Estimating human capital of graduates: the influence of institution, gender, year of entry into higher education and subject

Details of the project

The project is an academic project that has been designed to contribute to the existing economic evidence on graduate earnings. The work is entirely independent of the Government: the members of the research team are independent university researchers following their intellectual interests. To minimise conflicts of interest, the research team have not received any direct funding from the UK Government or any university mission group for this project.

What will the project do?

The project is a pilot study that will explore the use of linking tax data (from HMRC) and administrative data from the Student Loan Company to develop a model of the earnings of those who take out loans with SLC to undertake higher education (borrowers). The data, subject to some caveats, will allow researchers to observe how borrowers' earnings change through the years as they mature in the labour market. The aim of the research is to use these data to analyse the earnings of borrowers and contribute to the hitherto quite limited literature on the variation in graduate earnings by degree subject and institution. This is important from a social mobility perspective since graduates from different socio-economic backgrounds access different types of universities to a varying extent.

What are the potential benefits from the project?

The project seeks to improve our understanding of the earnings of different types of borrowers over a longer time span than has previously been possible. As such, the work has the potential to inform HMRC's modelling of graduate earnings. This work also has the potential to address some aspects of the issue of social mobility. Specifically, it has long been recognised that improving access to higher education is an important way one can potentially increase social mobility. However, access to higher education is, in and of itself, not necessary a good guide to a graduate's future earnings. It may be that the earnings profiles of graduates who experience different types of higher education (e.g. who study at different institutions) differ markedly. This research aims to produce a better understanding of the full distribution of graduate earnings by institution attended and subject area, and hence our understanding of the implications of improving access to higher education in terms of students' future earnings. The work also has the potential to inform our understanding of the value, management and design of the student loan system.

How will the research be conducted?

The research is innovative and the project is therefore necessarily a pilot study and great care will need to be taken in the analysis and interpretation of such complex data. The key issue that is being piloted is the practicalities of developing such a human capital measure using linked Student Loan Company and HMRC data. There are a number of methodological issues that need to be considered and the limits to both the data and the subsequent analysis will be explored and made clear by the research team.

Data protection

The identity of individuals must, of course, not be exposed when we analyse individual level data. The data we are linking is personal (e.g. the institution attended by a particular borrower) and sensitive (e.g. their earnings). Data confidentiality and protection is of paramount importance to the data owners, particularly HMRC and BIS, as well as the research team.

We have had extensive discussions with the data owners about how to deal with the confidentiality issues arising and concluded:

1. The analysis will be entirely carried out in the secure Datalab of the HMRC (http://www.hmrc.gov.uk/datalab/). This is a data enclave with significant data security

(e.g. the Datalab is a self-contained cluster of computers in a HMRC building, the computers have no external drives and no internet link). All of the researchers have been trained in the use of this secure environment and will be familiar with the data handling procedures required to protect individuals' confidentiality.

2. Team members who use this data have /will be subject to the same strict confidentiality and data protection requirements as HMRC staff and liable to legal penalties for breaches, including jail.

3. Any identifying material in the data will be anonymised. So although the data linkage will be undertaken using National Insurance numbers, these will not be revealed to our research team. Instead the researchers will see a scrambled National Insurance number, where the scrambling will be carried out by HMRC using their own in-house software.

4. Output from the Datalab is checked, line by line, by highly trained HMRC employees so that it is not disclosive.

Note that these data are only available for specific projects subject to approval by HMRC.

Data

The data for this project comes from two sources.

Student Loan Company

The (English) loan book is owned by the Secretary of State for Business, Innovation and Skills. This covers loans made to qualified English domiciled students who studied in UK universities. It does not cover students who were domiciled in Scotland, Wales or Northern Ireland.

The data would cover every student loan issued to students who started their studies from September 1998 onwards. Of course the data has no information about borrowers who did not borrow from the SLC. Typically these will be students who have alternative access to capital, generally from parents. This is an issue that that will be considered in the modelling.

The data is anonymised using the usual definitions (i.e. minimum cell count, excluding code, borrowing and voluntary repayments). An anonymisation protocol was agreed between SLC's chief statistician Dave Cartwright and Neil Shephard in the summer of 2011. This involves categorising ages and institutions where needed.

HMRC data

HMRC has tax year by tax year records of taxable earnings for each taxpayer (recall the tax year always ends on 5th April and starts on 6th April). The data we have available to us is:

1. A PAYE research database dating back to 2000/01. This is based on a simple random panel sample (selected through particular digits of the individual's national insurance number) of national insurance numbers which HMRC uses for research purposes. All the individuals who are in this database and who return pay as you earn (PAYE) forms for that year are then put in the PAYE research database for that year. We will see their aggregated PAYE annual data, recording their earnings and also their employment code (which tells us what sector of the economy they work in). The sample varies in size through time, starting as a 5% sample and moving up to 10% in more recent years.

2. Self-assessment (SA). A complete enumeration of the individuals who completed self-assessment records. This data goes back to 1998. SA forms in the UK are filled in by the self-employed, company directors, higher earners and those with more complicated forms of income.

Throughout our focus is on earnings through work, rather other forms of income.