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

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

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

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

June 14, 2018

Abstract

Funding Higher Education is a major policy challenge for governments worldwide. The use of income contingent loans is becoming increasingly common due to their relief of credit constraints and the provision of insurance against low returns from attendance. It is important to quantify who pays what to fund these loans and to understand how this varies with the loan contract and fee design. This is highly challenging because it requires the projection of earnings of graduates many years into the future. In this paper, we use linked English Student Loan Company records with information on borrowing, course, and institution to official tax records that give earnings for up to 12 years after graduation. Our innovative econometric methods fuse administrative tax records of graduates since they left university with survey data to allow us to extrapolate through the life cycle for the remaining 15 years of the loan contract. Our methodology and the richness of the data allows us to estimate government subsidies through unpaid loans at the subject and institution level for the first time. We find considerable heterogeneity in both, which has strong implications for higher education policy.

Keywords: Administrative data; Income contingent loans; Higher Education funding; Income dynamics.
JEL: H52, H81, I22, I26, C81

1 Introduction

There is considerable interest in how higher education should be funded, and in particular how the burden should be shared between taxpayers and graduates. While the private returns are high (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)). One of the ways in which higher education is subsidized is through the government provision of student loans. Income contingent student loans (ICLs), a particular type of student loan where the repayment depends on the borrower's income, are increasingly being adopted by governments across the world. ICLs have long been established in countries such as Australia, New Zealand and the UK, and are growing rapidly in the US. More than a quarter of graduates with federally managed loans in the US are on some type of income contingent repayment plan, and around 75% of new borrowers. These loans have the virtue of easing credit constraints and reducing the risk of an individual borrowing to fund their own education. Default rates are very low, but often a significant implicit government subsidy exists due to loans not being repaid before the end of the repayment period.

The first contribution of this paper is to produce the first estimates of government subsidies through unpaid loans at the subject and institution level and to quantify how these contributions would change if the design of these loans were to change. These estimates will be partially based on individual level tax data hard linked with the student loan book. This is crucially important in order to be able to make informed policy decisions on the optimal level and distribution of the government subsidy to higher education.

Our second contribution is to develop a new method to combine administrative records with survey data. We apply this model to the estimation of lifetime earnings paths based on a limited number of years of administrative data, but it has implications for many applications where administrative data exists, but either does not cover the entire period of interest or does not contain all desired information on the individual.

We find considerable heterogeneity in loan subsidy at both the major and institution level, which has strong implications for policy. Under the current English system we estimate that around 20% of the subsidy goes to creative arts majors, while those account for only 10% of graduates. The subsidy to many Science, Technology, Engineering and Maths (STEM) majors is negative. This distribution of the subsidy seems at odds with the stated objectives of many governments to encourage students to study STEM majors. We show that by allowing the maximum loan to vary depending on the performance of its graduates, and to a lesser extend lowering overall loans, means subsidies are more

evenly distributed across courses.

In order to estimate the government subsidy to income contingent loan schemes, earnings of graduates have to be projected many years in the future. Traditionally, the modelling of graduate lifetime earnings paths required for this estimation has been carried out using labour market survey data using relatively conventional panel models from the econometrics literature. Recent research (e.g. Britton et al. (2018)) however shows that administrative earnings data from tax collection records are quite different than that recorded in survey data, which has potentially significant implications for the estimated cost of providing these loans.

A further limitation of this approach has been the lack of information in survey data on the institutions graduates attended and the majors they took. This has meant it has not been possible to estimate the costs of income contingent student loans at the major and institution level, leading to very little understanding on exactly how the government subsidy to higher education is distributed. The large differences in graduate earnings between institutions (Chetty et al. (2017); Dale and Krueger (2002)) and majors (Kirkeboen et al. (2016); Hastings et al. (2013); Britton et al. (2016)) found in the literature indicate the loan subsidy will be very unevenly distributed across institutions. Given the very large subsidies implicit in many income contingent student loan systems - an estimated £8bn per year in the UK context Belfield et al. (2017) - knowing where this subsidy is targeted is of paramount importance to policy making.

We will study the English implementation of income contingent loans, where the scheme is run by the government and repayments are collected through the tax system. We use a unique administrative dataset built by Britton et al. (2018) which links the English student loan book to official tax records. This dataset covers the first 12 years of annual tax records after HE. To harness this data we develop new methods which fuse the administrative data to the corresponding survey data which provides a guide to the rest of the life cycle. The resulting new 'Admin-Survey fusing model' can value each person's loan and to see how the value of these loans changes with gender, HE Institution (HEI) and major studied. As well as valuing the loans, it can estimate other interesting quantities like the quantiles of discounted career earnings and also income tax take for each individual given their educational background and their past tax record. Hence these methods also measure human capital in the tradition of, for example, Jorgenson and Fraumeni (1989) and Jorgenson and Fraumeni (1992).

More abstractly, 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. There is also a stimulating literature on joining administrative and survey data. Leading references include Meyer and Mittag (2015) and Bricker et al. (2015).

This paper will have two significant empirical applications of the methods we develop here. First we use the Britton et al. (2018) database of actual annual income tax records from 2002 to 2013 on individuals from the HE 1999 cohort (as defined by year of entry to HE) to compute their repayments on the loans up to 2014. 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 to specify the Admin-Survey fusing model and then use this 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. We quantify using our fusing model 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 there are existing survey data based model estimates of the expected losses (e.g. Dearden et al. (2008), Chowdry et al. (2012) and Belfield et al. (2017)) and this second exercise does not use any individual data about the actual 2017 cohort. 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 more important from a policy viewpoint as the loans sizes are much larger and the results could potentially be used to redesign policy going forward.

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 **??** and **??** apply the model to estimating the government loan subsidy to the cohorts entering HE in 1999 and 2017 respectively. Section **??** then looks at the impact of changes to the design of the student loan contracts on the subsidy and its distribution across majors and institutions. Finally, Section **??** 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.

In this Section we build account for income contingent loan repayments through time for the i-th individual. This individual is assumed to be in group g of students, e.g. those studying Medicine or Creative Arts or those at Imperial College London. We model genders entirely separately and so gender is not referred to in our notation.

2.1 Notation using cash flows

Let

$$Y^*_{t,g,i} \geq 0, \quad t = 1, 2, ..., T, \quad g = 1, 2, ..., G, \quad i = 1, 2, ..., n_g,$$

be the taxable annual earnings of individual i in the g-th group in year t. Further, $X^*_{t,g,i}$ denotes the total annual repayment of the loan, $V^*_{t,g,i}$ be voluntary capital repayment the graduate decides to make and $L^*_{t,g,i}$ be the loan balance at the end of period t. All quantities are measured in cash, that is time t prices. Time $t \in \{0, 1, 2, ...\}$ is measured in years since the end of HE. Thus, for example, $L^*_{0,g,i} \geq 0$ is the loan balance at the end of HE.

2.2 Contract definition

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

Definition 1 *The income contingent loan contract is indexed by the payment rate β and time t payment threshold*

$$K^*_t = (1 + i_t)(1 + a) K^*_{t-1},$$

where i_t is the inflation rate in period t, and a is the real growth rate in the threshold. In period t, the i-th person in group g makes annual repayment and has loan balance governed by:

$$\begin{aligned} X^*_{t,g,i} &= \min\left\{\beta \max\left(Y^*_{t,g,i} - K^*_t, 0\right) + V^*_{t,g,i}, I^*_{t,g,i}\right\}, \\ L^*_{t,g,i} &= I^*_{t,g,i} - X^*_{t,g,i}, \end{aligned}$$

where

$$I^*_{t,g,i} = (1+i_t)\left\{1 + r(Y^*_{t,g,i}, K^*_t)\right\} L^*_{t-1,g,i}.$$

Here $L^*_{0,g,i}$ *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* $T \geq 0$.

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*

$$r(y,k) = r \min\left\{\frac{\max(y-k,0)}{k\alpha}, 1\right\},$$

where $k, \alpha > 0$ *and* $r \geq 0$.

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.

2.3 Notation using time t real prices

It is very convenient to convert everything to period t prices, using the "time t price index" $P^*_t = \prod_{s=1}^{t}(1+i_s)$, where $P^*_0 = 1$.

In terms of real payoffs at time s using prices at time 0, then $X_{t,g,i} = X^*_{t,g,i}/P^*_t$, real incomes $Y_{t,g,i} = Y^*_{t,g,i}/P^*_t$, real thresholds $K_t = K^*_t/P^*_t$, real voluntary repayments $V_{t,g,i} = V^*_{t,g,i}/P^*_t$ and real balances $L_{t,g,i} = L^*_{t,g,i}/P^*_t$. The following Lemma relates these terms, the trivial Proof is in Appendix A.1.

Lemma 1 *Assume* $r(\gamma y, \gamma k) = r(y,k)$ *for all* $\gamma, k > 0$ *and* $y \geq 0$, *then* $K_t = (1+a)K_{t-1}$. *Using the time* 0 *price level, then*

$$\begin{aligned} X_{t,g,i} &= \min\left\{\beta \max\left(Y_{t,g,i} - K_t, 0\right) + V_{t,g,i}, I_{t,g,i}\right\} \\ L_{t,g,i} &= I_{t,g,i} - X_{t,g,i}, \\ I_{t,g,i} &= \left\{1 + r(Y_{t,g,i}, K_t)\right\} L_{t-1,g,i}, \quad L_{0,g,i} = L^*_{0,g,i}. \end{aligned}$$

Sometimes it is useful to use the time t price level for items which appear at time s, these will be written using the transformation

$$X^*_{s,g,i}\frac{P^*_t}{P^*_s} = X_{s,g,i}P^*_t.$$

It is this form, expressed in real terms, together with the loan contract in Lemma 1, which we use throughout our paper.

2.4 Quantifying repayments, taxes and earnings

The total real value using period t prices of future repayments $X_{(T^A+1),g,i}, ..., X_{T,g,i}$ by the graduate over the remaining loan period will be

$$P_t^* \sum_{s=t+1}^{T} \left(\frac{1}{1+d}\right)^{s-t} X_{s,g,i},$$

where d is a real discount rate.

An (risk neutral) expected present value at time t of the repayment stream from the i-th individual is

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

where $X_{s,g,i}|\mathcal{F}_{t,g,i}$ is a forecast distribution with $\mathcal{F}_{t,g,i}$ being the information we use to perform the calculation at time 0 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

$$LOSS_{t,g,i}^* = L_{t,g,i}^* - \mathcal{PV}_{t,g,i}^*.$$

Remark 1 *The UK public accounts call $LOSS_{0,g,i}^* = LOSS_{0,g,i}^*$ the "Resource Allocation Budget" (RAB) charge and is often reported as a percentage $RAB_{i,g} = 100LOSS_{0,g,i}^*/L_{0,g,i}$.*

Of some interest is the probability the i-th individual fully repays their loans,

$$FULL_{t,g,i} = \mathrm{E}(1_{L_{T,g,i}} = 0|\mathcal{F}_{t,g,i}). \tag{2}$$

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

$$Y_{g,i}^* = P_t^* \sum_{s=1}^{T'} \left(\frac{1}{1+d}\right)^{s-t} Y_{s,g,i}.$$

Now $Y_{g,i}^*$ 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 report quantiles of 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

$$TAX_{t,g,i}^* = P_t^* \sum_{s=1}^{T'} \left(\frac{1}{1+d}\right)^{s-t} S_{s,g,i}, \quad S_{s,g,i} = \sum_{j=1}^{\mathcal{G}} \beta_j^{\mathcal{T}} \max(Y_{s,g,i} - \mathcal{K}_j^{\mathcal{T}}, 0), \tag{3}$$

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. Here $S_{s,g,i}$ is the real tax payments paid by the i-th individual at time s.

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* $\beta_3^{\mathcal{T}} = 0.05$.

Example 3 *The US Income based repayment (IBR) scheme, which has* $\beta = 0.1$, $T = 20$, K_t^* *is 150% of poverty guideline for household size (e.g. for a 1 person household the guideline is* $\$12,060$ *in 2017, while for a 4 person household it is* $\$24,600$*). For graduates working for the government or non-profit sector loans are forgiven after 10 years. The forgiven loan amount is taxed as income. These types of loans were made available to students taking out loans after July 2014 and around* 52% *of those taking out loans used this form of loan. Variants of this scheme is the Pay As You Earn loan, Income-Contingent Repayment and REPAYE.*

The papers estimands will be, for each group g using prices in year T^A and information available at time T^A, average loan losses, average rate of full repayments and average RAB charges

$$\overline{LOSS}^*_{T^A,g} = \frac{1}{n_g}\sum_{i=1}^{n_g} LOSS_{T^A,g,i}, \quad \overline{FULL}^*_{T^A,g} = \frac{1}{n_g}\sum_{i=1}^{n_g} FULL_{T^A,g,i}, \quad RAB_g = \frac{1}{n_g}\sum_{i=1}^{n_g} RAB_{i,g}.$$

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

$$TAX^*_{0,g,i}, \quad Y^*_{g,i},$$

given $\mathcal{F}_{T^A,g,i}$.

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 see

$$\mathbf{V}_{1:T^A,g,i} \quad \text{and} \quad \mathbf{Y}_{1:T^A,g,i}.$$

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 distributional properties of the cross-sectional marginals 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

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

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. (2018) 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

In this Section we build models for the earnings paths of groups of students, e.g. those studying Medicine or Creative Arts or those at Imperial College London. We model genders entirely separately and so gender is not referred to in our notation. As the Admin data only goes up to those aged T^A we need to extrapolate it to cover times $T^A + 1$ up to time T and our only available source of extrapolation is the survey data on graduate earnings. The challenge is the survey data is flawed as it often has the wrong levels of earnings and is top coded and so is uninformative in the right hand tail of earnings (e.g. Jenkins (2017)).

In particular, Britton et al. (2018) show there are important differences in the year by year marginal distributions of the admin data and the survey data. Student loan repayments are calculated exactly off the tax records and so in the context of our paper we regard the admin data as the truth when it is available, viewing the survey data as potentially biased although still valuable as it is our only source of latter life-cycle information about graduate earnings. These differences make the direct use of conventional missing data methods for joining different datasets problematic.

To overcome these issues we build a new type of model which we call the "Fusing model". Before we detail it we establish some notation.

Before we start we 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

$$Y_{t,g,i}, \quad t = 1, 2, ..., T^A, \quad g = 1, 2, ..., G, \quad i = 1, 2, ..., n_g,$$

where i is the i-th individual in the g-th group in year t. The raw Admin data is sometimes transformed through

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

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 Imperial College London students 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.

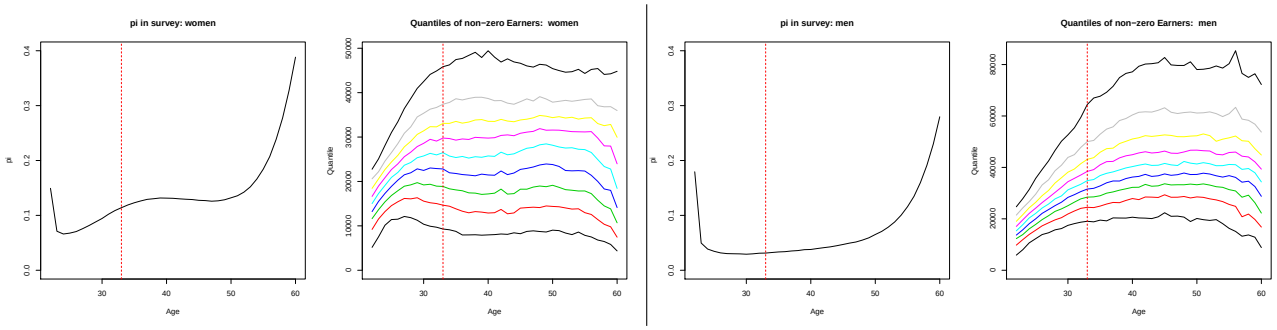

Figure 1: From the survey data. LHS: estimates of π_t^S and $Q_t^{*S}(q)$ plotted against time for women. For we plot quantiles for $q = 0.1, 0.2, ..., 0.9$. RHS: corresponding results for men. Source: `Basic.r`.

The next subsection deals with the margins, while subsection 3.4 tackles the copula. Subsection **??** is devoted to forecasting, while subsection **??** 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

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

  where $\pi_{T^A|g}$ is the proportion of zero earners and $Q^*_{t|g}(q)$ 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

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

  where π_t^S is the proportion of zero earners and $Q_t^{*S}(q)$ 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_t^S, $Q_t^{*S}(q)$ for the survey data for female graduate earnings, while the right hand side shows the corresponding results for men.

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.

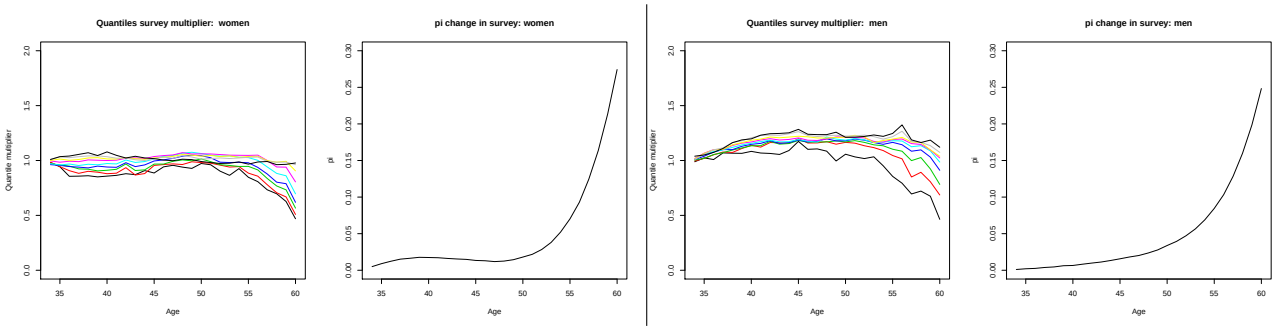

Figure 2: LHS: from the survey data estimates of the multiplier $Q_t^{*S}(q)/Q_{T^A}^{*S}(q)$ and shift $\pi_t^S - \pi_{T^A}^S$ plotted against time for women for $t = T^A + 1, ..., T$. RHS: corresponding results for men. Source: `Basic.r`.

The core marginal distributions of this extrapolation are carried out using the "Admin-Survey Fusing Skeleton," which are the quantile functions for the forecast distribution for group g,

$$Q_{(T^A+1)|T^A,g}(q), ..., Q_{T|T^A,g}(q), \quad q \in [0,1].$$

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

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

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*

$$Q_{t|T^A,g}(q) = \begin{cases} 0, & q \leq \pi_{t|T^A,g} \\ Q^*_{t|T^A,g}\left(\frac{q-\pi_{t|T^A,g}}{1-\pi_{t|T^A,g}}\right), & 1 > q > \pi_{t|T^A,g}. \end{cases} \tag{4}$$

Here

$$\pi_{t|T^A,g} = \min\left\{\max\left(\pi_{T^A|g} + \pi_t^S - \pi_{T^A}^S, 0\right), 1\right\},$$

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

$$Q^*_{t|T^A,g}(q) = \frac{Q_t^{*S}(q \wedge \overline{q})}{Q_{T^A}^{*S}(q \wedge \overline{q})} Q^*_{T^A|g}(q), \quad \textit{recalling } q \wedge \overline{q} = \min(q, \overline{q}). \tag{5}$$

The equation (5) 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. We do this as we have no significant quantity of survey data with group information.

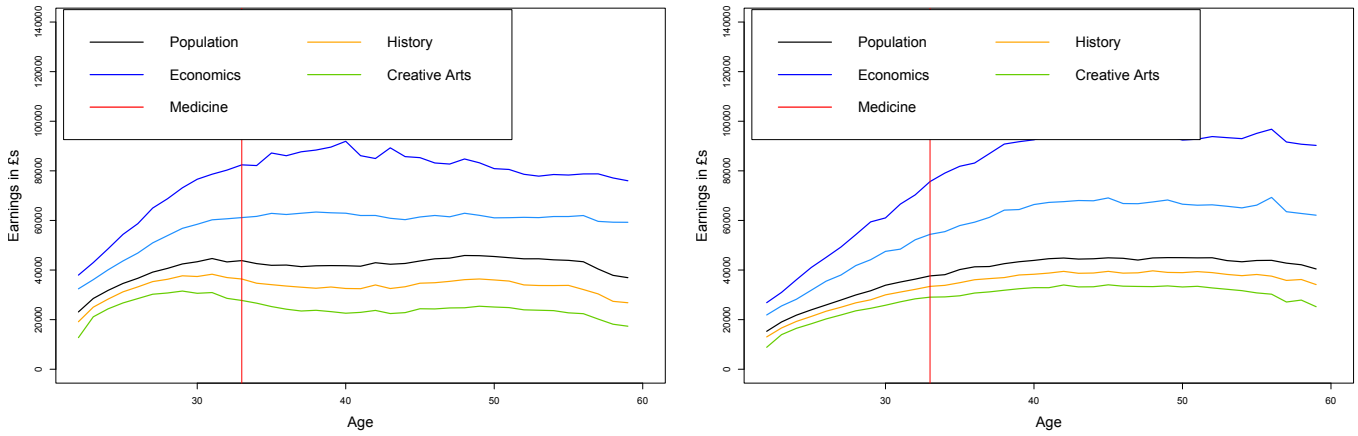

Figure 3: Admin-Survey fusing skeleton $Q_{t|T^A,g}(q)$ for $q = 0.5$: Median earnings for female (left) and male (right) borrowers that entered university in 1999, assuming they entered university at age 18. Earnings are in 2017 prices. [Synthetic data]

Figure 2 show survey estimates of the multiplier $Q_t^{*S}(q)/Q_{T^A}^{*S}(q)$ and shift $\pi_t^S - \pi_{T^A}^S$ 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 **??**. 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 2017 prices, with seven groups split up according to gender. Up to age 33, 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.

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 two copulas: one for men and one for women.

We note 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". It only has two parts: a model for men and a model for women.

Our approach is to take R simulations (e.g. $R = 200,000$) from the IFS graduate earnings model,

which we will write as $X^M_{t,i}$. Then we jitter each survey data point by adding stand uniform noise

$$X^S_{t,i} = X^M_{t,i} + U_{t,i}, \quad U_{t,i} \overset{iid}{\sim} U(0,1), \quad i = 1, 2, ..., R,$$

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

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

As R is massive there should be very little estimation error in this computation.

Then we compute

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

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

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

In practice $\overline{V}_t$ will be tiny by construction, while again as R is large there should be very little estimation error. Write the matrix of correlations as $\widehat{\boldsymbol{\rho}}$, with t, s-th entry $\widehat{\boldsymbol{\rho}}_{t,s}$.

Figure 4 plots $\left\{\widehat{\boldsymbol{\rho}}_{t,s}\right\}$ for the IFS graduate earnings model for each value of $t > s$. The lowest black line shows $\widehat{\boldsymbol{\rho}}_{t,1}$ 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.

3.4.2 Copula based forecasts

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

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

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

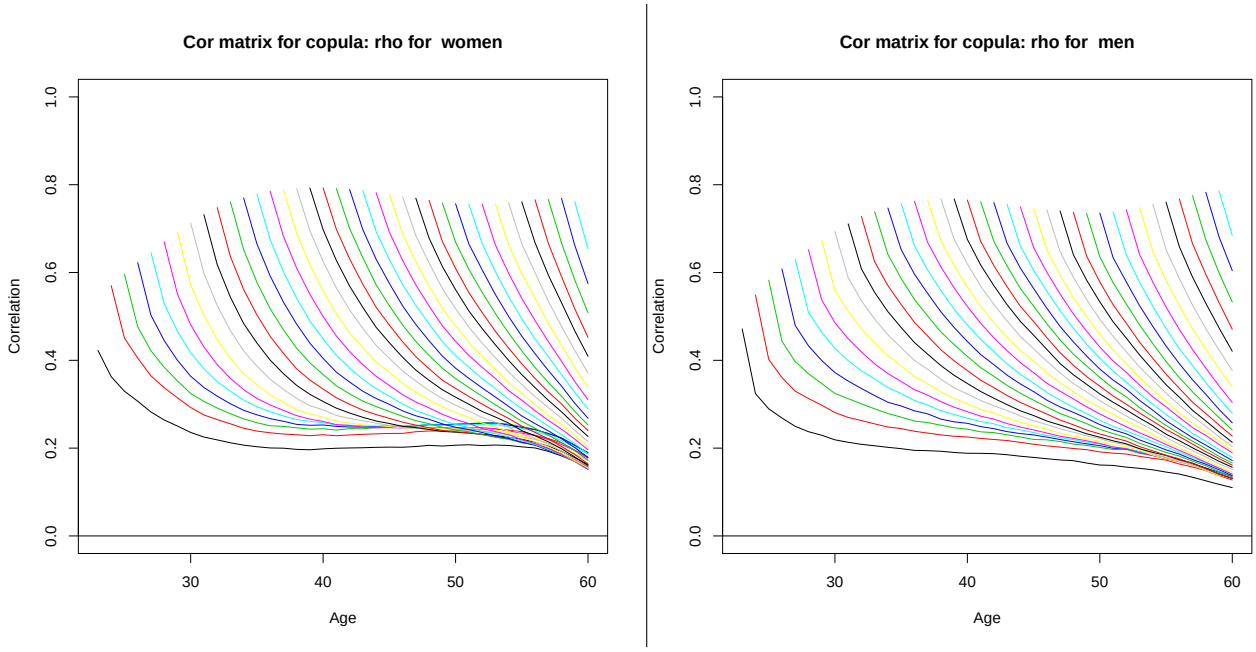

Figure 4: Measure of dependence in the Survey earnings model. LHS: the (cross-sectional) correlation $\widehat{\boldsymbol{\rho}}_{t,s}$ between transformed earnings at times s and t for women. RHS: shows the corresponding result $\widehat{\boldsymbol{\rho}}_{t,s}$ for men. Source: `Basic.r`.

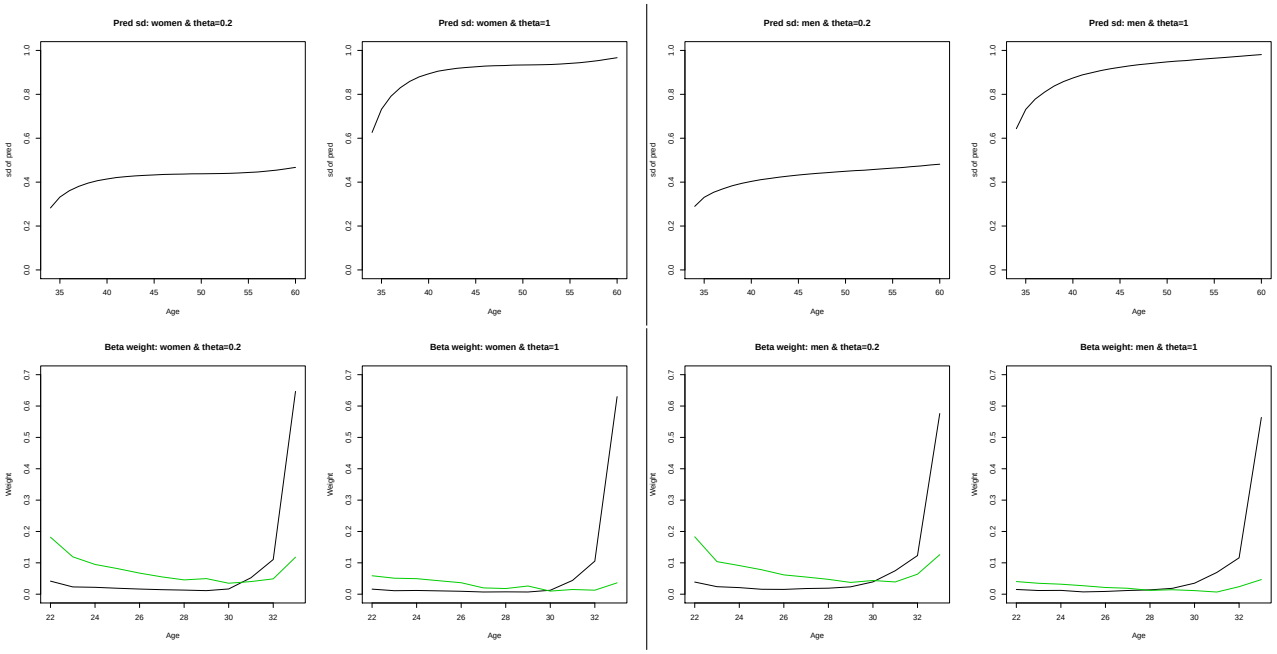

Figure 5: Top row: Measure of predictive uncertainty in Σ: square root of diagonal elements of Σ. LHS: results for women, first for $\theta = 0.2$ then for $\theta = 1$. RHS: shows the corresponding results for men. Bottom row: LHS: first (black line) and last row (red line) of β weights for $\theta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

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

Recall $v_{1:T^A,g,i}$ is the transformed admin data from group g. Then we have a simulated future earnings

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

Hence a single earnings path is

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

3.6 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:

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

where

$$\begin{aligned} V_{t,g,i} &= 0.15 \left(1_{t<10}\right) + V_{t-1,g,i} + 0.1 \varepsilon_{t,g,i}, \quad \varepsilon_{t,g,i} \overset{iid}{\sim} N(0, 1), \\ V_{0,g,i} &\sim N(\log(10000), 1). \end{aligned}$$

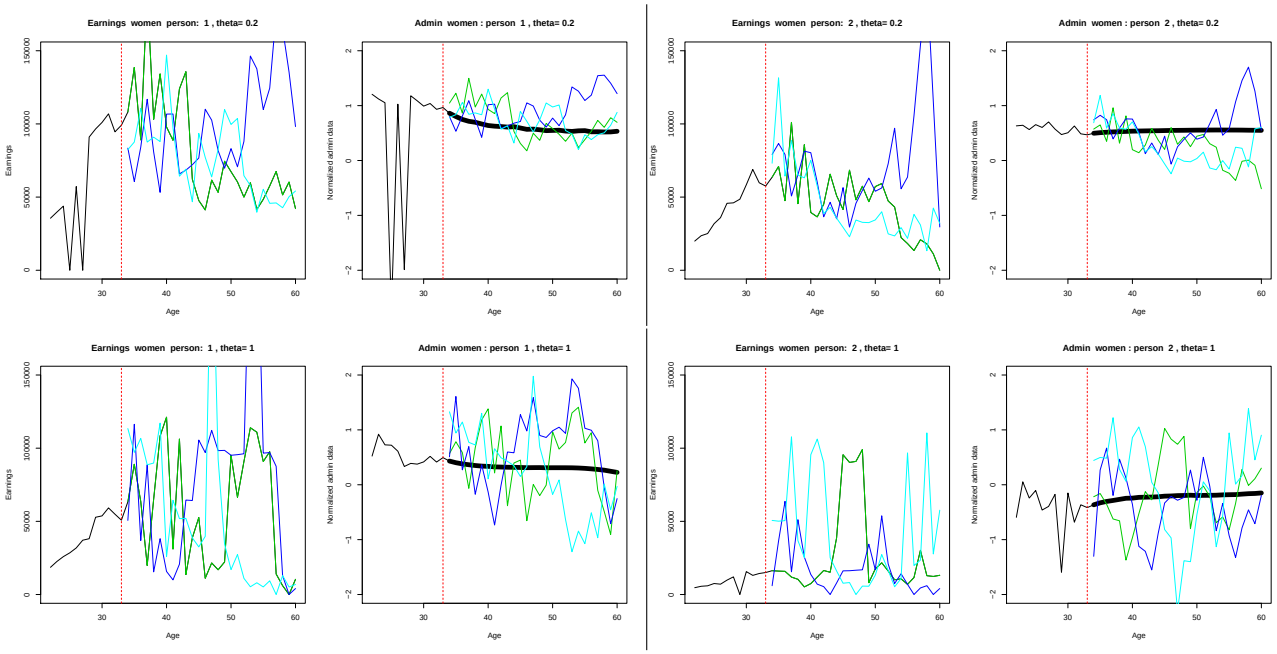

Figure 6: Two synthetic earnings paths for $T^A = 13$ years and 3 simulated paths from their future earnings. First plot is in terms of earnings $\mathbf{Y}_{(T^A+1):T}|\mathbf{Y}_{1:T^A}$, second is carried out on the Gaussian modelling scale $\mathbf{V}_{(T^A+1):T}|\mathbf{V}_{1:T^A}$. The bold lines 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`.

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

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 much lower throughout.

These 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

	$\overline{Y}_{1:t}$			L_t				D_t	$TAX_{1:t}$		
	.25	.50	.75	.25	.50	.75	Aver		.25	.50	.75
$t = T^A$	12.2k	24.1k	46.3k	19.9k	37.2k	40.0k	27.9k	27.9%	22k	57k	189k
$t = T$	21.3k	41.1k	72.8k	0.0	0.0k	19.5k	9.5k	60.0%	224k	519k	1,106k

Table 1: 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 of career averaged incomes up to time t, quantiles and average outstanding loan size at time t, percentage of former students who have repaid their loans fully by time t, quantiles of cumulative tax take up to time t.

$\beta v_{1:T^A,g,i}$. 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 1 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.

A Proofs

A.1 Lemma 1

Now

$$K_t^* = K_t P_t = (1+i_t)(1+a) K_{t-1}^* = (1+i_t) P_{t-1} (1+a) K_{t-1} = P_t (1+a) K_{t-1}.$$

Likewise

$$\begin{aligned} I_{t,i}^* &= (1+i_t)\left\{1 + r(Y_{t,i}^*, K_t^*)\right\} L_{t-1,i}^* \\ &= I_{t,i}^* P_t = (1+i_t)\left\{1 + r(P_t^* Y_{t,i}, P_t^* K_t)\right\} P_{t-1}^* L_{t-1,i} \\ &= \left\{1 + r(Y_{t,i}, K_t)\right\} P_t^* L_{t-1,i}, \quad \text{as} \quad r(\beta y, \beta k) = r(y,k), \end{aligned}$$

so

$$I_{t,i} = \left\{1 + r(Y_{t,i}, K_t)\right\} L_{t-1,i}.$$

Likewise

$$L_{t,i}^* = L_{t,i} P_t = I_{t,i}^* - X_{t,i}^* = P_t I_{t,i} - P_t X_{t,i},$$

so $L_{t,i} = I_{t,i} - X_{t,i}$. Then

$$X_{t,i}^* = \min\left\{\beta \max\left(Y_{t,i}^* - K_t^*, 0\right) + V_{t,i}^*, I_{t,i}^*\right\} = P_t \min\left\{\beta \max\left(Y_{t,i} - K_t, 0\right) + V_{t,i}, I_{t,i}\right\},$$

so

$$X_{t,i} = P_t \min\left\{\beta \max\left(Y_{t,i} - K_t, 0\right) + V_{t,i}, I_{t,i}\right\}.$$

B Institutional background and data

B.1 Administrative data

Our administrative dataset is a database we built and described in Britton et al. (2018), 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 2017 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 2). 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).

	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 2: 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 (2016)

B.1.3 Basic summaries of the Golden sample

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 2 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 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 12 years of earnings data after people left HE, typically from ages 22 to 33. 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. (2018) 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 12 years of data after people left HE, typically from ages 22 to 33.

For each individual, for each tax year, we have the admin record of real earnings, age, HEI, subject studied and gender. Britton et al. (2016) 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

$$p_{a1} = \Pr(e_1 = 1|\mathcal{F}_0), \tag{6}$$
$$p_{kj,t} = \Pr(e_t = k|e_{t-1} = j, \mathcal{F}_{t-1}), \quad k, j \in \{0, 1\}, \quad t > 1. \tag{7}$$

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

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

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:

$$y_t = \beta X_t + \alpha + \sigma_{a,\pi}\pi_t + \sigma_{a,\epsilon}\epsilon_t, \tag{8}$$
$$\pi_t = \rho_a \pi_{t-1} + \eta_t, \quad \eta_t \overset{iid}{\sim} WS(0, 1) \tag{9}$$
$$\epsilon_t = \theta\psi_{t-1} + \psi_t, \quad \psi_t \overset{iid}{\sim} WS(0, 1). \tag{10}$$

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

$$y_t = r(age) + \gamma D_{t-1} + \delta y_{t-D_{t-1}-1} + \sigma_\varepsilon \varepsilon_t, \quad \varepsilon_t \overset{iid}{\sim} WS(0, 1). \tag{11}$$

    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 (11) 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

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

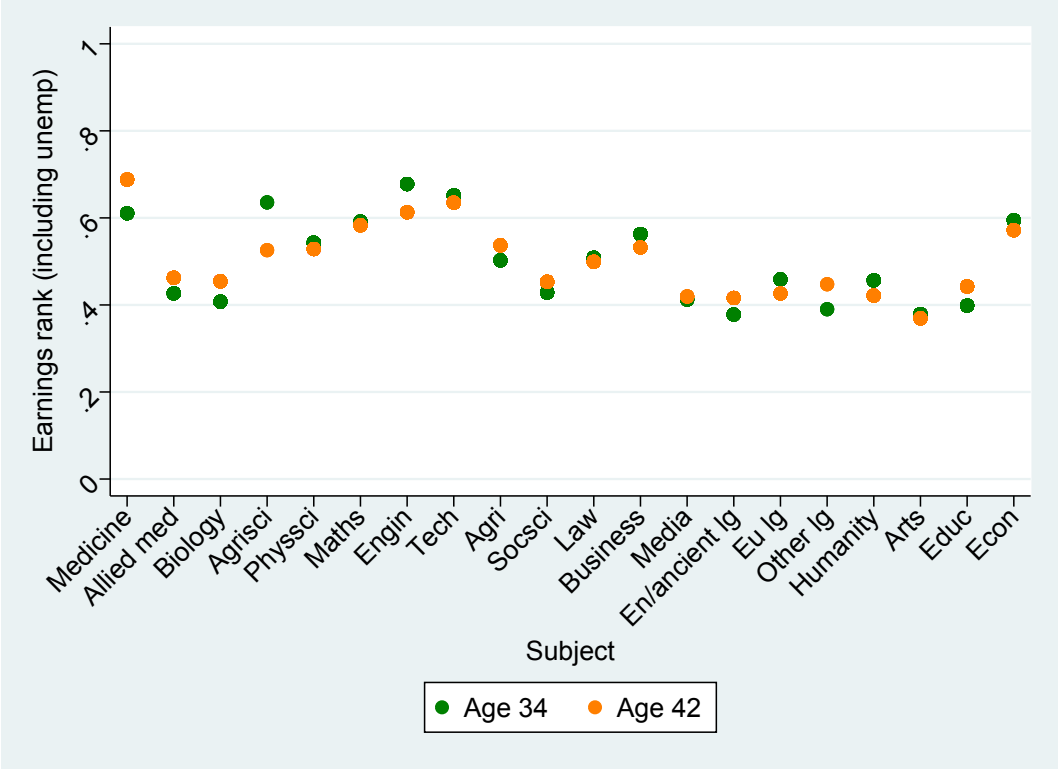

Figure 7: Scaled mean earnings ranks by subject in the BCS age 34 and 42

C.1 Checking time-invariance of individual effects

We use earnings data up to age 33, the last available age in the HMRC data, to construct the scaled ranks by subject. This would be a problem if earnings ranks by subject vary over the lifecycle. To investigate this issue, we have used earnings data from the British Cohort Study (BCS) which follows individuals born in the UK in 1970 throughout their life and records information on employment and earnings a regular intervals. Figure 7 shows the comparison between earnings ranks by subject at age 42, the last available year in the BCS and at age 34 (there was no survey at age 33). We see that subject earnings ranks stay remarkably constant throughout the lifecycle, particularly given we end up with small sample sizes when we look at graduates by subject in the BCS and hence would expect to see some noise. This provides support for the notion that estimating subject-specific fixed effects using earnings up to age 33 gives an accurate estimate of subject-specific fixed effects over the lifecycle.

For computational reasons we are limited in the number of parameters of the group-specific distribution we can try and estimate. We therefore only estimate the mean μ_g and standard deviation σ_g of each group-specific distribution. These are estimated by minimizing the distance between quantiles of the simulated data and HMRC data for each group. The quantiles we match on are the 25th, 50th, 75th, 90th, 95th, 99th percentiles, putting equal weight on each of those quantiles. As small differences at the bottom of the income distribution do not impact the estimates of the government cost of providing income contingent student loans due to the £21,000 threshold below which no repayments are made, we match on more quantiles at the top of the income distribution.

References

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

Belfield, C., J. Britton, and L. van der Erve (2017, Oct). Higher education finance reform: Raising the repayment threshold to 25,000 and freezing the fee cap at 9,250. Technical report.

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. Amsterdam: Elsevier.

Bricker, J., A. M. Henriques, J. A. Krimmel, and J. E. Sabelhaus (2015). Measuring income and wealth at the top using administrative and survey data. Finance and Economics Discussion Series 2015-030. Washington: Board of Governors of the Federal Reserve System.

Britton, J., L. Dearden, N. Shephard, and A. Vignoles (2016). How english domiciled graduate earnings vary with gender, institution attended, subject and socio-economic background. Institute of Fiscal Studies working paper W16/06.

Britton, J., N. Shephard, and A. Vignoles (2018). 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. Unpublished paper: Economics Department, Stanford University.

Chowdry, H., L. L Dearden, A. Goodman, and W. Jin (2012). The distributional impact of the 201213 higher education funding reforms in uppercaseEngland. *Fiscal Studies 33*, 21136.

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*. Princeton University Press.

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? NBER Working Papers 20913.

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.

Jenkins, S. P. (2017). Pareto models, top incomes, and recent trends in uk income inequality. *Economica 261–289*.

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.

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.

Meyer, B. D. and N. Mittag (2015, October). Using linked survey and administrative data to better measure income: Implications for poverty, program effectiveness and holes in the safety net. Working Paper 21676, National Bureau of Economic Research.

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.