Estimating a Model of Assortative Matching with Large Firms

With German and Spanish data

Part I: The Model & Estimation Strategy

Model

- The economy consists of workers with different skill x, and firms with different managerial skill y. Both are characterised by cumulative density functions $H_w(x)$, $H_f(y)$.
- They both match and produce output according to a production technology
$$F(x, y, l_x, r_x)$$
- Which depends on their skills, the l_x number of workers of type x the firm type y employs, and the managerial resources r_x (time the manager can spend supervising her workers of skill x). $\int r_x \, dx = 1$ (resources of the firm are normalized to one)
- Total output of the firm is the sum of output across worker types:
$$\int F(x, y, l_x, r_x) \; dx$$

Model

- The optimization problem of the firm is therefore

$$\max_{l_x, r_x} \int [F(x, y, l_x, r_x) - w(x)] \; dx$$

- Dividing by r_x and denoting $\theta = l_x / r_x$, the problem becomes

$$\max_{x, \theta} f(x, y, \theta) - \theta w(x)$$

- Where $f(x, y, \theta) = F(x, y, l_x / r_x, 1)$ is the production function in intensive form.

This means that optimally firms only hire one type of worker, and they have to decide which type and how many to hire.

Assortative Matching

- An equilibrium is therefore a feasible allocation $R(x, y, \theta)$ and a strictly positive wage $w(x)$ that solves the firms problem.
- We focus on assortative matching, that is, monotonic allocations that are monotonic in x, y. That is:
  - Positive Assortative Matching (PAM): High x matches with high y.
  - Negative Assortative Matching (NAM): High x matches with low y.
- Here is when the complementarities of the different inputs become important

Input Complementarities

Denote F_{ij} the cross partial derivate w.r.t inputs I and j:

- F_{xy} is *type complementarity* – good firms do better with good workers.
- F_{lr} is *quantities complementarity* – always positive with constant returns to scale.
- F_{yl} is *span of control complementarity* - good firms do better with more workers.
- F_{xr} is *managerial resource complementarity* – if positive (and large) more time spent with good workers is more productive than time spent with *bad* workers

Input Complementarities

Why are the cross-derivatives important?

- A necessary condition for PAM is:

$$F_{xy}F_{lr} \geq F_{yl}F_{xr}$$

- The opposite inequality is necessary and sufficient condition for NAM.
- The interpretation of the input complementarities relates to questions of skill-bias technological change vs quantity-bias technological change (or increases in F_{xy} vs increases in F_{yl}).

Solving the Model

The solution to the model is given by solving a system of two differential equations:

- Under PAM: $\theta'(x) = \frac{H(x)F_{yl} - F_{xr}}{F_{lr}}$; $\mu'(x) = \frac{H(x)}{\theta(x)}$;
- Under NAM: $\theta'(x) = \frac{H(x)F_{yl} + F_{xr}}{F_{lr}}$; $\mu'(x) = -\frac{H(x)}{\theta(x)}$;

Where $\mu(x)$ is the map $y^*(x)$.

- This system can be solved using numerical methods.

Choice of Function

A very nice special case arises when the production function F is multiplicative separable. In particular, the function we use is:

$$F(x, y, l, r) = A(x, y) * B(l, r)$$

Where:

$$A(x, y) = \left(\omega_A * x^{\frac{(\sigma-1)}{\sigma}} + (1 - \omega_A) * y^{\frac{(\sigma-1)}{\sigma}}\right)^{\frac{\sigma}{(\sigma-1)}}$$

$$B(l, r) = l^{\omega_B} * r^{(1-\omega_B)}$$

- Our goal is to estimate ω_B, ω_B and σ using real world data.

Targets for estimation

- That means that we have 3 unknown parameters. What can we target to get them?
- The solver takes as inputs the distributions of $H_w(x)$ and $H_f(y)$.
- The solver delivers matched vectors of $(x, \mu(x), \theta(x, \mu(x), w(x))$, which in turn can be used to calculate moments of the distributions of θ and w.

Estimation Strategy

- The idea is to estimate a distribution of worker skill and firm skill from the data, and get a distribution of firm size, wages and profits.
- The solver takes as inputs $\widehat{H}_w(x)$ and $\widehat{H}_f(y)$, calculates the moments and the distributions of $\hat{\theta}$ and $\hat{w}(x)$, and then calculates the distance between the actual firm size and wage distribution and the one implied by the model.
- An optimization routine does this over and over until it finds the parameter combo that produces the result that best fits the data.

Estimation Strategy

Some pending questions:

- Should it be best to target distribution moments of firm size and wages, or distance between distributions? → if using distance, we would need the distribution of profits as well: 3 parameters to estimate need 3 equations to solve.
- What is the best proxy for firm and worker skill? → This depends on the data…

Solving a Model of Assortative Matching with Large Firms

Cristina Lafuente

Julia Faltermeier

David Pugh

Python

- The solver has been built using 
- Some very useful libraries currently expanding, like quantecon – see the GitHub page of the project for more info! – make the code intuitive and easy to program.
- Much of the credit of the following goes to David Pugh, who started this project and now keeps updating it with quantecon functions.

The Setup

- There are 3 basic classes of objects we use:
  - Input class: like workers or firms, it needs a distribution, boundaries, and name. It contains functions for symbolic and numeric pdf and cdf distributions.
  - Model class: Needs two inputs, a production function (symbolic), a set of parameters for the production function and a type of assortativity (it will be check while solving). It contains all cross-derivatives, expressions for the wages and profits, numeric and symbolic. It uses them and the cdfs from inputs to build the system of differential equations.
  - Solver class: Which branches out into Shooting or Collocation. It contains the code to carry out the integration of the system and store the results.

A Shooting Solver

- Intuitively, the way the solver works is the following:
- We know from the solution that if PAM holds, the highest type worker matches with the highest type firm. The only bit we need to guess is the size of this top firm: $\bar{x} - \bar{y} - \theta^{guess}$
- Once we have those three, we can integrate one step the system of differential equations:

$$\theta'(x) = \frac{H(x)F_{yl} - F_{xr}}{F_{lr}} \; ; \; \mu'(x, \theta(x)) = \frac{H(x)}{\theta(x)}$$

- Which will give us the next combination of $x - \mu(x, \theta(x)) - \theta(x)$
  - We do not need to integrate for the wages, as when we have both other two variables (θ and y) it can be calculated as F_l.

A Shooting Solver

- We keep doing this until one of three things happen:
  - We run out of workers! – feasibility condition failed → update guess and start again
  - We run out of firms! – market clearing failed (-wage) → update guess and start again
  - Last firm matches with last worker. All firms and workers matched with positive wages → success!

A Shooting Solver

- A faster way of solving this problem is to approximate the two set of differential equations with polynomials – orthogonal collocation.
- This method is faster because we can use simple, ready implemented functions to search for the best coefficients that satisfy the optimality, feasibility and market clearing conditions.
- The initial guess in this case would be the solution from the shooting solver of $\mu^*(x)$, $\theta^*(x)$
- Work in progress to be done this week – ask David for details.

Generating distributions

- The shooting solver delivers matched vectors of $x - \mu(x) - \theta(x) - w(x)$
- We don't know a priori the distribution of $\theta(x)$ or $w(x)$, but from the distribution of x we can do a simple change of variables to get the distribution of θ and w.
- Unfortunately, this requires to invert these functions (so we can calculate $H_w(x(\theta))$ for example), something that is complicated if the function $x(\theta)$ is non-monotonic.
- A possible solution to this is just calculate the moments of the distribution of θ and w, which do not requite to evaluate the pdf.

Objective function

- In the end, the idea is to get an objective function that contains three (or more) equations that measure the distance between what the model gives and what we observe in the data.
- For example, we can have [mean(θ), std(θ), mean(w), std(w)], or [distance of θ distributions, distance of w distributions, distance of π distributions].
- Once we have that, apply a simple(x) minimization routine to iterate in search for the best parameters.
- When collocation is ready to use, we only need to use shooting the first time round, and use the previous parameters results as new initial guesses.

Contents

Number of observations

	Number of workers	Number of firms
1996	2,472,655	8,292
2005	2,397,092	14,870
2010	1,629,542	14,359

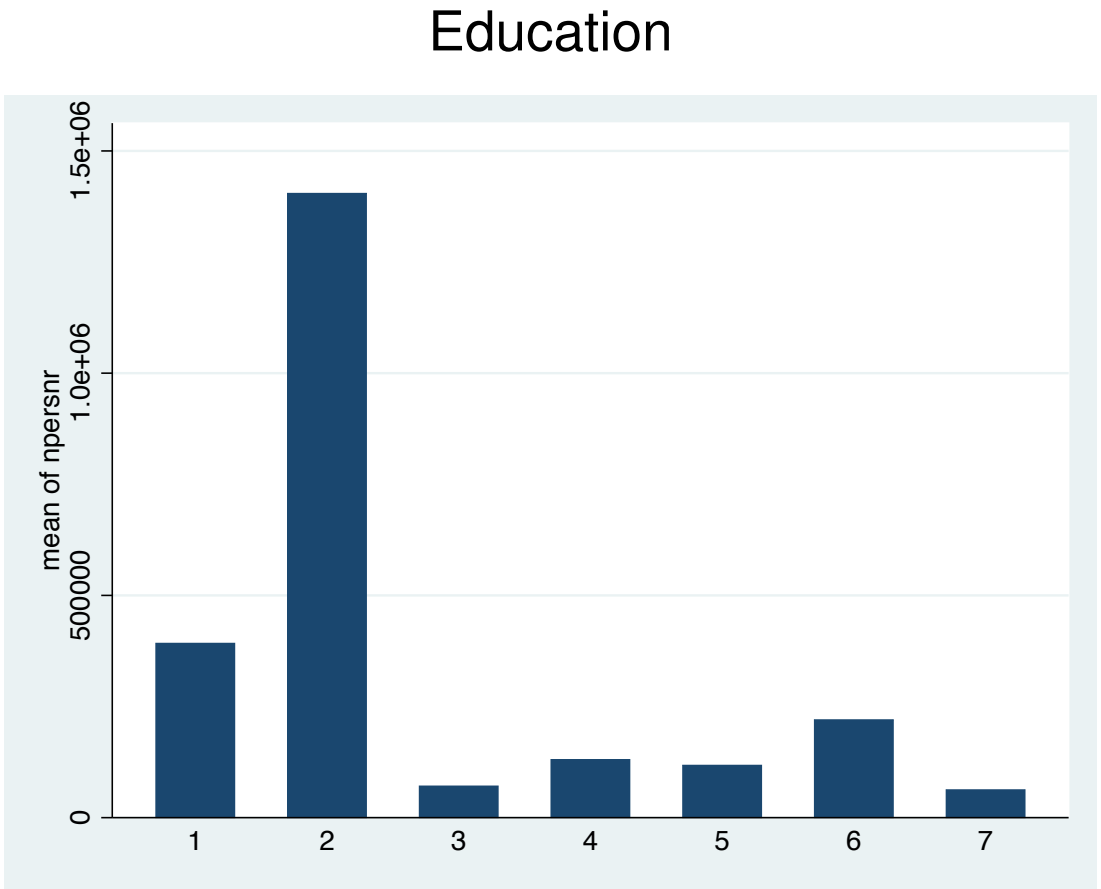
- parallel employments possible, same person can appear several times in the dataset
- nearly 7 times as many firms in the dataset, but IAB survey (needed for profit) only available for small subset

Worker type - Education

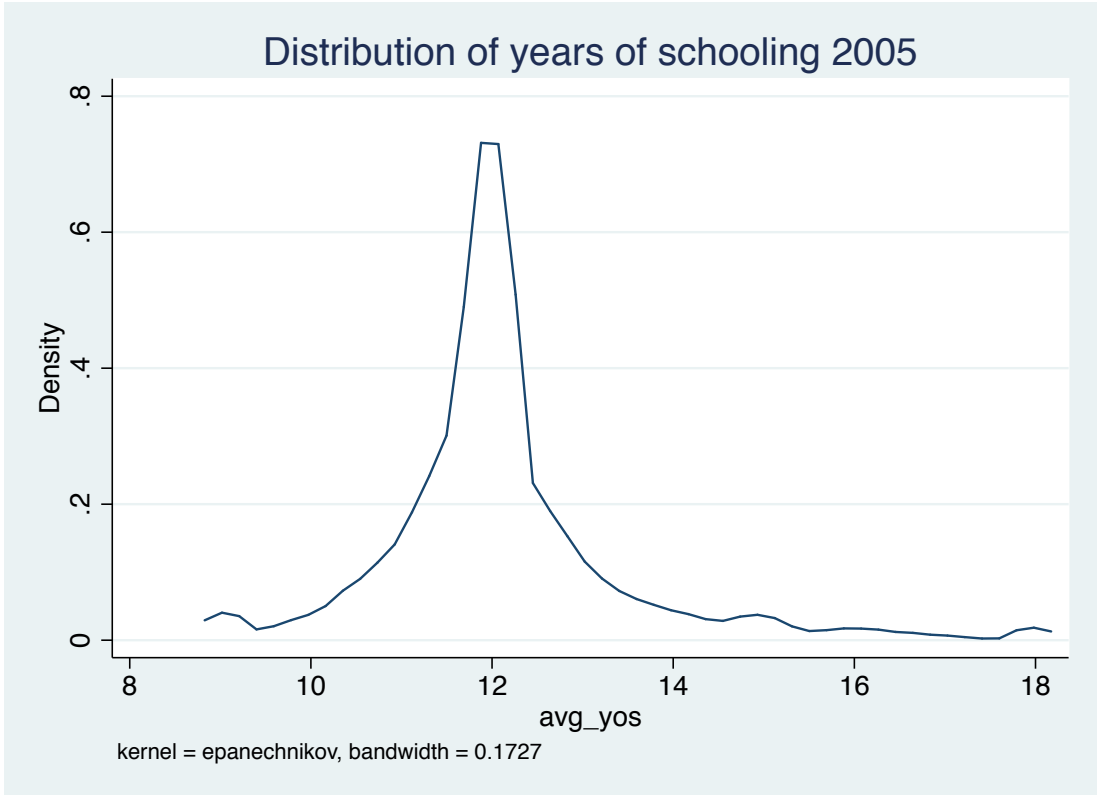
	Number of obs	Avg. daily wage	Years
Only general school leaving certificate	392,302	60.138	9
General school leaving certificate + Apprenticeship	1,404,520	90.867	12
Abitur w/o Apprenticeship.	71,124	45.739	13
Abitur with Apprenticeship	130,793	104.170	15
Uni. of applied science	117,834	130.408	16
University degree	220,259	128.936	18
Missing	231,878	42.104	-

- Need to assume years of schooling to make education variable numerical
- Ranking? Is an apprenticeship less than Abitur?
- We don't know if worker completed school
- No info about apprenticeship before/ after uni degree
- Alternatively: use difference between date of birth and year of first employment (but data is left-censored before 1975 west, 1990-1992 east)

Schooling - 2005



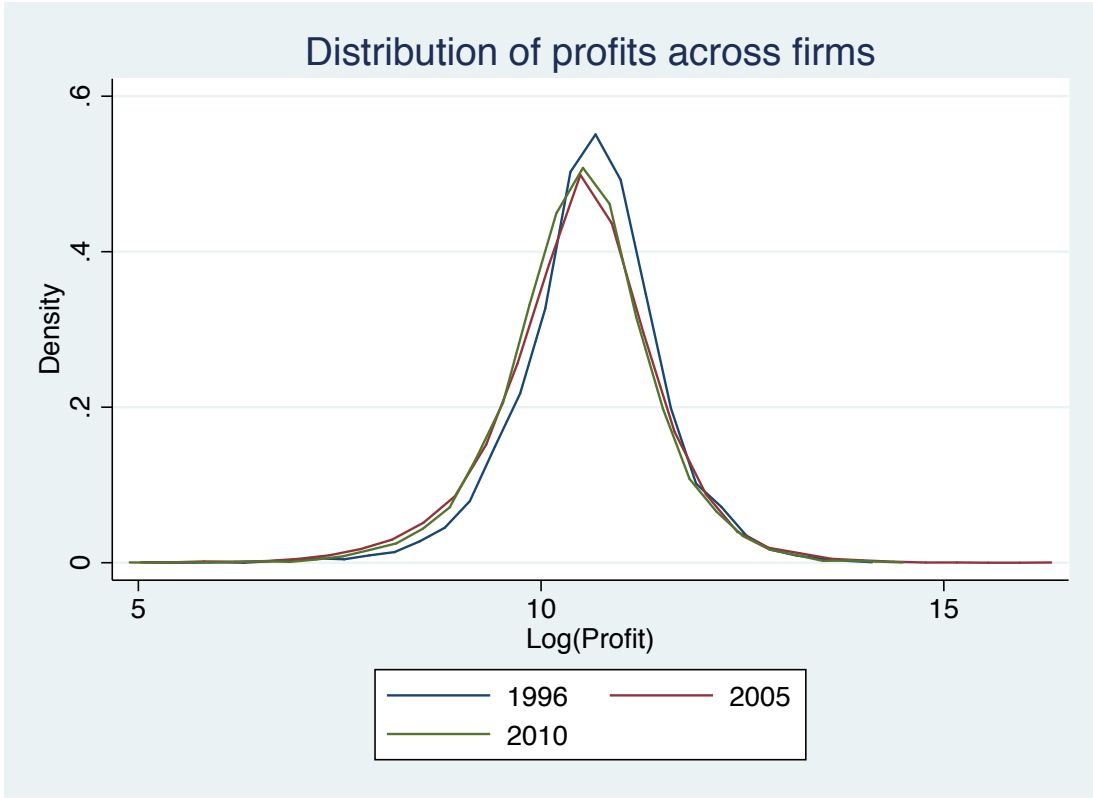
Schooling distribution



Firm type - Profit (previous year)

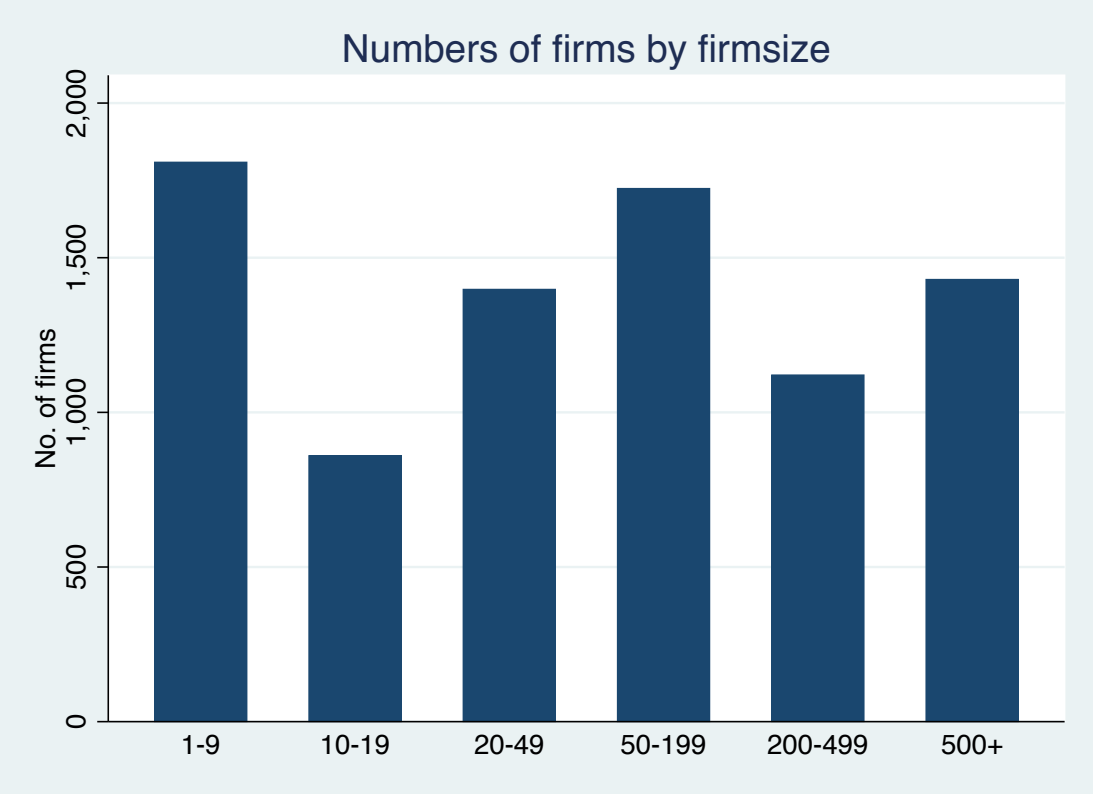
	Number of firms	Mean	s.d.	Min	Max
Revenue	13266	1.55e+08	4.06e+09	1000	3.40e+11
Profit per worker	9854	59153.99	158495	-88800	1.24e+07
Total profit	9854	1.22e+07	9.76e+07	-5720000	4.79e+09

Profit


Distribution of profits across firms
Density
0
.2
.4
.6
5
10
15
Log(Profit)
1996
2005
2010

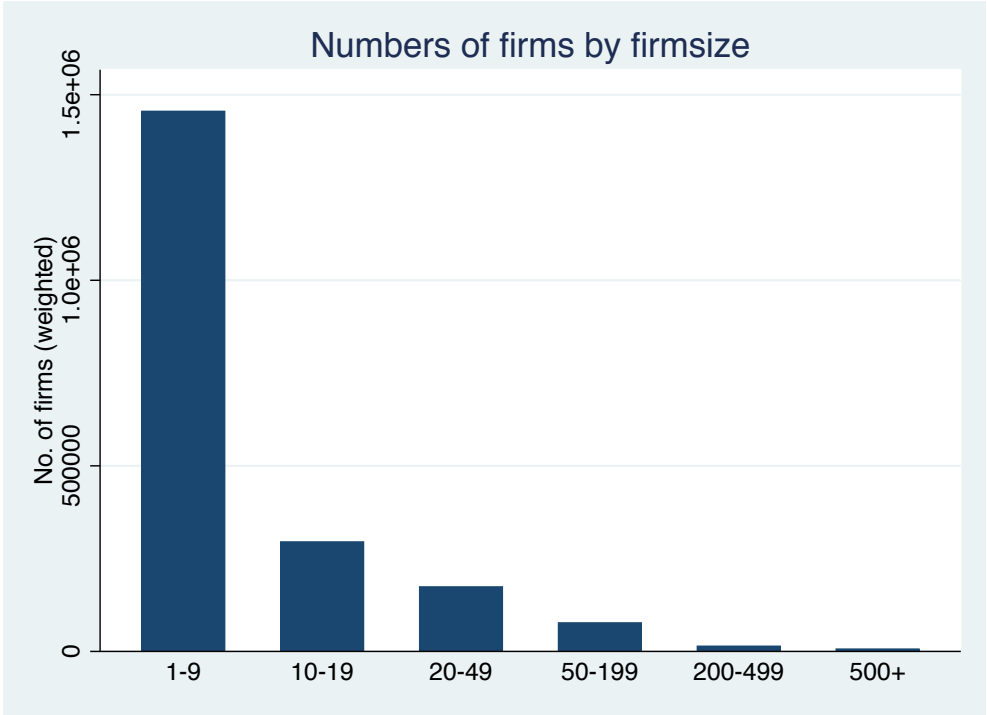
Firm size - 1996

Firm size distribution



Firm size - 1996

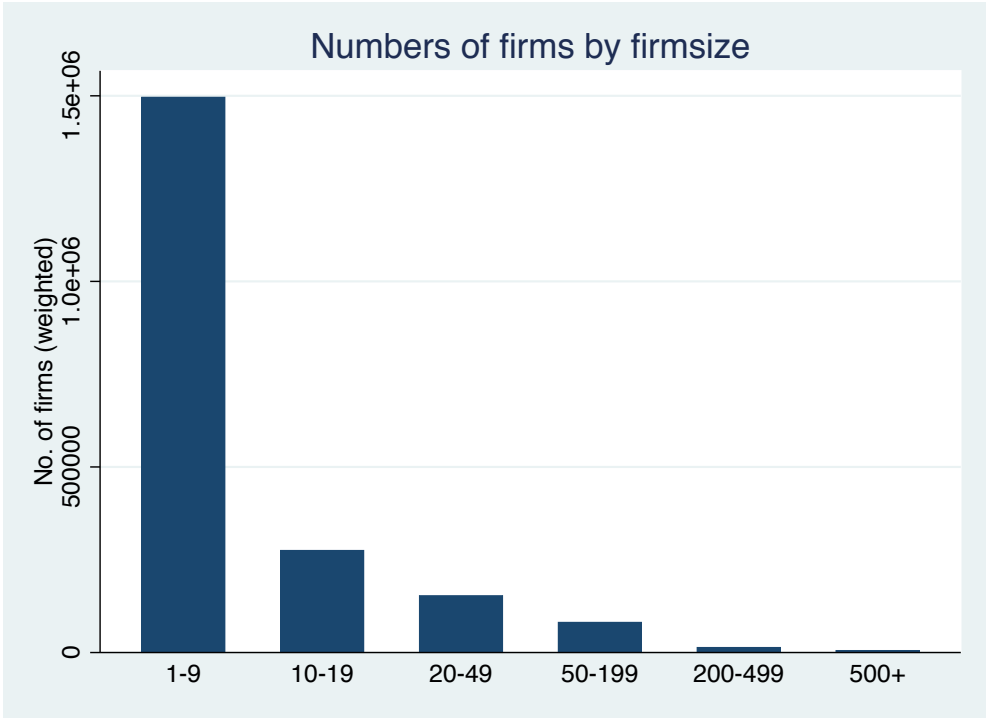
Firm size distribution - weighted



- need to consider weights in estimation
- variables comes from IAB survey, does not differentiate part-time, full-time
- alternatively use number of employees in the sample, but not necessarily the same

Firm size - 2005

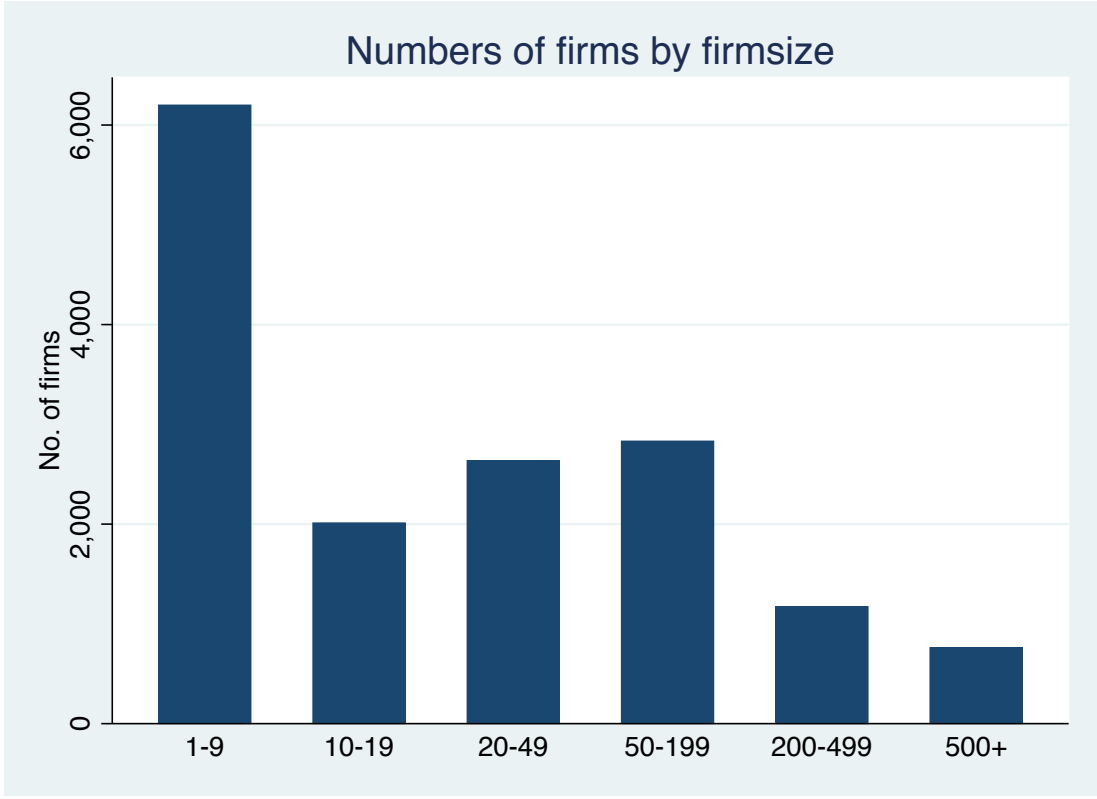
Firm size distribution - weighted



- need to consider weights in estimation
- variables comes from IAB survey, does not differentiate part-time, full-time

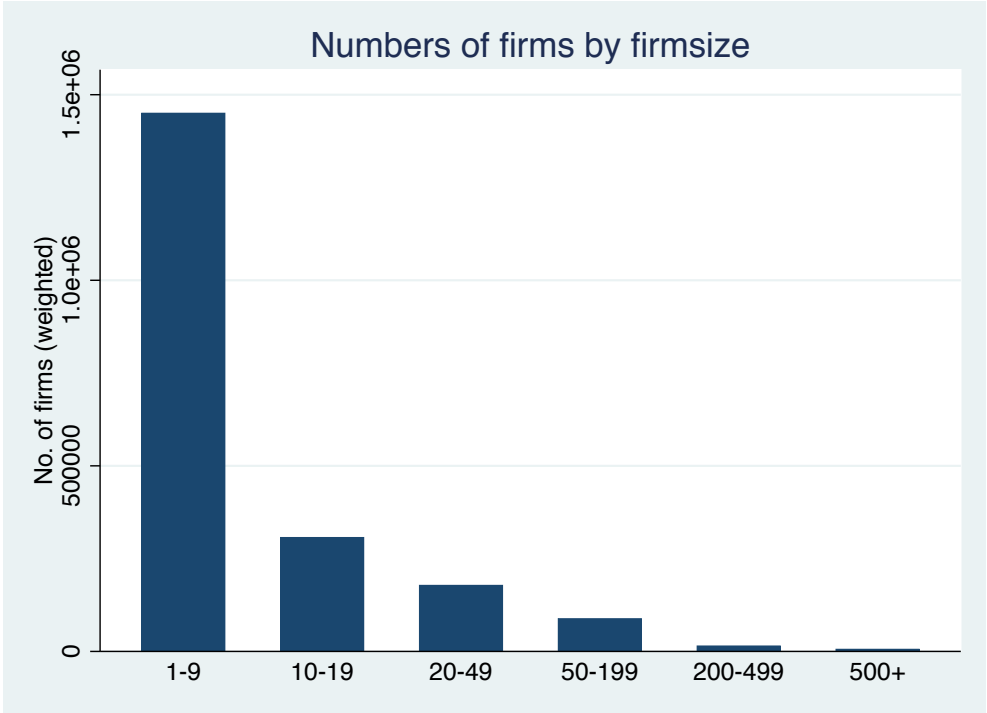
Firm size - 2010

Firm size distribution



Firm size

Firm size distribution - weighted



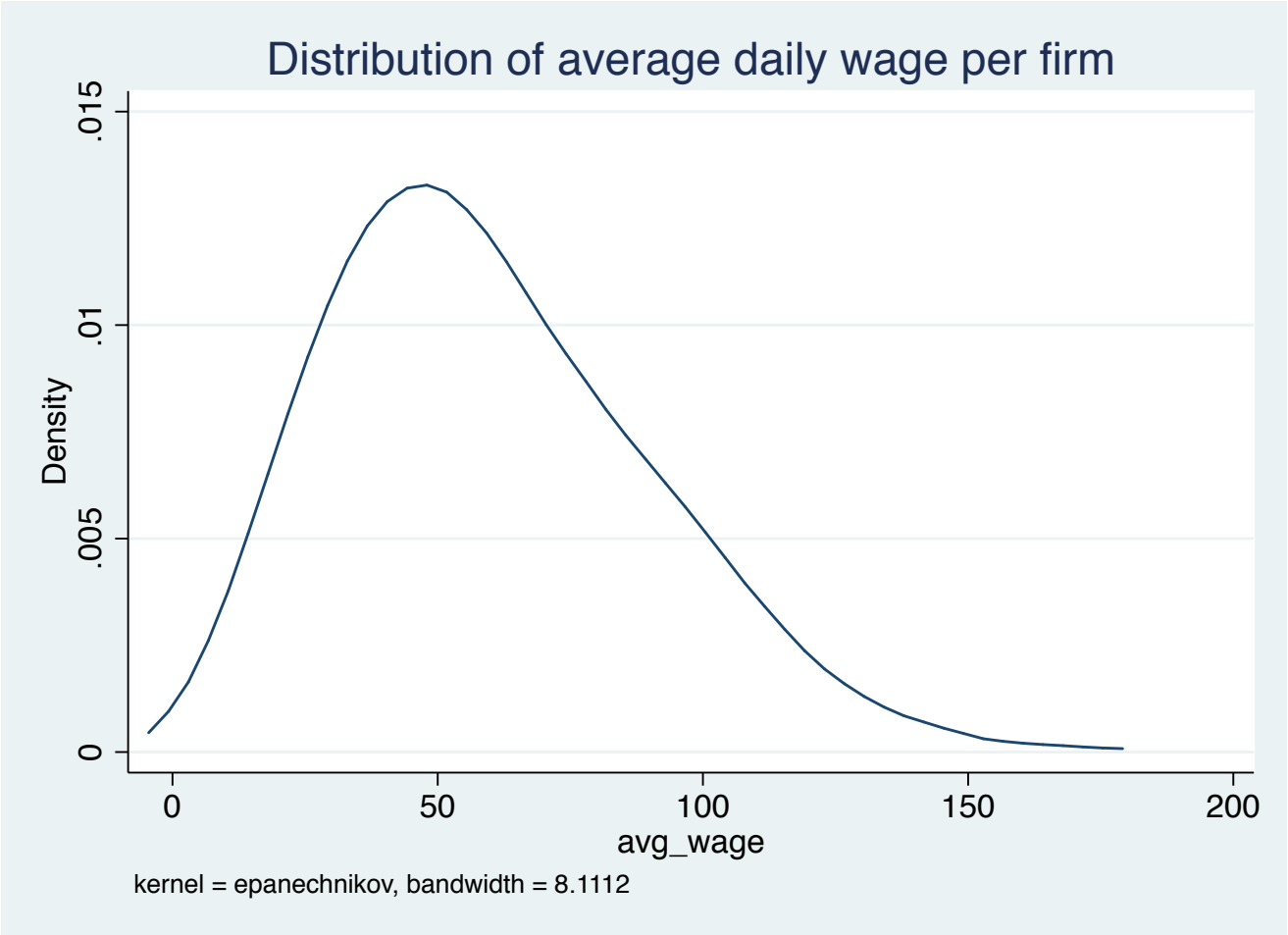
- need to consider weights in estimation
- variables comes from IAB survey, does not differentiate part-time, full-time
- alternatively use number of employees in the sample, but not necessarily the same

Descriptive statistics

	No. of observations	mean	sd	min	max
daily wage	2517835	85.62	47.01	0	170.96
Log tentgelt	2503258	4.18	0.94	−4.605	5.141

- wage=0 means intermittance of employment (illness, sabbaticals, maternity leave), legally counted as employed
- Wage and employment benefit saved in same variable
- Right-censored: censoring at the annual Social Security earnings maximum
- Before 1999 left-censored: only includes wages above marginal earnings threshold

Daily wage distribution if > 0

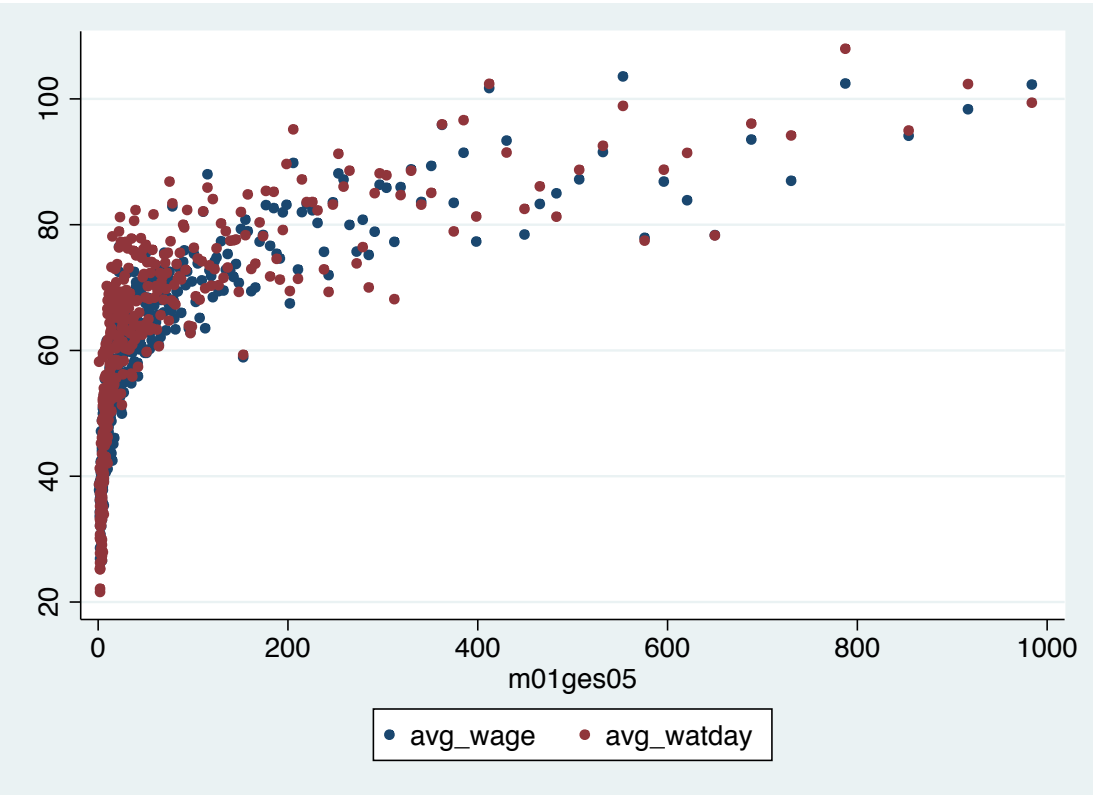


Average wage over firm and bin of 20 ordered by average wage

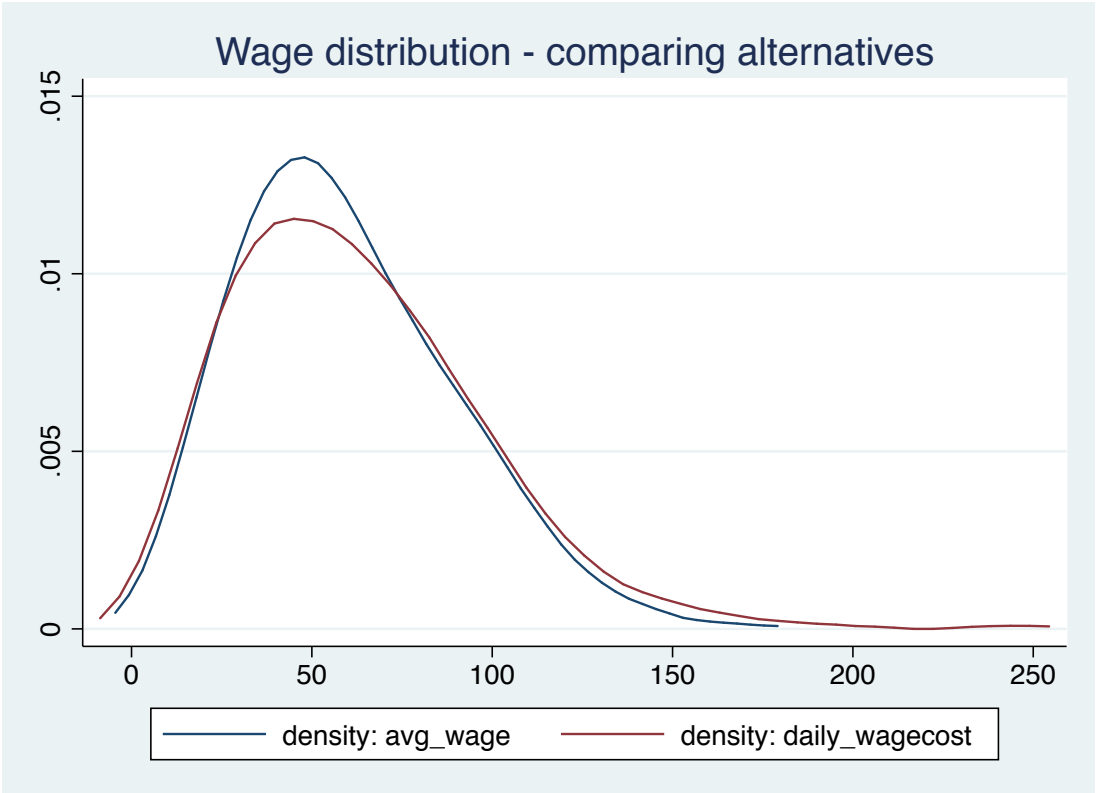
Alternative wage - for 2005

- Know total wage bill for month of June $\rightarrow$ self-reported, but not censored
- doesn't account for hours (part-time employment, on-leave)

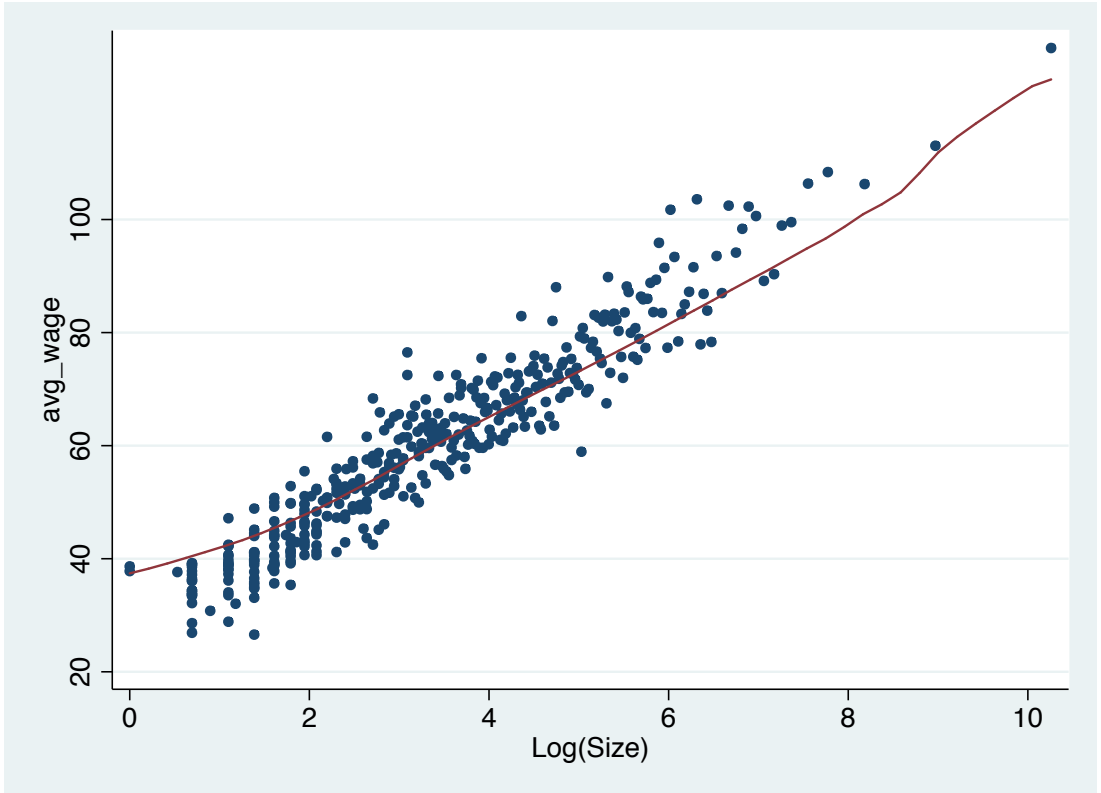
Difference in average wage by estimation method



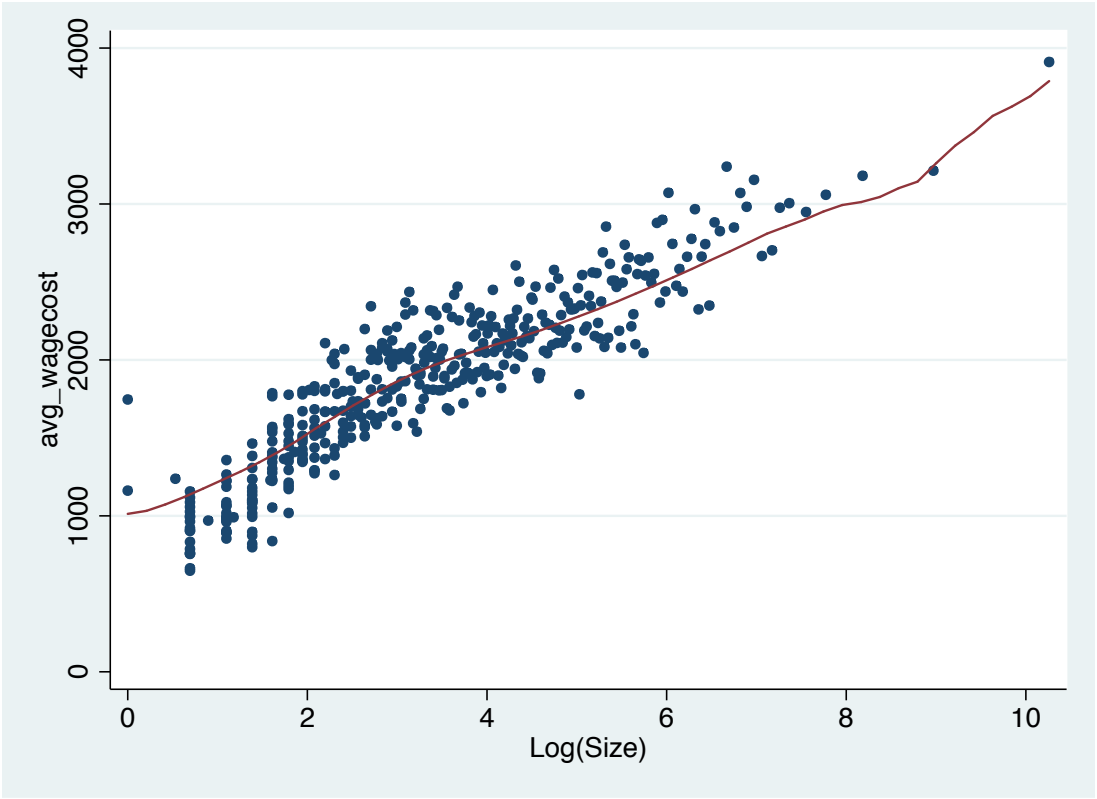
Comparing alternatives



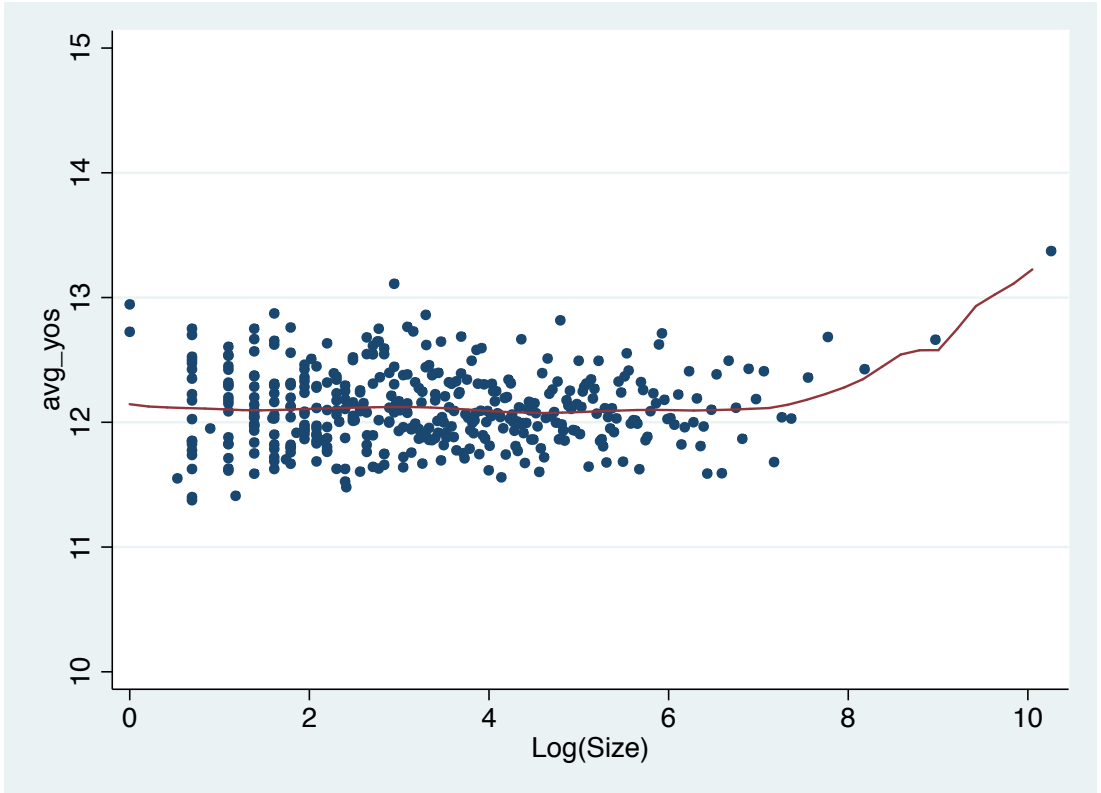
Firm size and wage - 2005



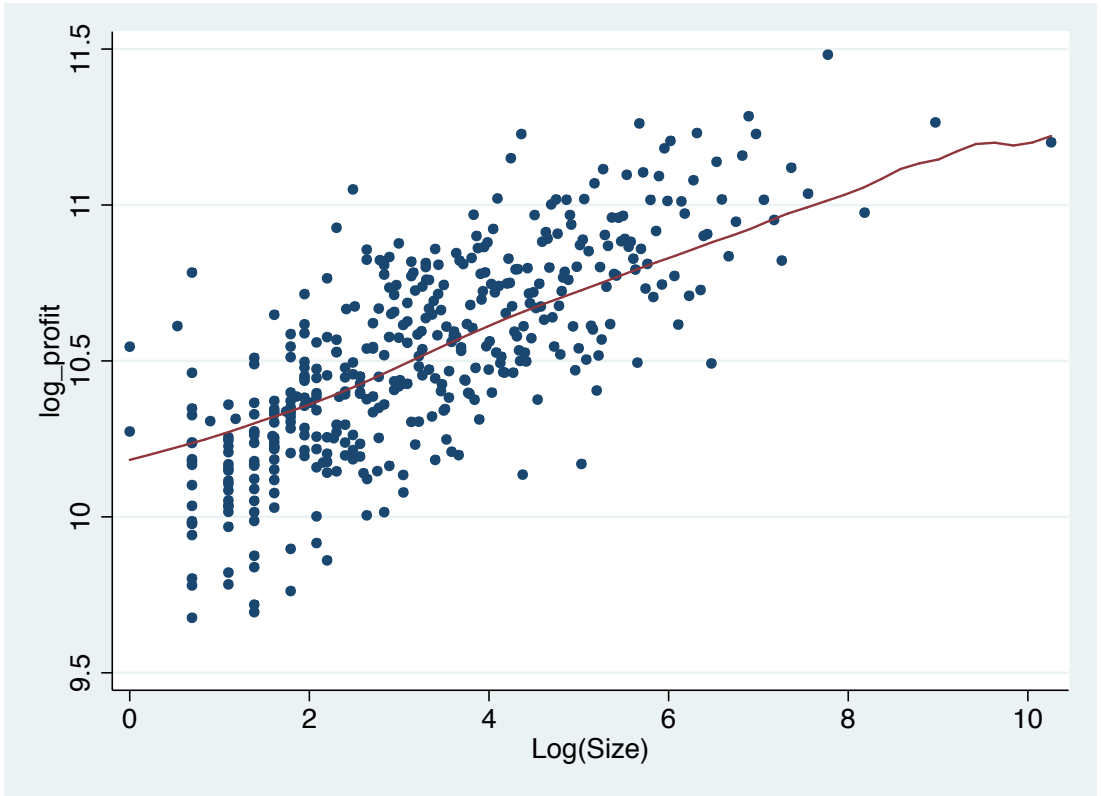
Firm size and wage - 2005



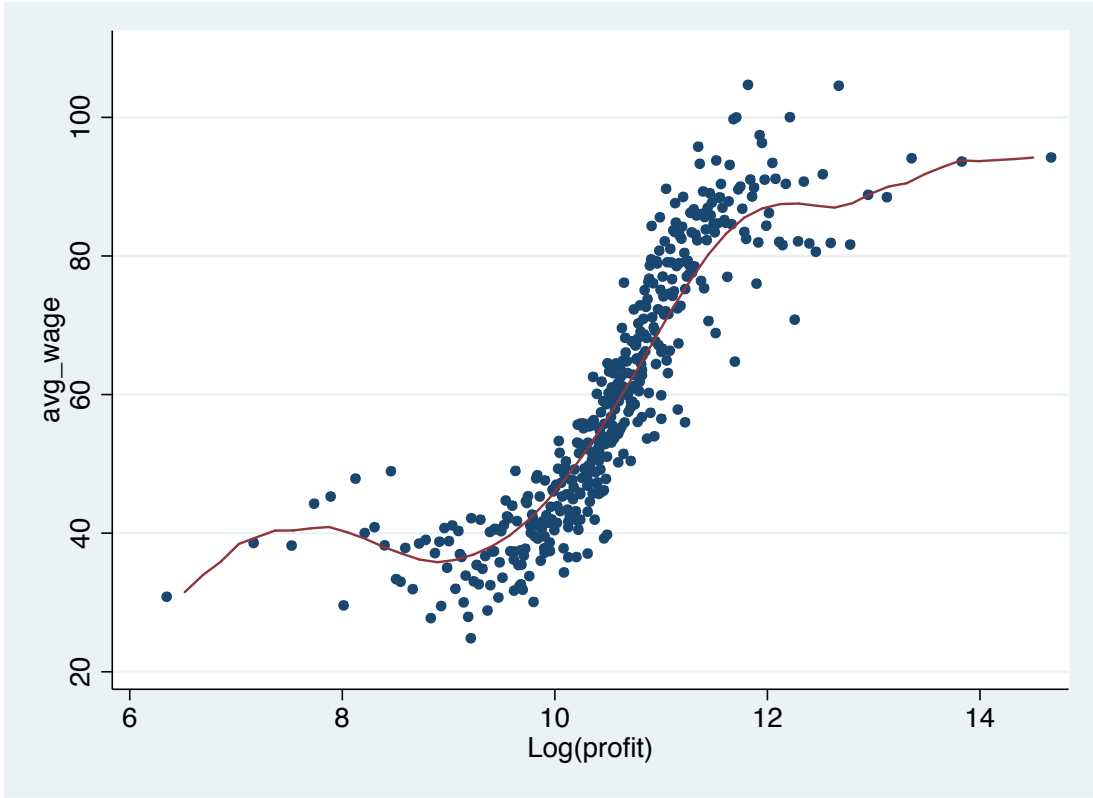
Firm size and education - 2005



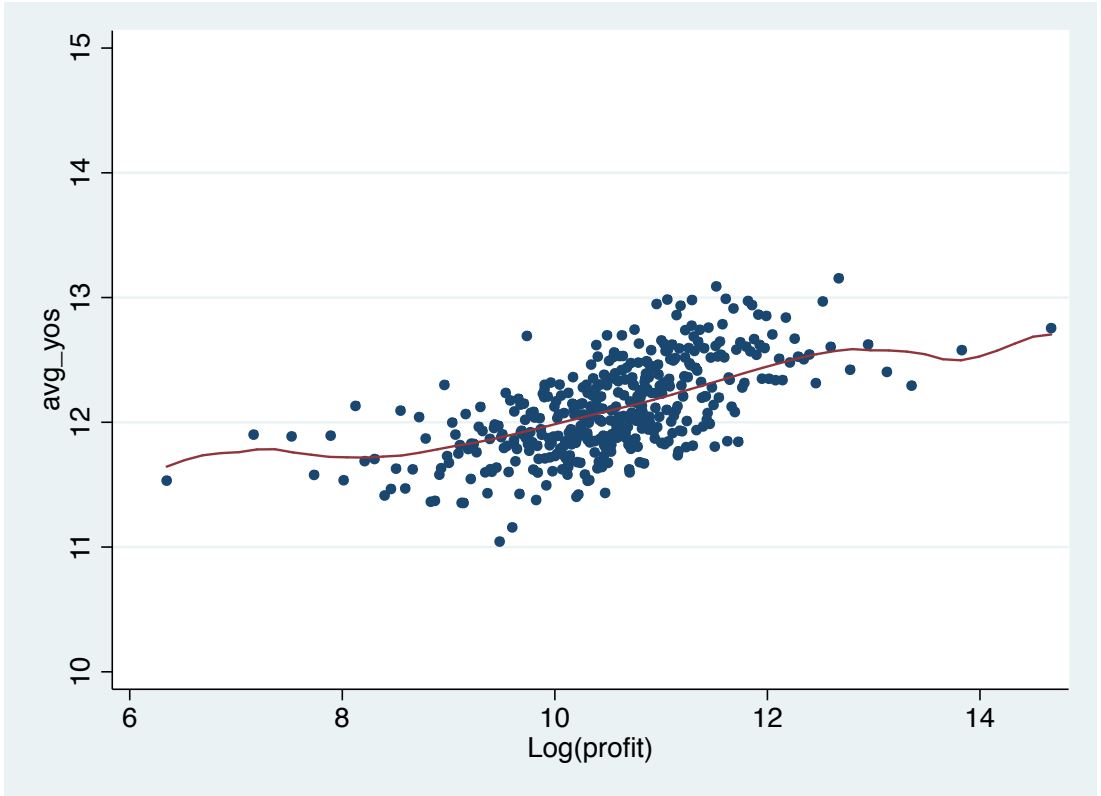
Firm size and profit - 2005



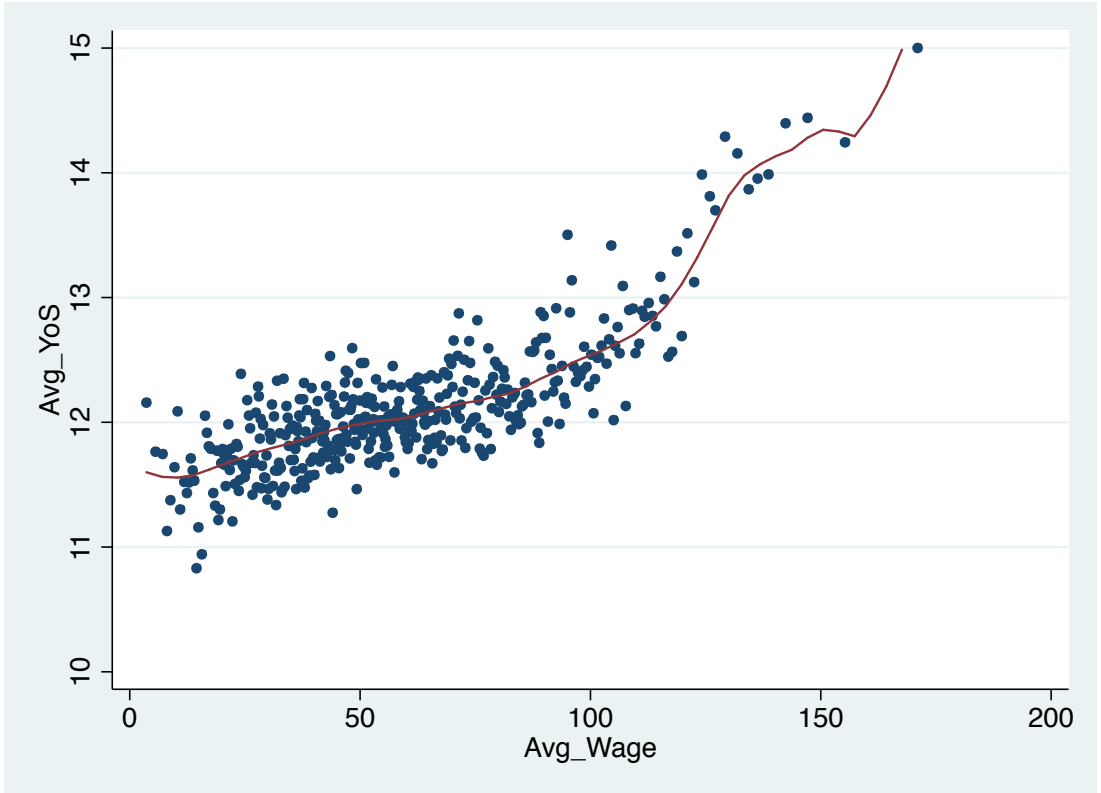
Profit and wage - 2005



Profit and education - 2005



Wage and Education - 2005



Preliminary regression

Source	SS	df	MS	Number of obs	=	7713
Model	2305.31577	29	79.4936471	F(29, 7683)	=	129.43
Residual	4718.73228	7683	.614178353	Prob > F	=	0.0000
				R-squared	=	0.3282
				Adj R-squared	=	0.3257
Total	7024.04805	7712	.910794612	Root MSE	=	.7837

log_profit	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
m01ges05	-.0000221	9.42e-06	-2.34	0.019	-.0000405	-3.60e-06
avg_wage	.0165564	.0004173	39.67	0.000	.0157383	.0173744
sd_wage	.0092325	.0009971	9.26	0.000	.0072779	.0111871
avg_yos	-.0088463	.009057	-0.98	0.329	-.0266004	.0089079
sd_yos	.0220476	.0113109	1.95	0.051	-.0001249	.0442201
2.wo2005	.1590555	.0698264	2.28	0.023	.0221767	.2959343

Part III: Spain

Data Source

- Data comes from the *Muestra Contínua de Vidas Laborales* elaborated by the Spanish Social Security.
- It is an administrative dataset that comprises a panel of a representative sample of the Spanish workforce.
- It also includes matched records form Income Tax declarations of workers – so we can observe their wages and profits.
- Currently it covers the years 2005-2013.

Data Source

- This is not an employer-employee matched data set by construction.
- But the unique identifiers for firms and workers allow us to build it that way – and add information about firm size, age, location, sector.
- This means that, unfortunately, we have to drop unmatched workers and firms.
- Good news is that the data set is already very big, so even when dropping unmatched data, it is still a big sample.

Descriptive statistics – by worker

	Mean	Standard Deviation	Min	Max	Median	90th percentile	Obs
Education	2.166434	1.017146	0	6	2	3	552,364
Av wage	15,519.59	16,691.64	.01	1,733,613	12,988.59	30,602.05	374,717
Av wage (daily)	58.06863	249.641	0.00003	68,373.84	43.39396	91.57422	319,919
Av profit	7,187.25	27,099.97	.01	3,005,061	1,829.724	17,724.17	27,225

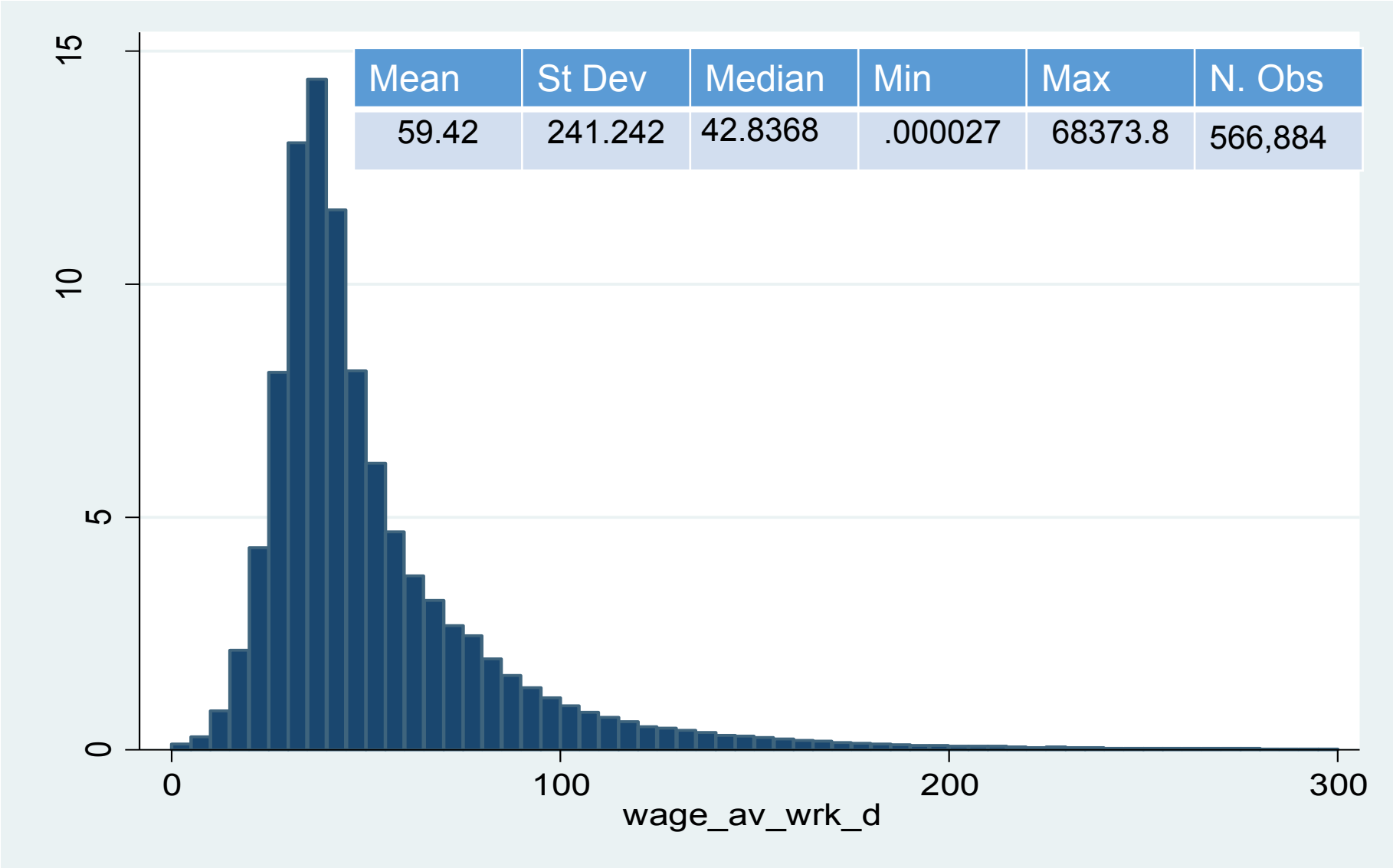
Descriptive statistics – by firm

	Mean	Standard Deviation	Min	Max	Median	90th percentile	Obs
Number of workers	19.10815	255.9593	0	96,402	5	37	376,891
Firm age	9.262611	9.61008	0	105.737	6.00274	21.51233	353,072
Av profit	3,601.598	16,809.16	.01	3,005,061	761.25	8,235.29	64,817
Av wage	8,865.947	10,613.06	.01	733,636.4	6,552.8	18,487.82	255,872
Av wage (daily)	41.5741	92.41269	.0001111	26,461.27	35.97185	63.53	250,344
Av worker education	1.987123	.87669	0	6	2	3	342,023
Number of firms with some income information				403,372			

Choice of variables

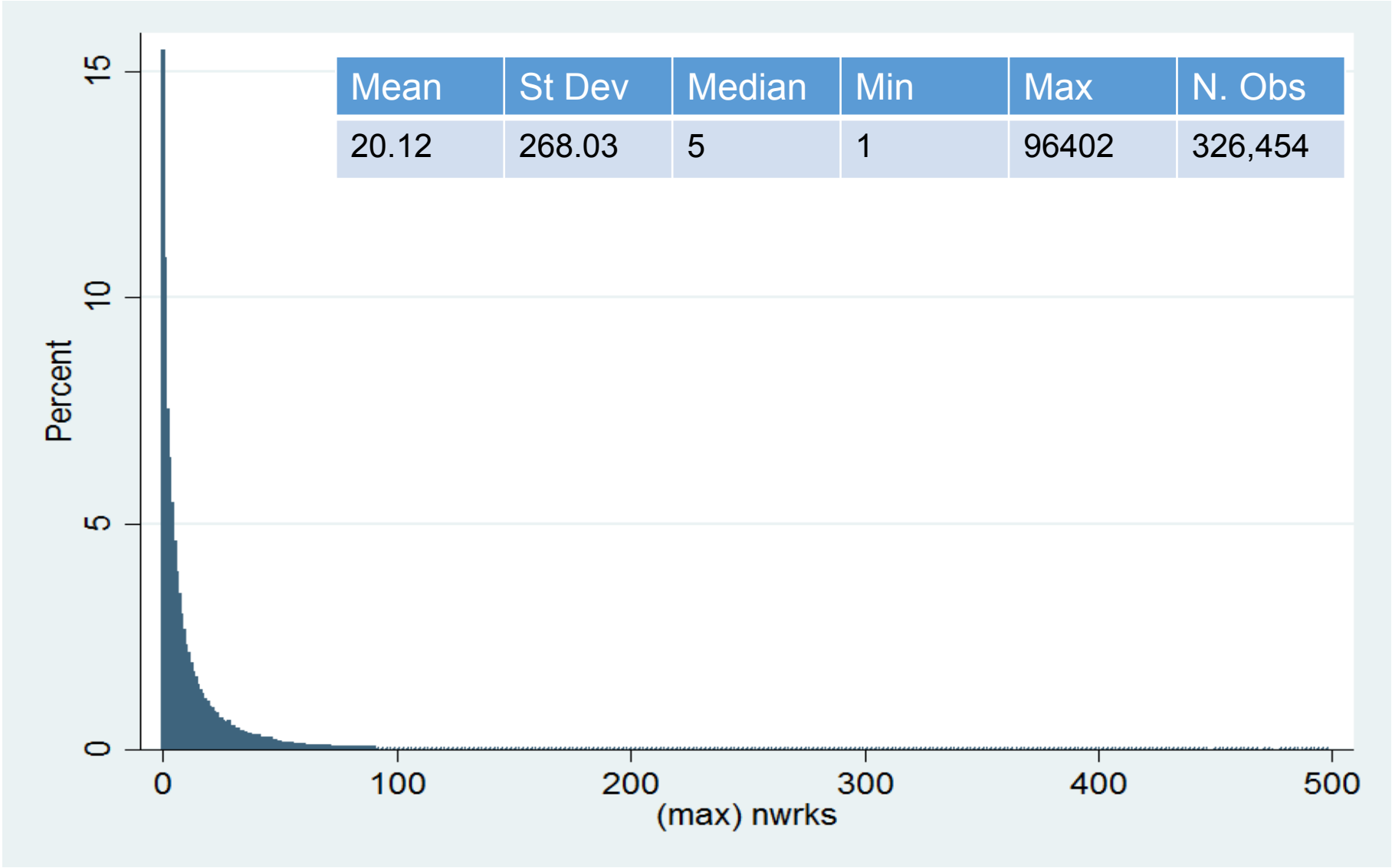
Model variable	Data variable	Alternatives
Worker Skill	Education level (0 to 6) (illiterate to PhD)	Log Wages + experience Wage regression residual
Manager Skill	Log Average Reported Profits (assumes same distribution as profits)	Profit regression residual Profit per worker
Firm Size	Firm size (in workers)	
Wages (return to worker skill)	Average Daily Wages	
Profits (return to manager skill)	Average Reported Profits	Average Reported Profits per worker

Average Daily Wage distribution by firm

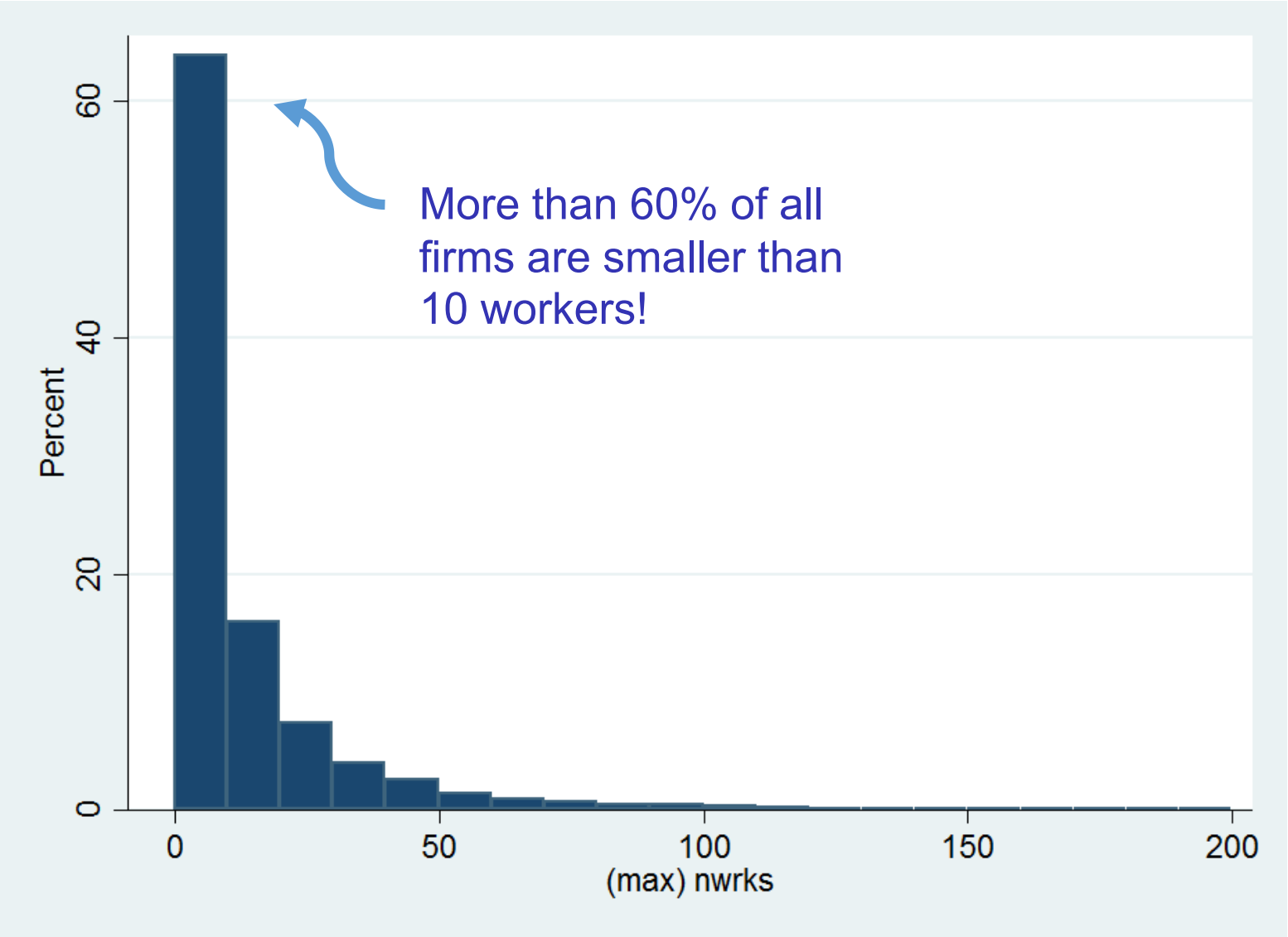


Firm Size Distribution

Mean	St Dev	Median	Min	Max	N. Obs
20.12	268.03	5	1	96402	326,454

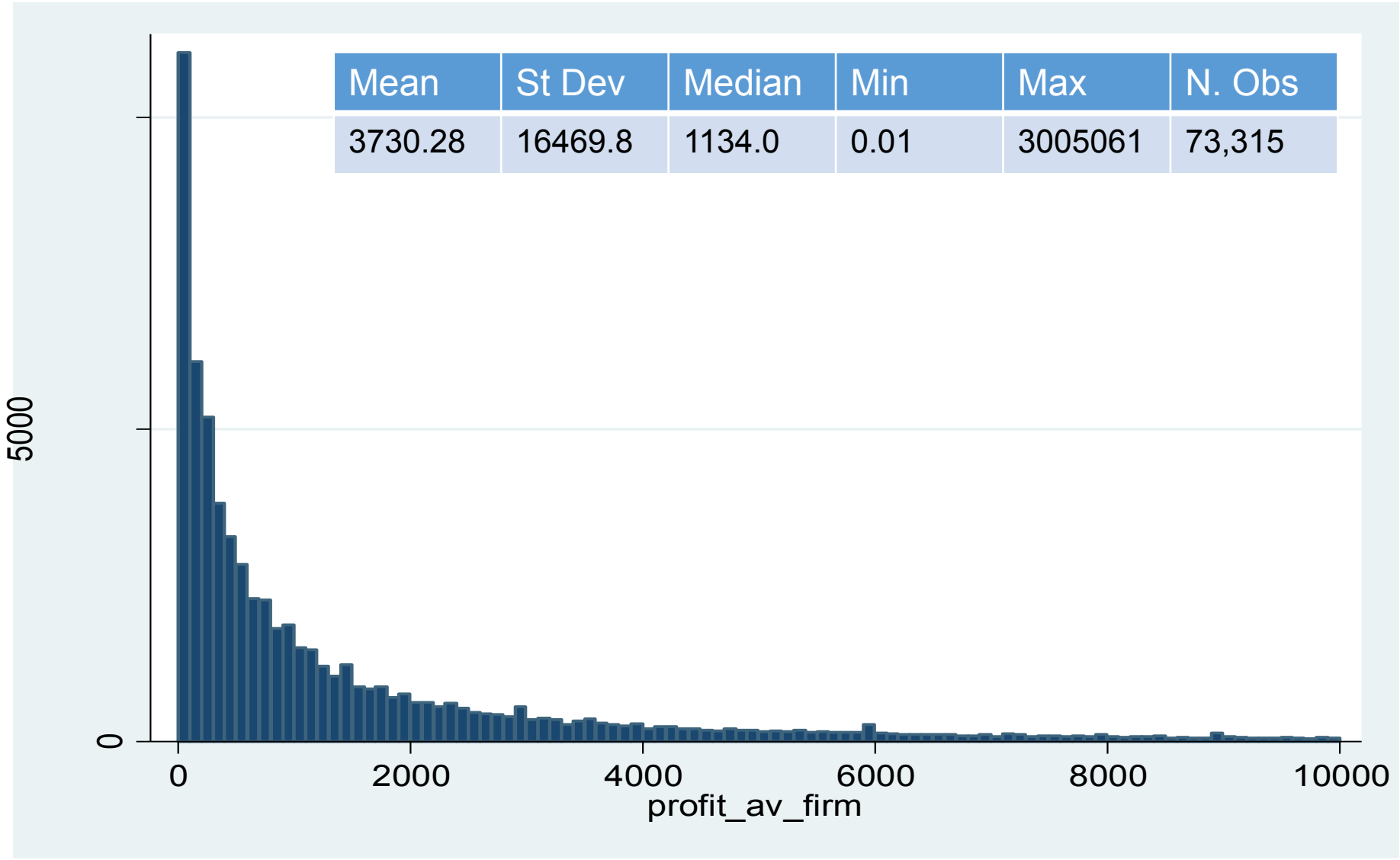


Firm Size Distribution (all)



Profit distribution

Mean	St Dev	Median	Min	Max	N. Obs
3730.28	16469.8	1134.0	0.01	3005061	73,315



Matched data

- For the estimation, we can only take in observations that have everything:

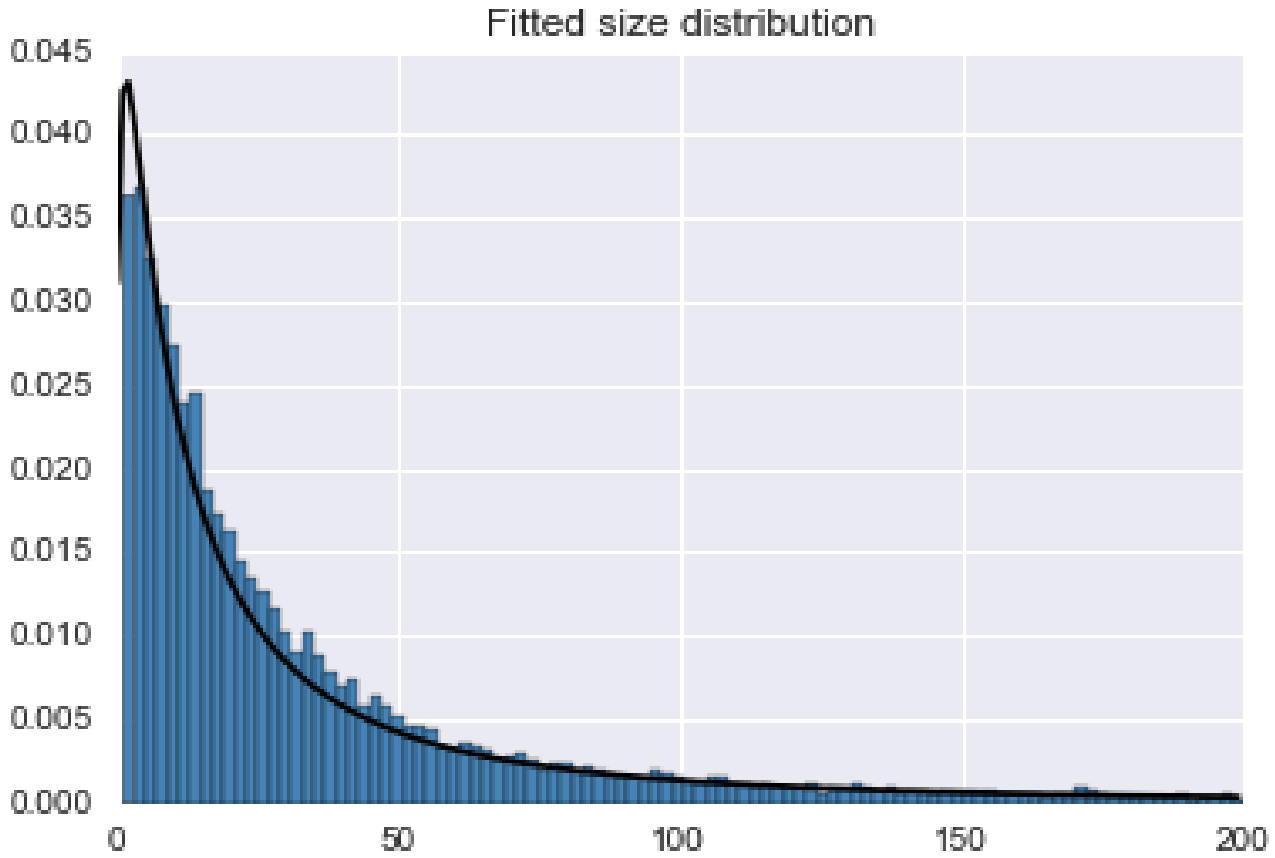
  Firm size – av. Wage – av.profit – av education

- This leaves us with around 20,000 observations:

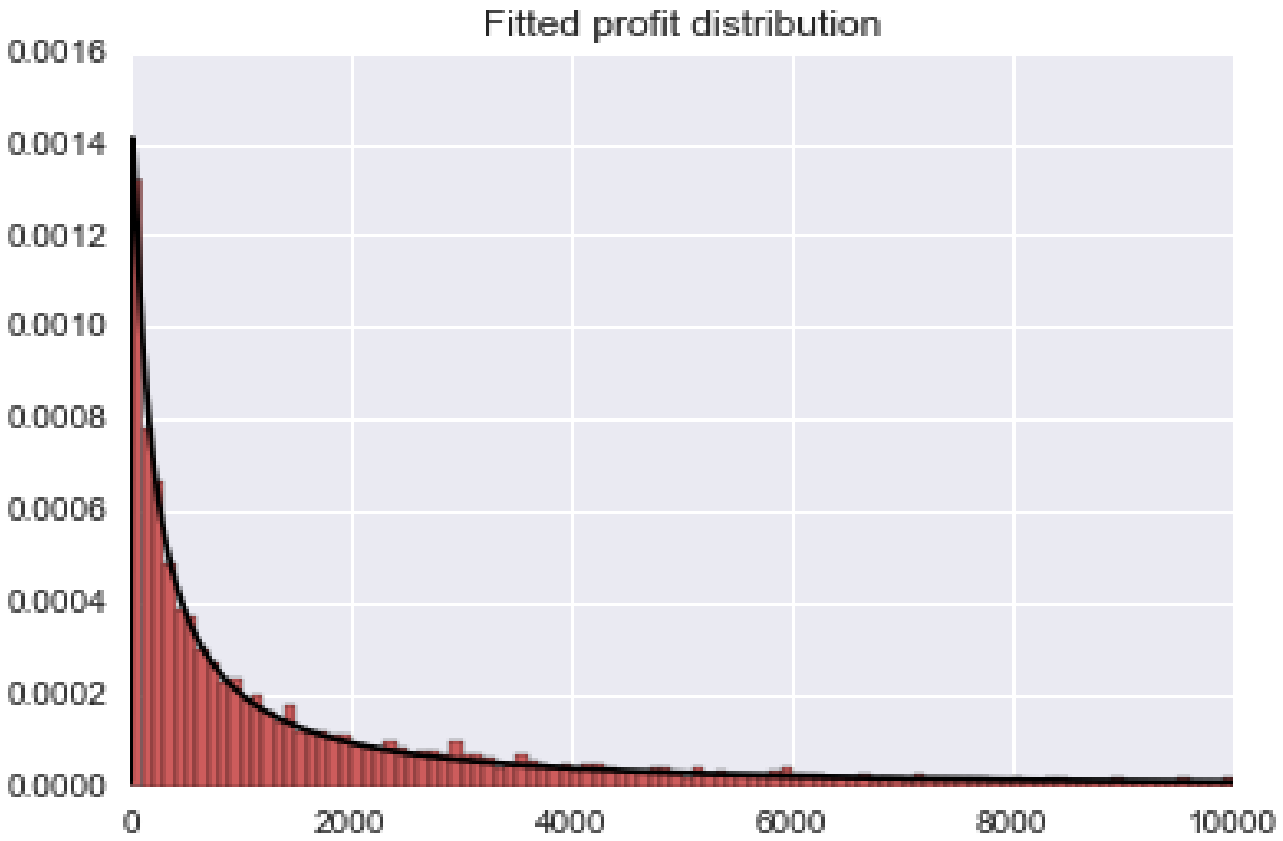
```
. reg profit_av_firm nwrks educ_av_firm wage_av_firm if firmID[_n]!=firmID[_n+1

      Source |       SS       df       MS              Number of obs =   21925
-------------+------------------------------           F(  3, 21921) =   40.79
       Model |  3.6716e+10     3  1.2239e+10           Prob > F      =  0.0000
    Residual |  6.5770e+12 21921   300034139           R-squared     =  0.0056
-------------+------------------------------           Adj R-squared =  0.0054
       Total |  6.6138e+12 21924   301667757           Root MSE      =   17321

------------------------------------------------------------------------------
profit_av_~m |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       nwrks |   .7055322   .3788993     1.86   0.063    -.0371378    1.448202
educ_av_firm |   355.0792   136.6959     2.60   0.009     87.14528     623.013
wage_av_firm |   .0856381   .0093068     9.20   0.000     .0673961    .1038801
       _cons |   2976.042   329.3207     9.04   0.000     2330.549    3621.534
------------------------------------------------------------------------------
```

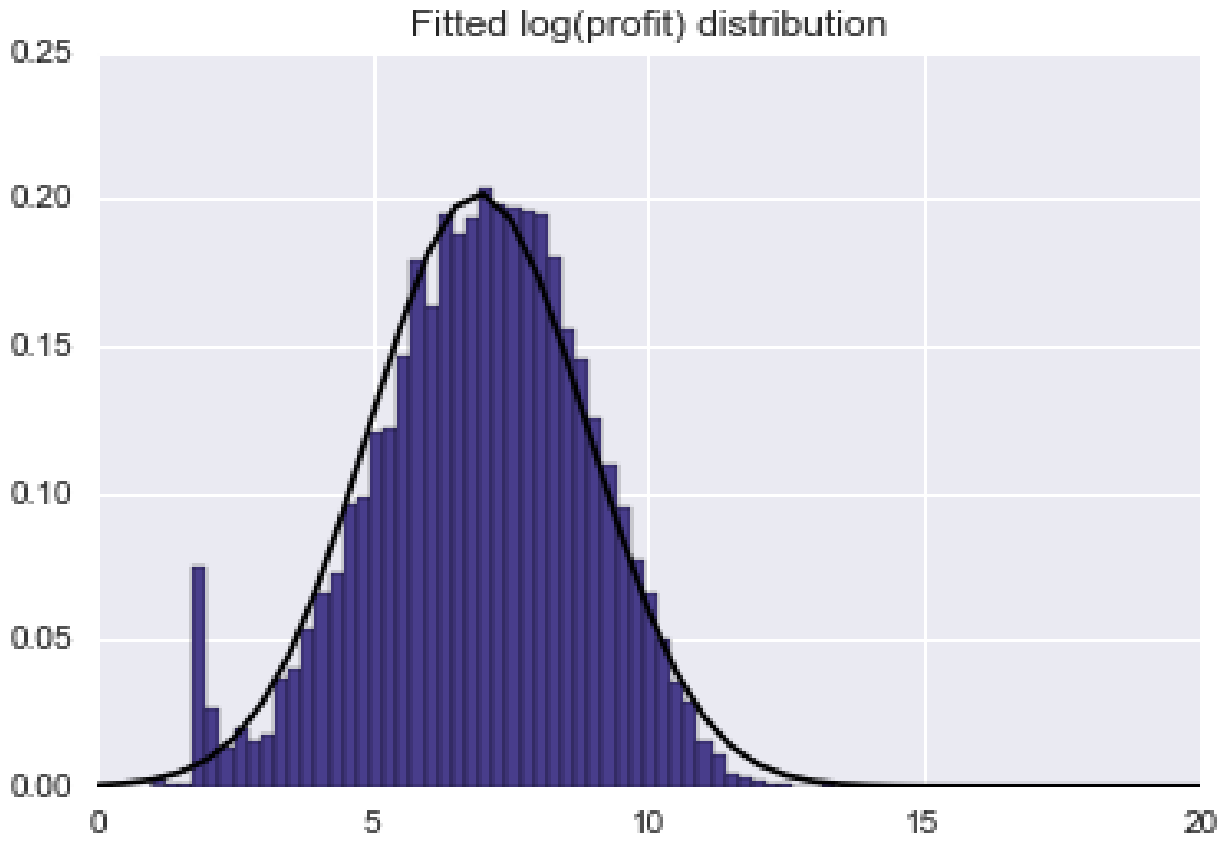
Fitted distributions: Average Wage by firm

Fitted distributions: Firm Size

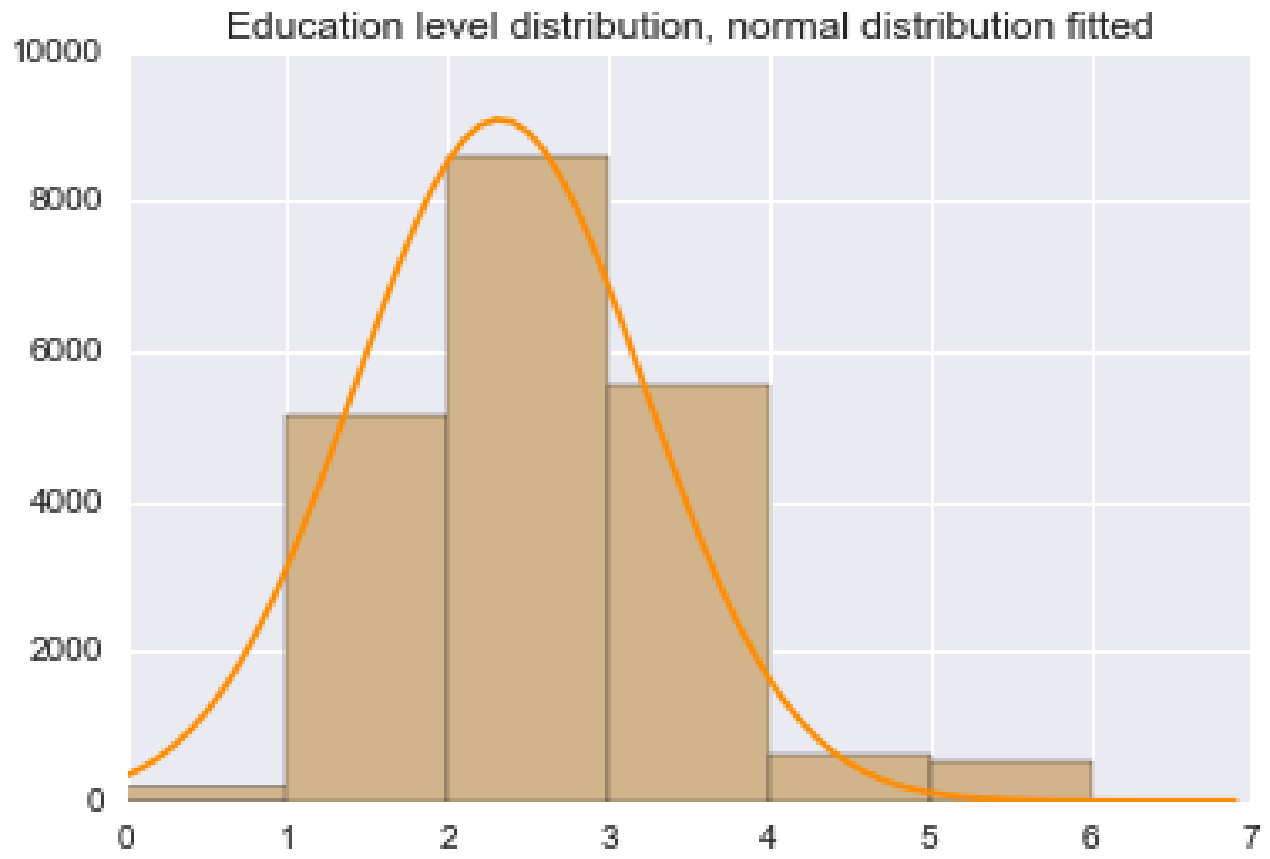


Fitted distributions: Profits

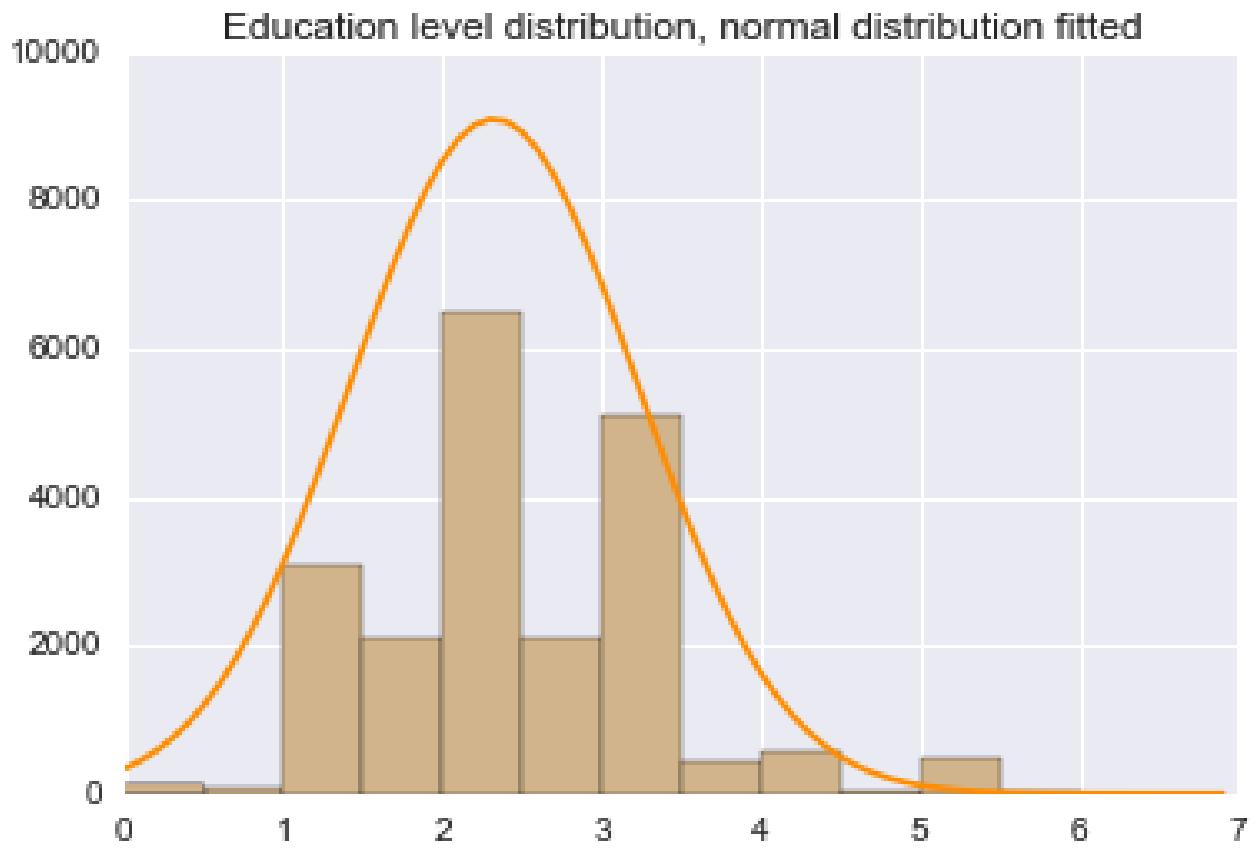




Fitted distributions: Average Worker Education by firm



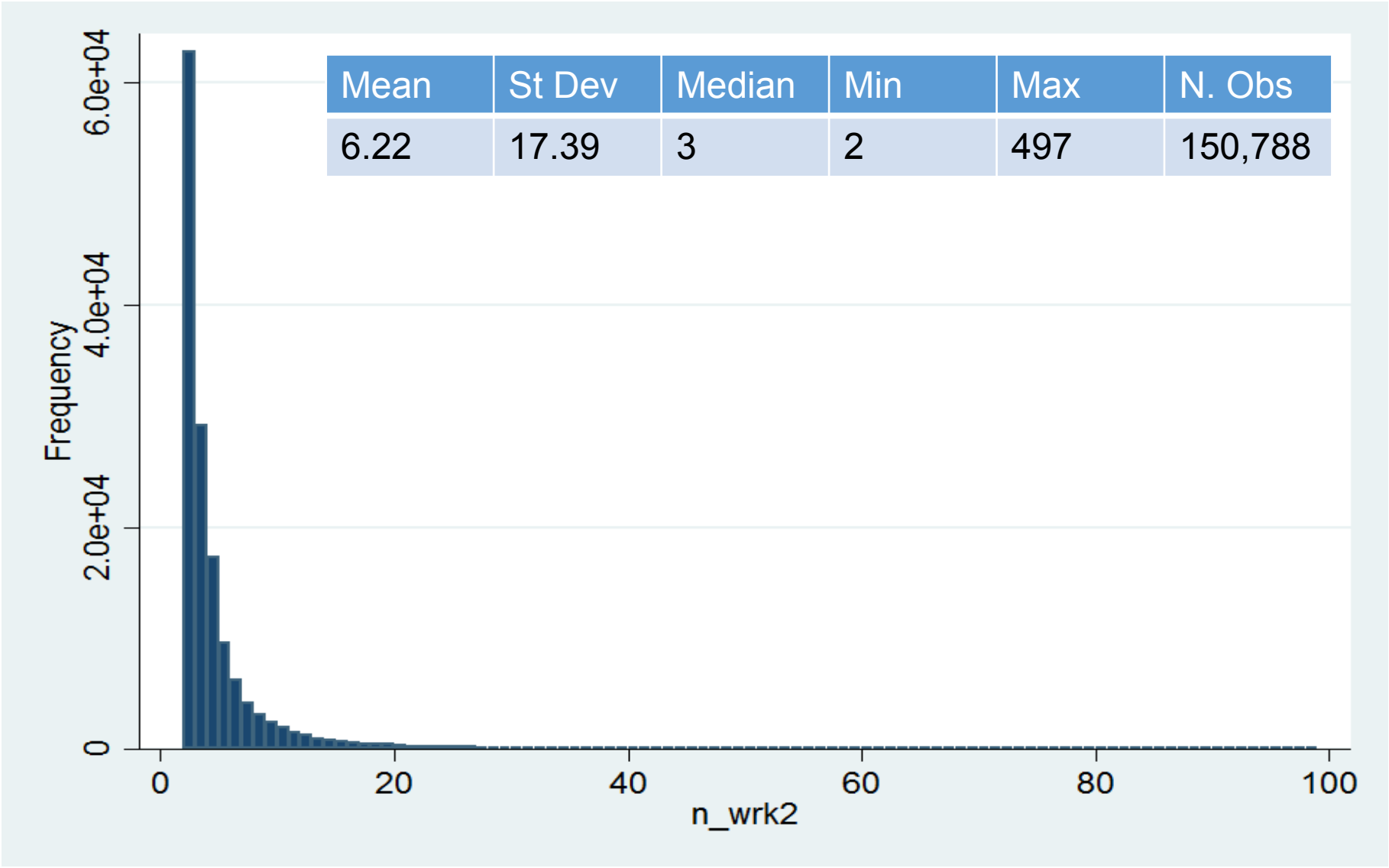
Fitted distributions: Average Worker Education by firm



Appendix: More Graphs

Number of observations per firm (>1)

Mean	St Dev	Median	Min	Max	N. Obs
6.22	17.39	3	2	497	150,788



Education Coding

0	Illiterate
1	Primary Education Completed
2	Secondary Education Completed
3	Pre-university Education (Bachillerato and equivalent) Completed
4	Short University Diploma (Diplomatura and Technical School equivalent)
5	Graduate (Licenciado)
6	Postgraduate