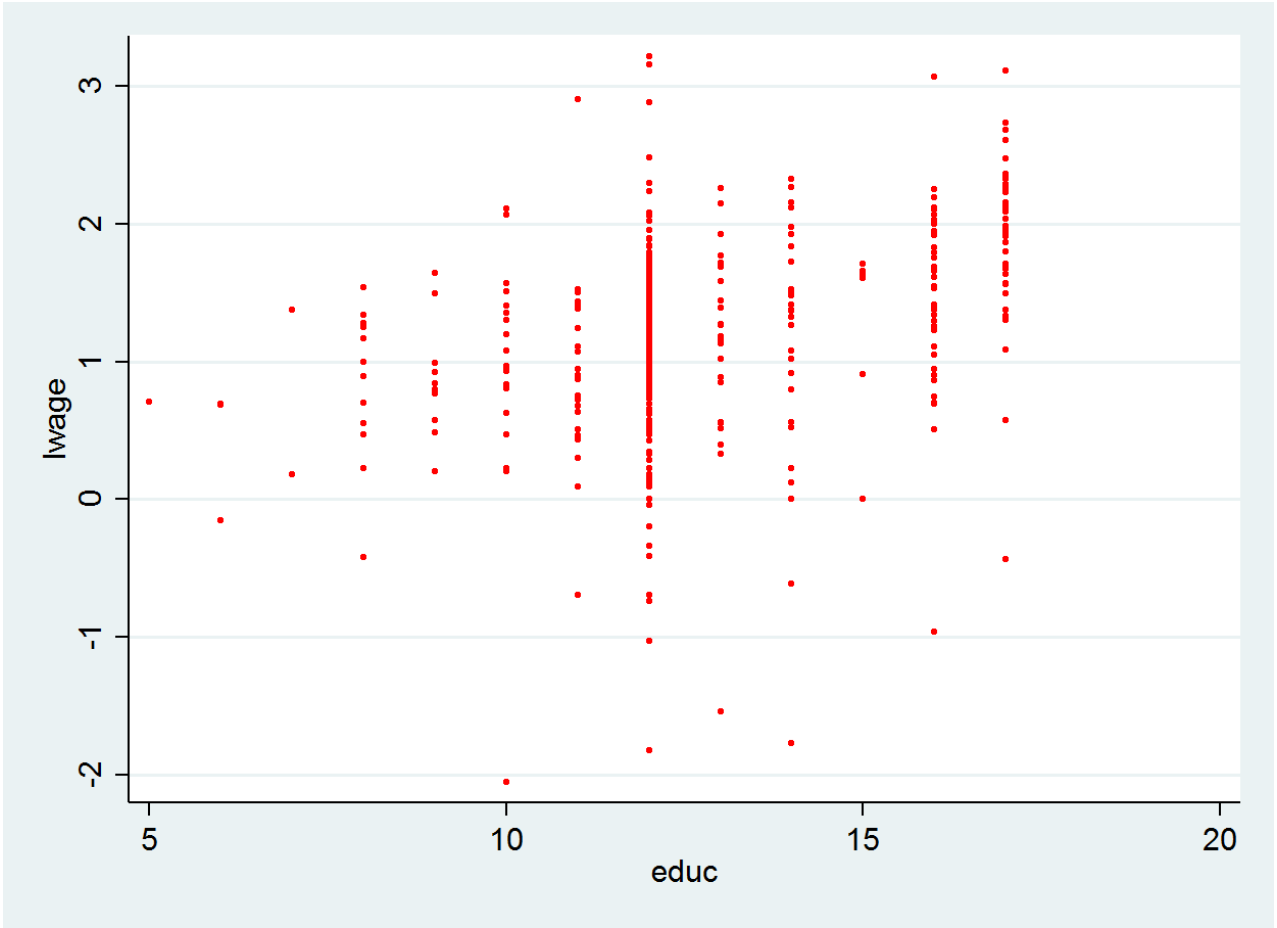
Gini Regressions

Advantages of Gini Regressions

- Gini regressions do not rely on
  - Symmetric correlation and variability measure
  - Linearity of the model.
  - Coefficients do not change after monotonic transformations of the explanatory or independent variables.
- GMD here definition has two asymmetric correlation coefficients, one can be used for the regression, the other can be used to test the linearity assumption.
- Summarized in Yitzhaki and Schechtman (2013); Yitzhaki (2015).

Example

- `mroz.dta` Dataset
- Estimate log wage using education. ▸ wage



Estimation in Stata

`ginireg` (Schaffer, 2015)

- Package to estimate gini regressions. Allows for extended and mixed Gini regressions and IV regressions.
- Post estimation commands allow prediction of residuals and fitted values, and calculation of LMA curve.
- Includes `ginilma` to graph Gini LMA and NLMA curves.

Example

```
. use http://fmwww.bc.edu/ec-p/data/wooldridge/mroz.dta , clear
. reg lwage educ
```

Source	SS	df	MS	Number of obs	=	428
				F(1, 426)	=	56.93
Model	26.3264237	1	26.3264237	Prob > F	=	0.0000
Residual	197.001028	426	.462443727	R-squared	=	0.1179
				Adj R-squared	=	0.1158
Total	223.327451	427	.523015108	Root MSE	=	.68003

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	.1086487	.0143998	7.55	0.000	.0803451	.1369523
_cons	-.1851969	.1852259	-1.00	0.318	-.5492674	.1788735

```
. ginireg lwage  educ
Gini regression
```

				Number of obs	=	428
				GR	=	0.321
				Gamma YYhat	=	0.319
				Gamma YhatY	=	0.450

lwage	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
educ	.105074	.0150097	7.00	0.000	.0756556	.1344924
_cons	-.1399459	.1928283	-0.73	0.468	-.5178824	.2379906

```
Gini regressors:           educ
Least squares regressors:  _cons
```

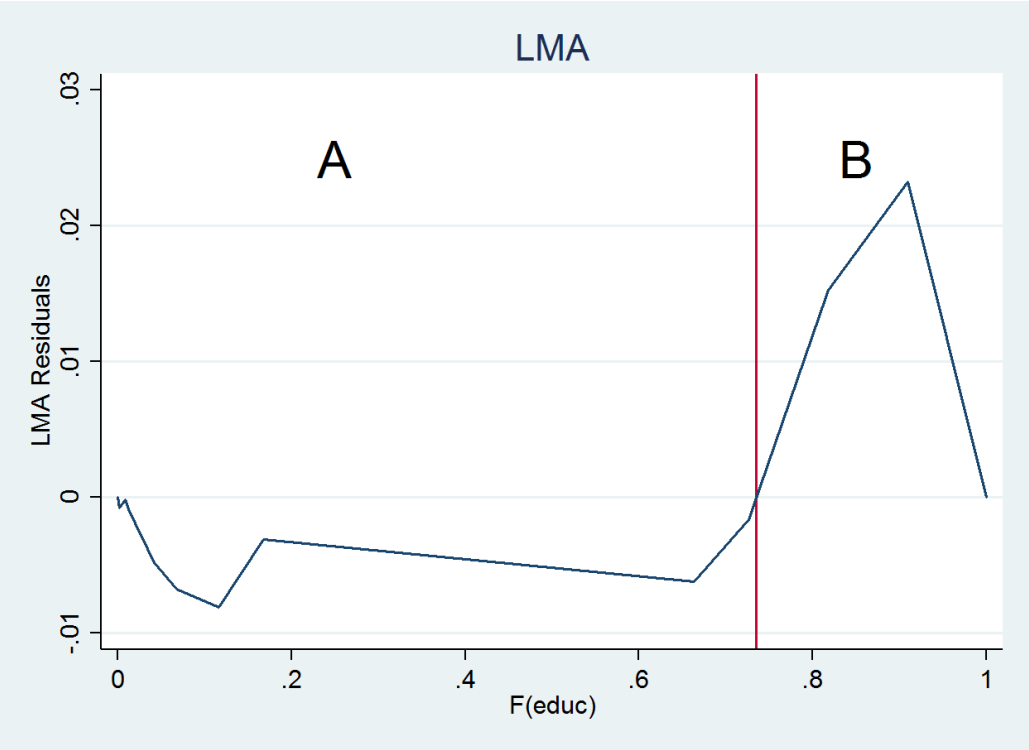
- One additional year of education increases the hourly wage by 10.9% (OLS) and by 10.5% (gini).

Gini Regressions

Line of independence minus absolute concentration curve (LMA)

- LMA defined as $LOI - ACC$:
  - Line of Independence (LOI) is a straight line from $(0, 0)$ to $(\mu_y, 1)$, represents statistical independence between X and Y. $LOI(p) = \mu_y p$.
  - Absolute concentration curve $ACC(p) = \int_{-\infty}^{x_p} g(t)dF(t)$, where $g(x)$ represents the regression curve.
- Properties:
  - Starts at $(0, 0)$ and ends at $(1, 0)$.
  - If it is above (below) the horizontal axis, section contributes positive (negative) to the regression coefficient.
  - If intersects the horizontal axis, then the sign of an OLS regression coefficient can change if there is a monotonic increasing transformation of X.
  - If curve is concave (convex, straight line), then the local regression coefficient is decreasing (increasing, constant).
- The LMA allows an interpretation of how the Gini covariance is composed and thus how the coefficients are effected as it includes the $Gcov(Y, X)$.

Example



- $cov(e, F(x)) = 0$ by construction, thus in the optimal case LMA fluctuates randomly around 0.
- Section A has a negative contribution to β, Section B has a postive contribution to β, or differently: a monotonic transformation that changes the sign of the OLS coefficient.
- This is not reflected by `ginireg` (or `reg`).

piecewise_ginireg

Introduction

Aim:

- Estimate regression which splits the data into sections determined by the LMA.
- Split the data until normality conditions of the error terms hold or the sections are "small".

Steps

1. Run Gini regression using the entire data.
2. Calculate residuals and LMA to determine sections.
3. Check if assumption for normality in the errors within the sections holds, or sections are small enough. If it does, stop; if not, continue.
4. Run a gini regression on each of the sections with the errors as a dependent variable and repeat steps 2 - 4.

- Iteration: Step 2 - 4.

piecewise_ginireg

syntax

Syntax

`piecewise_ginireg` *depvar indepvars* [*if*] , *maxiterations(integer) stoppingrule*
[`minsample(integer) restrict(`*varlist* `values) turningpoint(options)`
`ginireg(string) nocontinuous showqui noconstant showiterations`
`drawlma drawreg addconstant bootstrap(string) bootshow`
`multipleregressions(options)`]

where either `maxiterations(integer)` or `stoppingrule` have to be used.

piecewise_ginireg

options `stoppingrule` and `bootstrap()`

When to stop?

- If X and Y are exchangeable random variables, then the gini correlation of Y and X ($C(Y, X)$) and X and Y ($C(X, Y)$) are equal.
- Schröder and Yitzhaki (2016) suggest to split the dataset into two subsamples and test the gini correlations for equality:

$$H_0: \ C(Y, X) = C(X, Y)$$
$$H_A: \ C(Y, X) \neq C(X, Y)$$

with

$$C(Y, X) = \frac{cov\,(Y, F(X))}{cov\,(Y, F(Y))}$$

piecewise_ginireg

options `stoppingrule` and `bootstrap()`

- If option `stoppingrule` used, standard errors for gini correlation required.
- The difference between the two gini correlations, $D = C(X, Y) - C(Y, X)$, is bootstrapped and then tested with: $H_0 : D = 0$ vs. $H_A : D \neq 0$.
- Option `bootstrap(p(level) R(#))` sets the p-value and number of replications.
- Option `minsample(#)` Alternative rule: minimal size of a section. Default: _N/10

piecewise_ginireg

Example

```
. piecewise_ginireg lwage  educ, addconstant stoppingrule

Piecewise Linear Gini Regression.

Dependent Variable: lwage                          Number of obs      =       428
Independent Variables: educ _cons                  Number of groups   =         2
Groupvariables: educ                               Iterations         =         1
                                                   GR                 =     1.658
                                                   Gamma YYhat        =     0.321
                                                   Gamma YhatY        =     0.445
```

Final Results (sum of coefficients)

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Final Group Estimates for 5 <= educ <= 13 (N=311) in group 1						
educ	.086041	.047903	1.80	0.072	-.007846	.1799286
Final Group Estimates for 14 <= educ <= 17 (N=117) in group 2						
educ	.256339	.277607	0.92	0.356	-.2877608	.800438

```
Sections determined by LMA crossing line of origin (LMA(p) = 0).
Bootstrap performed with 50 replications. p-value for test of difference: .1
```

piecewise_ginireg

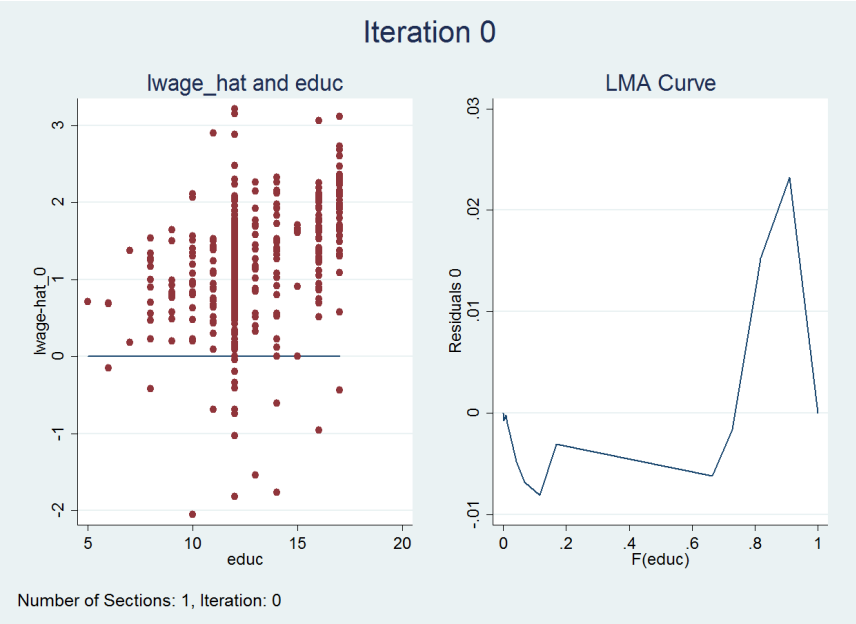
Example, including iterations

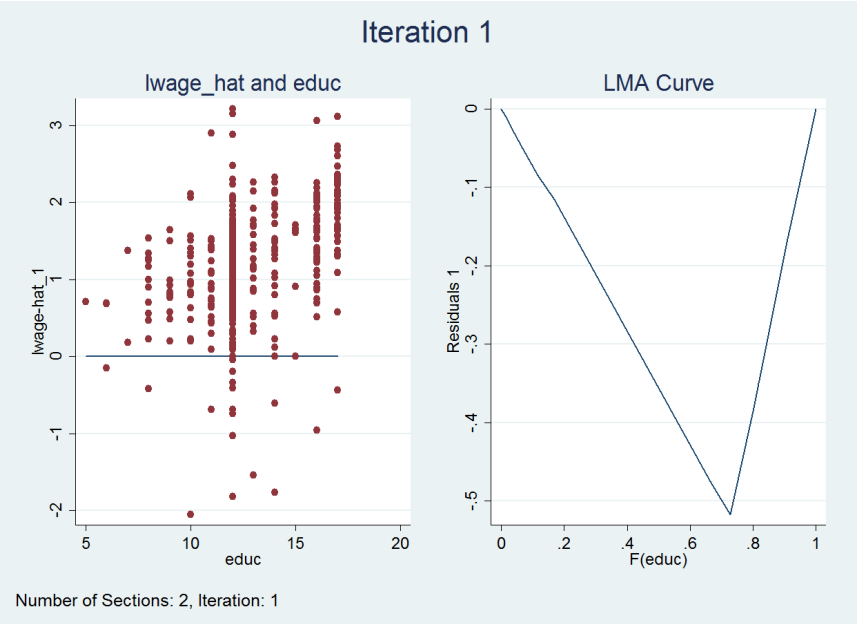
```
. piecewise_ginireg lwage  educ , addconstant stoppingrule showiterations
Piecewise Linear Gini Regression.
Dependent Variable: lwage                          Number of obs      =       428
Independent Variables: educ _cons                  Number of groups   =         2
Groupvariables: educ                               Iterations         =         1
                                                   GR                 =     1.658
                                                   Gamma YYhat        =     0.321
                                                   Gamma YhatY        =     0.445
Iteration: 0, with 1 groups
------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Estimates for         5 <= educ <= 17         (N=428)
-------------+----------------------------------------------------------------
        educ |    .105074    .01501     7.00   0.000     .0756556    .1344924
       _cons |   -.139946    .192828    -0.73   0.468    -.5178824    .2379906
------------------------------------------------------------------------------
Final Results (sum of coefficients)
------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Final Group Estimates for         5 <= educ <= 13         (N=311) in group 1
-------------+----------------------------------------------------------------
        educ |    .086041    .047903     1.80   0.072     -.007846    .1799286
-------------+----------------------------------------------------------------
Final Group Estimates for        14 <= educ <= 17         (N=117) in group 2
-------------+----------------------------------------------------------------
        educ |    .256339    .277607     0.92   0.356    -.2877608     .800438
------------------------------------------------------------------------------
Sections determined by LMA crossing line of origin (LMA(p) = 0).
Bootstrap performed with 50 replications. p-value for test of difference: .1
```

drawlma

back

```
. qui piecewise_ginireg lwage  educ  , addconstant  stoppingrule drawlma
. estat savegraphs , as(png) path("....")
Graph graph_0_educ saved as .../graph_0_educ.png
Graph graph_1_educ saved as .../graph_1_educ.png
```





piecewise_ginireg I

Options

- `maxiterations(integer)`: number of maximum of iterations
- `turningpoint(zero|maxmin)` specifies the turning point. Default is `turningpoint(zero)` and the sections are defined by intersections of the LMA with the origin. Alternative is `turningpoint(minmax)` or `turningpoint(maxmin)`. Then sections are defined by maxima and minima of the LMA curve.
- `restrict(`*varlist* `values)`: specifies group variables and values for sections. For example if the group variable is age and ranges from 10 to 20, 2 sections are wanted, from 10 to 15 and 16 to 20, then `restrict(age 15)` is used.

piecewise_ginireg II

Options

- nocontinuous no continuous piecewise regression. The constant is included and estimated in all estimations for sections > 2. If not specified, the constant is the predicted value of the last observation in the previous section. It is only included in regression of the first section. All regressions for the following sections are run without a constant.
- noconstant: suppresses the constant in the first initial regression and in the 1st section of the following iterations.
- addconstant: adds a constant for the section regressions in iterations > 1.
- showiterations displays in the output the regression results from all iterations. If not specified only the accumulated results are shown.

piecewise_ginireg

Further options and work in progress

Implemented

- `drawlma` and `drawreg` ▸ Example
  - Saves line graph of LMA and scatter plot of fitted values and independent variable for later use. Can be saved with `estat`.
- Postestimation
  - `predict`: calculation of linear prediction, LMA, residuals and coefficients.

Work in progress

- `multipleregressions`
  - Allows for more than one independent variable.
  - `order(`*varlist*`)` controls specifies order of variables for determining the sections.
  - `groups`: first the number of sections for each variable is calculated until convergence is achieved. Then the variables are ordered in as- or descending order of groups.
- Statistics such as Gini godness of fit

Conclusion

`piecewise_ginireg`...

- Extends `ginireg`
- Determines sections using the LMA.
- Estimates coefficients for each section.
- Several criteria for optimal number of sections possible.
- Alternative names:
  - pwginireg
  - pginireg
  - ...any other?

Definitions

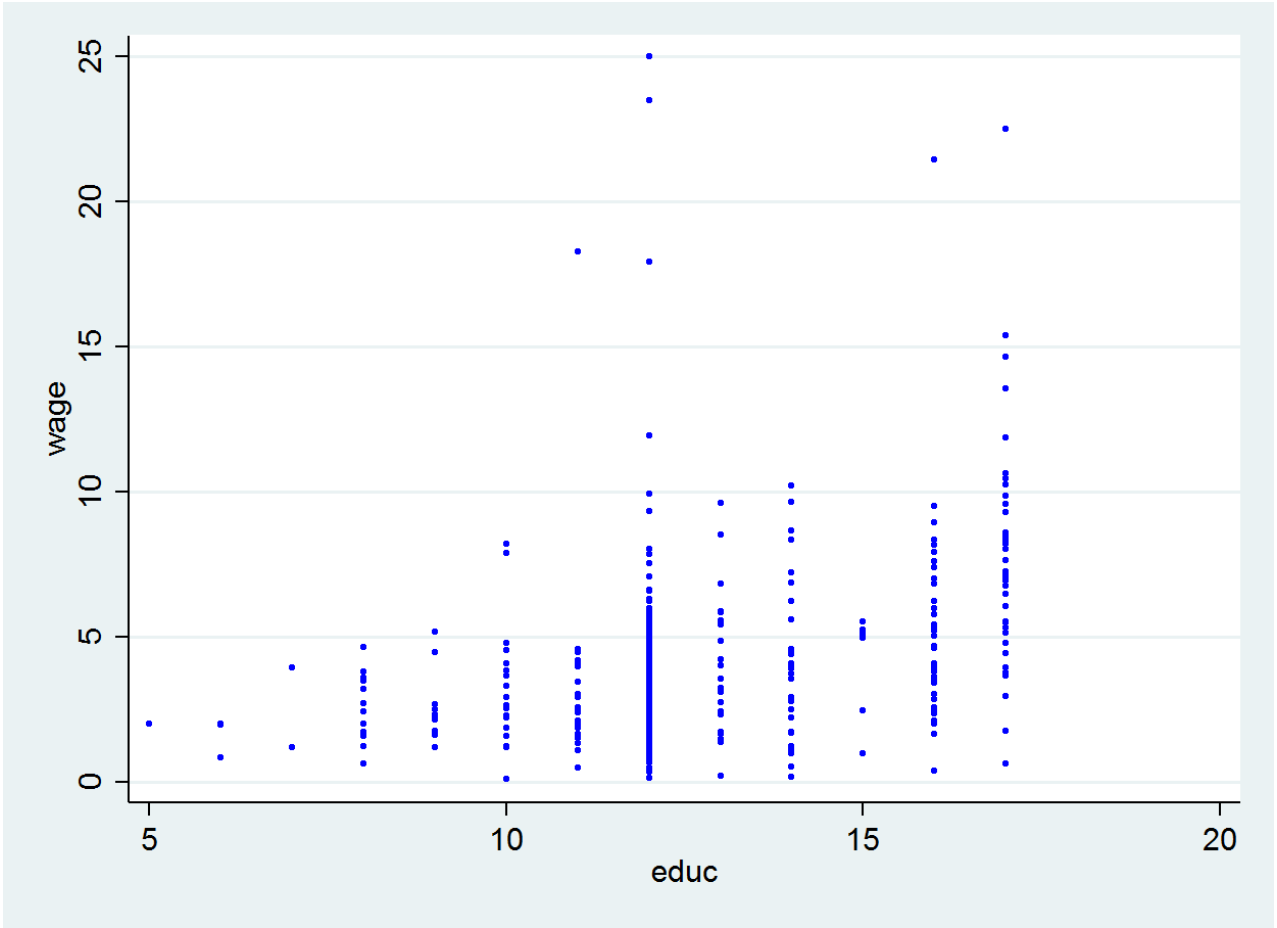
See Olkin and Yitzhaki (1992)

- Gini Mean Difference (GMD) of X and Y: $G_{XY} = E|X - Y|$
- $G_X = 4cov\,(X, F_X(X))$
- Gini Covariance: $Gcov(Y, X) = cov\,(Y, F_X(X))$, with F_X population cumulative distribution function.
- Gini Correlation: $C(X, Y) = \frac{Gcov(Y,X)}{Gcov(Y,Y)} = \frac{Cov(Y,F_X(X))}{Cov(Y,F_Y(Y))}$
- Properties of $C(X, Y)$:
  - If X and Y are exchangeable random variables, then $C(X, Y) = C(Y, X)$.
  - If (X, Y) has a bivariate normal distribution with means μ_x, μ_y and variances σ_x^2, σ_y^2 and correlation ρ then $C(X, Y) = C(Y, X) = \rho$
  - If X and Y are random variables, then $G_{X+Y} = C(X, X + Y)G_x + C(Y, X + Y)G_Y$.
  - If sample estimator of Gini covariance and the correlations are U-Statistics and asymptotically normal.

Example

back

- `mroz.dta` Dataset
- Estimate wage using education.



References I

Olkin, I. and S. Yitzhaki (1992): "Gini Regression Analysis," International Statistical Review/Revue Internationale de Statistique, 60, 185–196.

Schaffer, M. E. (2015): "ginireg: Program to estimate Gini regression." .

Schröder, C. and S. Yitzhaki (2016): "Reasonable sample sizes for convergence to normality," Communications in Statistics - Simulation and Computation, 0918, 1–14.

Yitzhaki, S. (2015): "Gini's mean difference offers a response to Leamer's critique," Metron, 73, 31–43.

Yitzhaki, S. and E. Schechtman (2013): The Gini Methodology, vol. 272.