Third lab

Linear and non-linear regression functions

Fixed effects regression

AKM regression

Bernardo Fanfani
bernardo.fanfani@unito.it

Multivariate linear regression in STATA

All regressors enter as a linear function of the dependent variable

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + e$$

β_1 : Predicted increase of Y for a unit increase of X, holding steady the value of the other regressors.

If we insert a group of mutually exclusive dummies (e.g., one dummy for each year) one of these dummies is excluded to avoid multicollinearity.

I n this case $\beta_{anno=2000}$ is the predicted increase in Y in 2000 with respect to the year category omitted from the regression (1999 in this case), holding steady the value of the other regressors.

```
. reg retrib03 eta tempo_d occ_manuale uomo n_dipendenti i.settore i.anno
```

Source	SS	df	MS
Model	1.0520e+14	9	1.1689e+13
Residual	1.9592e+14	515,404	380137771
Total	3.0113e+14	515,413	584244651

Number of obs	=	515,414
F(9, 515404)	=	30749.89
Prob > F	=	0.0000
R-squared	=	0.3494
Adj R-squared	=	0.3494
Root MSE	=	19497

retrib03	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
eta	562.7524	2.950287	190.74	0.000	556.97	568.5349
tempo_d	-13667.28	94.3575	-144.85	0.000	-13852.22	-13482.35
occ_manuale	-21026.54	61.51605	-341.81	0.000	-21147.11	-20905.97
uomo	12947.99	59.18251	218.78	0.000	12831.99	13063.98
n_dipendenti	3.655495	.0269801	135.49	0.000	3.602615	3.708375
settore						
Servizi	-5496.835	59.10781	-93.00	0.000	-5612.685	-5380.986
Altri settori	-5476.379	119.5129	-45.82	0.000	-5710.621	-5242.138
anno						
2000	-308.8896	66.75735	-4.63	0.000	-439.7319	-178.0473
2001	-466.7969	66.68498	-7.00	0.000	-597.4973	-336.0964
_cons	25071.9	127.1806	197.14	0.000	24822.63	25321.17

Non-linear relations between Y (dependent var.) and X (independent var.)

If the relationship between Y and X is nonlinear:

- The effect on Y of a change in X depends on the value of X - that is, the marginal effect of X is not constant.
- A linear regression is not a correct specification of the relationship between Y and X - the functional form is wrong!
- The estimator of the effect of X on Y is biased.

The solution is to estimate a regression function that is nonlinear in X:

$$Y = f(X_1, X_2, \ldots X_n) + e$$

The Expected Effect on Y of a Change in X_1 in the Nonlinear Regression Model (6.3)

The expected change in Y, ΔY, associated with the change in X_1, ΔX_1, holding $X_2, \ldots, X_k$ constant, is the difference between the value of the population regression function before and after changing X_1, holding $X_2, \ldots, X_k$ constant. That is, the expected change in Y is the difference:

$$\Delta Y = f(X_1 + \Delta X_1, X_2, \ldots, X_k) - f(X_1, X_2, \ldots, X_k). \tag{6.5}$$

The estimator of this unknown population difference is the difference between the predicted values for these two cases. Let $\hat{f}(X_1, X_2, \ldots, X_k)$ be the predicted value of Y based on the estimator $\hat{f}$ of the population regression function. Then the predicted change in Y is

$$\Delta \hat{Y} = \hat{f}(X_1 + \Delta X_1, X_2, \ldots, X_k) - \hat{f}(X_1, X_2, \ldots, X_k). \tag{6.6}$$

How to interpret the marginal effect?

Nonlinear functions of one independent variable

- There are two main approaches:
- 1. *Polynomials in X*
- • The population regression function is approximated by a quadratic, cubic, or higher-order polynomial
- 2. *Log transformations*
  • We transform either Y, X, or both using the natural logarithm
- • Logarithmic specifications allow estimation of percentage relationships of interest (elasticity)

Polynomials in X

La funzione di regressione della popolazione è approssimata da un polinomio:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \ldots + \beta_r X_i^r + u_i$$

- E' un modello di regressione con regressori multipli in cui i regressori sono potenze di X!

Cubic polynomial

The joint significance test on β_{eta2} e β_{eta3} tells us that – with a 1% significance level - we cannot reject the assumption that the correct specification is quadratic or cubic

The effect of unit age increase on income depends on age. We can calculate the average of the predicted values of Y at given ages (which we denote by $\hat{Y}$) and compute the difference:

$$\hat{Y}(età = x1) - \hat{Y}(età = x2)$$

To quantify and test the significance of the effect of a unit increase in age from 39 to 40:

```
reg retrib03 c.eta##c.eta##c.eta tempo_d occ_manuale uomo n_dipendenti i.settore i.anno
margins, over(eta) post
di _b[40.eta]-_b[39.eta]
test _b[40.eta]==_b[39.eta]
```

```
. * specificare il reddito come funzione cubica dell'età
.
. reg retrib03 eta eta2 eta3 tempo_d occ_manuale uomo n_dipendenti i.settore i.anno

      Source |       SS           df       MS      Number of obs   =   515,414
-------------+----------------------------------   F(11, 515402)   =  25374.36
       Model |  1.0579e+14        11  9.6170e+12   Prob > F        =    0.0000
    Residual |  1.9534e+14   515,402   379005318   R-squared       =    0.3513
-------------+----------------------------------   Adj R-squared   =    0.3513
       Total |  3.0113e+14   515,413   584244651   Root MSE        =     19468

------------------------------------------------------------------------------
    retrib03 | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         eta |   -1428.29   115.2209   -12.40   0.000     -1654.12   -1202.461
        eta2 |   63.96913   3.101201    20.63   0.000     57.89087    70.04738
        eta3 |  -.6341895   .0268471   -23.62   0.000    -.6868091   -.5815699
     tempo_d |  -13470.54   94.88028  -141.97   0.000     -13656.5   -13284.57
 occ_manuale |  -21003.45   61.56471  -341.16   0.000    -21124.11   -20882.78
        uomo |   12878.27    59.1266   217.81   0.000     12762.39    12994.16
n_dipendenti |   3.655694   .0269519   135.64   0.000     3.602869    3.708519
             |
     settore |
     Servizi |  -5437.054   59.04858   -92.08   0.000    -5552.788   -5321.321
Altri settori|  -5387.755    119.359   -45.14   0.000    -5621.695   -5153.815
             |
        anno |
        2000 |  -329.3663   66.66388    -4.94   0.000    -460.0254   -198.7072
        2001 |  -500.4556   66.60292    -7.51   0.000    -630.9953    -369.916
             |
       _cons |    44085.8   1379.985    31.95   0.000     41381.07    46790.53
------------------------------------------------------------------------------

.
. *verifica l'ipotesi nulla di linearita` contro l'ipotesi alternativa
. *che la regressione della popolazione sia quadratica o cubica
.
. test eta2 eta3

 ( 1)  eta2 = 0
 ( 2)  eta3 = 0

       F(  2,515402) =  771.00
            Prob > F =    0.0000
```

Logarithmic functions of Y or X

- $\ln(X)$ = logaritmo naturale di X
- Le trasformazioni logaritmiche consentono di modellare le relazioni tra variabili in termini "percentuali" (elasticità)

Ecco perché: $\ln(x+\Delta x) - \ln(x) = \ln\left(1+\frac{\Delta x}{x}\right) \approx \frac{\Delta x}{x}$

(quando $\frac{\Delta x}{x}$ è piccolo)

(analisi: $\frac{d\ln(x)}{dx} = \frac{1}{x}$)

Numericamente:

$x = 100$

$\Delta x = 1$

$\frac{\Delta x}{x} = 0{,}01 = 1\%$

$\ln(x+\Delta x) - \ln(x) = 0{,}00995$

$x = 100$

$\Delta x = 5$

$\frac{\Delta x}{x} = 0{,}05 = 5\%$

$\ln(x+\Delta x) - \ln(x) = 0{,}04897$

Logarithmic functions

Tre casi:

Casi	**Funzione di regressione della popolazione**
I. lineare-log	$Y_i = \beta_0 + \beta_1 \ln(X_i) + u_i$
II. log-lineare	$\ln(Y_i) = \beta_0 + \beta_1 X_i + u_i$
III. log-log	$\ln(Y_i) = \beta_0 + \beta_1 \ln(X_i) + u_i$

- L'interpretazione di β_1 è diversa nei tre casi
- Applicando la regola "prima e dopo" è agevole interpretare il significato di β_1 nei tre casi

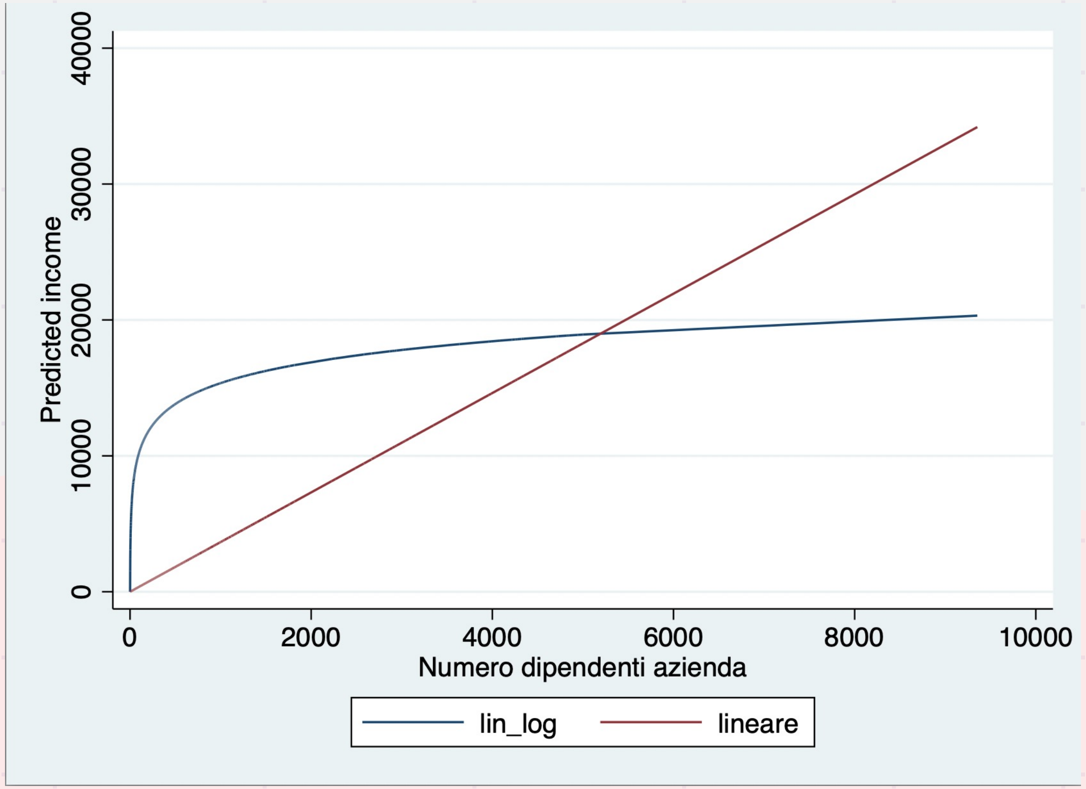
The lin-log case

$$Y = \beta_0 + \beta_1 \ln(X) + u_i$$

con ΔX *piccolo*,

$$\beta_1 \approx \frac{\Delta Y}{\Delta X / X}$$

$100 \times \frac{\Delta X}{X}$ = variazione percentuale in X, quindi **un aumento dell'*1% in X è associato ad una variazione in Y pari a* 0,01β_1**

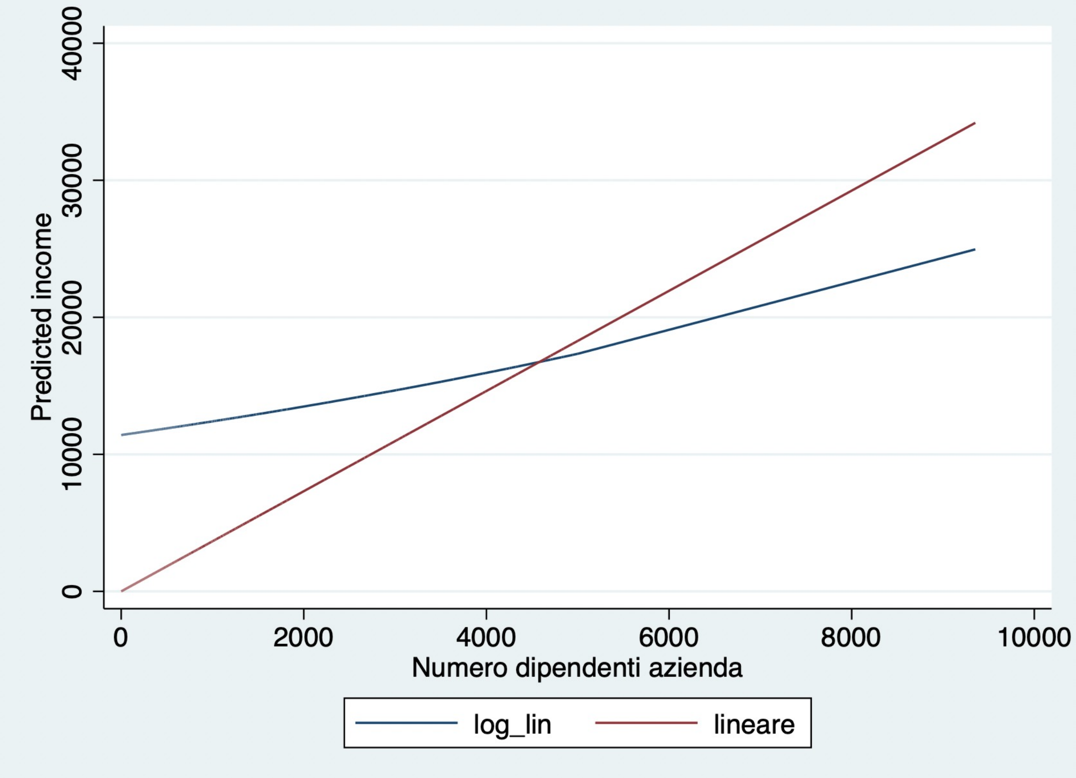


The log-lin case

$$\ln(Y) = \beta_0 + \beta_1 X + u$$

con ΔX *piccolo*, $\qquad \beta_1 \approx \dfrac{\Delta Y / Y}{\Delta X}$

- $100 \times \dfrac{\Delta Y}{Y}$ = variazione percentuale in Y, segue che **un aumento unitario di X è associato ad una variazione in Y pari a $100\beta_1\%$**



The log-log case

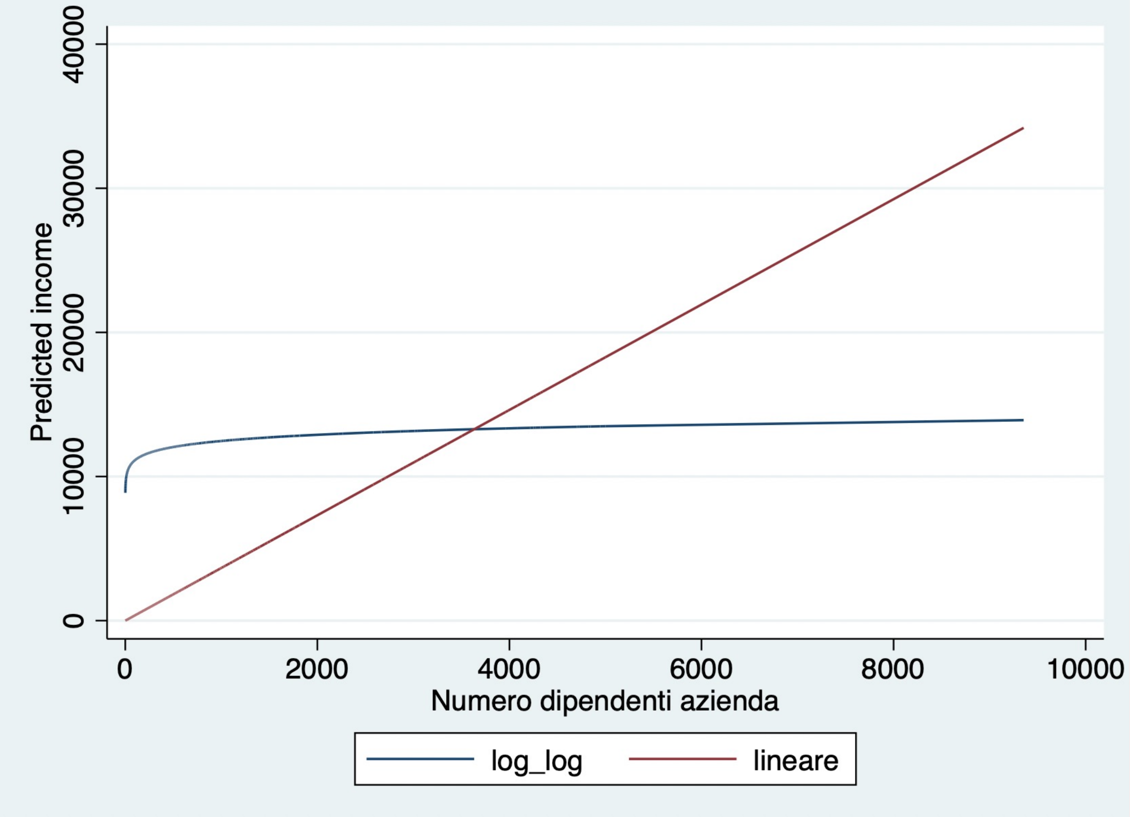
$$\ln(Y) = \beta_0 + \beta_1 \ln(X) + u$$

con ΔX piccolo,

$$\beta_1 \approx \frac{\Delta Y / Y}{\Delta X / X}$$

$100 \times \frac{\Delta Y}{Y}$ = variazione percentuale in Y, e $100 \times \frac{\Delta X}{X}$ = variazione percentuale in X, quindi **un aumento dell'*1% in X è associato ad una variazione in Y pari a β_1%***

Nella specificazione log-log, β_1 è interpretabile come elasticità



Interactions between independent variables

-It is possible that the effect of an independent variable X on Y depends on the value of a second independent variable Z. For example:

- --The effect of being in a manual occupation on wages is different between men and women
- --The effect of being in a larger firm is different between men and women.

--To estimate these heterogeneities in the effect of independent variables, interactions are used.

-Interactions are products of two (or more) independent variables. They are themselves independent variables that are entered into the regression in addition to the basic independents.

With interactions, a classic regression model with two independent variables

$$Y=a + b1\ X + b2\ Z + \text{residual}$$

becomes

$$Y=a + b1\ X + b2\ Z + b3\ (X*Z) + \text{residual}$$

Interpretation of interactions

```
. reg ln_income i.occ_manuale##i.uomo n_dipendenti c.eta##c.eta##c.eta tempo_d i.settore i.anno

      Source |       SS           df       MS      Number of obs   =   515,414
-------------+----------------------------------   F(12, 515401)   =  21422.30
       Model |   75194.192        12  6266.18266   Prob > F        =    0.0000
    Residual |   150758.64   515,401  .292507466   R-squared       =    0.3328
-------------+----------------------------------   Adj R-squared   =    0.3328
       Total |  225952.832   515,413  .438391799   Root MSE        =    .54084

------------------------------------------------------------------------------
   ln_income | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
1.occ_manuale |  -.4453301   .0027106  -164.29   0.000    -.4506428   -.4400174
      1.uomo |   .3688581   .0027086   136.18   0.000     .3635494    .3741669
             |
occ_manuale#uomo |
        1 1  |  -.0577615   .0033948   -17.01   0.000    -.0644152   -.0511078
```

Consider the model

Y=a + b1 X + b2 Z + b3 (X*Z) + residual

There are **three kinds** of interactions: two categorical variables;
categorical*continous;
two continous variablses.

- **Interactions between categorical variables** (dummy variables)

b1: effect of X on Y when Z=0

b2: effect of Z on Y when X=0

b1+b3: effect of X on Y when Z=1

b2+b3: effect of Z on Y when X=1

Intuitively: b3 is the additional effect of X on Y when Z=1

Example: is the gender wage gap greater in manual occupations or office jobs?

GWG in clerical jobs: 36.8%
GWG in manual jobs: 36.8% - 5.7%

Interpretation of interactions

```
. reg ln_income c.ln_dipendenti##i.uomo c.eta##c.eta##c.eta tempo_d occ_manuale i.settore i.anno

      Source |       SS           df       MS      Number of obs   =   515,414
-------------+----------------------------------   F(12, 515401)   =  21831.06
       Model |  76145.4533        12  6345.45444   Prob > F        =    0.0000
    Residual |  149807.379   515,401  .290661794   R-squared       =    0.3370
-------------+----------------------------------   Adj R-squared   =    0.3370
       Total |  225952.832   515,413  .438391799   Root MSE        =    .53913

--------------------------------------------------------------------------------
         ln_income | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------------+------------------------------------------------------------
     ln_dipendenti |   .0475542   .0006331    75.11   0.000     .0463133     .048795
            1.uomo |   .3181581   .0035421    89.82   0.000     .3112157    .3251006
                   |
uomo#c.ln_dipendenti
                 1 |   .0026467   .0007931     3.34   0.001     .0010923    .0042012
```

Consider the model

Y=a + b1 X + b2 Z + b3 (X*Z) + residual

- **Interaction between continous (Z) and dummy variable (X)**

b1: effect of X on Y when Z=0

b2: effect of unit increase of Z on Y when X=0

b1+b3*z: effect of X on Y when Z=z

b2+b3: effect of unit increase of Z on Y when X=1

Intuitively: b3 is the difference in the slope of the relationship between Y and Z when X=1

-Example: -does the gender wage gap increase or decrease as firm size increases?

- is the wage premium from firm size greater for men or women?

- GWG when ln(employees)=0: 31.8%
- GWG increase for 1% increase in number of employees: + 0.002%
- Income increase for 1% increase in no. of employees among women: + 0.047%
- Income increase for 1% increase in no. of employees among men: 0.047% + 0.002%

Fixed effects regression (I)

The classic regression model in which individual *i* is observed several times over time (*t*) can be written:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + e_{it}$$

β_0 and β_1 have a causal interpretation if $E(e_{it} | X_{ij})=0 \ \forall j, t$, that is, if there are no unobservables that influence both X_{ij} and Y_{it}

$E(e_{it} | X_{ij})=0$ can be difficult to defend. For example, working in non-manual occupations is correlated with education, and education has an independent effect on income, but is not observable in the data.

Fixed effects regression (II)

Education and skill are fairly constant over time for workers. We can divide the individual error term into time-constant elements (a_i) and idiosyncratic elements (u_{it}):

$$Y_{it} = \beta_0 + \beta_1 X_{it} + a_i + u_{it}$$

By including individual fixed effects among the independent variables in the regression, I am able to control for a_i.

These fixed effects control for all unobservable individual characteristics that have influence on Y_{it}, as long as these characteristics are constant over time.

The underlying assumption of the model becomes E(u_{it}| X_{it})=0, which is less demanding than the assumption E($e_{it} = a_i + u_{it}$| X_{it})=0

Fixed effects regression (III)

$$Y_{it} = \beta_0 + \beta_1 X_{it} + a_i + u_{it}$$

- The interpretation of β_1 does not change when using the fixed effects regression.
- Variables in X_{it} that are constant over time cannot be included because they are collinear with individual fixed effects.

The spread of so-called linked employer-employee data (LEED) has allowed labor economists to develop also *high-dimensional* fixed effects models…

(Bruno Contini from Univ. Torino has been one of the very first ever developing and analysing LEED since the 1980s)

AKM regression

$$Y_{it} = \beta_0 + \beta_1 X_{it} + a_i + f_{J(i,t)} + u_{it}$$

- $f_{J(i,t)}$ is a firm fixed effect, which measures firms' wage policy conditional on their employment composition
- $f_{J(i,t)}$ can have a causal interpretation if: $E(u_{ij} | X_{it}, a_i, f_{J(i,t)}) = 0 \ \forall j, t$
- This implies that temporary shocks in wages can't be a systematic reason driving worker mobility toward high or low-wage firms…
- The name AKM comes from Abowd, Kramarz and Margolis (1999 Econometrica) first using this method

Card, Heining and Kline (2013) use a variance decomposition method (see do file of the lecture) based on the AKM regression. They show that higher dispersion in firm fixed effects can explain a relatively large portion of the growth in wage inequality occurred in West-Germany from the 1980s to the early 2000s.

It's a very influential paper that has given rise to a large literature trying to estimate firm wage policies and to use them for several purposes (see papers provided in the lecture material)…

AKM-based variance decomposition

$$var(Y_{it}) = var(\beta_1 X_{it} + a_i) + var(f_{J(i,t)}) + 2 * cov(\beta X_{it} + a_i, f_{J(i,t)}) + var(u_{it})$$

This type of decomposition has been highly debated and attracted a lot of interest. If firms' heterogeneity explains a great proportion of inequalities and of their growth, non-competitive mechanisms and the underlying models of the labor market could be more credible than competitive ones.

The academic debate around this decomposition is highly technical, as there are some known sources of bias in this decomposition…

- The measurement error of $f_{J(i,t)}$ is negatively correlated with the measurement error of a_i. This induces an underestimation of cov($f_{J(i,t)}$, a_i). The problem is particularly relevant whenever the mobility of workers across firms is low, which increases the measurement error of $f_{J(i,t)}$
- The recent literature has developed methods for correcting the bias in the estimates of var($f_{J(i,t)}$) and cov($f_{J(i,t)}$, a_i). See in particular the papers by Bonhomme et al 2023 and Kline et al. 2020.