Britton, Jack; Shephard, Neil G.; van der Erve, Laura

Working Paper

Econometrics of valuing income contingent student loans using administrative data: Groups of English students

IFS Working Papers, No. W19/04

Provided in Cooperation with:
Institute for Fiscal Studies (IFS), London

Suggested Citation: Britton, Jack; Shephard, Neil G.; van der Erve, Laura (2019) : Econometrics of valuing income contingent student loans using administrative data: Groups of English students, IFS Working Papers, No. W19/04, Institute for Fiscal Studies (IFS), London, https://doi.org/10.1920/wp.ifs.2019.0419

This Version is available at:
https://hdl.handle.net/10419/200323

Institute for Fiscal Studies

Econometrics of valuing income contingent student loans using administrative data: groups of English students

IFS Working Paper W19/04

Jack Britton

Neil Shephard

Laura van der Erve

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

Econometrics of valuing income contingent student loans using administrative data: groups of English students

Jack Britton
Institute for Fiscal Studies, London
jack_b@ifs.org.uk

Laura van der Erve
University College London and Institute for Fiscal Studies, London
laura_v@ifs.org.uk

Neil Shephard
Department of Economics and Department of Statistics, Harvard University
shephard@fas.harvard.edu

March 1, 2019

Abstract

Income contingent loans are an increasingly popular tool for funding higher education. These loans have desirable features, but also potentially high overall government write-offs in the long run. This latter fact has been well documented, but little is known about how those write-offs vary by subgroups of borrowers. It is important to quantify this and also to understand how it is affected by the design of the higher education system. Estimation is challenging because it requires the projection of earnings of graduates many years into the future. In this paper, we use English Student Loan Company records with information on borrowing, course, and institution linked to official tax records that give earnings for up to 11 years after graduation. Our innovative econometric methods fuse administrative tax records of graduates since they left university with graduate survey data to allow us to extrapolate through the life cycle for the remainder of the loan

contract. Our methodology is potentially applicable in a wide range of settings that uses incomplete administrative data. We estimate government subsidies through unpaid loans in England at the subject and institution level for the first time, finding considerable heterogeneity in both. England is an interesting case study due to the significant reforms to higher education over the past twenty years. We show that these reforms have had strong implications for the distribution of government spending on higher education.

Keywords: Administrative data; Income contingent loans; Higher Education funding; Income dynamics.

JEL: H52, H81, I22, I26, C81

Britton would like to thank the British Academy for funding through a postdoctoral grant and van der Erve gratefully acknowledges funding from the Administrative Data Research Centre England (ADRC-E). The data creators, depositors, copyright holders and funders bear no responsibility for the analysis, inferences, conclusions or interpretation of the data presented here. Responsibility for interpretation of the data, as well as for any errors, is the authors' alone.

1 Introduction

There is considerable debate about how Higher Education (HE) should be funded, and in particular how the burden should be shared between taxpayers and graduates.[1] Governments commonly provide student loans to ease credit constraints students face when attending university. Income contingent student loans (ICLs) are a particular type of student loan where the period-by-period repayments depend on the borrower's income after graduation, and the loan balances at the end of the repayment period are absorbed by the lender, not repaid (or defaulted on) by the borrower. In addition to easing credit constraints, ICLs reduce the earnings risk of an individual borrowing to fund their education by removing the possibility of defaulting on the loan. However, the implicit government subsidy, which is due to the state being the lender and then absorbing the write-offs from the loan balances at the end of the repayment period, can be very large and extraordinarily opaque. This opaqueness can be extremely damaging for policy makers as well as students (who may equate these loans to mortgage style debt or not understand the subsidies they are in effect receiving from the state) and taxpayers (who pay for the subsidies).

Income contingent student loans are increasingly being adopted by governments across the world: they are well established in Australia, New Zealand and the UK; they have recently been introduced in the Netherlands, Vietnam, Brazil, Colombia and Japan; and they are growing rapidly in the US, where around 75% of new borrowers are now on some type of income contingent loan plan. The size of these loans has become very large, making up a substantial fraction of public sector debt.

The broad applied contribution of this paper is to produce for the first time estimates of government subsidies through unpaid loans at the subject and institution level and to quantify how these subsidies change with changes to the higher education funding system. This is carried out using English Administrative data, but the application of the results is potentially broad.

Previous work has investigated overall loan subsidies and their distribution amongst English graduates based on their lifetime earnings (e.g. Dearden et al., 2008; Chowdry et al., 2012; Belfield et al., 2017), but that work has relied entirely on labour market survey data for the simulation of graduate earnings, which is insufficiently rich for the estimation of ICL costs at the major and institution level. This is important: the current lack of information on variation in government spending on different majors and institutions means that the full implications of various HE funding systems are poorly understood.

[1] While the private returns can be very large (see e.g. Blundell et al. (2000) and Kirkeboen et al. (2016)), there are also considerable social returns to higher education (e.g. Milligan et al. (2004), Lochner and Moretti (2004) and Moretti (2004)).

We investigate this problem in England, where the ICL scheme is run by the government and repayments are collected through the tax system. It provides loans for both tuition fees and living costs during study. We will make use of a dataset that hard links individual level tax data to a 10% sample of all ICL borrowers from the English Student Loan Company (SLC). England is an ideal case study because of data availability, the limited number of alternative borrowing sources, and because of the significant reforms to higher education in the past 20 years. We compare the system in 1999 when income contingent loans were small (around £16,000 on average) to the system in 2017, by which point annual tuition fees had increased to £9,250 for almost all courses and the average loans had consequently increased significantly to more than £50,000.

ICL subsidies are determined by loan sizes and graduate earnings. In England, there is very little variation in loan sizes across institutions and majors due to the homogeneity in university tuition fees. The large differences in graduate earnings between institutions and subjects (for example, see Britton et al., 2019) indicate the loan subsidy will be very unevenly distributed across subgroups. Given the very large subsidies implicit in many income contingent student loan systems - an estimated £8 billion per year in the UK context (Belfield et al., 2017) - knowing where this subsidy is targeted is of paramount importance to policy making. Similar issues exist in other institutional settings: for example, it has been well documented that earnings vary considerably by institution in the US (Chetty et al., 2017; Dale and Krueger, 2002) and by major in Norway (Kirkeboen et al., 2016) and Chile (Hastings et al., 2013).

The broad methodological contribution of this paper is to develop a new panel data framework for combining administrative records (Admin data) with survey data. We apply this strategy to the estimation of student loan subsidies, but the simulation of lifetime earnings paths based on a limited number of years of administrative data has many potential applications. For example, it is increasingly of interest to estimate the long-run effects of policies and also social returns to policies that affect lifetime earnings, such as early education interventions, expansion of higher education or increased access to on-the-job training. Our model is useful when there is a limited number of years of earnings available in the Admin dataset and researchers want to simulate forward based on what they have observed for each individual.

We use a unique administrative dataset built specifically for this purpose and documented in Britton et al. (2019). It links the English Student Loan Book to official tax records and covers the first 11 years of annual tax records after HE. We develop new methods which fuse the administrative data to survey data to which provide simulation of the rest of the life cycle. The resulting new 'Admin-

Survey fusing model' can value each person's loan, which allows us to value of the loans by gender, HE Institution (HEI) and major studied.

Developing individual level models of earnings over the life cycle is an important area of microeconometrics as many decisions and policy choices play out through the life cycle. Important recent contributions to the econometric literature include Kaplan and Violante (2010), Guvenen et al. (2015) and Arellano et al. (2017). Our contribution to this more general literature is to fuse administrative and survey data, which are both useful information sources about the life cycle. We hope our new methodology may have uses outside the scope of our own applications.

This paper will have two specific empirical applications of the methods we develop here. First we use the Britton et al. (2019) database of actual annual income tax records from 2002 to 2014 on individuals from the HE 1999 cohort (as defined by year of entry to HE) from the SLC to compute their repayments on the loans up to 2014. The SLC dataset is a random 10% sample of every English student who borrowed to attend higher education. Approximately 85% of the population borrows, and selection is not an problem as the population of borrowers is exactly the one we need to estimate the value of the loan book.

We combine this with the use of the Admin-Survey fusing model to extrapolate their likely payments for the remaining years of the Income Contingent loan contract. Taken together, our analysis provides a first published estimate of expected losses on the loans based on actual repayment data. We provide all estimates broken up by gender and groups such as subject studied and HE institution attended.

Second we use the tax data and survey data the fusing model to study the HE 2017 cohort. Members of the 2017 cohort will leave HE with, on average, much larger loan balances than the 1999 cohort. Using our fusing model we quantify how large the expected losses of this cohort are likely to be and how they vary by group and gender. We also use our model to study how these losses would vary with the tuition policy and various growth assumptions implicit within the Admin-Survey fusing model.

This second exercise is less novel than the first as it does not use any individual tax data about the actual 2017 cohort, and relies on stronger assumptions about the population of borrowers. However, it is the first report of results of a model based estimate whose model is deeply influenced by Admin data on graduates and the first report which breaks up these model based estimators by groups and gender. Further, this second exercise is important from a policy viewpoint as the loans sizes are much larger and the results could potentially be used to redesign policy going forward.

We find considerable heterogeneity in loan subsidy at both the major and institution level. The subsidy to many Science, Technology, Engineering and Maths (STEM) majors as well as subjects like Economics, Management and Law, are relatively low compared to that of cheaper to teach arts and humanities courses. This distribution of the subsidy seems at odds with the stated objectives of many governments to encourage students to study STEM majors.

Section 2 gives a formal definition of an Income contingent loan contract and discusses the broad econometric challenges in estimating these quantities. Section 3 defines the fusion model for survey and Admin data and discusses its implementation. Sections 4 and 5 apply the model to estimating the government loan subsidy to the cohorts entering HE in 1999 and 2017 respectively. Finally, Section 6 concludes. The econometric model used to drive the copula path as a component of the model and the particular survey and Admin data structures we use are discussed in the Appendix.

2 Income contingent contract

Here we will detail the income contingent loan contract, which will motivate the econometric developments we give in this paper. The rules of the contract will be derived from the English student loan scheme, but it will be expressed abstractly. Broadly, borrowers pay a fraction of their annual income above some threshold to the lender until the loan is repaid in full or the repayment period finishes. Lenders absorb losses on the loans. In the UK repayments are collected through the tax system and so borrowers never default.

We now formalize this, which includes the role of inflation and interest rates.

2.1 Notation using cash flows

Use $i \in \{1, 2, ..., n_g\}$ to label the i-th individual in the $g \in \{1, 2, ..., G\}$ group in year $t \in \{0, 1, 2, ...\}$ since the end of HE. Groups could be groups of students who study subjects like Medicine or Creative Arts, or attend institutions like Imperial College London.

It is very convenient to convert financial items from period t prices, using the time t price index $P_t^* = \prod_{s=1}^{t} (1 + i_s)$, where $P_0^* = 1$ and i_t is the t-th period inflation rate. Then the core notation for earnings, loan repayments, etc. are given in Table 1.

In period t	In cash terms	In period $t=0$ prices
Taxable annual earnings	$Y^*_{t,g,i}$	$Y_{t,g,i}$
Annual repayment on loan	$X^*_{t,g,i}$	$X_{t,g,i}$
Voluntary capital repayment	$Vol^*_{t,g,i}$	$Vol_{t,g,i}$
Loan balance at end of period t	$L^*_{t,g,i}$	$L_{t,g,i}$
Income contingent threshold	K^*_t	K_t
Interest payments of loan balance	$I^*_{t,g,i}$	$I_{t,g,i}$

Table 1: Notation for earnings $Y^*_{t,g,i}$, repayments $X^*_{t,g,i}$, loan balances $L^*_{t,g,i}$, etc through time for each group and individual. Deflating them by the price index, delivers the terms without superscipt stars.

Here real incomes are $Y_{t,g,i} = Y^*_{t,g,i}/P^*_t$, real repayments are $X_{t,g,i} = X^*_{t,g,i}/P^*_t$, real thresholds are $K_t = K^*_t/P^*_t$, real voluntary repayments are $Vol_{t,g,i} = Vol^*_{t,g,i}/P^*_t$, real balances are $L_{t,g,i} = L^*_{t,g,i}/P^*_t$ and real interest payments of loan balance $I_{t,g,i} = I^*_{t,g,i}/P^*_t$.

Finally, $L^*_{0,g,i}$ is the original loan amount, r is the real interest rate for the student loan contract. The loan is forgiven at the end of time period $T \geq 0$.

2.2 Contract definition

Here we define the type of income contingent loan we study in this paper.

	In cash terms	In period $t=0$ prices
Repayments	$X^*_{t,g,i} = \min\left\{\beta \max\left(Y^*_{t,g,i} - K^*_t, 0\right) + Vol^*_{t,g,i}, I^*_{t,g,i}\right\}$	$X_{t,g,i} = \min\left\{\beta \max\left(Y_{t,g,i} - K_t, 0\right) + Vol_{t,g,i}, I_{t,g,i}\right\}$
Loan balances	$L^*_{t,g,i} = I^*_{t,g,i} - X^*_{t,g,i}$	$L_{t,g,i} = I_{t,g,i} - X_{t,g,i}$
Threshold	$K^*_t = (1+i_t)(1+a)K^*_{t-1}$	$K_t = (1+a)K_{t-1}$
Interest pay	$I^*_{t,g,i} = (1+i_t)\left\{1 + r(Y^*_{t,g,i}, K^*_t)\right\} L^*_{t-1,g,i}$	$I_{t,g,i} = (1+i_t)\left\{1 + r(Y_{t,g,i}, K_t)\right\} L_{t-1,g,i}$

Table 2: Income contingent loan balances are defined by iterations. Here the main iterations are given. Key quantities are the time-t annual repayment $X^*_{t,g,i}$, loan balance $L^*_{t,g,i}$, payment threshold K^*_t and interest payment $I^*_{t,g,i}$. Also given are the recursions based on real prices.

Definition 1. *The income contingent loan contract is indexed by the payment rate β and a is the real growth rate in the threshold. We assume throughout that the real interest rate $r(y,k)$ only depends upon the level of interest and the repayment threshold and that $r(\gamma y, \gamma k) = r(y,k)$ for all $\gamma, k > 0$ and $y \geq 0$. Then Table 2 shows how $X^*_{t,g,i}$, $L^*_{t,g,i}$, K^*_t, $I^*_{t,g,i}$ are computed by iteration, together with the corresponding real terms. The loan is forgiven at the end of time period $T \geq 0$. In addition some countries augment income contingent loans with grants. We write the cost to the Government of this as $G_{g,i}$.*

Example 1. *In most Income Contingent Loan contracts $r(y,k) = r \geq 0$. However, from 2012 onwards, the UK uses an interest rate which also depends upon the level of earnings and payment*

threshold. Mathematically it can be expressed as $r(y,k) = r\min\left\{(\alpha k)^{-1}\max(y-k,0),1\right\}$*, where* $k,\alpha>0$ *and* $r\geq 0$*. Hence the interest rate is 0 if earnings are at or below the threshold* k *and linearly increase with earnings until they hit* $(\alpha+1)k$ *at which time the interest rate is* r*.*

Those not paying off the loan after T years are not regarded as having defaulted, instead they have fully complied with their income contingent contract and the lender writes off $L^*_{T,g,i}$.

2.3 Expected present value and estimands

A (risk neutral) expected present value at time t of the repayment stream $X_{(T^A+1),g,i},...,X_{T,g,i}$ from the i-th graduate over the remaining loan period will be

$$PV^*_{t,g,i} = P^*_t \sum_{s=t+1}^{T}\left(\frac{1}{1+d}\right)^{s-t}\mathrm{E}\left(X_{s,g,i}|\mathcal{F}_{t,g,i}\right), \tag{1}$$

where d is a real discount rate of the lender (e.g. the government's cost of borrowing) and $X_{s,g,i}|\mathcal{F}_{t,g,i}$ is a forecast distribution with $\mathcal{F}_{t,g,i}$ being the information we use to perform the calculation at time t about the i-th person in the g-th group. The information includes person i's past earnings and group g.

The expected net present loss on the loan, at time t, is $L^*_{t,g,i} - PV^*_{t,g,i}$.

Remark 1. *The UK public accounts call* $L^*_{t,g,i}-PV^*_{t,g,i}$ *the "Resource Allocation Budget" (RAB) charge and is often reported as a percentage for the group* $RAB_g = 100\sum_{i=1}^{n_g}\left(L^*_{0,g,i}-PV^*_{0,g,i}\right)/\sum_{i=1}^{n_g}L_{0,g,i}$*.*

To place these numbers in context we also compute summarizes of the career sum of discounted real earnings of an individual

$$\overline{Y}^*_{g,i} = \frac{1}{T'}P^*_t\sum_{s=1}^{T'}\left(\frac{1}{1+d}\right)^{s-t}Y_{s,g,i}. \tag{2}$$

Now $T'\overline{Y}^*_{g,i}$ is an economic measure of human capital in the spirit of, for example, Jorgenson and Fraumeni (1989) and Jorgenson and Fraumeni (1992). In addition we will compute the real present value at time t of career income tax from the i-th individual. This is implemented using a simple income tax schedule

$$TAX^*_{T,g,i} = P^*_t\sum_{s=1}^{T}\left(\frac{1}{1+d}\right)^{s-t}\left\{\sum_{j=1}^{\mathcal{G}}\beta^{\mathcal{T}}_j\max(Y_{s,g,i}-\mathcal{K}^{\mathcal{T}}_j,0)\right\}, \tag{3}$$

where $\left\{\mathcal{K}_j^{\mathcal{T}}, \beta_j^{\mathcal{T}}\right\}$ are the income tax real thresholds and the steps in the marginal income tax rates.

Example 2. *In 2017 the English loan scheme had* $\beta = 0.09$, $K_0 = £21,000$, $r = 0.03$, $T = 30$, $a = 0.02$. *The UK Government currently uses* $d = 0.007$ *in its public accounts in the relevant present value calculations. In the same year,* $\mathcal{G}$ =3, $\mathcal{K}_1^{\mathcal{T}} = £11,500$, $\mathcal{K}_2^{\mathcal{T}} = £45,000$ *and* $\mathcal{K}_3^{\mathcal{T}} = £150,000$ *and* $\beta_1^{\mathcal{T}} = 0.2$, $\beta_2^{\mathcal{T}} = 0.2$ *and* $\beta_3^{\mathcal{T}} = 0.05$.

Table 3 lists core estimands for this paper, all computed using the information available at time T^A.

Group g property	Estimand
Outstanding loan at time t	$\overline{L}_{t,g} = \frac{1}{n_g}\sum_{i=1}^{n_g} L_{t,g,i}$,
Average grant level	$\overline{G}_g = \frac{1}{n_g}\sum_{i=1}^{n_g} G_{g,i}$,
Share with outstanding loans at time t	$\overline{D}_{t,g} = \frac{1}{n_g}\sum_{i=1}^{n_g} 1_{L_{t,g,i}>0}$,
Average RAB charge	$RAB_g = \frac{\sum_{i=1}^{n_g}\left(L^*_{0,g,i} - PV^*_{0,g,i}\right)}{\sum_{i=1}^{n_g} L^*_{0,i,g}}$,
Share of student	$\omega_g = 100\frac{n_g}{\sum_{i=1}^{n_g} n_i}$
Share of loan subsidy	$S_g = 100\frac{\sum_{i=1}^{n_g} L_{T,g,i}}{\sum_{g=1}^{G}\sum_{i=1}^{n_g} L_{T,g,i}}$
Share of total subsidy	$S_g^+ = 100\frac{\sum_{i=1}^{n_g}\left(L_{T,g,i}+G_{g,i}\right)}{\sum_{g=1}^{G}\sum_{i=1}^{n_g}\left(L_{T,g,i}+G_{g,i}\right)}$
Quantiles of career average earnings	$Q_{0.25}(\overline{Y}^*_{g,i}), Q_{0.5}(\overline{Y}^*_{g,i}), Q_{0.75}(\overline{Y}^*_{g,i})$
Quantiles of career tax	$Q_{0.25}(TAX^*_{g,i}), Q_{0.5}(TAX^*_{g,i}), Q_{0.75}(TAX^*_{g,i})$

Table 3: Key estimands in this paper include average loan outstanding by time t, average grants for group g, average RAB charge, etc. Here $\overline{Y}^*_{g,i}$ is the career averaged earnings of the i-th individual within the g-th group.

Here the 0.25, 0.50 and 0.75 cross-sectional (over the n_g people in group g) quantiles are computed on career averaged earnings (i.e. economic measure of human capital) and individual career tax takes.

3 Econometrics of Admin-survey data fusion for group g

3.1 Challenges of income contingent contracts

Our Admin data only goes up to those aged T^A, so we observe earnings data up to time T^A

$$\mathbf{Y}_{1:T^A,g,i} = (Y_{1,g,i}, ..., Y_{T^A,g,i}), \tag{4}$$

as well as voluntary repayments $\mathbf{Vol}_{1:T^A,g,i} = (Vol_{1,g,i}, ..., Vol_{T^A,g,i})'$. We need to extrapolate these into the future to cover times $T^A + 1$ up to time T.

Although $PV^*_{t,g,i}$ is an expectation, the repayments $\mathbf{X}_{(T^A+1):T,g,i} = (X_{(T^A+1),g,i}, ..., X_{T,g,i})'$ are functionally a stream of payoffs from a basket of path dependent real options (e.g. Hull (2017), Dixit and Pindyck (1994) and Cochrane (2005)) written on the discrete time "underlying" $\mathbf{Y}_{(T^A+1):T,g,i} = (Y_{(T^A+1),g,i}, ..., Y_{T,g,i})'$ and $\mathbf{V}_{(T^A+1):T,g,i} = (V_{(T^A+1),g,i}, ..., V_{T,g,i})'$. Thus the valuation depends upon the entire conditional distribution of the future earnings path $\mathbf{Y}_{1:T,g,i}$ and voluntary repayments $\mathbf{V}_{1:T,g,i}$. This is also true of $LOSS^*_{t,g,i}$, $RAB_{g,i}$ and $FULL_{t,g,i}$, while $TAX^*_{g,i}$ and $Y^*_{g,i}$ depends not on the dynamics of the earnings path but solely on the properties of the cross-sectional marginal distribution of earnings at each time t. In practice, for English loans, voluntary repayments are quite modest so we will record $\mathbf{V}_{1:T^A,g,i}$ and focus on the joint distribution of

$$\mathbf{Y}_{(T^A+1):T,g,i}|\mathcal{F}_{T^A,g,i}. \tag{5}$$

The econometrics of model building for the path of earnings $\mathbf{Y}_{(T^A+1):T,g,i}|\mathcal{F}_{T^A,g,i}$ is the focus of, for example, Kaplan and Violante (2010), Guvenen et al. (2015) and Arellano et al. (2017). Traditionally modelling the joint distribution of $\mathbf{Y}_{(T^A+1):T,g,i}|\mathcal{F}_{T^A,g,i}$ has been carried out just using labour market survey data using relatively conventional panel models from the econometric literature. Arellano et al. (2017) is a move away from this, using a quantile based model on the assumed Markovian dynamics. Guvenen et al. (2015) use panels of Admin data to study earnings dynamics in the US, using a very flexible empirical model. Prominent examples of the use of traditional econometric survey models in the context of graduate earnings paths include Dearden et al. (2008), Chowdry et al. (2012) and Belfield et al. (2017) where the conditioning variable $\mathcal{F}_{0,g,i}$ in these studies is just gender.

Recent research by Britton et al. (2019) shows that Admin data from UK tax collection records for graduates are somewhat different than that recorded for UK survey data. Further, the Admin data has additional information not available in UK survey studies, such as HEI attended and subject studied. It is important to understand how the valuations varies over these groups. Policy makes may wish to (i) sell particular parts of the loan book, (ii) change the fee structure to have fee levels varying across institutions or subjects.

All told it is vital that the econometrics behind loan valuing is modernized to exploit the relevant Admin data, giving a deeper analysis of loan values. Unfortunately the survey data cannot be entirely discarded. The Admin data only covers around the first 1/3 of the loan period. The only information we have about the remaining parts of the lifecycle of graduates is through surveys.

This gives an urgency in developing the econometrics of fusing survey data into Admin data to complete the lifecycle. This is not trivial, as we care about the entire distribution of the path

$\mathbf{Y}_{(T^A+1):T,g,i}|\mathcal{F}_{T^A,g,i}$, whose dimension is quite large and the levels of earnings seen in survey data is often higher than we see in Admin data.

We develop a first econometric model of earnings paths for this data fusion of survey and Admin data, using the Admin data whenever possible and filling in the gaps with survey data. We then use it to produce valuations of individual loans using information on past earnings, group and gender.

3.2 Groups and data sources

To start we establish some notation. We first jitter each Admin data point by adding an independent standard uniform random variable to the earning number (UK tax forms leave off pennies to earnings, so adding standard uniform noise is a form of imputation, but also has the desired effect of allowing us to uniquely cross-sectionally rank the i-th person's earnings). The resulting raw Admin data will be written as $Y_{t,g,i}$. This raw Admin data is sometimes transformed through

$$v_{t,g,i} = \Phi^{-1}\left(\frac{rank_{t,g}(Y_{t,g,i}) - 1/2}{n_g}\right), \quad i = 1, 2, ..., n_g, \quad t = 1, 2, ..., T^A, \tag{6}$$

where $rank_{t,g}(Y_{t,g,i})$ is the rank of the i-th individual in the g-th group at time t (e.g. the rank of earnings within Cambridge University in year t). Here Φ^{-1} is the quantile function of a standard normal random variable.

To form forecasts from the Admin data through the lifecycle, we will separately model the year by year cross-sectional distributions by group and model the copula dependence between years through a high dimensional microeconometric model which was estimated using many waves of survey data.

The next subsection deals with the margins, while subsection 3.4 tackles the copula. Subsection 3.4.2 and 3.5 are devoted to forecasting, while subsection 3.7 details how we implement these methods in practice.

3.3 Marginal distributions via the fusing model

First consider the cross-sectional marginal distributions separately at each time point. The inputs into the Fusing model are:

- Group g Admin quantities will be written as

$$\pi_{T^A|g}, \quad Q^*_{T^A|g}(q), \quad q \in [0, 1), \tag{7}$$

where $\pi_{T^A|g}$ is the proportion of zero (or very close to zero) earners and $Q^*_{T^A|g}(q)$ is the q-quantile of non-zero earners. We define zero earners as those with earnings which are less than £2,000.

- Survey data quantities will be written as

$$\pi^S_t, \quad Q^{*S}_t(q), \quad t = 1, ..., T, \quad q \in [0, \overline{q}), \quad \overline{q} \leq 1, \tag{8}$$

where π^S_t is the proportion of zero (or very close to zero) earners and $Q^{*S}_t(q)$ is the q-quantile of non-zero earners. Here $1 - \overline{q}$ is the fraction of the data which is top coded in the survey data. The left hand side of Figure 1 shows π^S_t and $Q^{*S}_t(q)$ for the survey data for female graduate earnings, while the right hand side shows the corresponding results for men.

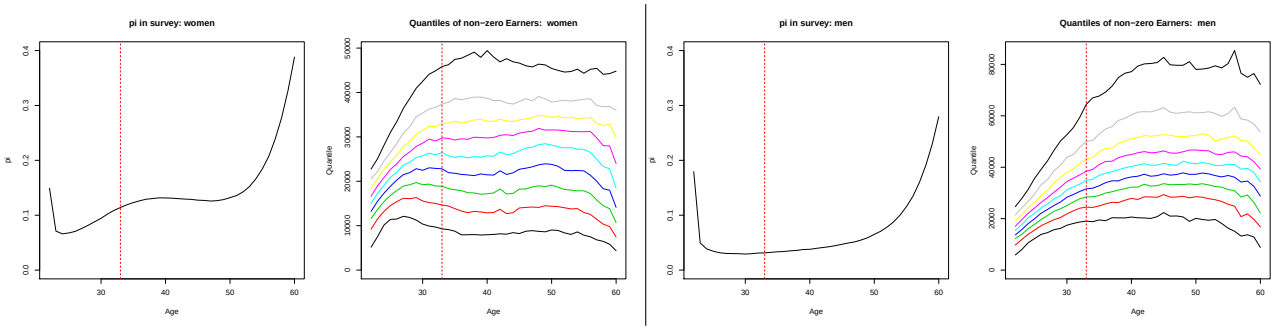


Figure 1: From the survey data. LHS: estimates of π^S_t and $Q^{*S}_t(q)$ plotted against time for women. For we plot quantiles for $q = 0.1, 0.2, ..., 0.9$. RHS: corresponding results for men. Source: `Basic.r`.

3.3.1 Admin-Survey fusing skeleton

To deal with the problem of having no Admin data beyond time T^A we extrapolate the quantiles of the non-zero earners in the Admin data using the growth rates of the quantiles of non-zero earners in the survey data. The percentage of zero earners is extrapolated by using the change in the rate for the survey data.

The core marginal distributions of this extrapolation are carried out using the "Admin-Survey Fusing Skeleton" . It is the introduction of this model which is the sole econometric innovation in this paper. It is needed to join the Admin and Survey data together in an environment where the average earnings levels in the Survey and Admin data are importantly different.

Of course it will need a copula path to complete a probabilistic "Admin-Survey Fusing Model" for

$$\mathbf{Y}_{(T^A+1):T,g,i} | \mathcal{F}_{T^A,g,i}. \tag{9}$$

We will turn to the copula in the next section.

Definition 2 (Admin-Survey Fusing Skeleton)**.** *For time $t = T^A + 1, ..., T$ and group g, we define for $q \in [0, 1]$, the skeleton as*

$$Q_{t|T^A,g}(q) = \begin{cases} 0, & q \leq \pi_{t|T^A,g} \\ Q^*_{t|T^A,g}\left(\frac{q-\pi_{t|T^A,g}}{1-\pi_{t|T^A,g}}\right), & 1 > q > \pi_{t|T^A,g}. \end{cases} \tag{10}$$

Here

$$\pi_{t|T^A,g} = \min\left\{\max\left(\pi_{T^A|g} + \pi^S_t - \pi^S_{T^A}, 0\right), 1\right\}, \tag{11}$$

is the predictive model of the probability of zero earners, while the predictive quantile function model of non-zero earners is

$$Q^*_{t|T^A,g}(q) = \frac{Q^{*S}_t(q \wedge \overline{q})}{Q^{*S}_{T^A}(q \wedge \overline{q})} Q^*_{T^A|g}(q), \quad \textit{recalling } q \wedge \overline{q} = \min(q, \overline{q}). \tag{12}$$

The equation (12) says that future q-quantiles in the Admin data for non-zero earners grow at the rate of growth of the q-quantiles in the survey data for non-zero earners. It would have been beneficial to have been able to use group growth rates but we have no significant quantity of survey data with group information.

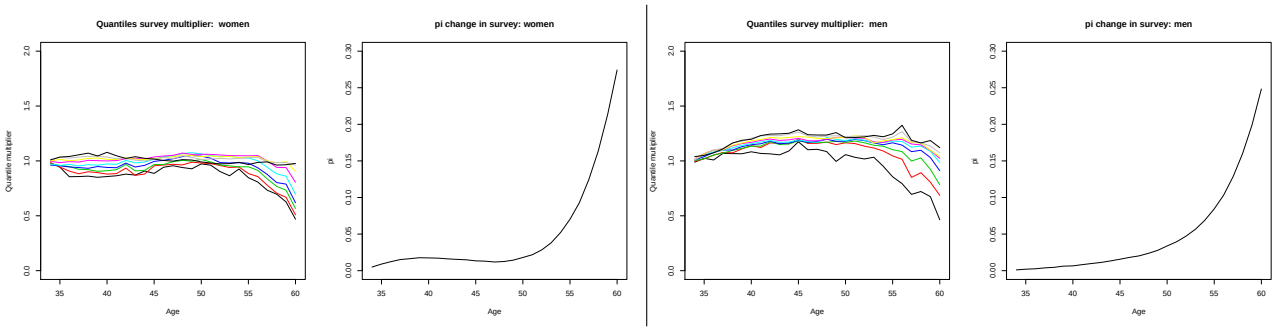
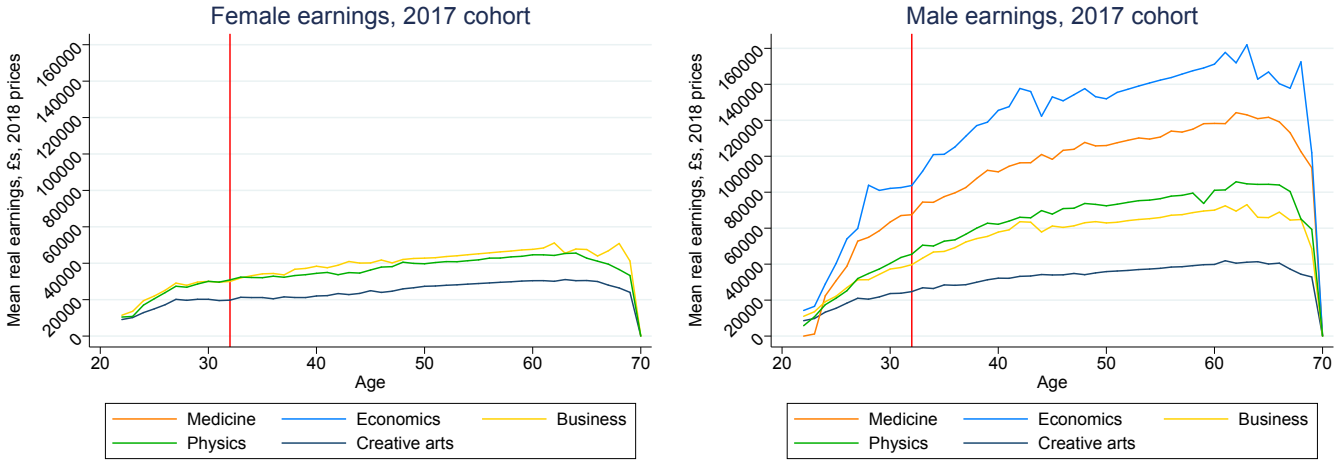


Figure 2: LHS: from the survey data estimates of the multiplier $Q^{*S}_t(q)/Q^{*S}_{T^A}(q)$ and shift $\pi^S_t - \pi^S_{T^A}$ plotted against time for women for $t = T^A+1, ..., T$ and for $q = 0.1, ..., 0.9$. RHS: corresponding results for men. The graphs show quantiles falling as graduates approach 60 as does the share of earnings which are close to zero.

Figure 2 show survey estimates of the multiplier $Q^{*S}_t(q)/Q^{*S}_{T^A}(q)$ and shift $\pi^S_t - \pi^S_{T^A}$ plotted against time $t = T^A + 1, ..., T$, for women and for men. The exact details of how this is implemented will be given in Section 5. Further, the Admin-Survey Fusing Skeleton is illustrated in Figure 3, which shows $Q_{t|T^A,g}(1/2)$ (median earners) for 18 year old individuals who entered HE in 2017 using 2018 prices,

with five groups split up according to gender. Up to age 32, where $T^A = 13$, the results are from past administrative data $Q_{1|g}(1/2), ..., Q_{T^A|g}(1/2)$, the rest are extrapolated using survey data through the fusing skeleton $Q_{t|T^A,g}(1/2)$. Again the implementation details will be discussed shortly.



Women Men

Figure 3: Admin-Survey fusing skeleton $Q_{t|T^A,g}(q)$ for $q = 0.5$: Mean earnings for female (left) and male (right) borrowers, assuming they entered university at age 18. Earnings are in 2018 prices. Earnings profiles based on admin data increased with earnings growth up to age 32 and using our admin-survey fusion model thereafter. Earnings profiles are not shown when sample sizes are small and so risk disclosure.

3.4 Copula path for i-th individual

3.4.1 Copula construction

To complete the Admin-Survey Fusion Model we need to augment the Skeleton in Definition 2 with a copula for earnings. Here we will build copulas entirely separately for the two genders.

Copulas have been used before for models of earnings, such as Bonhomme and Robin (2006) who employed parametric statistical copulas. Our fusion model will be generally agnostic about the particular form of the copula. Here we implement one in our applied work using a survey based econometric model which we will convert into a copula of earnings. We will refer to this as the "IFS graduate earnings model". Previous papers which have used model implied copulas include Smith and Maneesoonthorn (2018) and Loaiza-Maya and Smith (2019). Work which uses copulas to model serial dependence include pp. 280 of Joe (1997), Lambert and Vandenhende (2002), Frees and Wang (2006), Beare (2010) and Smith et al. (2010). An important area of work is copula modelling for partially discrete data, which is relevant for us as a substantial amount of our data has zero earnings.

A review of that literature is provided by Genest and Nelehov (2007).

Our approach is to take R simulations, where R is vast, from the IFS graduate earnings model, which we will write as $Y_{t,i}^M$, $i = 1, ..., R$ and $t = 1, ..., T$. Then we jitter each survey data point by adding standard uniform noise

$$Y_{t,i}^S = Y_{t,i}^M + U_{t,i}, \quad U_{t,i} \overset{iid}{\sim} U(0,1), \quad i = 1, 2, ..., R,$$

so $Y_{t,i}^S$ has a unique rank. We use these simulations to estimate

$$\pi_t^S, \quad Q_t^{*S}(q), \quad t = T^A, T^A + 1, ..., T, \quad q \in [0, \overline{q}]. \tag{13}$$

As R is massive there should be only trivial estimation error in computing the terms in (13). They should be regarded as population quantities implied by the "IFS graduate earnings model."

Then we compute

$$V_{t,i} = \Phi^{-1}\left(\frac{rank_t(Y_{t,i}^S) - 1/2}{R}\right), \quad i = 1, 2, ..., R,$$

where $rank_t$ denotes the cross-sectional rank at time t. $V_{t,i}$ is Gaussian over the cross-section i, although there is no reason to think that the path

$$\underset{T\times 1}{\mathbf{V}_i} = (V_{1,i}, ..., V_{T,i})',$$

is jointly Gaussian. We now compute

$$\widehat{\rho}_{t,s} = \frac{\sum_{i=1}^R \left(V_{t,i} - \overline{V}_t\right)\left(V_{s,i} - \overline{V}_s\right)}{\sqrt{\sum_{i=1}^R \left(V_{t,i} - \overline{V}_t\right)^2 \sum_{i=1}^R \left(V_{s,i} - \overline{V}_s\right)}}, \quad t, s \in \{1, 2, ..., T\}.$$

In practice $\overline{V}_t$ will be tiny by construction, while again as R is enormous $\widehat{\rho}_{t,s}$ should again be regarded as the population quantities implied by the "IFS graduate earnings model." Write the matrix of correlations as $\widehat{\rho}$, with t, s-th entry $\widehat{\rho}_{t,s}$.

Figure 4 plots $\{\widehat{\rho}_{t,s}\}$ for the IFS graduate earnings model for each value of $t > s$. The lowest black line shows $\widehat{\rho}_{t,1}$ plotted against t. It does not approach zero as t increases, presumably because of the impact of individual effects on earnings. As s increases the correlations, as a function of t, shift upwards — showing the increase in serial dependence in earnings with age. The left hand side shows the results for women. The right hand side shows the corresponding results for men.

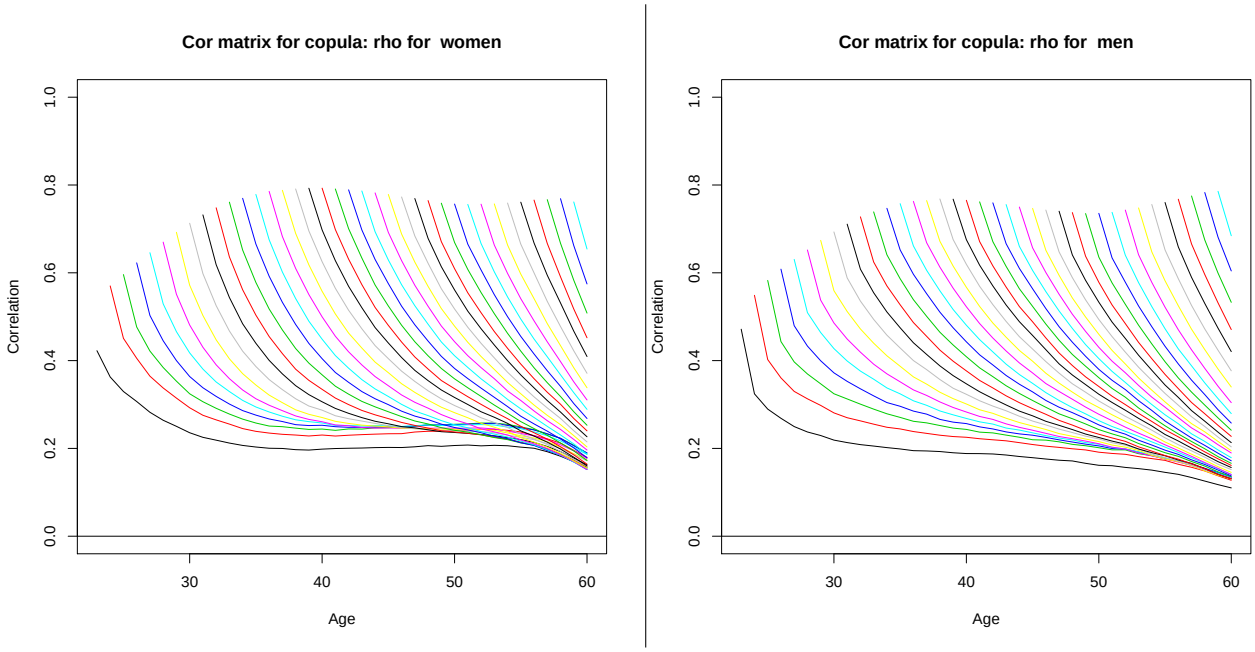


Figure 4: Measure of dependence in the Survey earnings model. LHS: the (cross-sectional) correlation $\widehat{\rho}_{t,s}$ between transformed earnings at times s and t for women. RHS: shows the corresponding result $\widehat{\rho}_{t,s}$ for men. Source: `Basic.r`.

3.4.2 Copula based forecasts

To add some flexibility, we now introduce a parameterized correlation model

$$\underset{T\times T}{\Psi} = \theta\widehat{\rho} + (1-\theta)\iota\iota', \quad \theta \in (0,1],$$

where ι is a $T \times 1$ vector of ones. As θ falls then the element by element dependence in Ψ rises monotonically. We will select θ using statistical methods later in Section 3.6 based on features of the Admin data.

Write the blocks of Ψ in the usual way

$$\underset{T\times T}{\Psi} = \begin{pmatrix} \Psi_{1:T^A,1:T^A} & \Psi_{1:T^A,T^A+1:T} \\ \Psi_{T^A+1:T,1:T^A} & \Psi_{T^A+1:T,T^A+1:T} \end{pmatrix},$$

while partitioning the $\mathbf{V}_i = \left(\mathbf{V}'_{1:T^A,i}, \mathbf{V}'_{(T^A+1):T,i}\right)'$.

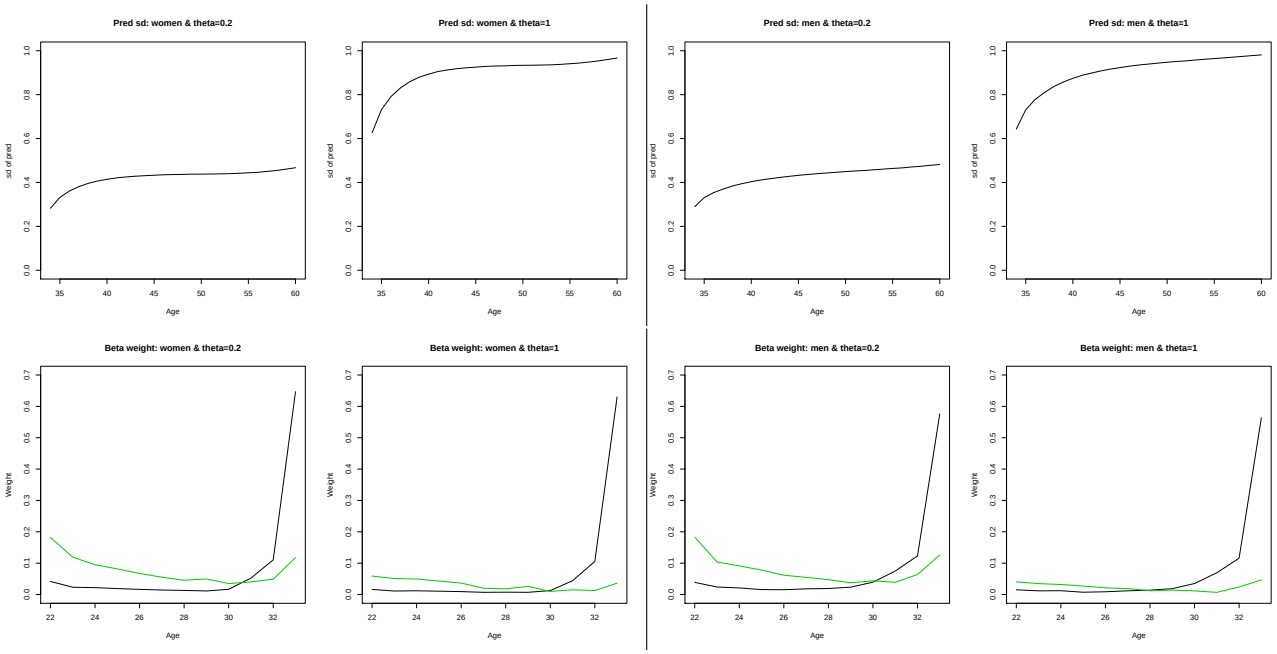


Figure 5: Top row: Measure of predictive uncertainty in Σ: square root of diagonal elements of Σ. LHS: results for women, first for $\theta = 0.2$ then for $\theta = 1$. RHS: shows the corresponding results for men. Bottom row: LHS: first (black line) and last row (red line) of β weights for $\theta = 1$ and $\theta = 0.2$ for women. β. RHS: shows the corresponding results for men. Source: `Basic.r`.

If Ψ is positive definite and $\mathbf{V}_i \sim N(0, \Psi)$, then

$$\mathbf{V}_{(T^A+1):T,i} | \mathbf{V}_{1:T^A,i} \sim N(\beta \mathbf{V}_{1:T^A,i}, \Sigma),$$

where

$$\begin{aligned} \underset{(T-T^A)\times T^A}{\beta} &= \Psi_{T^A+1:T,1:T^A} \Psi^{-1}_{1:T^A,1:T^A} \\ \underset{(T-T^A)\times(T-T^A)}{\Sigma} &= \Psi_{T^A+1:T,T^A+1:T} - \Psi_{T^A+1:T,1:T^A} \Psi^{-1}_{1:T^A,1:T^A} \Psi_{1:T^A,T^A+1:T}. \end{aligned}$$

The top row of Figure 5 shows the square root of the diagonal elements of Σ for $T^A = 11$ using $\theta = 1$ and $\theta = 0.2$. This shows the uncertainty in the forecast of futures values in the series.

The bottom row of Figure 5 shows the coefficients of the first and the last rows of β matrix for $T^A = 11$ using $\theta = 1$ and $\theta = 0.2$. Again, the left hand side results are for women.

3.5 Overall forecasting

For the group g, we use the gendered copula to form a predictive sample, which is

$$\mathbf{V}_{(T^A+1):T,g,i} | \left\{ \mathbf{V}_{1:T^A,i} = v_{1:T^A,g,i} \right\} \sim N(\beta v_{1:T^A,g,i}, \Sigma). \tag{14}$$

Recall $v_{1:T^A,g,i}$ is the transformed Admin data from group g.

We then introduce a parameter η which allows for additional idiosyncratic noise in the probability of becoming unemployment by letting

$$q_{t,i} = 1_{e_{t,i} \leq \eta} \eta e_{t,i} + 1_{e_{t,i} > \eta} \left\{ \eta + (1-\eta) V^*_{t,i} \right\}, \quad e_{t,i} \overset{iid}{\sim} U(0,1), \quad \eta \in [0,1]. \tag{15}$$

Then by construction $q_{t,i} \overset{iid}{\sim} U(0,1)$ over i.

Then we have a simulated future earnings

$$Y_{t,g,i} = Q_{t|T^A,g}(q_{t,g,i}), \quad q_{t,g,i} = \Phi(V_{t,g,i}), \quad t = T^A + 1, ..., T. \tag{16}$$

Hence a single earnings path for the i-th person is given by

$$Y_{1,g,i}, ..., Y_{T^A,g,i}, Y_{T^A+1,g,i}, ..., Y_{T,g,i}. \tag{17}$$

3.6 Selecting θ and η

The conditional (forecast) distribution of $Y_{T^A+1,g,i}, ..., Y_{T,g,i} | Y_{1,g,i}, ..., Y_{T^A,g,i}$ is indexed by $\widehat{\rho}$ (determined by the IFS graduate earnings model), θ (which impacts Ψ) and η. Here we provide empirical methods for selecting θ and η. As usual we selecting them separately by gender.

As earnings are more volatile early in the career, we focus on the last couple of years of earnings Admin data. We measure the quantiles of the cross-sectional difference in the normalized ranks in the Admin data at time T^A from the normalized ranks at time $T^A - 1$.

$$Q^{\Delta,q}_{T^A} = Q_q \left\{ \frac{rank_{T^A}(Y_{T^A,i}) - 1/2}{n} - \frac{rank_{T^A-1}(Y_{T^A-1,i}) - 1/2}{n} \right\}$$

We denote the q-quantile in short hand as $Q^{\Delta,q}_{T^A}$.

Now do the same operation for the future simulated normalized rank for time T^A+2 compared to time T^A+1. This is based on the path (17). This is computed as

$$Q^{\Delta,q}_{T^A+2} = Q_q\left\{\frac{rank_{T^A+2}(Y_{T^A+2,i})-1/2}{n} - \frac{rank_{T^A+1}(Y_{T^A+1,i})-1/2}{n}\right\}.$$

We choose to do the comparison between times T^A+2 and T^A+1 so it is entirely based on simulated future earnings.

We select θ to roughly match quantiles $Q^{\Delta,q}_{T^A}$ with $Q^{\Delta,q}_{T^A+2}$. We have implemented this using the error criterion

$$\begin{aligned} C(\theta) = \quad & \Big\{(Q^{\Delta,.1}_{T^A} - Q^{\Delta,.1}_{T^A+2})^2 + (Q^{\Delta,.25}_{T^A} - Q^{\Delta,.25}_{T^A+2})^2 + (Q^{\Delta,.5}_{T^A} - Q^{\Delta,.5}_{T^A+2})^2 \\ & + (Q^{\Delta,.75}_{T^A} - Q^{\Delta,.75}_{T^A+2})^2 + (Q^{\Delta,.9}_{T^A} - Q^{\Delta,.9}_{T^A+2})^2\Big\}^{1/2}. \end{aligned}$$

In our Admin data we selected $\theta = 0.05$ for men and $\theta = 0.035$ for women minimize this error criterion. This puts a great deal of weight on the individual effect within the model for earnings, much more than is seen in panel based econometrics models which are implemented using many snippets of labour market survey data.

3.7 Illustration: simulation using synthetic Admin data

To illustrate this approach we use a simple simulator to generate some synthetic Admin data from a single group g:

$$Y^*_{t,g,i} = 1_{U^*_{t,g,i}>0.1} e^{V_{t,g,i}}, \quad t = 1,2,...,T^A, \quad i = 1,2,...,n_g, \quad U^*_{t,g,i} \overset{iid}{\sim} U(0,1), \tag{18}$$

where $V_{0,g,i} \sim N\{\log(10000), 1\}$ and

$$V_{t,g,i} = 0.15\,(1_{t<10}) + V_{t-1,g,i} + 0.1\varepsilon_{t,g,i}, \quad \varepsilon_{t,g,i} \overset{iid}{\sim} N(0,1). \tag{19}$$

Throughout $\left\{U^*_{t,g,i}\right\}$, $\{\varepsilon_{t,g,i}\}$ and $\{V_{0,g,i}\}$ are independent. Here $T^A = 12$ and $n_g = 200$.

Hence there is a 10% chance an individual has zero earnings at each time period but there is no dependence between these periods of zero earnings. Latent earnings $\exp(V_{t,g,i})$ then grow at around 15% a year for the first 10 years, while initially after higher education earnings are log-normal with median earnings of £10,000.

Again we jitter this data, to produce $Y_{t,g,i} = Y^*_{t,g,i} + U_{t,g,i}$, where $U_{t,g,i} \overset{iid}{\sim} U(0,1)$, and it is this dataset we regard as our simulated Admin data. For each of the n_g individuals we compute $v_{1:T^A,g,i}$, their normalized ranks on the jitter data.

The top row's first plot and third plot in Figure 6 shows the earnings path of the synthetic earnings data for $t = 1, 2, ..., T^A$. The year T^A is indicated by the vertical dotted red line. The first synthetic person has earnings starting out at about £15,000 rising to £35,000 by time T^A. The second person's earnings are lower throughout. The second and fourth plots use normalized ranks $n^{-1}\left\{rank_t(Y^*_{t,i}) - 1/2\right\}$.

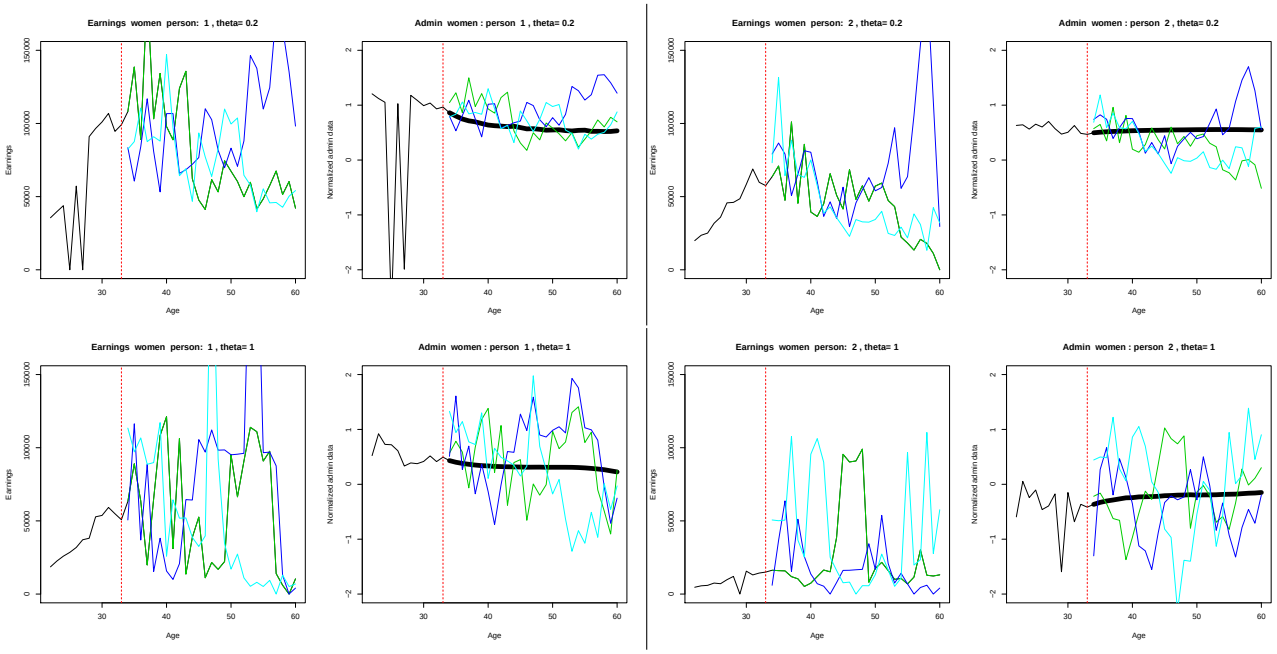


Figure 6: Two synthetic earnings paths for $T^A = 13$ years and 3 simulated paths from their future earnings. First plot is in terms of earnings $\mathbf{Y}_{(T^A+1):T}|\mathbf{Y}_{1:T^A}$, second is carried out on the Gaussian modelling scale $\mathbf{V}_{(T^A+1):T}|\mathbf{V}_{1:T^A}$. The bold lines shows the conditional mean $\beta v_{1:T^A,g,i}$. Left hand side shows the first synthetic earnings path, right hand side the corresponding results for the second person. Top row uses $\theta = 0.2$, bottom uses $\theta = 1.0$. Source: `Basic.r`.

The normalized ranks are projected into the future using $N(\beta v_{1:T^A,g,i}, \Sigma)$. This is illustrated in Figure 6 for these two synthetic earnings paths. The bold lines shows the conditional mean of the forecast distribution $\beta v_{1:T^A,g,i}$. The top row shows the results for $\theta = 0.2$ and the bottom row shows the results for $\theta = 1.0$. The earnings paths with $\theta = 1$ are much rougher.

Table 3.6 provides a summary of the results for the synthetic Admin data. The form of these synthetic results will be given in the identical manner as the real data we will see in a moment.

	$\overline{Y}_{1:t}$			L_t				D_t	TAX_t		
	.25	.50	.75	.25	.50	.75	Mean		.25	.50	.75
$t = T^A$	12.2	24.1	46.3	19.9	37.2	40.0	27.9	82%	22	57	189
$t = T$	21.3	41.1	72.8	0.0	0.0	19.5	9.5	40%	224	519	1,106

Table 4: Economic summaries of loans computed through synthetic admin data, which is then simulated forwarded using the fusion model. Here $n_g = 200$. Given here are quantiles 0.25, 0.5 and 0.75 of career averaged incomes up to time t, quantiles and average outstanding loan size at time t in £1,000s. D_t is the percentage of former students who have not repaid their loans fully by time t, TAX_t shows the quantiles 0.25, 0.5 and 0.75 of cumulative tax take up to time t, also in £1,000s.

4 The 1999 cohort's loans

Our first empirical work will focus on valuing existing income contingent loans made by the UK government to English domiciled students in the 1999 HE cohort. These students started HE around September 1999 and typically graduated around June 2002. We begin by outlining the details of their borrowing and the repayment rules that they faced. We then use the Admin-Survey Fusing Model to highlight their lifetime earnings, combining the actual administrative data in early years with simulated data subsequently. We then give estimates of final loan balances, the proportion with debt written off, the loan subsidy by different groups, and the proportion of the overall HE subsidy that is allocated to each group.

4.1 Type of loans in 1999

Definition 3 gives the rules for their ICLs. Following the previous section, β is the repayment rate, K is the repayment threshold, r is the real interest rate, a is real earnings growth and d is the real discount rate applied by government to value future repayments. T is the time at which the loan is written off, which for the 1999 cohort is age dependent, with all outstanding debt written off at 65. Finally, $t = 0$ here is set to 2003 for all individuals.

Definition 3. *The 1999 English cohort has income contingent loan rules with* $\beta = 0.09$, $K_0 = £10,000$, $K_t = 1.02K_{t-1}$,[2] $r = 0$, $T_i = 65 - age_{i0}$, $a = 0.02$ $d = 0.007$.

The size of these loans are relatively modest for in 1999 they were intended to solely cover living costs. Tuition fees at this time were set at £1,000 and students were expected to fund this up-front from other resources, although in practice the living cost loans could be used to cover the tuition costs. A significant number of students qualified for fee waivers based on their family income. Larger

[2]The rule for K_t is not quite as simple as this. It was held fixed at £10,000 between 2002 and 2006, before being increased to £15,000, where it was fixed until 2011. It has then gone up roughly in line with RPI inflation since.

income contingent loans that also covered tuition fees - which by that point were £3,000 - were only introduced in 2006.

4.2 Earnings and outstanding debt

Our administrative data is a 10% sample of all the English domiciled people who took out loans in 1999. The sample contains 22,621 individuals, made up of 10,590 men and 12,031 women (see Britton et al. (2019) for more details). For each of these individuals we know the HEI they attended, the subject they studied, their gender, their loan size, their year by year taxable earnings (including Self-Assessed earnings) from 2002/03 to 2013/14 and their voluntary repayments from 2002/03 to 2011/12. The latter are repayments made by the borrower to the SLC over and above those they are obligated to make based on their income.

Our valuation of the loans will be carried out using information dated April 2014, which typically corresponds to $t = 12$, and be expressed using 2018 prices. So we will report group averages of quantities like the outstanding loan at $t = 12$, which for the i-th individual we have written as $L_{12,i}$. The extrapolation of the taxable earnings beyond 2014 will be made using the forecasting leg of the Admin-Survey fusing model which was detailed in Section 3.4.2.

We now turn to some descriptive statistics from the repayment data, broken down by subject of study. We also show results for two of the 170 institutions for which we have data. These institution choices were constrained as we only have permission to name a subset of relatively high-status universities.[3] Table 5 displays debt and earnings information based on the administrative data for men and women separately.

Here $\overline{Y}$ denotes the cross-sectional and time-series quantiles (specifically, the 25th, 50th and 75th) of real earnings (in £1,000s in 2018 prices) averaged across the first 11 years since HE, and including zeros. V is the cross-sectional group average of the sum (not average) of voluntary real repayments (in £1,000s in 2018 prices) over the period we observe them. L_t denotes the cross-sectional average of the real loan balances at time t (in £1,000s in 2018 prices). For $t = 12$ this is estimated based on the earnings and voluntary repayments we observe, applying interest on outstanding debts.[4] Errors in this will be generated by:

- Not observing all of the income student loan deduction are made from: we make calculations based on PAYE earnings or, for those who return Self-Assessment forms, earnings from self

[3] In order to report results by institutions each institution has to give us permission to report the results on their own former students even though they did not participate in any aspect of the data collection.

[4] These are publicly available: for example see https://en.wikipedia.org/wiki/Student_loans_in_the_United_Kingdom

employment, employment or profits from partnerships. For those returning Self-Assessment forms and with unearned income above £2,000, repayments would also be taken over this income. As we do not observed unearned income, we effectively assume this to be zero.

- Not observing voluntary repayments in 2012/13 or 2013/14: we assume these to be zero in these years.
- Not observing timing of earnings or voluntary repayments during the tax year: we assume that both voluntary repayment and income contingent repayments are made at the end of the year.

Finally, D_t is the proportion of individuals who have not repaid the loans by time t. This is also estimated through the same process used to calculate L_{12}.

	Women							Men						
	$\overline{Y}$			$\overline{V}$	$\overline{L_0}$	$\overline{L_{12}}$	D_{12}	$\overline{Y}$			$\overline{V}$	$\overline{L_0}$	$\overline{L_{12}}$	D_{12}
Quantiles	.25	.5	.75				%	.25	.50	.75				%
Agri & vetsci	7	16	28	623	14	8	.	.	25	.	698	13	6	.
Allied to med	12	22	34	785	13	5	62	16	31	46	775	14	5	62
Architecture	.	.	.	783	14	6	.	13	29	46	539	15	6	58
Biosci	11	22	33	845	14	6	72	12	29	45	763	14	7	76
Business	9	22	39	791	13	6	67	14	31	52	670	13	5	59
Comms & media	8	17	30	652	13	7	80	8	21	37	372	12	6	70
Creative arts	5	14	26	585	14	8	85	8	19	32	478	14	8	83
Economics	.	.	.	807	13	2	.	18	49	83	627	14	3	.
Education	12	23	34	586	16	8	81	16	32	43	438	16	7	69
Engineering	9	21	37	1,252	13	5	.	13	35	55	831	14	5	55
English	11	20	34	735	13	6	75	10	24	41	864	13	6	70
Foreign lang	9	22	43	1,609	16	6	66	.	25	.	1,883	15	6	.
History & phil	11	22	36	846	13	6	67	11	29	48	984	13	5	60
Law	10	25	43	658	14	6	64	12	33	62	579	14	5	58
Maths & compsci	10	20	36	826	13	5	64	13	32	51	650	13	5	58
Medicine	25	41	56	1,201	25	7	48	38	65	88	1,665	24	5	.
Missing	8	19	32	570	14	7	76	10	25	39	521	14	7	70
Other LEM	.	23	.	697	13	5	71	10	31	48	567	12	4	58
Other STEM	7	19	33	717	13	6	72	9	24	41	650	13	6	70
Physsci	11	23	35	965	14	6	72	16	35	54	1,136	15	5	62
Social studies	9	21	33	709	14	7	76	12	28	43	754	14	6	66
Average	9	20	34	721	14	7	.	12	29	47	695	14	6	.
Uni of Cambridge	15	26	44	1,182	15	4	.	21	40	71	1,614	15	3	.
Uni of Leeds	14	24	37	942	14	5	68	16	33	53	825	14	5	53

Table 5: Results for the 1999 cohort by group in 2018 prices and £1,000s using administrative data up to 2014. $\overline{Y}$ shows the time series and average yearly earnings between age 22 and 32 for individuals at the 25th, 50th and 75th percentile of the earnings distribution. $\overline{V}$ is the cross-sectional average of total voluntary repayments between ages 22 and 32. $\overline{L_t}$ is the cross-sectional average loan balance at time t years after leaving HE, including any accumulated interest and repayments. D_{12} is the proportion of individuals that have not completely repaid their loans by age 32. A dot indicates that statistics have been suppressed due to small sample sizes.

We see from Table 5 that loan balances are around £14,000 on average. Loans were around £3,000

per year at this point, and most people borrowed for three years (students of medicine and education borrowed for longer). The figures include interest accrued during study, but most importantly have been uprated to 2018 prices. The table also shows that men have substantially higher earnings quantiles in most subjects and correspondingly lower loans (and share of people still with loans) at the end of period 12. We see that a large fraction - around 70% - still have loans outstanding after 11 years. Reflecting the higher average earnings, particularly for men, this share is much lower at the University of Leeds - a relatively highly ranked university.

Likewise, higher earnings in STEM subjects as well as equally high earnings in Economics, Law and Business, lead to lower fractions. A very high percentage of creative arts students have loan balances at time T^A.

4.3 Empirical results on loan repayments

We now turn to the estimation of lifetime repayments and implied government subsidies by group. As discussed, this involves the use of our Admin-Survey fusing model which gives us lifetime earnings profiles at the individual level.

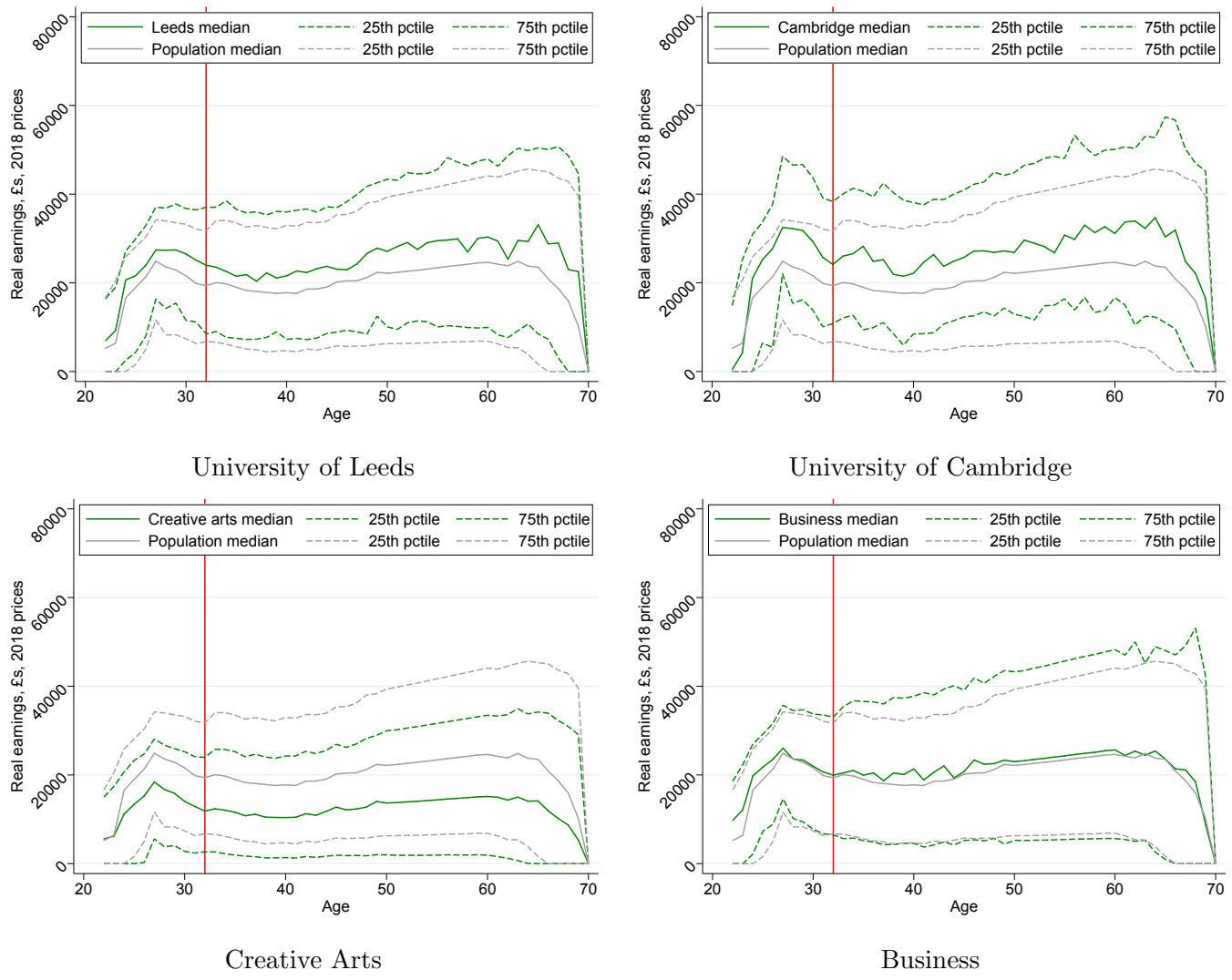


Figure 7: Distribution of annual earnings (in 2018 prices) for the 1999 cohort of women taking out loans. The results from 2002 to 2013 use administrative data. Results from 2014 onwards come from the Admin-Survey fusing model.

Figure 7 shows estimated 25%, 50% and 75% quantiles of real earnings (in 2018 prices) over the lifecycle for women who entered university in 1999 to study business and creative arts, and for those attending the Universities of Leeds and Cambridge. As indicated by the vertical line in each plot, the first 11 years of the plots come directly from the tax data, while the remaining years are the extrapolated versions using the Admin-Survey fusing model. To benchmark each plot we show quantiles for all women in the 1999 cohort in grey.

The figures show the dramatic differences in earnings between creative arts and business, with vastly lower quantiles for creative arts. The differences between institutions is a little less severe here, although the graduates from the University of Cambridge have quite high quantiles throughout.

For estimating future repayments we use the loan rules as outlined in Section 4.1. The sole remaining issue is to forecast future voluntary repayments. In this exercise we will take them all to be zero. Figure 8 shows the implied mean repayments at each age. Due to the relatively small loan sizes for this cohort we see that average repayments start declining soon after graduation as some graduates start to repay the full amount of their loan, although some students are still repaying until the end of the repayment period. The large peak in repayments for medicine is driven by a combination of their high earnings and higher initial debts.

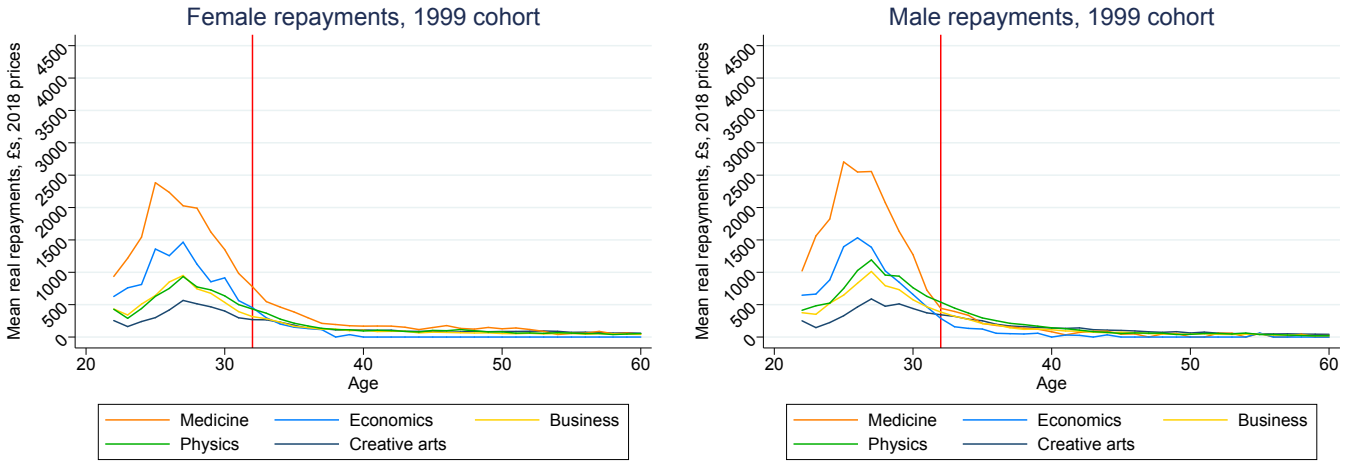


Figure 8: Mean repayment profiles for female (left) and male (right) graduates. Repayment profiles shown are mean yearly repayments for graduates that entered university in 1999. Repayments are in 2018 prices. Repayments based on earnings profiles and debt levels. The repayment period is until age 65, after which the debt is written off and no further repayments are made, regardless of the amount repaid until then.

Table 6 shows estimates of the write-off of debt at the end of the repayment period and government costs by group. Recall $\widehat{L}_T$ is the estimate of the group's cross-sectional average of the real loan balances at the end of the loan period T (in £1,000s in 2018 prices). This partially depends upon the Admin-Survey fusing model and the discount parameter d. Likewise $\widehat{D}_T$, the share of former students with loans which were paid off by the end of the T-th period, when the loan is forgiven. As defined in Section 2.3, the RAB is the net present value of the loan subsidy expressed as a percentage of the loan value, using the government's real discount rate of 0.7%. Alternatively put, this is the share of the loans that the government expects to write off. The percentage of graduates in group g is ω_g. Also given is the group g loan subsidy and overall subsidy. The overall subsidy includes both the loan subsidy and any subsidy through fee subsidies and teaching and maintenance grants. It is important

to note that this does not take into account any cross subsidisation that may occur within universities across subjects, and is instead the subject-specific subsidy implied by the loan system.

	Women								Men							
	$\widehat{\overline{G}}$	$\widehat{\overline{L}}_0$	$\widehat{\overline{L}}_T$	$\widehat{\overline{D}}_T$ %	RAB %	ω_g %	S_g %	S_g^+ %	$\widehat{\overline{G}}$	$\widehat{\overline{L}}_0$	$\widehat{\overline{L}}_T$	$\widehat{\overline{D}}_T$ %	RAB %	ω_g %	S_g %	S_g^+ %
Agri & vetsci	31.5	13.8	4	58	55	1	1	2	32.8	12.5	2.5	44	40	1	1	1
Allied to med	15.4	12.8	3	44	38	7	6	7	15.1	13.6	2.4	34	34	4	4	4
Architecture	19.9	13.6	3	.	43	0	0	1	20.2	14.7	2.8	38	36	2	2	2
Biosci	15.6	13.6	3	51	39	6	5	6	15.8	14.3	2.9	42	37	4	4	4
Business	8.3	13.2	3	48	38	10	9	7	8.3	12.8	2.2	37	34	11	10	8
Comms & media	13.2	12.6	3	61	46	3	3	3	13.2	12.3	3.1	46	44	2	2	2
Creative arts	12.5	13.6	4	67	55	10	13	10	12.5	13.6	4.1	60	51	10	13	9
Economics	7.6	13.2	1	.	16	0	0	0	7.7	13.6	1.5	23	25	1	1	1
Education	10.4	16.4	4	54	40	13	14	12	10.5	16.0	2.9	36	36	5	6	4
Engineering	23.2	13.5	2	45	36	1	1	1	23.5	14.4	2.1	32	31	8	7	10
English	7.5	13.5	3	51	39	4	4	3	7.5	13.5	2.8	42	38	2	2	1
Foreign lang	10.6	16.1	3	49	34	2	1	1	10.3	14.5	2.8	.	35	1	1	1
History & phil	7.8	13.3	3	50	38	3	3	2	7.9	13.1	2.2	37	34	3	3	2
Law	7.7	13.6	3	42	37	5	5	4	7.7	13.6	2.1	34	32	4	4	3
Maths & compsci	14.9	13.3	3	49	38	4	3	4	14.8	13.0	2.1	35	33	14	12	14
Medicine	62.7	25.5	2	23	37	2	3	7	61.4	24.4	1.9	17	30	2	3	6
Missing	16.3	13.8	4	54	46	15	16	17	16.2	13.9	3.3	46	42	11	13	12
Other LEM	7.7	12.5	3	42	39	1	1	1	7.7	12.2	2.2	36	35	2	1	1
Other STEM	21.8	13.4	3	55	44	2	2	2	21.7	13.4	2.7	41	40	4	4	5
Physssci	21.8	13.9	3	50	37	2	2	3	22.3	14.8	2.0	32	29	5	5	7
Social studies	9.9	13.8	3	53	44	8	8	7	9.7	13.9	2.7	39	38	5	5	4
Average	13.7	14.1	3	.	42	100	100	100	15.1	13.9	2.6	.	37	100	100	100
Uni of Cambridge	14.2	14.8	2	36	24	1	1	1	15.7	14.9	1.0	.	21	1	1	1
Uni of Leeds	13.5	14.0	3	45	32	2	2	2	15.6	14.2	2.0	33	30	2	2	2

Table 6: Results 1999 cohort by group in 2018 prices and £000s using Admin-Survey fusing model. G is the average grants given out by government, which includes teaching grants, maintenance grants and fee subsidies. $\widehat{\overline{L}}_0$ is average debt at graduation, $\widehat{\overline{L}}_T$ is average debt written off at time T, $\widehat{\overline{D}}_T$ is the percentage of individuals with some debt written off. ω_g is the percentage of the cohort studying each subject. The RAB charge is the total value of the loan minus the net present value of repayments made by the graduate as a proportion of the loan value, using the government discount rate of 0.7% real. S_g is the percentage of the loan subsidy doing to group g. S_g^+ is the percentage of the total subsidy going to group g. A dot indicates that statistics have been suppressed due to small sample sizes.

We find that overall the government loan subsidy is 37% for men, and higher for women at 42%. For the University of Leeds the loan subsidy is comfortably lower than average (32% for women; 30% for men), while for the University of Cambridge it is well below average (24% for women; 21% for men). Across subjects, we see significant variation in the RAB of around 20% for economics at one end, and a RAB above 50% for creative arts at the other. This means that the government can expect to write off around 50% of the value of the loans that it issued to creative arts students.

Across subjects, even though there are sizeable differences in earnings, the overall government

subsidy is relatively evenly distributed across subjects, meaning it corresponds closely to subject size. For example while the highest-RAB subject area - creative arts courses - receive a larger-than-proportional share of the loan subsidy (it has around 10% of the students but accounts for around 13% of the overall loan subsidy), they also receive lower teaching grants than other courses, and so the share of *overall* spending is roughly proportional to the share of students.

5 The 2017 cohort's loans

We now turn to the estimation of the loans and subsidies for the cohort of students that started HE in Autumn 2017. These students entered HE under a dramatically different system compared to the 1999 cohort, and therefore are an interesting point for comparison. In this section, we provide details on the loans before estimating the lifetime earnings and repayments for students of different subgroups. We also highlight the lifetime tax contribution and compare this to student loan repayments and subsidies.

5.1 Type of loans in 2017

The 2017 cohort will typically end HE in June 2020. Tuition fees for this cohort were up to £9,250 per year, real interest rates of 3% while studying had been introduced, and grants for maintenance had been abolished. Combined, will result in students graduating with significantly higher debts of around £50,000 for a 3 year degree on average, with students from the poorest backgrounds graduating with the largest debts (Belfield et al., 2017). The 2017 version on the loan contracts is detailed in Definition 4. In this section, 2020 will correspond to $t = 0$, while we will report results using 2018 prices.

Definition 4. *The 2017 English cohort has income contingent loan rules with* $\beta = 0.09$, $K_0 = £25,000$, $K_t = 1.015K_{t-1}$, $T = 30$, $a = 0.015$. *The UK Government currently uses* $d = 0.007$ *in its present value calculations and has an interest rate which varies with the level of earnings and payment threshold as* $r(y,k) = 0.03 \min\{\max(y-k,0)/\alpha k, 1\}$, *where* $\alpha = 45/25$.

The 2017 income contingent contracts are more complicated than the 1999 version given in Definition 3. The length of the contract is 30 years from entering repayment in the April after graduation. The threshold is now formally indexed by average earnings growth, which we assume (based on forecasts from the OBR[5]) to be 4.5% (or 1.5% in real terms). The most significant new complexity is that the interest rate inside the loans has been set to vary between 0% and 3% plus RPI depending on

[5]The OBR (Office for Budget and Fiscal Responsibility) is an independent body that provides forecasts for economic indicators. We draw on the 2018 OBR Fiscal Sustainability Report and use the long term earnings projections of 1.5% above RPI inflation.

earnings once the individual has left higher education. Prior to leaving, the rate is equal to 3% plus RPI.

5.2 Admin-Survey fusing model

We have no past tax data for the 2017 cohort as individuals. Instead we use the 1999 cohort and uprate their earnings to match 2017 levels. We then assign student loans based on major, parental income and whether they study in London. We then reweight the population so as to match the proportion studying each major at each university in 2017.

This type of analysis is more susceptible to misspecification in the Admin-Survey fusing model than our analysis of the 1999 cohort study which used observed taxable earnings of the individuals for their first 11 years and the fusing model for the remaining years. Now we are using the lifecycle behaviour of the 1999 cohort to proxy the 2017 cohort.

Table 7 gives the career average of earnings and debt on graduation for all subjects for the 2017 cohort, by gender. Debts are considerably higher in 2017 compared to 1999. Earnings for each subject are taken from the 1999 cohort, but uprated to reflect earnings growth. Note that the numbers here are averages of lifetime earnings, rather than just earnings over the first 11 years, as in Table 5 above.

We estimate very large differences in lifetime median earnings by subject. For example, median lifetime earnings for female medical students are around three times that of creative arts students. More generally, science, business and economics graduates perform relatively well in the labour market, while humanities graduates have much lower earnings. These large differences in earnings by subject are not simply a feature of using the median and exist at other points in the distribution, highlighting large differences in the variance of lifetime earnings by subject. As highlighted above, we estimate debt on graduation to be around £50,000 for all subjects. The exception to this is medicine which has average debts on graduation of around £70,000.[6] To calculate debt at graduation we assume that all borrowers take out the maximum loan available to them. This will overstate our estimated RAB charges.

[6]Medicine, dentistry and veterinary science are 5 year courses in the UK, while most other subjects are 3 year course. The exception are some science courses which often offer integrated masters making the course length 4 years. As we cannot observe whether individuals take integrated masters in our data we assume all subjects other than medicine, dentistry and veterinary science are 3 year courses.

	Women				Men			
		$\widehat{\widehat{Y}}$		$\widehat{\widehat{L_0}}$		$\widehat{\widehat{Y}}$		$\widehat{\widehat{L_0}}$
Quantiles	.25	.5	.75		.25	.50	.75	
Agri & vetsci	9	23	39	64	.	35	.	66
Allied to med	17	30	47	54	22	43	65	53
Architecture	.	.	.	52	18	40	64	52
Biosci	14	30	46	52	16	40	63	52
Business	13	30	54	53	19	43	73	52
Comms & media	11	23	42	51	11	29	52	50
Creative arts	7	19	36	51	11	26	44	51
Economics	.	.	.	52	24	68	116	52
Education	17	31	47	53	22	44	60	53
Engineering	11	30	51	61	18	48	77	61
English	15	28	47	51	13	34	57	51
Foreign lang	13	30	60	58	.	35	.	59
History & phil	14	30	50	50	15	41	67	50
Law	13	34	60	52	16	46	86	52
Maths & compsci	13	27	50	56	18	44	72	56
Medicine	35	58	79	72	53	93	124	72
Other LEM	.	31	.	53	14	43	67	52
Other STEM	10	26	46	57	13	33	58	57
Other arts & hum	11	26	44	51	14	34	55	52
Physsci	15	31	49	57	22	49	76	57
Social studies	13	29	45	51	16	39	60	51
Average	13	28	47	53	16	40	67	54
Uni of Cambridge	21	36	62	50	29	56	97	53
Uni of Leeds	18	33	54	51	19	46	79	53

Table 7: Results for the 2017 cohort by group in 2018 prices, and £1,000s using administrative data up to 2014. $\widehat{\widehat{Y}}$ shows the career average yearly earnings between ages 22 and 70 for individuals at the 25th, 50th and 75th percentile of the earnings distribution in £1,000s in 2018 prices. $\widehat{\widehat{L_0}}$ is the cross-sectional average loan at graduation in £1,000s in 2018 prices. A dot indicates that statistics have been suppressed due to small sample sizes.

5.3 Empirical results on loan repayments

We now turn to the implied repayments from our model for the 2017 cohort, and subsequently, the implied long run government subsidy. Figure 9 shows the pattern of lifetime repayments for medicine, economics, business, education and creative arts students, by gender. Repayments are considerably larger than for the 1999 cohort, peaking much later. Due to the repayment threshold of K_t, below which no repayments have to be made, repayments increase more than proportional to earnings. Peaks appear where sufficient numbers of borrowers are clearing their loans to outweigh increases in earnings that result in increased repayments from everybody else in that group. For economics, this occurs early in around the late 20s, while for medicine - due to the larger initial debts - this occurs during individual's late 30s. This is also true for business, while for creative arts and education the peaks happen later for men and not at all for women, as a result of so few people clearing their loans.

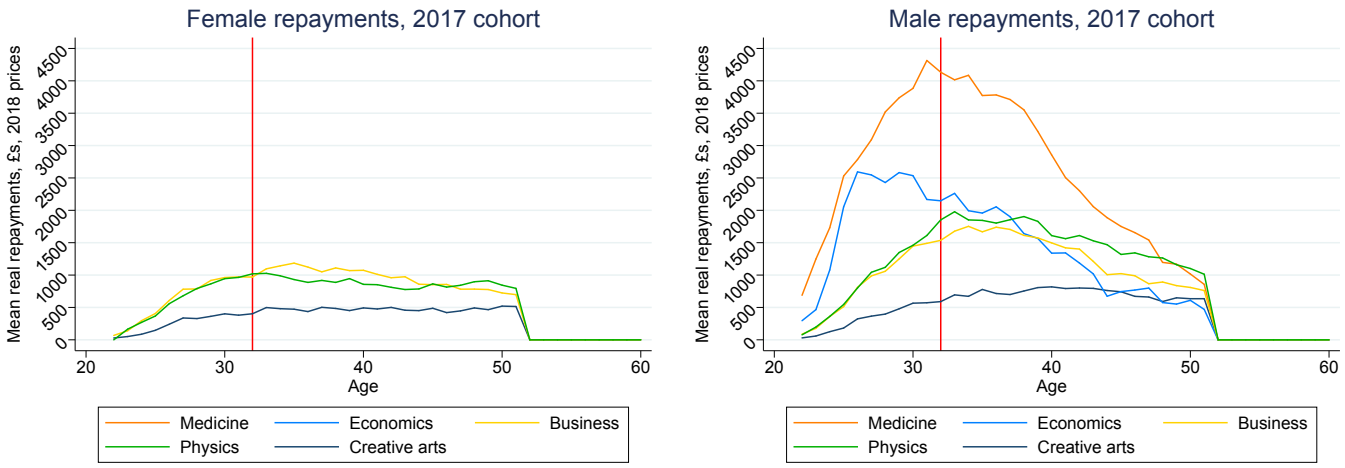


Figure 9: Mean repayment profiles for female (left) and male (right) graduates. Repayment profiles shown are mean yearly repayments for graduates that entered university in 2017, assuming they entered university at age 18. Repayments are in 2018 prices and 000s. Repayments based on earnings profiles and debt levels. The repayment period is 30 years, after which the debt is written off and no further repayments are made, regardless of the loan outstanding. Repayments are not shown when sample sizes are too small.

The top row of Figure 10 shows the estimated year by year repayments quantiles by creative arts and business graduates, as well for the population as a whole. The repayments by the creative arts graduates are very modest, with average repayments not going much above £500 per year even in the peak earnings years. Repayments by business graduates are much stronger. Figure 10 show average repayments peak at over £1,000 per year for much of the lifecycle.

Figure 10 also gives results focusing on female graduates at two HEIs, the University of Leeds and the University of Cambridge. Both universities have repayments above the population average, with a significant proportion of University of Cambridge graduates repaying their loan relatively early in their career, as shown by the decrease in mean repayments after age 30.

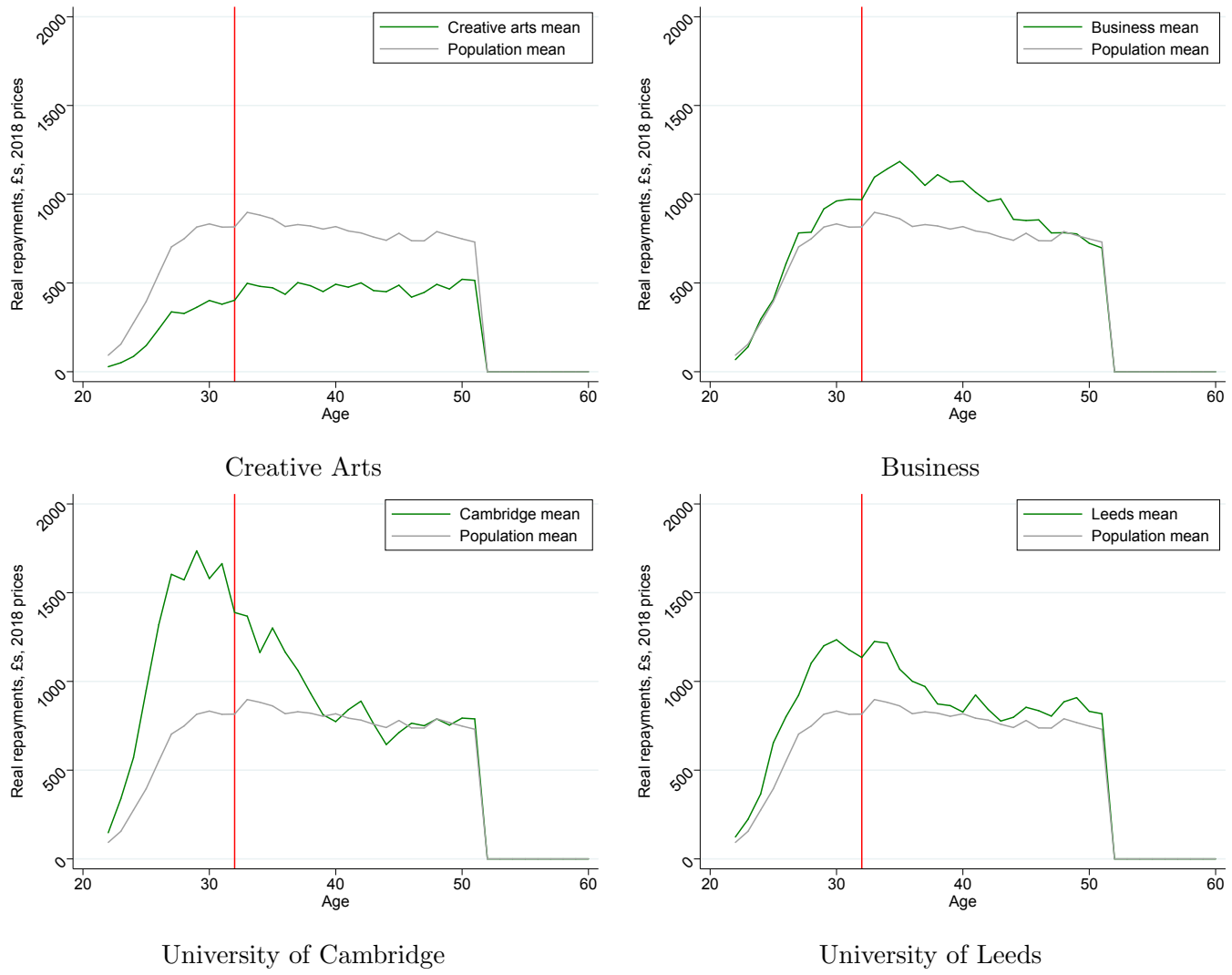


Figure 10: Yearly mean repayment for female graduates for creative arts (top left) and business (top right) and the University of Cambridge (bottom left) and the University of Leeds (bottom right). Benchmark results for the graduate population as a whole are given in grey.

Table 8 shows the estimated RAB charge, share of graduates who do not clear their debt before it is written off and the estimated government subsidy by subgroup. We can see that loans are much higher than they were in 1999, and as a result the share of loans written off is much higher. Overall we see that the government expects to write off around 44% of all student loans issued to men and 63% of those issued to women. This translates to around 54% overall.

There is considerable variation in the RAB charge, both across genders and within genders across subject. For women who studied medicine, the RAB is around 30%, while the equivalent figure for men is 4%, which means that men repay more over their lifetime than the present value of the loan. This is particularly impressive given the size of the initial loan, reflecting the large disparities in lifetime

earnings.

	Women								Men							
	$\widehat{\overline{G}}$	$\widehat{\overline{L}}_0$	$\widehat{\overline{L}}_T$	$\widehat{\overline{D}}_T$ %	RAB %	ω_g %	S_g %	S_g^+ %	$\widehat{\overline{G}}$	$\widehat{\overline{L}}_0$	$\widehat{\overline{L}}_T$	$\widehat{\overline{D}}_T$ %	RAB %	ω_g %	S_g %	S_g^+ %
Agri & vetsci	12.3	64.3	45	.	80	1	2	2	12.8	65.9	39.3	.	59	1	1	1
Allied to med	1.7	53.8	33	93	64	15	16	15	1.5	52.5	23.7	75	40	5	4	4
Architecture	4.9	52.1	33	.	72	1	1	1	4.9	52.3	26.1	78	46	2	2	3
Biosci	1.6	52.0	31	91	65	12	12	12	1.6	51.5	25.5	76	48	10	10	10
Business	0.1	52.8	29	83	57	10	9	9	0.1	52.1	22.8	68	41	15	13	12
Comms & media	0.1	51.0	32	.	73	2	2	2	0.1	49.6	26.8	85	57	2	2	2
Creative arts	0.8	51.1	34	96	79	9	11	10	0.8	51.2	31.7	91	70	6	9	9
Economics	0.1	51.7	19	.	29	2	1	1	0.1	52.2	15.2	45	22	2	1	1
Education	0.1	52.6	33	96	64	4	4	4	0.1	52.7	26.3	82	42	1	1	1
Engineering	5.6	61.0	35	.	57	2	2	2	5.7	61.0	27.5	72	36	10	9	10
English	0.1	50.6	30	89	63	3	3	3	0.0	50.7	26.6	81	54	1	2	2
Foreign lang	0.1	57.9	31	83	55	1	1	1	0.1	59.4	26.4	.	43	1	1	1
History & phil	0.1	49.9	27	81	59	3	3	2	0.1	50.5	23.0	72	43	3	3	3
Law	0.1	52.4	26	77	51	4	4	3	0.1	51.8	21.0	62	37	3	2	2
Maths & compsci	0.8	55.6	32	86	60	3	3	2	0.8	55.5	25.5	72	40	12	12	11
Medicine	34.1	71.9	30	76	30	2	1	3	33.4	71.5	14.4	38	4	2	0	2
Other LEM	0.1	53.1	31	.	61	1	1	1	0.1	52.0	23.1	73	40	2	2	1
Other STEM	5.2	57.0	36	.	67	3	3	3	5.2	56.8	31.2	84	54	4	5	6
Other arts & hum	0.3	51.3	32	92	68	11	11	10	0.3	51.7	28.0	84	55	8	10	9
Physsci	5.3	56.8	34	88	63	3	3	4	5.4	57.3	25.8	70	37	5	5	5
Social studies	0.1	51.0	31	92	67	7	7	7	0.1	50.6	24.9	77	48	6	6	5
Average	1.9	53.2	31	.	63	100	100	100	2.2	53.9	25.5	.	44	100	100	100
Uni of Cambridge	2.0	50.2	25	.	46	1	0	0	2.4	53.5	17.7	58	23	1	0	0
Uni of Leeds	2.0	51.3	27	85	54	2	2	2	2.8	52.6	21.2	65	35	2	1	1

Table 8: Information on the 2017 cohort by group in 2018 prices and 000s using Admin-Survey fusing model. $\widehat{\overline{L}}_T$ is average debt written off at time T, $\widehat{\overline{D}}_T$ is the % of individuals who do not fully repay their loans by the end of the repayment period. 100ω is the % of the cohort studying each subject based on 2017-18 HESA student numbers (https://www.hesa.ac.uk/data-and-analysis/students/what-study). The RAB charge is the total value of the loan minus the net present value of repayments made by the graduate as a proportion of the total loan value, using the government discount rate of 0.7% real. S_g is the percentage of the loan subsidy doing to group g and S_g^+ is the percentage of the total subsidy going to group g, both make use of the HESA student numbers above to calculate these proportions. A dot indicates that statistics have been suppressed due to small sample sizes.

Meanwhile, the RAB charge is more than 70% for women studying communications & media, agriculture and creative arts. For men, creative arts has by far the highest RAB charge at 70%, while the next highest is communications & media.

The latter has a high RAB charge, but accounts for a relatively low share of overall government spending. This is because it is a fairly small course. The same is not true for creative arts, which accounts for around more than 10% of the overall government subsidy.

5.3.1 Loan repayments and tax contributions

Table 9 shows the average tax take by subject discipline, including income taxes and national insurance charges (NICs) compared to lifetime student loan repayments. We provide these to give an impression of the size of the losses on loans compared to the general income tax take from these graduates.

We see the variation in the tax take across subjects outweighs the outright losses on loans. Overall loss rates are vastly less important for the government's finances than subject choice. Moving some students over from creative arts into business or STEM dramatically helps the Government's finances. Or perhaps more usefully, HEI's may help their students by combining creative arts subjects with other fields like business or computer science.

	Women						Men					
	Repayments			Tax take			Repayments			Tax take		
Quantiles	.25	.5	.75	.25	.5	.75	.25	.5	.75	.25	.5	.75
Agri & vetsci	0	3	15	17	116	296	.	14	.	.	267	.
Allied to med	2	9	30	84	207	400	4	28	48	131	356	647
Architecture	.	.	.	.	.	.	4	22	48	112	315	620
Biosci	1	9	29	65	205	396	2	20	46	84	319	618
Business	1	11	41	58	210	512	3	25	49	101	357	763
Comms & media	0	2	21	37	133	352	0	11	36	40	203	452
Creative arts	0	2	12	18	95	269	0	4	23	30	158	370
Economics	.	.	.	.	.	.	16	44	55	200	673	1,413
Education	2	10	30	84	225	407	6	27	49	144	364	557
Engineering	1	9	40	53	216	454	5	33	61	106	425	813
English	1	9	29	66	189	417	1	14	43	76	251	548
Foreign lang	1	13	47	75	221	578	.	20	.	.	283	.
History & phil	1	10	39	75	227	470	3	21	47	79	328	673
Law	2	17	45	60	257	604	5	33	48	88	381	969
Maths & compsci	1	8	37	60	184	447	3	26	52	103	363	744
Medicine	13	52	70	157	494	774	44	67	84	395	970	1,399
Other LEM	.	10	.	.	229	.	2	26	49	66	365	712
Other STEM	0	6	27	37	162	393	1	14	46	63	250	532
Other arts & hum	0	6	25	47	170	370	1	12	42	57	252	503
Physsci	2	9	33	78	227	432	6	34	53	141	413	786
Social studies	0	6	26	49	196	389	2	18	44	80	304	579
Average	1	8	30	54	187	406	2	21	48	81	322	668
Uni of Cambridge	4	17	45	113	275	601	14	42	62	226	510	1,088
Uni of Leeds	2	12	42	89	258	488	3	30	49	98	379	817

Table 9: Estimated lifetime repayments and tax take for women from the 2017 cohort by group in 2018 prices, in £1,000s. A dot indicates that statistics have been suppressed due to small sample sizes.

6 Discussion and conclusions

This paper is the first to use administrative data to estimate the value of the student loan book in England. Creating a new type of panel model that fuses administrative and survey data, we simulate the lifetime earnings of all graduates using a method that enables us to value every individual's student

loan. This enables us to break down the overall cost of the student loan book by sub-groups such as subject and university.

We provide estimates by gender and for 21 major subject groups finding considerable heterogeneity in the long-run cost to government. Across all subjects, earnings of male graduates are considerably higher than that of their female counterparts. Men are hence expected to repay a larger part of their loans and receive a smaller government subsidy. For both male and female graduates earnings, and hence the government subsidy, vary hugely across degree subjects. Medicine and economics students on average repay more than 75% of the value of their student loans, while creative arts, agriculture and communications graduates on average repay less than 35% of the value of their loans.

It is useful to place this in the context that it is well established that, although women graduates earn on average less than men, the average economic return to HE for women is much higher than that for men — a significant majority of women who do not go through HE tend to perform particularly poorly in the UK labour market, e.g. Britton et al. (2019). These higher returns greatly benefit the Governments fiscal position, as increases in earnings also increase the income taxes women pay in later life.

The shift from teaching grants to loans between 1999 and 2017 has meant that the overall government subsidy is increasingly geared towards lower earning subjects instead of expensive to teach subjects. As a result, the share of the government subsidy going to arts and humanities subjects, which have the lowest graduate earnings, has increased over this period, while the share going to science subjects has decreased.

References

Arellano, M., R. Blundell, and S. Bonhomme (2017). Earnings and consumption dynamics: A nonlinear panel data framework. *Econometrica 85*, 693–734.

Beare, B. K. (2010). Copulas and temporal dependence. *Econometrica 78*, 395–410.

Belfield, C., J. Britton, and L. van der Erve (2017). Higher education finance reform: Raising the repayment threshold to £25,000 and freezing the fee cap at £9,250. Unpublished technical report: Institute of Fiscal Studies.

Blundell, R., L. Dearden, A. Goodman, and H. Reed (2000). The returns to higher education in Britain: Evidence from a British cohort. *Economic Journal 110*, 82–99.

Bonhomme, S. and J. M. Robin (2006). Modelling individual earnings trajectories using copulas: France, 1990-2002. In H. Bunzel, B. J. Christensen, G. R. Neumann, , and J.-M. Robin (Eds.), *Structural Models of Wage and Employment Dynamics. Contributions to Economic Analysis*, Volume 275, Chapter 18, pp. 441–478. Amsterdam: Elsevier.

Britton, J., L. Dearden, N. Shephard, and A. Vignoles (2019). Is improving access to university enough? socio-economic gaps in the earnings of English graduates. *Oxford Bulletin of Economics and Statistics*. Forthcoming.

Britton, J., N. Shephard, and A. Vignoles (2019). A comparison of sample survey measures of earnings of English graduates with administrative data (with discussion). *Royal Statistical Society, Series A (Statistics in Society)*. Forthcoming.

Chetty, R., J. N. Friedman, E. Saez, N. Turner, and D. Yagan (2017). Mobility report cards: The role of colleges in intergenerational mobility. Working paper 23618, National Bureau of Economic Research.

Chowdry, H., L. L Dearden, A. Goodman, and W. Jin (2012). The distributional impact of the 2012-13 higher education funding reforms in England. *Fiscal Studies 33*, 211–36.

Cochrane, J. H. (2005). *Asset Pricing* (2 ed.). Princeton: Princeton University Press.

Dale, S. B. and A. B. Krueger (2002). Estimating the payoff to attending a more selective college: an application of selection on observables and unobservables. *Quarterly Journal of Economics 117*, 1491–1527.

Dearden, L., E. Fitzsimons, A. Goodman, and G. Kaplan (2008). Higher education funding reforms in England: the distributional effects and the shifting balance of costs. *Economic Journal 118*, F100–125.

Dixit, A. K. and R. S. Pindyck (1994). *Investment under Uncertainty*. Woodstock, Oxfordshire, United Kingdom: Princeton University Press.

Frees, E. and P. Wang (2006). Copula credibility for aggregate loss models. *Insurance: Mathematics and Economics 38*, 360–373.

Genest, C. and J. Nelehov (2007). A primer on copulas for count data. *ASTIN Bulletin 37*, 475–515.

Guvenen, F., F. Karahan, S. Ozkan, and J. Song (2015). What do data on millions of U.S. workers reveal about life-cycle earnings risk? Working Papers 20913, National Bureau of Economic Research.

Hastings, J. S., C. A. Neilson, and S. D. Zimmerman (2013). Are some degrees worth more than others? Evidence from college admission cutoffs in Chile. Working Paper 19241, National Bureau of Economic Research.

Hull, J. (2017). *Options, Futures, and other Derivative Securities* (10 ed.). New Jersey: Prentice-Hall International Editions.

Joe, H. (1997). *Multivariate Models and Multivariate Dependence Concepts*. London: Chapman and Hall.

Jorgenson, D. W. and B. M. Fraumeni (1989). The accumulation of human and nonhuman capital, 1948-1984. In R. E. Lipsey and H. S. Tice (Eds.), *The Measurement of Savings, Investment and Wealth*, pp. 227–282. Chicago, I.L.: The University of Chicago Press.

Jorgenson, D. W. and B. M. Fraumeni (1992). Investment in education and U.S. economic growth. *Scandinavian Journal of Economics 92*, 51–70.

Kaplan, G. and G. Violante (2010). How much consumption insurance beyond self-insurance. *American Economic Journal 2*, 53–87.

Kirkeboen, L. J., E. Leuven, and M. Mogstad (2016). Field of study, earnings, and self-selection. *Quarterly Journal of Economics 131*, 1057–1111.

Lambert, P. and F. Vandenhende (2002). A copula-based model for multivariate non-normal longitudinal data: Analysis of a dose titration safety study on a new antidepressant. *Statistics in Medicine 21*, 3197–3217.

Loaiza-Maya, R. and M. S. Smith (2019). Real-time macroeconomic forecasting with a heteroskedastic inversion copula. *Journal of Business and Economic Statistics*. Forthcoming.

Lochner, L. and E. Moretti (2004). The effect of education on crime: Evidence from prison inmates, arrests, and self-reports. *American Economic Review 94*, 155–189.

Milligan, K., E. Moretti, and P. Oreopoulos (2004). Does education improve citizenship? Evidence from the United States and the United Kingdom. *Journal of Public Economics 88*, 1667–1695.

Moretti, E. (2004). Workers' education, spillovers, and productivity: Evidence from plant-level production functions. *American Economic Review 94*, 656–690.

Smith, M. S., A. M. C. Almeida, and C. Czado (2010). Modeling longitudinal data using a pair-copula decomposition of serial dependence. *Journal of the American Statistical Association 105*, 1467–1479.

Smith, M. S. and W. Maneesoonthorn (2018). Inversion copulas from nonlinear state space models with an application to inflation forecasting. *International Journal of Forecasting 34*, 389–407.

A Proofs

B Institutional background and data

B.1 Administrative data

Our administrative dataset is a database we built and described in Britton et al. (2019), using National Insurance Numbers (NINOs) to hard link three datasets: data from the SLC and Pay As You Earn (PAYE) and Self-Assessment (SA) databases from Her Majesty's Revenue and Customs (HMRC). This provides us with a large longitudinal database on UK earnings for individuals domiciled in England upon application to HE, who received loans from the SLC.

The two HMRC datasets arise because the UK has two types of income tax forms. The significant majority of tax payers use the PAYE system, which is operated by employers who withhold income and other employment taxes and report the earnings and deductions made to HMRC. This means the majority of UK citizens do not themselves file tax forms; around 90% of UK income tax is collected through the PAYE system. For those with more complicated tax affairs (e.g. high incomes, self-employed, owning a business, having significant investment accounts, being in a professional partnership) HMRC requires them to file a set of SA forms. Individual taxpayers can also opt to submit SA forms.

When we have both PAYE and SA earnings we use the SA data, as HMRC regard the SA records as definitive (noting that a SA form will include PAYE income). If an individual has no reported earnings then we take their earnings as zero. This is likely to miss some earnings for very low earners who do not have to return a PAYE form and who may not be asked to complete a SA form (although note that they have a legal responsibility to report this income). All earnings are converted into October 2018 prices using the Consumer Price Index (CPI).

A drawback of our database is that when former students become non-resident for UK tax purposes, HMRC may lose contact with them and generally will only record earnings from UK sources as these are their UK taxable earnings. We will express the earnings of such students as 0 in our reports if HMRC records it as 0, which clearly may underestimate their true earnings, and therefore their subsequent loan repayments.

B.1.1 Earnings data

Our focus is on earned labour income, so we defined this as the sum of employment income, profits from partnerships and profits from self-employment declared to HMRC. Clearly some aspects of the returns from a partnership are due to the capital risk a partner is exposed to, but we cannot break that component out here and so take profits from partnerships as earnings.

The SA databases also contain information on trust income, profits on share transactions, profits from land and property, UK dividends, pension income, life policy gains, "other" income, bank and building society interest and total income, all of which we exclude from earned income as they measure non-employment income. We wanted to include foreign income from employment and savings, but the calculation involved various delicate deductions, so we excluded it.

We do not make a record of any deductions tax payers make, e.g. capital losses on investments, nor of any tax free allowances individuals may have. We also do not account for employers' and employees' tax free pension contributions as labour earnings as UK tax forms only record pension income and

not pension contributions.

B.1.2 Student Loan Company (SLC) data

The SLC has offered income contingent loans to all UK domiciled HE students since 1998. The take-up rate amongst eligible students during this period is around 85-90% overall, a rate that has remained relatively stable (author's own calculations based on overall students numbers from the SLC "Student Support for higher education in England" archived series). Not all individuals receiving a loan from the SLC will be studying for first degrees, as individuals can access loans for foundation degrees, Higher National Diplomas (HNDs) and lower undergraduate qualifications. The dataset we received from SLC does not have any indicators to split individuals into these different groups. We observe the subject of and university of the final degree for which an individual qualifies for a loan. So, for example, for someone attending a HE institution for a term before dropping out and re-starting at a different institution sometime in the future, only their second degree is observed so long as they borrowed again (though the date they started in HE is the first degree start date).

The dataset only includes individuals who borrowed from the English part of the SLC - meaning they were domiciled in England upon application - between 1998 and 2010 and covers around 2.6M former borrowers who are qualified to be in repayment, which happens in April of the year after they leave HE. We have no data on those who are still in HE and have insufficient earnings to qualify them for repayment, which results in a decline in our cohort sizes for more recent student cohorts (see Table 10). Note that we only observe borrowers and not whether individuals graduate, resulting in individuals who borrow from the SLC but subsequently drop out being inaccurately defined as graduates (throughout, we use the terms "borrowers" and "graduates" interchangeably, but note that dropping out does not prevent people from having to make repayments on their student loans). During this period the drop out rate from UK universities for those who enroll was around one in ten, including mature entrants (taken from HESA performance indicators data series, where HESA measures drop out by those who attended for at least 90 days before dropping out).

B.1.3 Basic summaries of the SLC-HMRC linked sample of graduates

We work with a 10% sample of the SLC data, each of whom was carefully traced through the tax databases to link through their NINOs their earnings in each year. Our 10% sample has 263,052 members, covering cohorts from 1998 to 2011. We focus on the 2008-09 to 2012-13 tax years. It should be noted that this was a financially difficult period. The sample is detailed for the tax year 2011/12 in Table 10 to provide a snapshot of the data.

There are around 24,000 students in each cohort, with the smaller 1998 figure reflecting slow uptake of the new income contingent student loans and the decline at the end reflecting the fact that individuals have not entered repayment (i.e. left HE) by 2011/12. The student numbers align with HESA statistics for 2007/08, which state that around 325,000 UK domiciled students were studying in England. Our 10% sample is 25,000 students in this year, meaning a cohort size of around 250,000 borrowers. Around 15% of the English students do not borrow (taking us to 295,000), while the remaining students would be non-English UK students studying in England.

Each individual potentially has a SA and a PAYE tax record in each tax year, but may have neither. By construction, we are able to state that if they have neither a SA nor a PAYE record then they have no UK tax return at all - note that unlike the US, in the UK it is not legally necessary to file a tax form if your income is indeed zero, although it is required for any amount above 0. We will

	All				Male				Female			
Cohort	Golden	PAYE	SA	Either	Golden	PAYE	SA	Either	Golden	PAYE	SA	Either
1998	14,487	11,646	2,310	12,226	6,927	5,528	1,351	5,875	7,560	6,118	959	6,351
1999	22,621	18,410	3,447	19,354	10,590	8,529	1,912	9,063	12,031	9,881	1,535	10,291
2000	23,506	19,214	3,425	20,176	10,853	8,761	1,908	9,322	12,653	10,453	1,517	10,854
2001	23,924	19,921	3,108	20,818	11,025	9,060	1,759	9,625	12,899	10,861	1,349	11,193
2002	23,891	20,104	2,814	20,906	11,060	9,156	1,576	9,642	12,831	10,948	1,238	11,264
2003	23,972	20,387	2,447	21,097	11,024	9,315	1,314	9,726	12,948	11,072	1,133	11,371
2004	23,577	20,367	2,266	20,997	10,767	9,163	1,251	9,526	12,810	11,204	1,015	11,471
2005	25,103	21,800	2,085	22,397	11,439	9,822	1,141	10,183	13,664	11,978	944	12,214
2006	25,383	22,149	1,864	22,589	11,340	9,749	992	10,024	14,043	12,400	872	12,565
2007	25,352	22,303	1,527	22,694	11,292	9,746	774	9,981	14,060	12,557	753	12,713
2008	20,847	18,154	1,039	18,430	8,990	7,704	531	7,872	11,857	10,450	508	10,558
2009	6,510	5,386	426	5,485	3,029	2,452	215	2,509	3,481	2,934	211	2,976
2010	2,993	2,477	152	2,511	1,334	1,082	72	1,101	1,659	1,395	80	1,410
2011	851	721		724	360	291		294	491	430		430
All	263k	223k	27k	230k	120k	100k	15k	105k	143k	123k	12k	126k

Table 10: Number of Golden sample (10% sample of loan database) borrowers and tax data in 2011-12. PAYE (Pay As You Earn) and SA (self-assessment) denotes databases. Either denotes being in either PAYE or SA or both. Cohort denotes the first year the borrower received a loan from the SLC. Data from Britton et al. (2019)

.

record such non-filers as having zero earnings. We end up with the GS for whom we have earnings data from the PAYE database, the SA database or both.

The sample covers the first 11 years of earnings data after people left HE, typically from ages 22 to 32. We find the first 3 years of earnings in the administrative data to be very noisy and so our estimation strategies will highly downweight those data points.

For the rest of life, we model from surveys, using both the British Household Panel survey (BHPS) and Labour Force Survey (LFS). This identically follows Chowdry et al. (2012), which also provides a description of the survey data - the model is detailed in Appendix C. It covers the life-cycle from ages 22 to 65. Chowdry et al. (2012) used the model to estimate the long run cost of English income contingent loans. That work did not:

1. use any administrative data, so potentially suffers from some bias;
2. allow us to see how the values of these loans varies with HEI and subject;
3. place an individual value on the loan book for each person in the actual SLC loan book.

Our approach will be able to deal with each of these problems.

B.2 Details of Admin and survey earnings data

Our Administrative data comes from a database built by Britton et al. (2019) which links HMRC tax records with the SLC's English student loan book. This Admin data contains individual data on former students who took out student loans from 1998 onwards. It covers the first 11 years of data after people left HE, typically from ages 22 to 32.

For each individual, for each tax year, we have the Admin record of real earnings, age, HEI, subject studied and gender. Britton et al. (2019) document how the distribution of earnings varies by cohort, gender, HEI and subject studied. Our analysis of men and women is carried out entirely separately.

We roughly have ten thousand men in each Admin cohort in our Admin data, and around 10% more women in each cohort.

Our survey model comes directly from Dearden et al. (2008) and is detailed in Appendix C. It covers the life-cycle from ages 22 to 65. This model has been used to value English income contingent loans in Dearden et al. (2008). That work did not:

1. use any Admin data.
2. allow us to see how the values of these loans varies with HEI and subject.
3. place an individual value on the loan book for each person in the actual SLC loan book.

Our approach will be able to deal with each of these problems.

C IFS survey model for earnings paths

C.0.1 IFS parameterized model for $Y_{1:T}|\alpha, \mathcal{F}_0$

In this subsection we will document the specific details of the Dearden et al. (2008) survey model for the earnings path $Y_{1:T,i}|\alpha_i, \mathcal{F}_0$. To remove clutter we suppress dependence upon i in the notation in the rest of this subsection. The details are cumbersome, reasonably conventional in modelling earnings dynamics and can be skipped without loss of understanding.

Earnings model with periodic employment Earnings paths are built using the model of real earnings with

$$Y_t = \begin{cases} \exp(y_t), & \text{if } e_t = 1, \\ 0 & \text{if } e_t = 0, \end{cases}$$

where $\{e_t\}$ is a binary employment series and $\{y_t\}$ is a potential log-earnings series. The length of unemployment up to and including time t is recorded by the recursion

$$D_t = (1 - e_t)(D_{t-1} + 1), \quad \text{where} \quad D_0 = 0.$$

For the employment series we write

$$p_{a1} = \Pr(e_1 = 1|\mathcal{F}_0), \tag{20}$$
$$p_{kj,t} = \Pr(e_t = k|e_{t-1} = j, \mathcal{F}_{t-1}), \quad k, j \in \{0, 1\}, \quad t > 1. \tag{21}$$

We use the functional forms, for $t > 1$,

$$p_{01,t} = \Phi\{-g_{01}(age) - \gamma y_{t-1}\}, \qquad p_{10,t} = \Phi\{-g_{10}(age) - \gamma D_{t-1}\},$$

where Φ is the distribution function of a standard normal, $g_{01}(age)$ and $g_{10}(age)$ are 4-th order polynomials in age at time t (e.g. $g_{01}(age) = \sum_{j=0}^{4} \beta_{01,j} (age)^j$). Of course

$$age = t - 1 + a.$$

Here the age of the individual at time $t = 1$ is denoted a.

Steady log-earnings process Throughout we write $WS(m, s^2)$ to denote a distribution with a mean m and a variance s^2. Sequences of steady log-earnings $\{y_t\}$ are determined by observable predetermined characteristics X_t (age, year, region and ethnicity), a persistent AR(1) shock π_t, and a transitory MA(1) shock ϵ_t:

$$\begin{aligned} y_t &= \beta X_t + \alpha + \sigma_{a,\pi}\pi_t + \sigma_{a,\epsilon}\epsilon_t, && (22) \\ \pi_t &= \rho_a \pi_{t-1} + \eta_t, \quad \eta_t \overset{iid}{\sim} WS(0,1) && (23) \\ \epsilon_t &= \theta\psi_{t-1} + \psi_t, \quad \psi_t \overset{iid}{\sim} WS(0,1). && (24) \end{aligned}$$

where $(\{\eta_s\} \perp\!\!\!\perp \{\psi_s\}) | \alpha, \mathcal{F}_0$. Recall α is the individual effect. The $\sigma_{a,\pi}$, $\sigma_{a,\epsilon}$ and ρ_a are quadratic, quadratic and cubic functions of age, respectively, while θ is assumed to be fixed across ages.

Initializing steady log-earnings Every time employment periods are newly initialized we need a way of starting or restarting the autoregression and moving average terms in the log-earnings dynamic. Throughout the moving average model is initialized from its stationary distribution.

Suppose we need to initialize at time t, so we need a π_{t-1}. What value do we use?

Initialization happens in two different ways:

- Immediately after HE, so $t = 1$, initialized into employment. We then take $\pi_0 = 0$.
- Two steps
  - Following a period of unemployment, employment is achieved at time t with realized log-earnings of y_t. This is modelled using a separate reentry log-wage model

$$y_t = r(age) + \gamma D_{t-1} + \delta y_{t-D_{t-1}-1} + \sigma_\varepsilon \varepsilon_t, \quad \varepsilon_t \overset{iid}{\sim} WS(0,1). \tag{25}$$

    where ε_t is mixed Gaussian and $r(age)$ is a 4-th order polynomial in age. Here $y_{t-D_{t-1}-1}$ is log earnings when last employed. In this recursion we take $y_0, y_{-1}, ...$ as $\log \kappa$, a dummy for never having been employed. Note (25) does not depend upon α.
  - If the person is employed the next period, $t + 1$, we go back to the continual employment log-earnings process imposing

$$\pi_t = 0.35 \frac{(y_t - \beta X_{at} - \alpha)}{\sigma_{a,\pi}} + 0.65\eta_t, \quad \eta_t \overset{iid}{\sim} WS(0,1).$$