Using administrative data to investigate graduate earnings and beyond

Laura van der Erve*

October 2017

Joint work with Jack Britton (IFS), Lorraine Dearden (IFS & UCL), Neil Shephard (Harvard University), Anna Vignoles (University of Cambridge)

*Institute for Fiscal Studies

Introduction

- There is a rich literature on the power of big data, particularly administrative tax records
- Used to better understand earnings distributions of sub-populations (Chetty et al. (2014a,b))
- Advantages include:
  - Comprehensive coverage
  - Clearly defined income measures with low measurement error
  - Multiple years of data allowing tracking of individuals over time
  - Ability to break down to sub-populations
- Disadvantages include:
  - Lack of background characteristics
  - More difficult to get feedback on work!

Overview

- HMRC - SLC linked data
  - Data and linkage issues
  - Comparison of earnings in survey and admin data
  - Earnings differences by subject and institution
  - Differences by parental income
- Future work

Data I: Student Loan Company (SLC) Database

- 2.6 million students who borrowed from the SLC from 1998-2012
- English domiciled students only
- Data:
  - Higher Education Provider (HEP).
    - If <1000 loans grouped as ‘other’
  - Subject studied
  - Gender, region, cohort (first year of study)
  - Amount borrowed (first year and total)
  - Voluntary repayments
  - Flag for being abroad while still in repayment
- SLC loan take up of 85%-90%
  - Loan as proxy for being a graduate - misclassifies the $<$ 10% dropouts

Data II: UK tax data (HMRC databases)

- HMRC records:
  - Pay As You Earn (PAYE) - 10% sample
  - Self-assessment (SA) - all forms
  - Also observe gender, age and SIC code (employment sector)
- HMRC always treat SA records as definitive, we follow this
- Some issues with pre-2008 data - COP
- Highly restricted, we access anonymized data in a secure datalab

Data Linkage: “Golden Sample”

- Databases are hard-linked through NINO
- Can only link members of the SLC dataset also in the 10% NINO sample
    - Golden Sample members may have, each year: (i) no HMRC record, (ii) one or both of PAYE and SA
    - If (i) we set earnings equal to £0 - this mistreats those moving abroad (flag)
- Student characteristics at the **subject-institution level** merged in using HESA data:
    - Tariff (ATAR) scores
    - Ethnic mix
    - Various SES measures - POLAR, share living at home, share privately educated and average parental occupational class.

Golden Sample Summary Statistics I (2011/12 data)

	Male				Female			
	Golden	PAYE	SA	Either	Golden	PAYE	SA	Either
1998	6,927	5,528	1,351	5,875	7,560	6,118	959	6,351
1999	10,590	8,529	1,912	9,063	12,031	9,881	1,535	10,291
2000	10,853	8,761	1,908	9,322	12,653	10,453	1,517	10,854
2001	11,025	9,060	1,759	9,625	12,899	10,861	1,349	11,193
2002	11,060	9,156	1,576	9,642	12,831	10,948	1,238	11,264
2003	11,024	9,315	1,314	9,726	12,948	11,072	1,133	11,371
2004	10,767	9,163	1,251	9,526	12,810	11,204	1,015	11,471
2005	11,439	9,822	1,141	10,183	13,664	11,978	944	12,214
2006	11,340	9,749	992	10,024	14,043	12,400	872	12,565
2007	11,292	9,746	774	9,981	14,060	12,557	753	12,713
2008	8,990	7,704	531	7,872	11,857	10,450	508	10,558
2009	3,029	2,452	215	2,509	3,481	2,934	211	2,976
2010	1,334	1,082	72	1,101	1,659	1,395	80	1,410
2011	360	291		294	491	430		430
All	120k	100k	15k	105k	143k	123k	12k	126k

Golden Sample Summary Statistics II (2011/12 data)

		% No tax form			% Earnings = £0 (or no form)			% Earnings $<$ £8,000 (includes 0s & missings)		
Median age	Cohort	All	Male	Female	All	Male	Female	All	Male	Female
31	1998	13.0	12.6	13.3	15.6	15.2	16.0	27.3	26.7	27.9
30	1999	11.7	11.4	11.9	14.4	14.4	14.5	26.2	25.7	26.7
29	2000	11.4	11.2	11.5	14.2	14.1	14.2	26.1	25.7	26.5
28	2001	10.1	9.9	10.3	13.0	12.7	13.2	25.0	24.5	25.5
27	2002	9.6	9.9	9.3	12.5	12.8	12.2	25.3	25.5	25.0
26	2003	9.0	8.9	9.0	12.0	11.8	12.2	25.8	25.4	26.1
25	2004	8.0	8.3	7.7	10.9	11.5	10.5	25.9	26.8	25.2
24	2005	7.5	7.4	7.5	10.8	11.0	10.6	29.1	30.3	28.2
23	2006	7.5	7.8	7.2	11.0	11.6	10.5	34.3	36.3	32.6
22	2007	7.0	7.8	6.3	10.5	11.6	9.6	43.2	45.1	41.8
21	2008	8.4	9.1	7.8	11.6	12.4	11.0	61.6	63.2	60.4
21	2009	10.9	11.6	10.4	15.8	17.2	14.5	61.1	64.6	58.0
20	2010	11.0	12.0	10.2	16.1	17.5	15.0	67.9	72.0	64.6
18	2011	10.1	13.1	7.9	14.9	18.3	12.4	90.6	90.6	90.6

The “Silver Sample”

- We can also observe non-borrowers. This is anyone who is in the 10% NINO sample, but not in the SLC database
- Includes pre-1998 graduates and students who didn’t borrow from SLC: wealthy English and regular non-English UK students
- As before, could be in SA, PAYE or both
  - Unlike the Golden Sample, we have no way of identifying those who never file a tax receipt
- Only observe age, gender, income and occupational code
- Unlike Golden Sample, includes rest of UK other than England
- Sample to produce database with same age-profile as Golden Sample
- Sample quite large - randomly select from the population so we have 2 observations for every one GS observation

The “Silver Sample”: Summary stats

			% No tax form			% Earnings < £1			% Earnings < £8,000		
Med. age	LFS age	Cohort	All	Male	Fem	All	Male	Fem	All	Male	Fem
31	30-31	1998	22.1	21.5	23.0	27.3	26.7	27.9	46.3	43.3	49.9
30	29-30	1999	22.6	21.3	24.2	27.7	26.6	29.0	47.5	43.8	51.9
29	28-29	2000	23.5	21.8	25.5	28.5	27.0	30.4	48.8	45.2	53.2
28	27-28	2001	24.3	22.4	26.5	29.1	27.6	31.0	49.7	46.1	54.0
27	26-27	2002	24.8	23.1	26.8	29.7	28.3	31.4	51.2	47.9	55.1
26	25-26	2003	25.0	23.2	27.2	29.9	28.2	31.9	51.9	48.5	55.8
25	24-25	2004	24.9	22.7	27.5	30.1	28.1	32.5	52.9	49.8	56.6
24	23-24	2005	24.2	21.8	27.0	29.3	27.3	31.7	53.8	51.2	56.9
23	22-23	2006	23.7	21.4	26.4	29.0	26.9	31.4	55.8	53.4	58.6
22	21-22	2007	22.8	20.3	25.6	28.2	25.9	30.9	58.6	55.7	61.9
21	20-21	2008	21.6	19.4	24.1	27.8	25.4	30.5	61.6	59.0	64.5
21	20-21	2009	20.4	19.5	21.3	26.4	25.6	27.3	64.2	62.0	66.7
20	19-20	2010	18.4	17.1	19.9	24.4	23.1	25.8	68.8	66.0	71.8

Table: Silver Sample database for 2011-12. Shows percentage of individuals with no filed income tax form. Also shows numbers with no or low earnings.

The “NonHE Sample” - correcting the Silver Sample

- Correct for non-English UK and non-borrowing graduates:
  - $\omega =$ share of SS that went to HE. Roughly 14% for men and 21% for women.
  - By construction

$$F_{SS}(y) = \omega F_{HE}(y) + (1-\omega)\, F_{HE^c}(y), \quad \omega \in [0,1].$$

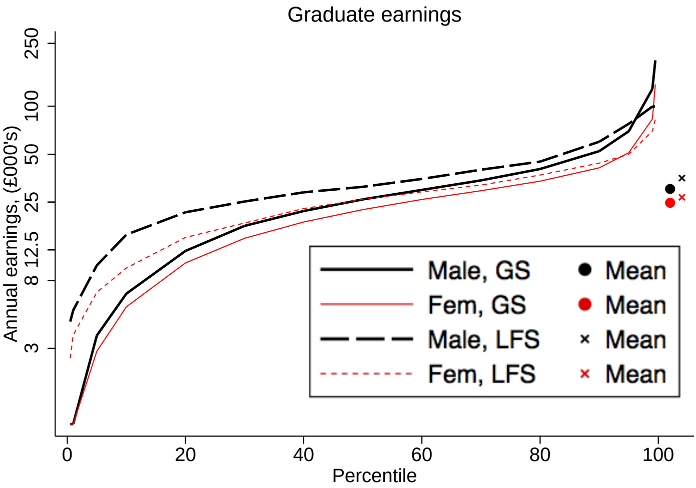
  - Assume

$$F_{HE}(y) = F_{GS}(y),$$

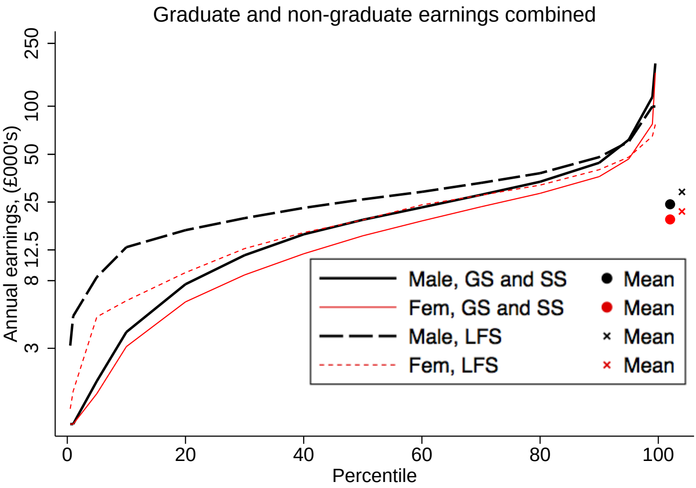
    where F_{GS} is the distribution function from the GS.
  - Then

$$F_{HE^c}(y) = \frac{F_{SS}(y) - \omega F_G(y)}{(1-\omega)}.$$

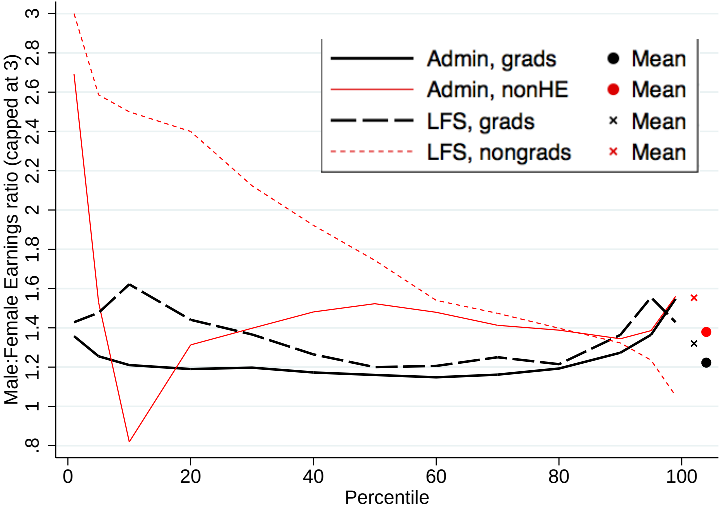
Earnings comparison I: LFS vs Golden Sample



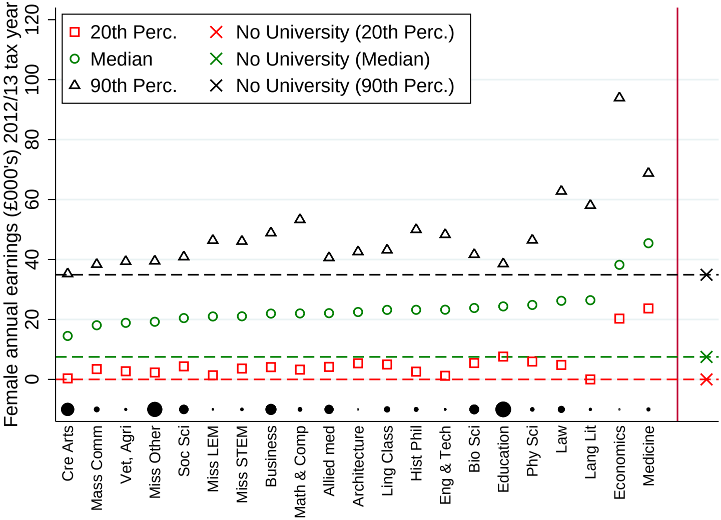
Earnings comparison II: LFS vs whole sample



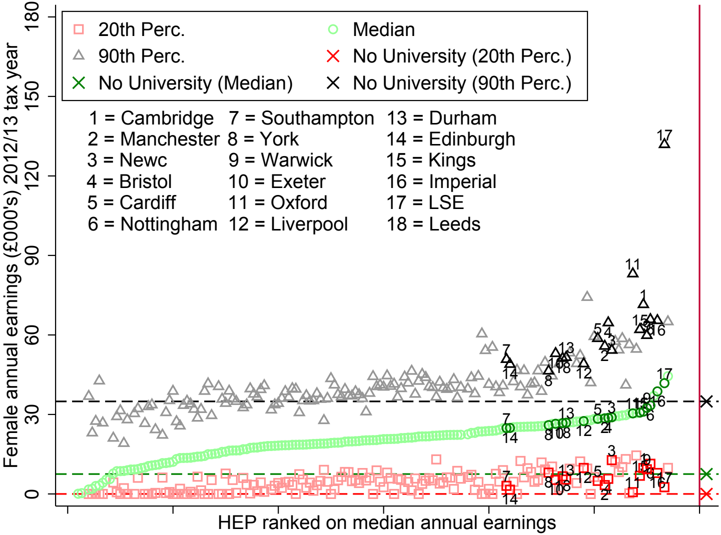
Implications of earnings differences I: Gender wage gap



Earnings differences by subject



Earnings differences by institutions



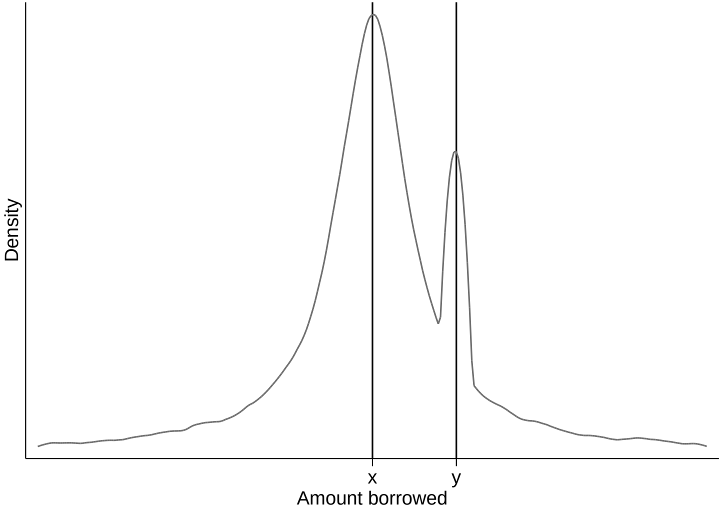
Identifying socio-economic background

- Infer a binary measure for parental income:
  - Individuals from high income households could borrow less from the SLC
  - Define as high income household if borrow rich maximum amount

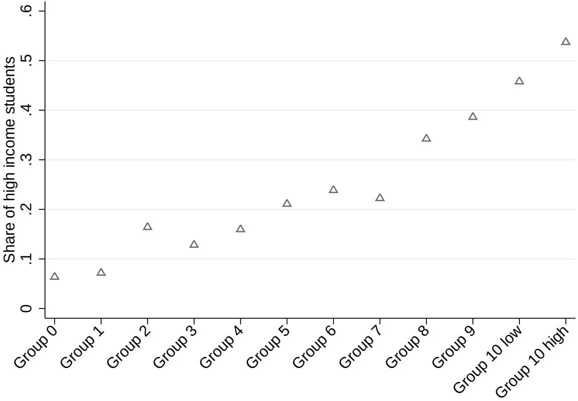
	Min Parental	Loan Amount		% borrowing = x		
Cohort	Income (£)	Non-London (£)	London (£)	Overall	Male	Female
1999	35,000	2,795	3,445	14.6	15.2	14.1
2000	36,000	2,795	3,445	18.9	20.2	17.8
2001	38,500	2,860	3,525	21.4	22.6	20.3
2002	40,000	2,930	3,610	21.8	23.2	20.5
2003	40,000	3,000	3,695	23.8	25.7	22.2
2004	40,950	3,070	3,790	24.8	26.1	23.6

- This is a blunt measure - those eligible for higher loans may borrow high-income maximum

Identifying socio-economic background



Identifying socio-economic background: Validation



Women

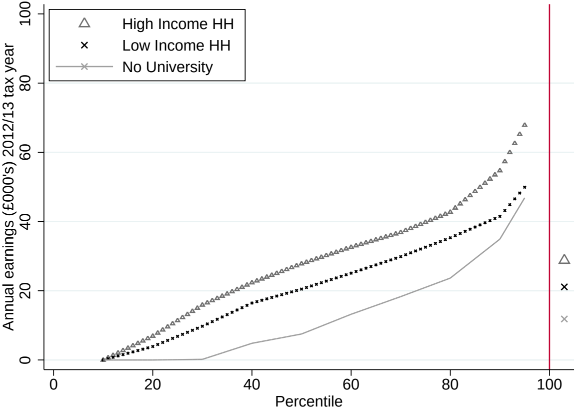
Figure: Share of individuals in each university group at the High Income HH borrowing amount. Includes the 1999-2005 cohorts. Shares incorporate borrowers only.

Identifying socio-economic background: Validation

	Repayments [1]	Add Gender [2]	Add Earnings [3]	Add HESA [4]
High Income HH	976.9***	958.5***	966.9***	614.5***
	(55.8)	(55.8)	(55.9)	(59.9)
Female		-384.6***	-395.0***	-334.2***
		(51.3)	(51.4)	(53.9)
Earnings			-0.004**	-0.008***
			(0.001)	(0.001)
Constant	2624.6***	2849.8***	2949.1***	2502.9***
	(72.5)	(78.4)	(85.5)	(720.3)
N	22,176	22,176	22,176	22,176
Adjusted R^2	0.067	0.069	0.070	0.095

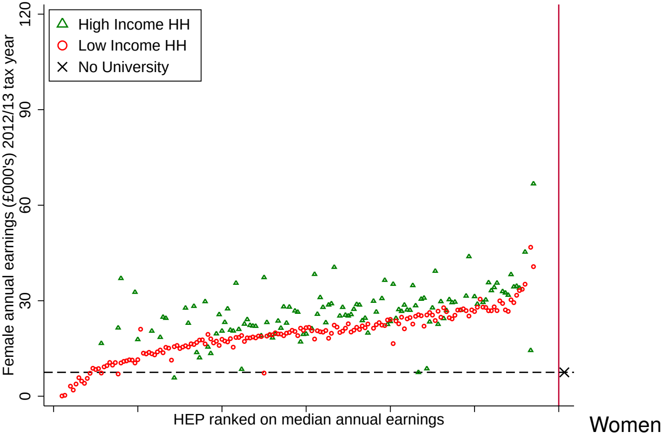
Table: Size of total voluntary repayments (£), conditional on making them. *** indicates significant at the 1% level; ** the 5% level. Female is a dummy set equal to one for women. Controls for cohort, age and year are included in all columns.

Earnings differences by socio-economic background



Women

Earnings differences by socio-economic background



Summary of findings

- Admin data shows lower mean earnings and greater proportion with very low earnings
  - This has important implications for empirical findings such as gender wage gap and earnings inequality
- Lots of variation in earnings by institution and subject
  - Medicine and economics very high earnings, creative arts very low
  - This has strong implication for where the government's HE subsidy is focussed
- Earnings of graduates from high income households considerably higher, even conditional on subject and institution
  - Unclear why - different occupational choices, better contacts, better non-cognitive skills

Future work I - Lifetime earnings

- In order to know student loan repayment we need to know lifetime income
- Earliest cohort (1998) earnings until age 33, less for later cohorts
- No HE indicator for pre-1998 cohorts, hence can't create synthetic cohorts
- Need to match HMRC earnings to modelled earnings from survey data for later years
- Developing new matching method to do so

Future work II - LEO data

- Higher Education Longitudinal Education Outcomes (LEO) dataset
  - NPD: prior attainment and background characteristics
  - HESA: subject, institution, level of degree, grade, dropout, POLAR
  - HMRC/DWP: employment, benefit receipt and earnings (PAYE; SA only for last 2 years)
- Can’t use NINO - perform fuzzy match based on name, DOB, postcode, gender
- NPD data available only if graduating after 2006/7

Future work II - LEO data

- *What is the causal return to subjects/institutions?*
- *How does this return vary by gender, ability or socio-economic background?*
- Data on prior attainment and individual/family characteristics allows us to control for selection and estimate causal return