EVALUATING AND DESIGNING STUDENT LOAN SYSTEMS: AN OVERVIEW OF EMPIRICAL APPROACHES

Lorraine Dearden
University College London and Institute for Fiscal Studies

Abstract

To understand and design student loan systems, it is important to have appropriate earnings and/or income projections for current and future graduates. In this paper, Current Population Survey (CPS) data from the US is used to demonstrate empirical approaches that can be exploited to simulate lifetime income and earnings profiles that are needed to understand and design effective and sustainable student loan systems. The crucial element in getting this analysis right, is having reliable simulations of the whole distribution of future graduate earnings and income. Typically in this literature, the repayment burdens (RBs) of student loans are calculated at different percentiles of the graduate income or earnings distribution. Often unconditional quantile regression (UQR) are used to calculate age earning/income profiles for different quantiles. The paper shows that this approach has its limitations when evaluating student loans and simple raw quantile estimates by age with some age smoothing is preferable. This approach can be used even in countries where income is censored and recorded in income bands. The paper shows a simple way of incorporating dynamics using these quantile profiles even when individuals only have access to very short panel data. This involves using copula functions drawing on earlier work by Dearden et. al (2008) and Bonhomme and Robin (2009). Having reliable dynamic estimates turns out to be important in assessing not only the taxpayer costs of designing an ICL but for correctly assessing the extent of loan repayment hardship for individuals.

Acknowledgements

I am grateful for extensive and insightful comments from Bruce Chapman, Jean-Marc Robin and two anonymous referees. Financial assistance received from the ESRC funded Centre for the Microeconomic Analysis of Public Policy at IFS (grant RES-544-28-5001) and the HEFCE and ESRC funded, Centre for Global Higher Education at UCL – IOE (Grant no. ES/M010082/1) is gratefully acknowledged.

JEL Codes: H28, I22, I28, J24
Keywords: Student loans, Student loan design, Repayment burdens, Copula functions.
Email: l.dearden@ucl.ac.uk

1 Introduction

To understand and design student loan systems, it is important to have appropriate earnings and income projections for current and future graduates. In this paper, Current Population Survey (CPS) data from the US is used to demonstrate some empirical approaches that can be used to simulate lifetime income and earnings profiles that are needed to both understand and design effective and sustainable student loan systems. The crucial element in getting this analysis right, is having reliable simulations of the *whole* distribution of current and future graduate earnings or income. Using income or earnings of example or average or median graduates is almost always not sufficient and often misleading.

Chapman and Lounkaew (2015) in their paper on US Stafford loans showed the importance of looking at loan repayment burdens (RBs) across the entire income distribution of US graduates by age to understand the current crisis in the US student loan system. The RB of a loan is the proportion of income[1] each period that is used by the debtor to repay a loan. Chapman and Lounkaew (2015) used Current Population Survey (CPS) data from 2009 (uprated to 2015) and showed convincingly that the RBs involved with the current US Stafford Loan system are very high and unsustainable for low earning graduates. The innovation in their paper was to calculate the entire distribution of graduate incomes by age and sex. For this, they used the unconditional quantile regression (UQR) method proposed by Firpo, Fortin and Lemieux (2009).

Many other authors now routinely use quantile methods to calculate RBs at different parts of the income distribution and to design alternative loan systems including income contingent loans (ICLs). Most of the papers in this special issue have used national cross-sectional data to estimate RBs for a particular country *across the distribution of income or earning*s using different methods including UQRs. These papers then use the result from this exercise to design ICL systems and to estimate the potential subsidies involved in such schemes.

This paper contributes to the literature in three important ways. First it shows that whilst UQRs are generally fine in this context, there can problems for the unwary, particularly if the UQR

[1] In this paper graduate income *and* earnings by age are simulated. Income is generally used for repayment burden analysis whereas income contingent loan systems generally apply to labour earnings.

model is incorrectly specified. This contrasts with most other applications using quantile regression approaches where UQR is more appropriate. This is demonstrated using pooled 2014-2017 US CPS data and shows that calculating raw percentiles conditional on age coupled with some age smoothing procedure gives much more reliable estimates of the quantiles of the income and earnings distribution that need to be used for RB analysis and student loan design. Moreover, the UQR will underestimate the RB problem.

In some countries, such as Japan (see Dearden and Nagase (2017)), income and earnings data are censored and reported in income groups or bands which makes distributional analysis problematic. The paper uses CPS data to show that typical grouped income and earnings data can be easily turned into reliable percentile estimates by age using either interval regression or simply midpoints *combined* with age smoothing. It is crucial for loan design to have income or earnings measured by percentile to capture distributional features of the loan system and for estimating the budgetary implications of different loan designs. The age smoothed predictions from this modelling exercise track the actual raw percentiles by age well, except for high incomes/earnings for males where the bands are wide or right censored. Fortunately, these high-income groups are generally the least important for estimating RBs and designing student loan systems.

The paper shows that to get RB analysis and student loan design right, incorporating earning or income dynamics is important. This generally requires panel data with reasonably large sample sizes (N) which follows individuals over a reasonable time (T). In most countries with ICLs, dynamic simulation models are used to estimate taxpayer costs (see for example Crawford, Crawford and Jin (2014), Higgins and Sinning (2013) and Britton *et. al* (2018)). In a lot of countries this type of panel data is not available or sample sizes are very small. In most countries, however, there are surveys which are the source of International Labor Organisation (ILO) labor force statistics, which may be used for this purpose. In Australia, Japan, France and the UK, it is the labour force survey (LFS). In the US it is the Current Population Survey (CPS) and in Colombia it is the Gran Encuesta Integrada de Hogares (GEIH).

Most of these surveys have rotating panels which mean that individuals are in the survey for several months or quarters. This means the data can be used to look at employment, income and wage transitions over a year (very short T). This panel element can be exploited to simulate

earnings dynamics in a simple but sophisticated way which involves directly using the percentile earnings and income estimates by age used in typical RB analysis. This is done using copula functions. Previous examples of using this copula function approach to estimate dynamics include Dearden et. al. (2008) for England and Bonhomme and Robin (2009) for France. The novelty of what is proposed in this paper is that by using a simple copula function method one can easily recast the data used to calculate RBs into dynamic income or earnings predictions which appear to match the dependence structure of the observed panel data well. This can then be used to simulate income and earnings dynamics for future graduates, crucial for loan design and RB analysis.

Estimating income and earnings dynamics has two important implications. First, it allows one to look at RBs at different percentiles of income over the *term of the loan* which adds an extra dimension to understanding the RB problem. It turns out that RB problems are likely to affect a much larger proportion of graduates than one would estimate using cross-sectional data which implicitly assumes no mobility. Second, it allows one to more reliably estimate the costs of an ICL system as incorporating appropriate dynamics is crucial to getting this right as shown in papers such as Higgins and Sinning (2013) and Dearden et. al. (2008). This finding is confirmed using *earning* simulations based on US CPS data and shows that the estimated costs of income contingent loan schemes are over-estimated if dynamics are not built into these earning simulations.

Indeed, assuming graduates stay in the same percentile of the earnings or income distribution over their working life will provide an *upper* bound on likely costs of an ICL scheme and a *lower* bound on RB problems with a TBRL. In a country like the US with high earnings mobility, this is especially important. In a low mobility country like Japan this will be less important.

Section 2 describes the CPS data used in the paper. In section 3 unconditional quantile regression methods are compared to smoothed raw percentiles by age and the paper shows that UQR is not always appropriate and probably should not be used in this context. The section also demonstrates how researchers can effectively deal with earnings or income survey data that has been banded (or partially banded) using either interval regression techniques or midpoints, coupled with age smoothing to get age earning profiles across the distribution that

match the actual age earning profiles well. Section 4 shows a simple but sophisticated approach to estimating dynamic lifetime earnings/income profiles when good longitudinal data is not available which involves a simple extension of the approach developed by Chapman and Lounkaew (2015) and copula functions. Section 5 shows the implications of incorporating income dynamics for RB analysis and ICL loan design. Section 6 concludes.

2 US Current Population Survey (CPS) data

All the analysis in this paper uses data from the March income supplement of the US Current Population Survey (CPS) from 2014, 2015, 2016 and 2017. From the CPS a sample of individuals who have completed a 4 year BA degree or higher degree who are aged 23 to 65 are chosen. Data on individual income (from all sources) and labour earnings are used in the analysis in the paper.

The total CPS sample across these 4 years consists of 142,385 observations of whom 64,376 (46 percent) are males and 78,009 (54 percent) are females. A panel is constructed for those subset of BA graduates that are observed in two consecutive years, that is March 2014 and 2015, March 2015 and 2016 or March 2016 and 2017, removing all clearly anomalous cases.[2] This results in a panel of 30,917 individuals (61834 observations) of whom 13,979 are males (46 percent) and 16,938 are females (54 percent). Summary statistics for the income and earnings variables for the whole CPS sample and the CPS panel are given in Table 1 below.

The sample sizes mean that for the panel, we have an average of 330 individuals per age transition for men and 400 per age transition for women. For the CPS cross-sectional data we have around 1500 observations per age for men and 1800 for women. There is of course variation by age in both datasets with the lowest numbers concentrated among young male BA graduates aged 23 to 25 and female graduates aged 60 and above.

[2] This includes change in sex, age going up by more than two years, change in ethnicity, change in where individual and/or parents were born.

Table 1: Summary Statistics for BA graduates: CPS 2014-2017

Gender	Income				Earnings			
	Whole sample		Panel		Whole sample		Panel	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Females	54496	62315	54527	50445	48651	60297	48849	48330
Males	96585	112710	97553	95012	87831	109654	88858	91996
Observations	142385		61834		142385		61834	

2 Estimating Age earnings profiles across the income and earnings distribution

Typically, in the repayment burden/student loan literature, unconditional quantile regressions (UQRs) have been use to estimate smoothed age earnings profiles across the distribution of income and/or earnings. UQRs are important for lots of important questions where causal impact is at the heart of the question. But this is not true for RB analysis or student loan design. To estimate RBs at each age across the distribution of earnings, knowledge of how the q^{th} quantile of actual or 'raw' earnings or income *conditional on age* $Q_q(y|A)$ changes by age. This is because the repayment burden is measured as the loan repayment at age *t* as a proportion of *actual* income at age *t*. An UQR instead identifies the impact of the population aging by one year on the q^{th} quantile of the unconditional earnings $Q_q(y)$ distribution *across all ages*.

When regression (mean) techniques are used then E[$y|A$] averages up to the unconditional mean E[y] over the range of A because of the law or iterated expectations i.e. E[E[$y|A$]]= E[y]. The estimated coefficient on age (and any polynomials) gives the impact of a change in age on both E[y] and E[y|A]. However, for quantile regression this does not hold. Firpo et. al. (2009) show that to get the unconditional effect of your variable of interest (A) on the outcome of interest (y) you need to perform conditional quantile regression (CQR) and then *integrate out* over all the conditioning variables to get the unconditional effect. They show how this can be done using the re-centered influence function (RIF). This however is not needed for RB analysis.

With only one regressor (age and polynomials in age), it turns out not to be generally critical, as long as the polynomial in age is correctly specified. This is sometimes not true at low and high quintiles of the earnings or income distribution and is particularly unstable for percentiles which at some age have zero income or earnings and at other ages non-zero observations.

To illustrate potential problems, age earnings profiles for BA graduates are estimated for every year of the CPS data used in the paper, that is for 2014, 2015, 2016 and 2017 from the age of 23 until 65.[3] The data is used to calculate the raw percentiles of income and earnings by age, sex and year.[4] This is all that is needed to calculate RBs but typically in the RB literature, these profiles are smoothed by using polynomials in age which can help reduce measurement error in income and earnings. Because zero earnings and income are included in these quantile estimates, a quintic in age is necessary to capture the drop-off in earnings during child-rearing ages for women but also for earnings/incomes at the bottom of the distribution where fluctuations in earnings and income are more likely. This is compared to the UQR method advocated by Chapman and Lounkaew (2015). UQR methods turn out to be very sensitive to the functional form used (whether log income/earnings are used or levels and the polynomial in age used) as well as the age range over which the model is estimated. For low and high quantiles it proves to be very unstable. [5]

Most papers in the RB literature use log linear UQR models and set zero earnings to one so that logs can be obtained and then obtain predictions from these models. To do this, one can simply exponentiate the predictions (cf the case with linear regressions) since unlike with regression, there is equivariance of quantiles under log linear transformations (or indeed any weakly increasing monotonic transformations) i.e. $\exp(Q_q(logy|A)) = Q_q(y|A)$. whereas $\exp(E[logy|A]) \neq E(y|A)$.

[3] Chapman and Lounkaew (2015) started their earning profiles at the age of 22. The number of 22 year old graduates in the CPS data is relatively small, so in this paper the base category is 23 year olds and 22 year old graduates are recoded as being 23. Having a sufficiently large sample size is important when calculating percentiles of the income distribution. Like Chapman and Lounkkaew (2015) military personnel are excluded but self-employed and BA graduates who go onto to do further post-graduate study are retained. This makes very little difference to the analysis but increases sample sizes for the dynamic simulation methods employed later in the study.

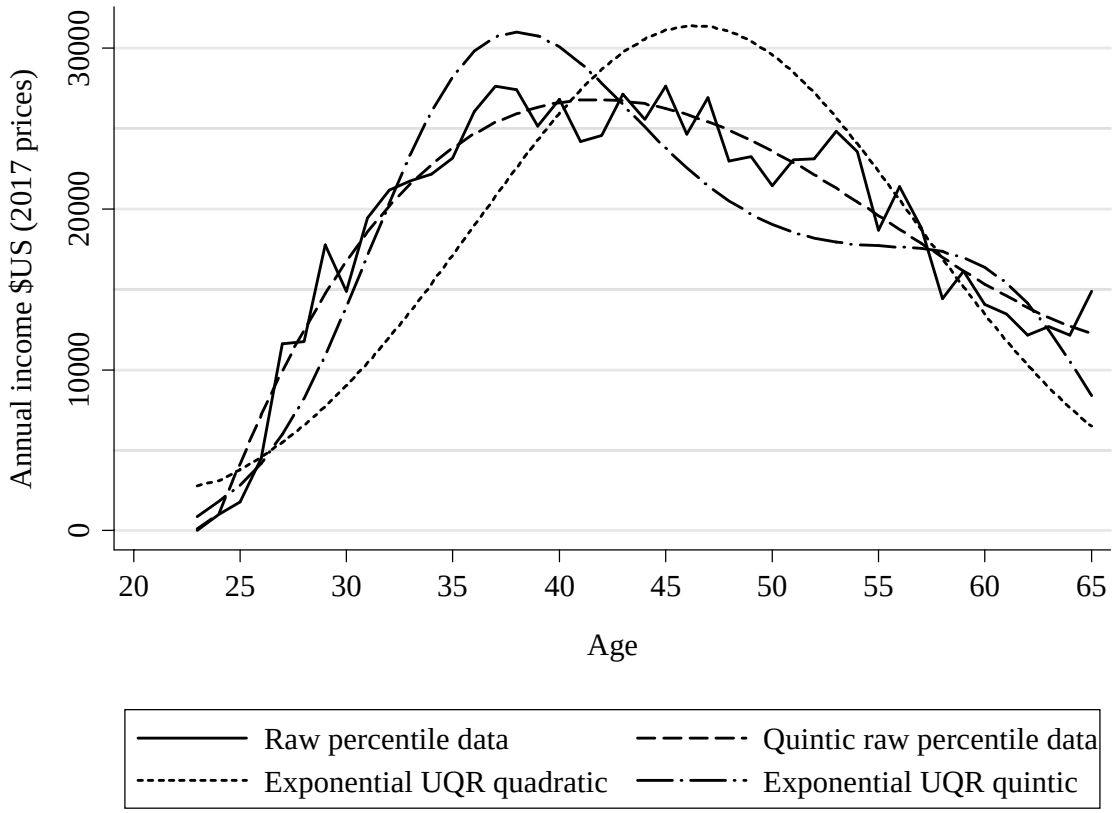
[4] For all the work either Stata 15 and R are used. The CPS data requires weighting so the *_pctile* function in Stata with sample weights is used to calculate the raw percentiles of income and earnings.

[5] The CPS data shows that using UQR with incorrectly specified polynomials in age performs particularly badly at low and high quintiles and is extremely unstable at low and high incomes if mis-specified. For instance, the estimates at the 5th centile for male income varies hugely by CPS year with 3 of the 4 years not producing credible estimates (predicted income way too high). With RB analysis it is crucial to get estimates of profiles at low quantiles correct, hence why the observed instability in these estimates is a major issue for UQR methods.

In Figure 1 estimates from UQR and the preferred approach are compared for the 10th percentile of the male income distribution using 2014-2017 US CPS data in 2017 prices. 'Raw percentile data' is the q^{th} quantile of income at each age. 'Exponential UQR quadratic' is the model used by Chapman and Lounkaew (2015). 'Exponential UQR quintic' is the same model, but includes a much more flexible polynomial in age (quintic). 'Quintic raw percentile data' is the predictions from a linear regression of the raw percentile level data on a quintic in age.

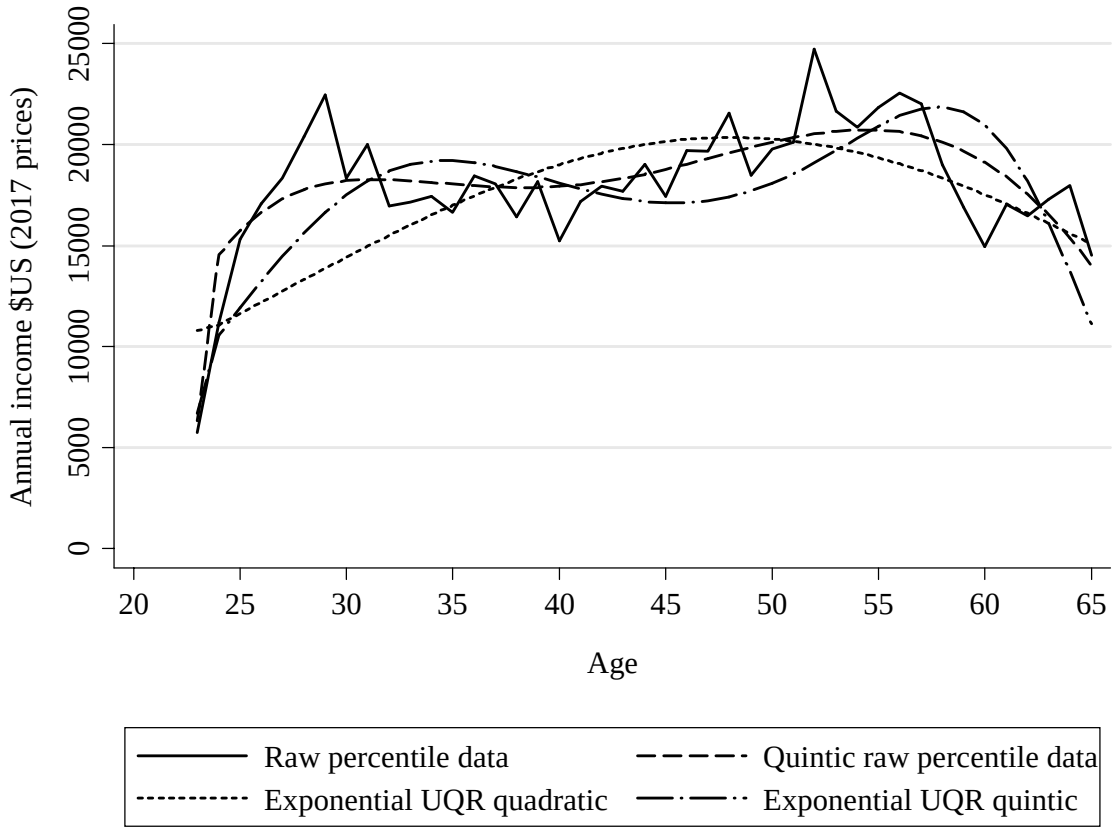
Figure 1 shows that all of the UQR approaches approximate the raw 10th percentile data very poorly over the full range of ages. The best fit is given by running a regression with the raw conditional quantile data as the dependent variable and a quintic polynomial in age as the independent variable and obtaining the prediction from this regression. A quadratic performs equally as well in this case (not shown).

Figure 1: Male BA Graduate 10th percentile of income distribution: comparing methods



In Figure 2, the same exercise is repeated for female BA graduates in the 25th percentile of the income distribution.[6] Again the UQR with either a quadratic or quintic does not replicate the raw percentile data well and the preferred model does much better. The importance of having a quintic specification is also evident here.

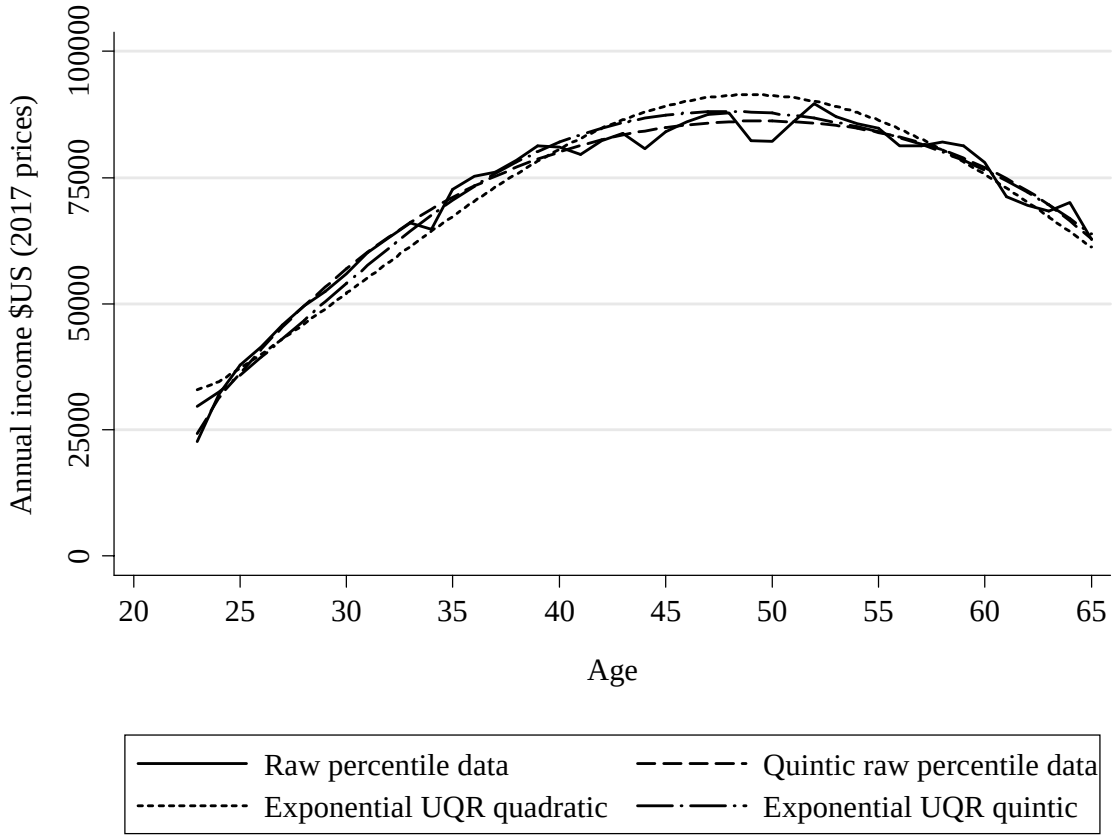
Figure 2: Female BA Graduate 25th percentile of income distribution: comparing methods



In Figures 3 the estimates of median income (50th percentile) for male BA graduates in the sample are shown. Whilst UQR with both quadratic and quintic specification performs quite well for most ages it overestimates income at young ages which is crucial for RB work. For example, at age 23 the overestimate with the quadratic UQR approach is just over £10,000 or 45 percent.

[6] Females below the 20th centile have a very high proportion of zero incomes due to being out of the labour market so these diagrams are not particularly instructive. The UQR approach is also highly unstable at this percentile and produces unrealistically high predictions of smoothed income.

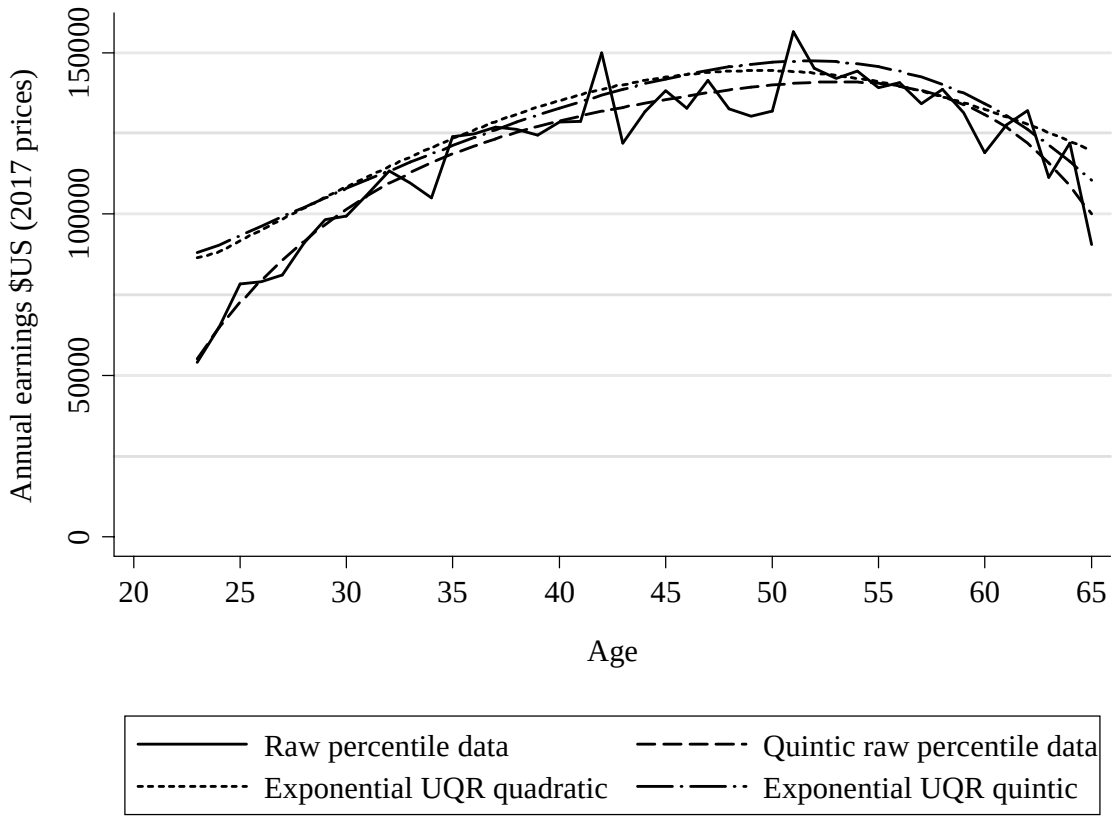
Figure 3: Male BA Graduate 50th percentile of income distribution: comparing quantile approaches



In Figure 4 the implications for high earning graduates are considered by looking at estimates for women in the 95th percentile of graduate *earnings*. Both specifications of the UQR model over-estimate earnings up until about the age of 35 and at low ages this is substantial (by around 60% or just over $32,000 at the age of 23 with the quadratic UQR specification).

This process has been repeated for every percentile of the US income distribution for men and women and the approach which best approximates the raw percentile data, essential for RB analysis, involves smoothing the raw percentile estimates by age using a flexible polynomial in age. UQR methods are unreliable for this exercise. Moreover, these preferred estimates approximate the empirical marginal distribution at each age by percentile well and this feature can be used in conjunction with Copula methods to simulate earnings and income dynamics. This is discussed further in section 4. Using UQR methods also means that the cross sectional RB problem is understated. This is also shown in Section 4.

Figure 4: Female BA Graduate 95th percentile of earnings distribution: comparing quantile approaches



Does this mean that all the studies using UQR methods to calculate RBs are wrong? The answer is not necessarily. The analysis presented here and papers such as Borah and Basu (2013) show that with only one conditioning variable (in RB analysis age), the results from using UQR and conditional quantile methods are less likely to vary if the model is correctly specified. The analysis undertaken using CPS data for this paper shows that a UQR approach with a sufficiently flexible functional form in age generally gives a reasonable estimate of the conditional quantiles of the income or earnings distribution needed for RB analysis and student loan design - except at low and high percentiles of the income and earnings distribution and young ages. [7] The sensitivity of the UQR to model specification means that without careful exploration of the data, estimated earnings profiles used for RB analysis and student loan design may be incorrect.

[7] Where earnings and incomes are zeroes this is not the case but it is true for low values of positive income or earnings and relatively high values of income or earnings. For instance the UQR estimates for the 20th centile of female income were not sensible (way too high) in all years except 2016 and simply not credible.

The US CPS data has all its income and earnings data measured without banding, although to preserve confidentiality there is income swapping procedure applied to prevent the identification of individuals with extremely low or high incomes. However, this is not true for all countries. In Japan for example, all Labour Force Survey earnings and income data is banded into around 20 bands.

Banded data, particularly when the number of bands is small, may limit the ability to look at the entire distribution of income across ages for BA graduates with any accuracy and the estimates will be heavily influenced by the distribution of respondents within each band by age. With appropriate age smoothing of raw quantile data, this problem may be ameliorated but it is an open question. In the Japanese LFS data, like the CPS data, there are lots of rich covariates which should be able to reliably position individuals within their known (log) income band and this can easily be done using interval regression (see Stewart (1983)), and then predicting income, *conditional on the band that the individual is in.* This can be compared with simple age smoothing of midpoint estimates of income for the case where data does not have rich background characteristics.

This is tested by banding the full CPS's income data into 20 income groups. The income groups and the proportion of males and female BA graduates falling into each category is shown in Appendix A Table A1. For non-zero incomes, logs of the lower and upper bands are taken and interval regression[8] performed and then the predicted log income *conditional on being within the observed band* gives us our log income prediction. These predictions are then converted into income levels. [9] The covariates include a cubic in age, year dummies, dummies of grouped total family income, a quadratic in hours of work, detailed industry and occupational dummy variables, ethnicity dummy variables, whether the individual was US born, whether their father and mother were US born, regional dummy variables as well as a metropolitan dummy variable. Clearly these variables are highly endogenous, but the sole purpose of this exercise is to get good predictions of income within bands so endogeneity is highly desirable for this exercise unlike most applications.

[8] Stewart (1983) calls this type of estimation grouped dependent variable estimation.

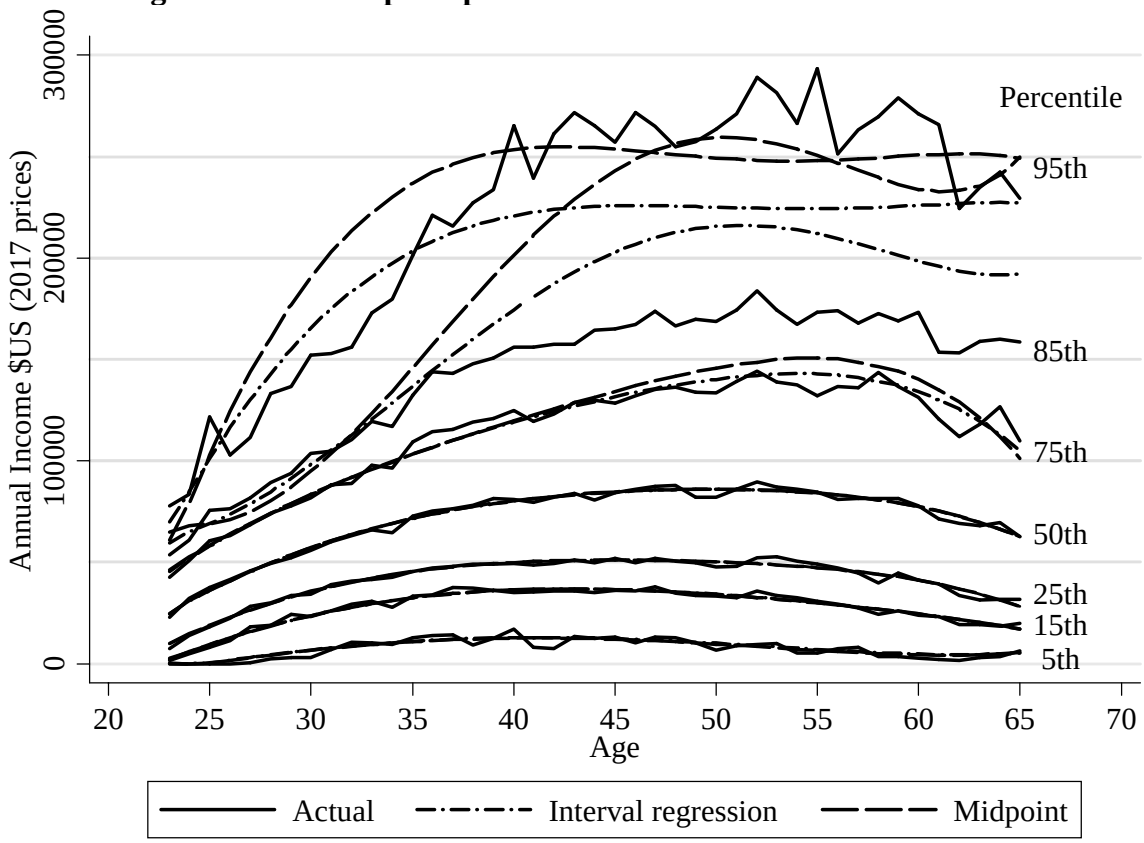
[9] This prediction includes an estimated residual therefore one can simply exponentiate this within band prediction.

From these predictions and the midpoint estimates raw percentiles by age are calculated and a quintic polynomial in age is used to smooth these raw percentile estimates by age, gender and year as was done earlier in this section of the paper.[10] Figure 5 shows the raw and smoothed quantile earnings profiles for males at the 5th, 15th, 25th, 50th, 75th, 85th and 95th percentile based on the interval regression and midpoint age smoothed predictions. Figure 6 shows the corresponding diagram for females.

Figure 5 shows that the age smoothed profiles perform well at all quintiles up until the 75th percentile. With the interval regression approach, earnings are too high for the 85th from the age of 35 and too low for the 95th percentile from the age of 45. The midpoint age smoothed estimates are too high for the 85th percentile and fine for the 95th percentile but this is purely because income is arbitrarily set to be $250,000 if men earned above $150,000, the top income group. The US Stafford Loan generally has to be repaid within 10 years so the estimated income profiles are accurate for the terms of these loans and RB analysis. Further, as Barr et. al (2018) show, high earning graduates actually pay off loans quicker with an ICL than with Stafford Loan so again this will have no implications for ICL loan design work. The problem at high percentiles arises because there is a large proportion of men earning above $100,000 (see Table A1 in Appendix 1) and there are only 3 income bands for this group covering just over 30% of male graduates in the CPS sample. This, however, is less of a problem at lower ages.

[10] For the top band Females are assigned $150,000 and Males $250,000.

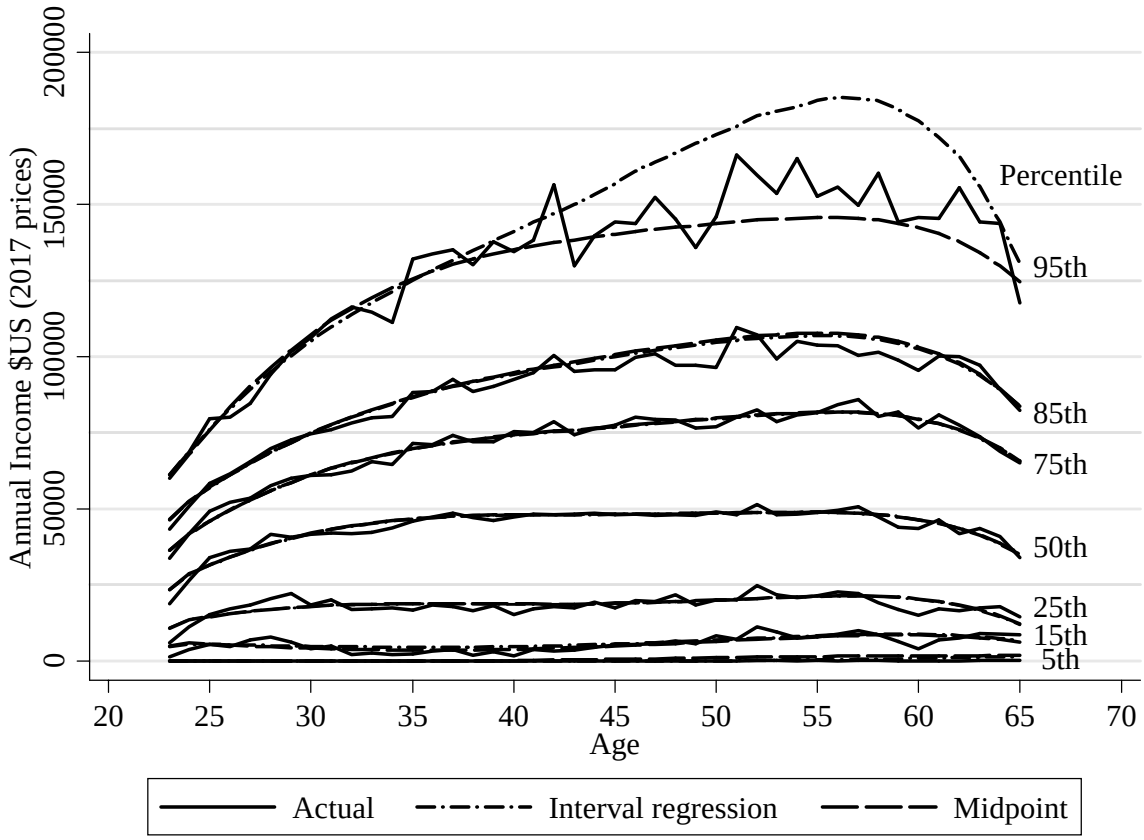
Figure 5: Male BA graduates earnings by quantile using aged smooth interval regression and midpoint predictions



For females the interval regression procedure works well for all but the 95th percentile. This reflects the fact that women are more equally distributed in the constructed income bands as can be seen in Table A1 in Appendix A. Again this will have little implication for either RB analysis or ICL design. The midpoint estimates for the 95th percentile perform well, but this again reflects the fact that the midpoint was arbitrarily set at $150,000 for this group.

The success or otherwise of using aged smoothed interval regression predictions or midpoint predictions by quantile will depend crucially on the distribution of individuals within each band and in the case of interval regression, the richness of the background data used in the regressions. But in most countries with banded data it is unlikely to cause any significant problems.

Figure 6: Female BA graduates earnings by quantile using aged smooth interval regression and midpoint predictions



4 Estimating and simulating income dynamics when good panel data is not available

The next innovation of the paper is to come up with a relatively simple way of simulating income and earnings dynamics for graduates which can easily use the quantile estimates of income and earnings by age, year and gender discussed in the last two sections. Including dynamics is essential to get the costs of student loan design right and to understand RBs over the loan term of a TBRL such as a Stafford loan. The more dynamics there is in income or earnings then the greater probability that *an individual* will experience a bad labour market outcome at some point in time (and hence a high RB) and the higher probability that they will also experience a good outcome (and hence have to make a repayment on their income contingent loan). This means that the RB problem will necessarily be underestimated and the cost of an ICL overestimated, with no mobility. In fact with no mobility you get a *lower bound* on the extent of RB problem and an *upper bound* on the cost of an ICL. This is demonstrated empirically in section 5. This means estimating dynamics in a reliable way is important. The

corollary of this is that if simulated earning or income dynamics have too much mobility they will exaggerate RB problems and underestimate the cost of designing an ICL.

With long panels, sophisticated methods can be used to get dynamics correct such as the approach outline in Britton et. al (2018) section 3.1. But these methods are not feasible with short panels and simple egression models will not be reliable as they assume linear dependence across the income or earnings distribution which is simply not realistic. It is better to use methods that rely on estimating rank dependence that allow for dependence to vary across the income or earnings distribution and can partially overcome issues with measurement error. This can be done using Copula functions and involves modelling the joint cumulative distribution of the two marginal cumulative distribution functions of income or earnings (including zeros) at adjacent ages. This is a simplified version of the approach used by Dearden et. al. (2008) and Bonhomme and Robin (2009) and provides a simple parametric way of estimating income transition matrices (of any dimension). Hence it is related to the to the dynamic simulation approach used by Higgins and Sinning (2013) using rich Australian longitudinal data. Of course, the approach involves the assumption that an individuals' rank in the income or earnings distribution next period only depends on their current rank (i.e. is first order Markov). Bonhomme and Robin (2009) show that for French LFS data with three income observations this assumption is reasonable and matches the observed transitions over one and two years well, despite the first order Markov assumption.

The copula function approach is so named as it defines the way two (or indeed many) continuous univariate marginal distributions can be 'coupled together' to form their joint bivariate (or multivariate) distribution F. It is assumed that earnings and income are continuous and observed for every individual at age t (y_{it}) and age $t+1$ (y_{it+1}). These earnings and incomes are turned into their cumulative distribution function (cdf) at each age, u_{it} and u_{it+1}. These, by definition, are standard uniform and can easily be mapped onto the percentile estimates from the previous section by rounding. From Sklar's theorem (see Sklar (1959)) if these cdfs are continuous and have joint distribution $F(u_t,u_{t+1})$ and marginal distributions $F(u_t)$ and $F(u_{t+1})$ there is a unique copula function C_t, such that:

$$F(u_t,u_{t+1}) = C_t\,(F(u_t),F(u_{t+1})) = C_t\,(u_t,u_{t+1}),\ \ t=23, 24, \ldots 64 \qquad (1)$$

noting that in the setting of this paper, $F(u_t) = u_t$ and $F(u_{t+1}) = u_{t+1}$ since u_t and u_{t+1} are the cdfs of the income or earnings variable at each age *t* and hence the marginals are also standard uniform. More generally the marginal distributions of income can be modelled at each age using any distribution or mixture but the approach used in this paper is to use the empirical marginal distribution by age[11] estimated in the previous section. In this example C_t is a 2-dimensional copula but the method extends to higher dimensions.[12] Another attraction of the Copula function is that it makes simulation very easy. This involves:

1. Estimating the conditional distribution of u_{t+1} given u_t which is given by:
$$c_{u_t}(u_{t+1}) = \frac{\partial}{\partial u_t} C_t(u_t, u_{t+1})$$
2. Generating a random standard uniform variable r with the same dimension as u_t
3. Generate $u_{t+1} = c_{u_t}^{-1}(r)$ to get our uniformly distributed predicted rank at age *t+1* which has a stochastic element due to the rank prediction being determined by the draw from the random uniform

Typically, parametric copula functions are used and different copula functions allow for different types of dependence (including symmetric and non-symmetric tail dependence cf regression). Goodness of fit criterion, such as the Akaike information criteria (AIC) can be used to choose the model that best fits the data.

To operationalize the copula estimation the CPS panel of BA graduates from 2014-2017 is used which contain the weighted cumulative distributions of incomes and earnings by age, year and gender as well as actual income and earnings including zeros.

The basic *dependence* characteristics of our panel data are shown in Table 2. It shows the rank correlation of the cumulative distribution function at adjacent ages, measured by Kendall's tau,

[11] In section 2, 100 percentiles of the income and earnings distribution at each age have been estimated by gender for BA graduates which can be mapped onto the cdf by rounding up the cdf to the nearest percentile. This appears to fit the continuous data well as shown later in the section.

[12] Moreover, this joint distribution can be decomposed as a function of the Copula function and the marginal densities, that is $f(u_{it}, u_{it+1}) = c_t(F(u_{it}), F(u_{it+1}))\, f_t(u_{it})\, f_t(u_{it+1}) = c_t(u_{it}, u_{it+1})$ where c_t is the copula density and f_t and f_{t+1} are the marginal densities of the copula which are equal to one as the marginals are standard uniform.

varies by age groups in our sample as well as the correlation of income and log income. Kendall's tau correlation is used to measure rank dependence as this can be easily estimated from the estimated parameters of our Copula model for comparative purposes and is less prone to bias due to earnings or income/earnings measurement error which is not true with correlation parameters[13] (see Appendix B for full details of the Kendall tau). It is important to emphasise, that if a person has zero earnings or income, then they are randomly distributed at the bottom of the cumulative income distribution at each age. For comparison, income and log income correlations for those with non-zero income in both periods are also shown.

Table 2: Measures of income dependence in CPS panel.

Age Group first year	**Kendall's tau**		**Income correlation including zero incomes**		**Income correlation for non-zero income**		**Log income correlation for non-zero income**	
	Males	Females	Males	Females	Males	Females	Males	Females
All ages	0.414	0.489	0.435	0.441	0.430	0.420	0.526	0.634
< 25	0.286	0.273	0.148	0.144	0.096	0.125	0.200	0.358
25-29	0.379	0.446	0.256	0.321	0.237	0.262	0.491	0.522
30-34	0.411	0.511	0.391	0.405	0.384	0.371	0.457	0.639
35-39	0.443	0.544	0.355	0.480	0.351	0.446	0.501	0.706
40-44	0.448	0.526	0.411	0.507	0.407	0.486	0.516	0.680
45-49	0.415	0.508	0.475	0.450	0.479	0.428	0.547	0.669
50-54	0.413	0.492	0.414	0.361	0.408	0.351	0.517	0.603
55-59	0.387	0.463	0.435	0.466	0.425	0.454	0.492	0.599
60-65	0.433	0.443	0.447	0.483	0.463	0.472	0.505	0.572

What is evident from the table is that dependence varies by age and there is a lot more mobility at younger ages. This will need to be captured in the estimation and simulations. The life cycle patterns of correlation exhibited for men and women are also different. It is also evident that the (linear) income and log income correlations are quite different, though show similar patterns by age as the rank correlation. The difference between the income and log income correlations strongly suggests non-linear dependence. Hence observed dependence is better captured by a rank correlation measure such as kendall tau (τ) which does not impose linearity

[13] The calculations use Kendall's tau-b where ties are counted as concordant rather than discordant, see Appendix B for more details. This makes no difference in reality with continuous marginal CDFs, but does if these are made discrete, e.g. turned into 100 percentiles.

and just evaluates the monotonic relationship between the ranks of two adjoining income or earnings variables (see Appendix B for details). Not requiring linear dependence highlights the huge advantage of the copula approach with short panels.

The estimation strategy involves finding a copula function which best captures the dynamics between the cdfs of marginal income or earnings at adjacent ages from 23 to 64.[14] For almost all ages, the t-Copula provides the best fit for the CPS data and this is true whether modelling earning or income dynamics.[15] Dearden et. al. (2008) also found that the t-Copula worked best with UK earnings data from the UK Labour Force Survey. The t-copula has the dependence structure implicit in a bivariate t-distribution.[16] It has two parameters – the correlation parameter, ρ, and the degrees of freedom parameter, v. These can be broadly interpreted as describing the overall level of immobility in the distribution (higher ρ) and the excess immobility in the tails of the distribution (lower v). From the model estimates Kendall tau can be estimated which in the case of the t-Copula is given by $\tau = 2v(\arcsin(\rho))$.

To take account of the observed change in dependence, the t-copula model is estimated separately by gender as well for every age transition from 23 to 64.[17] The estimates of the two t-Copula parameters rho (ρ) and degrees of freedom (v) and the associated confidence intervals by age, gender and for both income and earnings are shown in Figures 7, 8, 9 and 10. The estimates of rho are shown in Figures 7 (Males) and Figure 8 (Females), and the degrees of freedom estimates in Figure 9 (Males) and Figure 10 (Females). Smoothed estimates by age are also shown and it is these smoothed estimates that are used in the simulations.

[14] The R 'copula' and 'VineCopula' packages are used to do this. Transitions are modelled at every age and then goodness of fit tests are used to see which Copula best fits the data using 'fitCopula' from the 'copula' package.

[15] E.g. For male income dynamics, the t-Copula is best for 33 age transitions, the Frank copula for 6 age transitions and BB1, BB7 and survival BB8 for one each.

[16] For detailed information on the t-copula, including a formal definition, see Demarta and McNeil (2005).

[17] In Dearden (2008) this was explicitly built this into the Maximum Likelihood Estimation procedure but this is not available the R copula packages used. Instead separate estimates are obtained for each age transition and then these are smoothed before simulation.

Figure 7: Estimates of rho (ρ) from t-Copula: Males

Figure 8: Estimates of rho (ρ) from t-Copula: Females

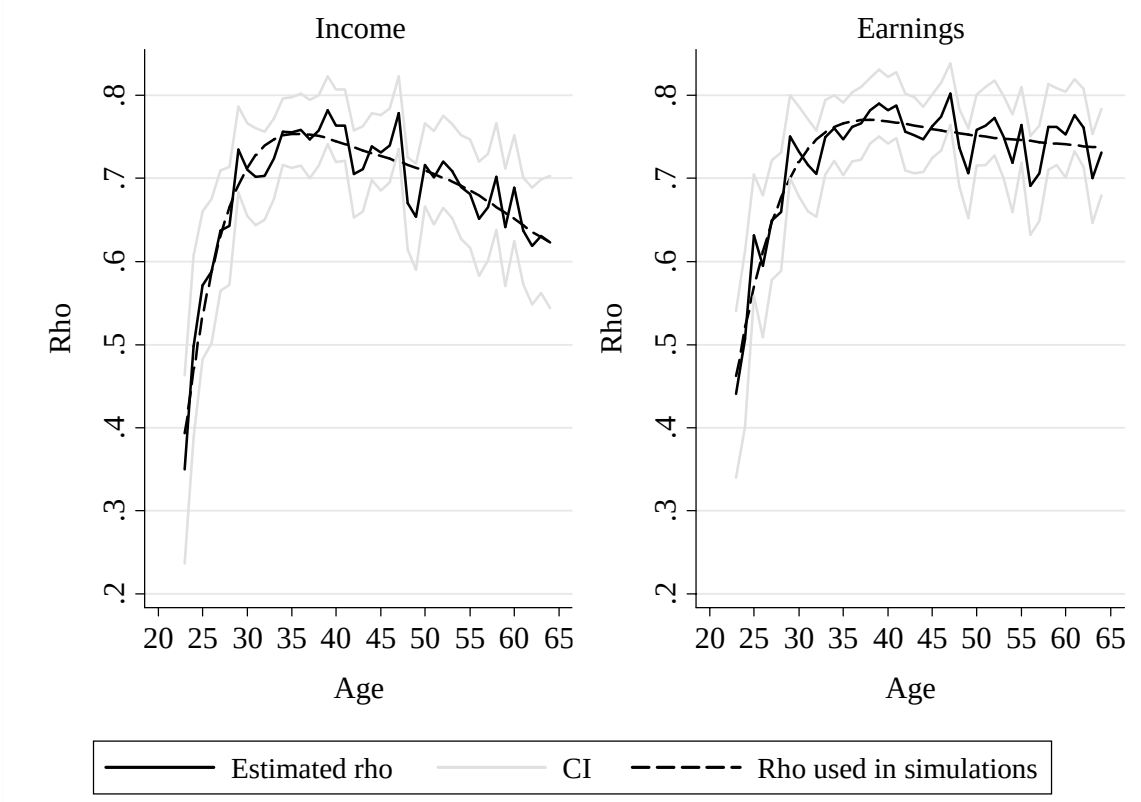


Figure 9: Estimates of degrees of freedom (*v*) from t-Copula: Males

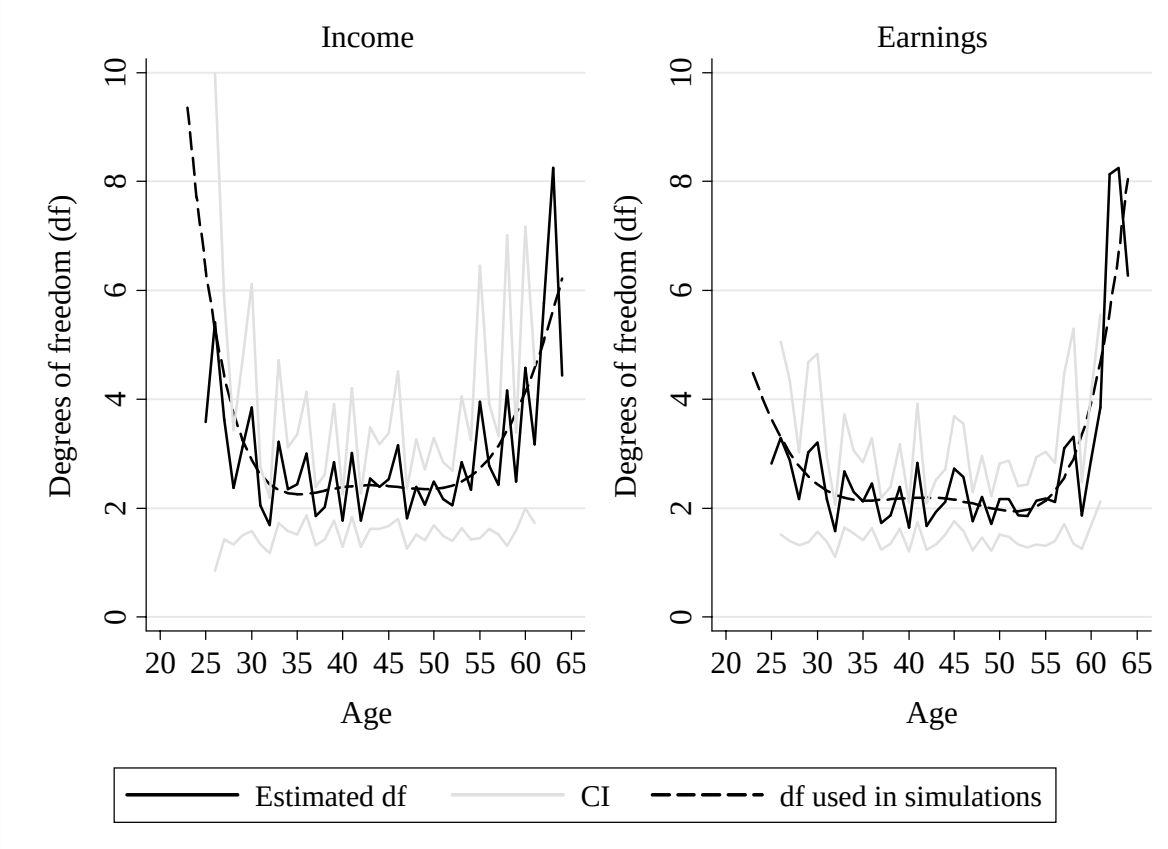
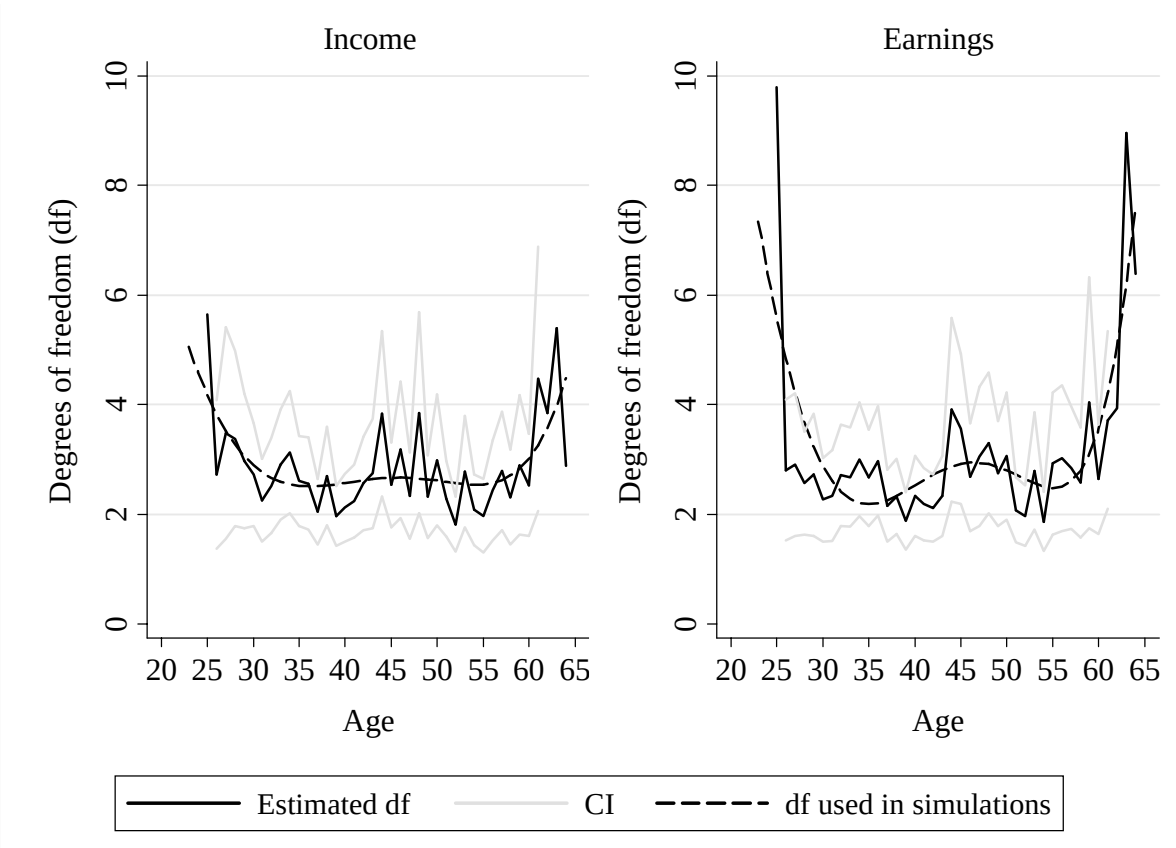


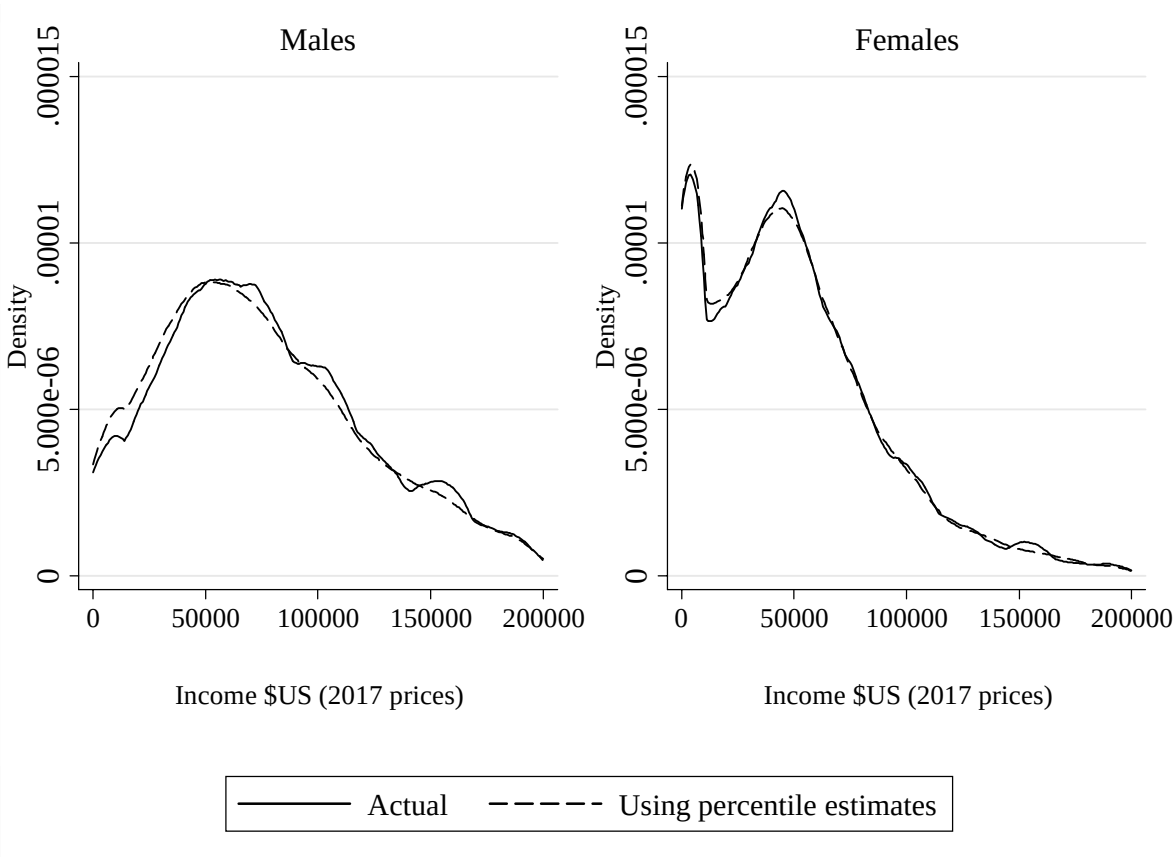
Figure 10: Estimates of degrees of freedom (*v*) from t-Copula: Females



The figure shows that whilst the estimates are reasonably similar for income and earnings and by gender, there are important differences, particularly at older ages but also to a lesser extent at younger ages.

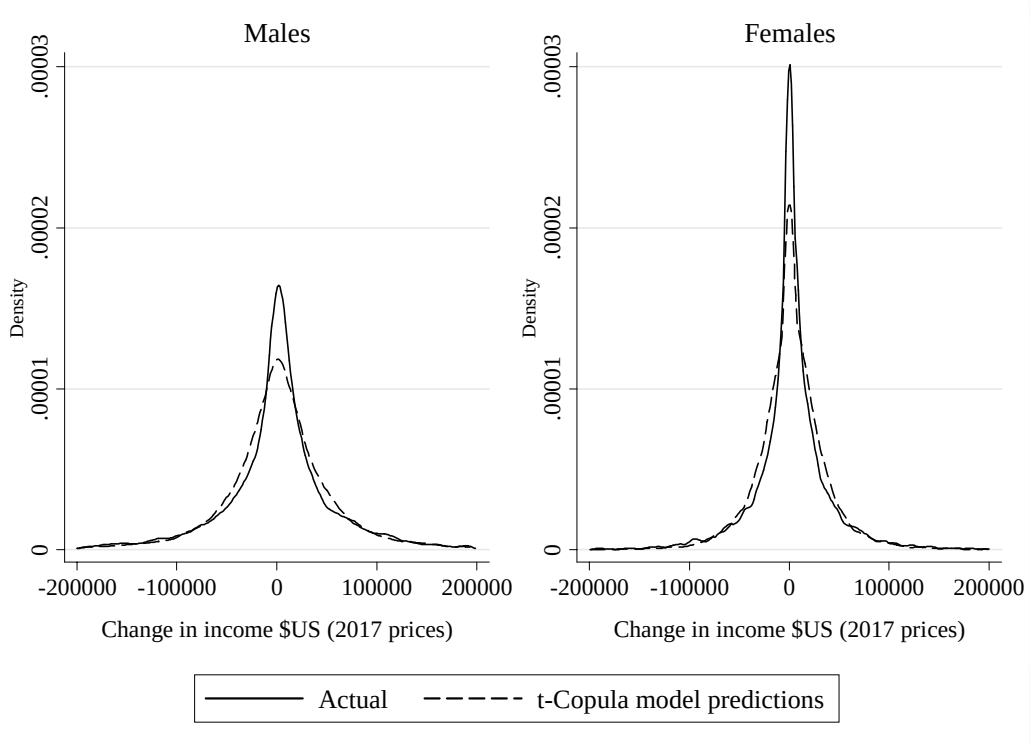
The model performance is tested for the CPS panel where the one age ahead predictions of income and earnings from the t-Copula model are compared with actual earnings and income outcomes. Quantile transition matrices are also compared. The simulation method proposed involves mapping the age earning/income profiles by percentile, age, year and gender to the estimated percentile from the t-Copula model which is constructed by rounding the standard uniform cdf to the respective percentile. Figure 11 shows how this performs when comparing actual income at age *t*, with the profiles matched onto actual percentile at time *t* for Males and Females. Figure 11 shows that the distribution of income is replicated closely using this simple approximation and the quantile age earning profiles from the previous section. It performs equally well for earnings (not illustrated).

Figure 11: Kernel density estimates of Actual vs Percentile approximated income (Ages 23 to 64).



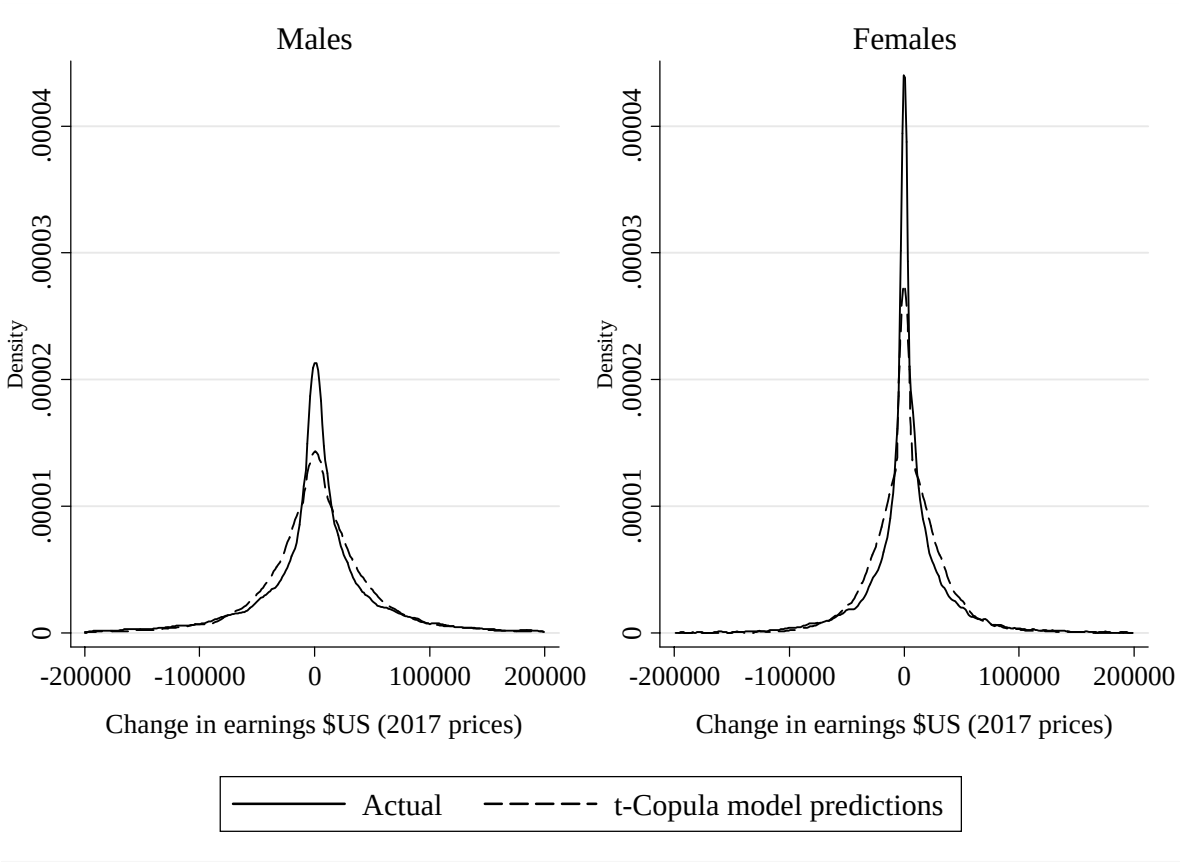
But the way dependence is modelled matters crucially for the distribution of the *difference* in income between age *t* and age *t+1* and from Figures 12and 13 it appears the t-Copula model performs well for both income and earnings for men and women respectively.[18] This is illustrated for income in Figure 12 and earnings in Figure 13.

Figure 12: Predicted difference in incomes at adjacent ages



[18] If the panel member was observed in March 2014 and March 2015, March 2014 income or earnings profiles by percentile are merged to their actual percentile observed in 2014 and March 2015 income or earnings profiles are merged to both their actual and predicted percentile (from t-Copula estimate) in March 2015. This is repeated for panel members observed in March 2015 and March 2016 and for those observed in March 2016 and March 2017. Figure 10 just shows the mapping of their actual income and mapped percentile income in the base year.

Figure 13: Predicted difference in earnings at adjacent ages



The simulations to not capture small changes in earnings or income completely but interestingly the overall rank correlation of the transitions in the simulations are always slightly higher than that observed in the actual data.

Next quintile transition matrices from the t-Copula model and the observed CPS panel data are compared for income in Tables 3 for men and Table 4 for women. The model replicates the observed income transitions well, although does not quite get the slight asymmetry observed in the female income transition matrix which show higher dependence at lower incomes than higher incomes.

Table 3: Male income transition matrices: Actual vs Predicted

Quintile at age *t*	Actual					Predicted from t-Copula model				
	Quintile at age *t+1*					Quintile at age *t+1*				
	1	2	3	4	5	1	2	3	4	5
1	54.79	21.22	10.13	7.49	6.11	53.21	22.22	12.34	6.49	5.46
2	20.61	39.59	21.11	10.20	8.70	22.84	32.60	23.15	14.59	7.05
3	10.20	19.47	37.45	21.75	11.29	10.74	24.32	29.39	25.46	10.24
4	7.72	11.70	18.58	40.41	21.82	7.18	14.41	24.75	30.92	22.97
5	6.68	8.02	12.73	20.15	52.08	6.03	6.46	10.38	22.54	54.28

Table 4: Female income transition matrices: Actual vs Predicted

Quintile at age *t*	Actual					Predicted from t-Copula model				
	Quintile at age *t+1*					Quintile at age *t+1*				
	1	2	3	4	5	1	2	3	4	5
1	65.76	18.55	6.77	4.77	4.00	58.89	22.35	9.34	5.21	4.03
2	18.14	43.34	19.85	10.81	7.92	22.55	34.39	25.33	12.69	5.19
3	7.01	19.96	40.56	22.06	10.50	8.91	25.41	30.96	25.50	10.32
4	4.76	10.22	22.71	41.90	20.58	5.62	12.31	24.45	36.04	21.77
5	4.32	7.83	10.10	20.47	57.00	4.02	5.54	9.93	21.55	58.69

The age smoothed model estimates are then used to recursively simulate lifetime income and earnings ranks for 10,000 male and 10,000 females who are assumed to have started college in 2017. The earnings and income age earning profiles averaged over all four years are merged to these ranks by percentile, age and gender. The simulated sample is re-weighted by gender to reflect latest US BA completions.[19] This is important when working out the budgetary implications of different ICL systems. In Figure 13 the estimates of Kendall tau (τ) from the CPS panel, from the actual model estimates (where $\tau = 2\nu(\arcsin(\rho))$ and from the simulated income sample (where smoothed model estimates of ρ and ν were used) are compared.

[19] BA degrees conferred in 2015 were 812,669 men and 1,082,253 women (see https://nces.ed.gov/programs/digest/d16/tables/dt16_301.10.asp).

Figure 14. Comparison of Kendall tau from CPS sample, t-Copula estimates and Simulated sample

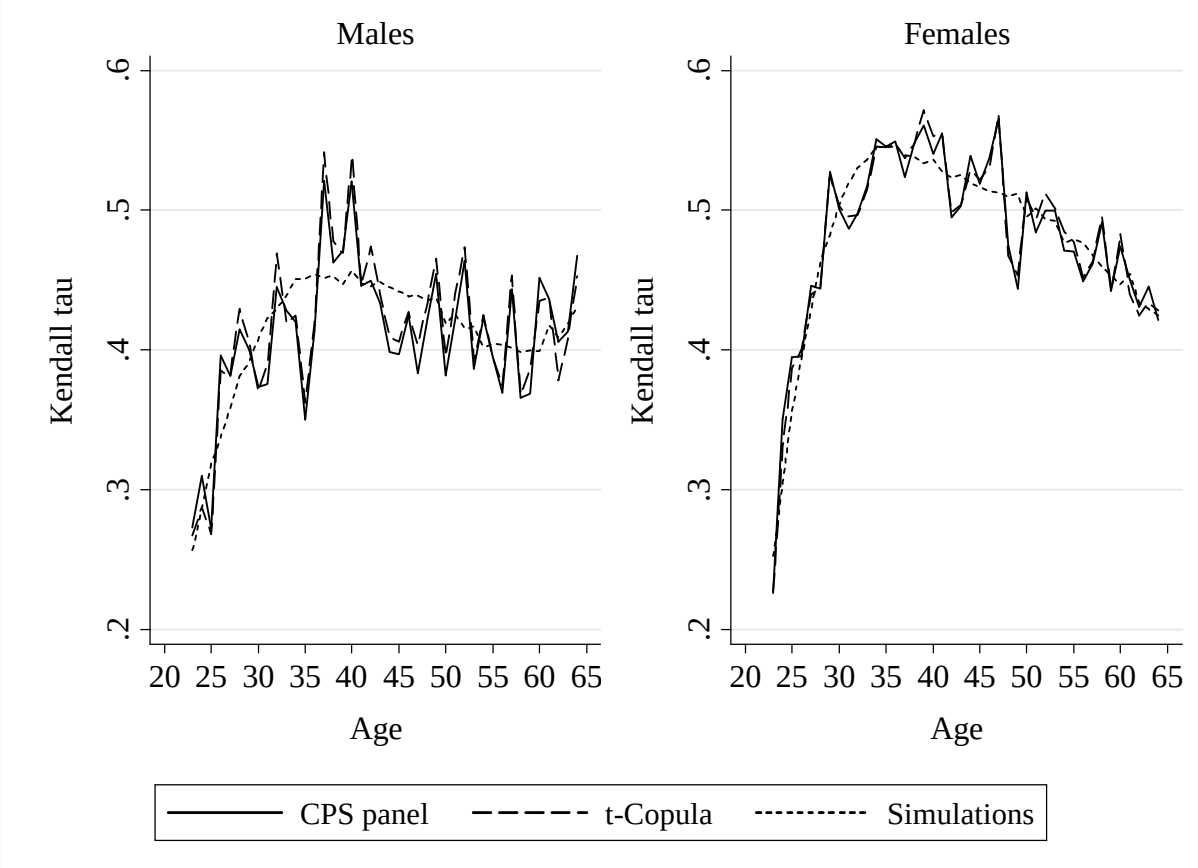


Figure 14 illustrates that the model estimates replicated the raw CPS rank dependence well. As a result, the dependence structures over adjacent ages of the simulated sample which use smoothed parameter estimates of ρ and ν also mirror the CPS panel rank dependence well.

5. Implications for estimating RBs and analysing the design and costs of an ICL

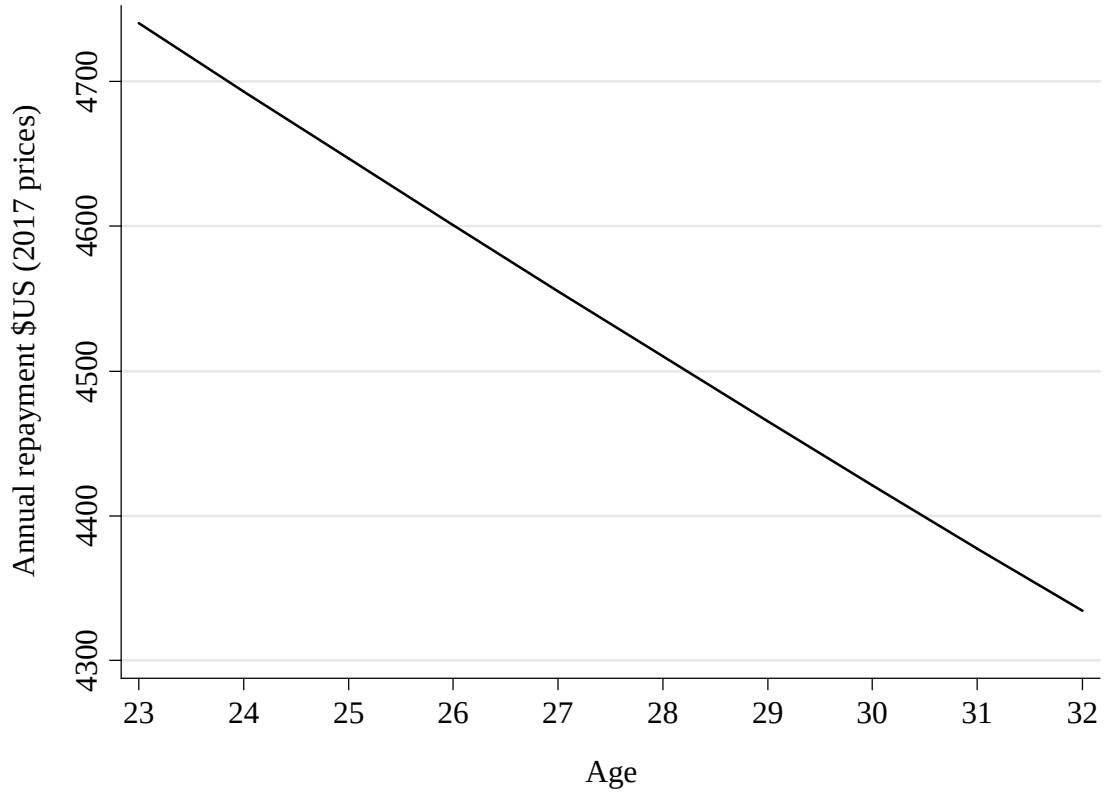
The final section of the paper considers how including income dynamics impacts on the analysis of RBs as well as estimating the taxpayer costs of an ICL. In the illustrations it is assumed that the average US graduate takes out a loan of $35,000 over a 4 year BA degree which is just under the current average of all US student loans for 2016 graduates of $37,172.[20] These loans are assumed to be log-normally distributed and have a standard deviation of

[20] See https://studentloanhero.com/student-loan-debt-statistics/

$20,000. It is also assumed that the amount of loan a BA student takes out is positively correlated with the first 10 years of their total graduate earnings. In these simulations a correlation of 0.3 is assumed but the sensitivity of results to this assumption are tested. Papers which have access to US administrative student loan data and tax records confirm that there is a positive correlation between debt levels and later earnings (see Looney and Yanellis (2018)).

Figure 15 shows the average repayment schedule for a Stafford student loan of £35,000 in $2017 US prices (the average in the simulated sample). Stafford loans are time based repayment loans (TBRLs) that in 2017/18 have a nominal interest of 4.45% and the majority must be paid within 10 years. In the first year of the loan an average student must pay back just over $4,700 regardless of income as shown in Figure 15. Typically in RB analysis, the repayment burden is shown by age and percentile. However this implicitly assumes that a graduate stays in the same percentile of the income distribution at each age, i.e. it assumes no income rank mobility. This is the approach taken in Chapman and Lounkaew (2015) and has been followed in some of the papers in this issue. This approach clearly demonstrates that with most TBRLs there is generally a significant problem at low incomes particularly at young ages.

Figure 14: Loan Repayment Schedule for Stafford Loan of $35,000



There is another way of looking at this problem when simulated incomes/earnings of individuals over their lifetime are available. The simulations from section 4 are used and it is assumed that there will be 1% real income growth per year over their lifetime.[21] The loan amount is assigned to individuals using our baseline correlation of 0.3 with the first 10 years of graduate earnings. Using these simulations, the number of times over the ten year term of the Stafford loan an individual will face RBs of more than 18% and more than 40% is calculated (so this could be 0,1,2,...10 times). Males and females are pooled together using the weight constructed earlier which reflect current BA graduation rates and in Table 5 the percentage of the cohort of borrowers falling into each category is shown. A comparison is made with the case where no income dynamics are assumed with the case where income dynamics are included. A comparison is also made with the no dynamic case using the quadratic UQR approach.

[21] Chapman and Laenkow (2015) assumed 1.5%. In fact there is evidence in the US that income growth varies across the distribution of earnings with higher growth for those in the higher part of the income distribution and lower (or even negative) real growth at the bottom of the distribution. Hence for this illustration a lower figure of 1 pecent is assumed.

As pointed out at the beginning of Section 4, assuming no mobility will necessarily underestimate the RB problem. Table 5 shows that if no mobility is assumed, it is estimated that around 48% graduates would never face RBs of greater than 18% and 70% would never face RBs greater than 40%. If the UQR approach was used then these estimates would be even higher – 60% and 78% respectively. The estimates from the dynamic model suggest that the correct figures are closer to 15% and 32% respectively. Further, just under 50% of graduates are likely to face 3 or more years of having RBs greater than 18% and just under 25% are likely to face 3 or more years of having RBs greater than 40%.

Table 5: Measures of Years of excessive RBs for average $35,000 US Stafford Loan

Number of years of excessive RBS	**RB > 18% (Percentage)**			**RB > 40% (Percentage)**		
	No dynamics UQR quadratic	**No dynamics raw smoothed quintic**	**Dynamics raw smoothed quintic**	**No dynamics UQR quadratic**	**No dynamics raw smoothed quintic**	**Dynamics raw smoothed quintic**
0	60.15	47.93	14.71	78.24	69.77	31.90
1	1.70	13.08	17.93	0.85	10.14	25.72
2	2.68	5.40	17.71	1.19	2.65	17.69
3	2.33	4.12	14.68	1.14	1.96	10.30
4	2.30	3.22	11.35	1.06	1.21	6.76
5	2.03	2.39	8.55	0.79	0.90	3.62
6	1.79	1.71	6.12	0.69	0.60	1.94
7	1.62	1.36	3.87	0.65	0.49	1.25
8	1.55	1.15	2.54	0.59	0.41	0.50
9	1.33	0.77	1.48	0.51	0.32	0.22
10	22.52	18.87	1.08	14.30	11.55	0.10

This nuanced picture is not captured if income dynamics are not included and helps explain the current default and delinquency problems with student loans in US which is the topic of the paper by Looney and Yannelis (2018).

Of course, all approaches fail to account for other factors that will determine whether a person faces financial hardship in repaying their loan. A more sophisticated RB analysis would look at RBs by household and consider other factors that affect the ability to pay such as number of children, and household taxes and benefits. This should be addressed in future work looking at student finance.

Including earning dynamics also impacts on the estimated costs of an ICL scheme. This has been previously shown in Higgins and Sinning (2013) and Dearden et. al. (2008). To illustrate this the Stafford loan interest and government cost of borrowing parameters are used in conjunction with other ICL parameters. It is assumed that there is:

(i) A first income repayment threshold of $17,000 per year, and a second threshold of $35,000 (in a policy reality these would both uprated annually with inflation). The $17,000 threshold is similar to that used with the current IBR scheme currently operating in the US.

(ii) A marginal 3 percent repayment rate on earnings above the first threshold and 10 percent marginal for earnings above the second threshold. Again the 10% marginal rate is similar to that used with the current IBR scheme.

(iii) A zero real interest rate whilst a student is at college and below the first income threshold (i.e. debt increases with inflation only); and then a real interest rate equal to the current Stafford Loan rate which is 4.45% nominal or 2.45% real. With the means tested component of the Stafford Loans, a zero real interest rate applies whilst students are at college. No means testing is applied for this simulation.

(iv) An inflation rate of 2% and a government cost of borrowing of 2.4% nominal or 0.4 % real.

(v) A loan write-off after 25 years.

To compare the full distributional implications of this ICL as well as the size of the taxpayer subsidy, the *earnings* simulations from Section 4 are used. Taxpayer costs are estimated under the assumption of no mobility (which provides an upper bound of costs, see Barr et. al. (2018)) and realistic dynamics from the earnings simulations described in Section 4. The taxpayer subsidy is calculated by pooling the male and female results using current BA enrolment proportions as highlighted earlier in the paper. All costs and repayments are discounted back to

when the student takes out the loan at age 18 and are in $US 2017 prices. The taxpayer subsidy is calculated by comparing the net present value (NPV) of repayments (which depend on future earnings simulations and ICL parameters) to the NPV of providing the loans (which depends on the amount of loans taken out). Real earnings growth of 1 per cent is assumed for all graduates throughout their working life. [22] Inflation is assumed to be 2 per cent and the government cost of borrowing is set to the current 10-year US bond rate. This is currently used to determine the Stafford Loan interest rate which is set at the government cost of borrowing (currently 2.4% nominal or 0.4% real) plus 2.05 percentage points.[23]

Figure 15 shows the distributional impact (by deciles of the male and female college *lifetime income* distribution[24]) under the assumption of no mobility and with dynamics. The analysis shows that when income dynamics are ignored, the estimated taxpayer subsidy for this ICL is around 15% whereas when dynamics are included the estimated taxpayer subsidy is -7%. All graduates receive a taxpayer subsidy in this scheme whilst they are at college and whilst they earn below the first threshold. However, once they are above the first threshold, they receive no subsidy as the interest rate is 2.05% above the government cost of borrowing so they are net contributors. Those who do not repay their loan within 25 years may also receive a subsidy due to the loan write-off. The difference in estimates of the taxpayer subsidy is large as there is high earnings mobility in the US for BA graduates and with the ICL operating over 25 years, there is a much higher chance of individuals making some repayments. The extent of the difference in estimated taxpayer subsidy however, depends crucially on the size of the ICL loan and the ICL loan parameters (see Barr et. al. (2018) and Britton et. al. (2018)) as well as the correlation between the total student loan taken out and future earnings. For example, if there is no correlation between earnings in the first 10 years and loan amount the estimate of the taxpayer subsidy increases to 16% (no mobility) and -6% (mobility). If there is perfect negative correlation between loan amount and gross earnings in the first 10 years the taxpayer

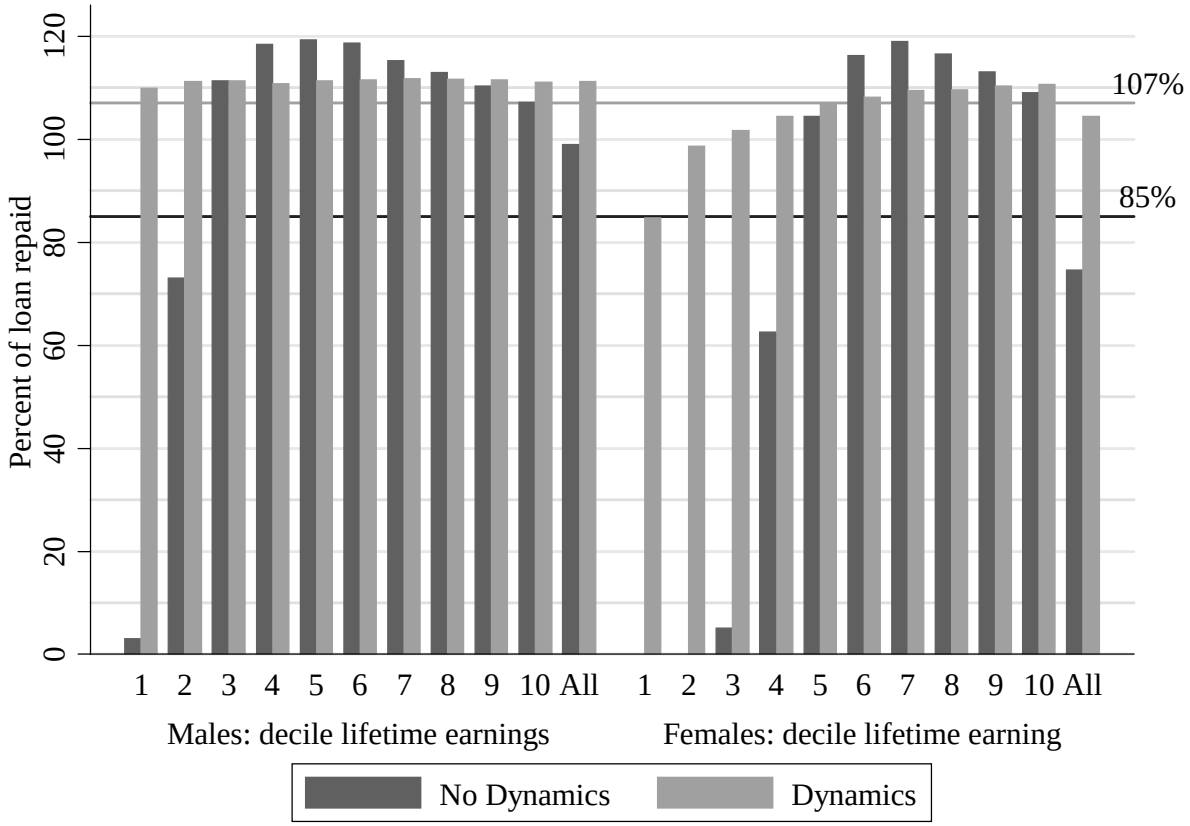
[22] This can be easily adjusted. There is evidence that real earnings growth is higher at the top end of the earnings distribution than at the bottom which will therefore underestimate costs.

[23] The current Stafford interest rate of 4.45 percent per annum nominal is the government cost of borrowing or 10 year bond rate (2.4 percent nominal) plus 2.05 percentage points and hence 0.4 percent real with our assumption of 2 percent inflation.

[24] The lifetime earnings of all individuals from the ages of 23 to 65 are summed without discounting and including assumed real earnings growth of 1% and deciles constructed from this measure.

subsidy estimates increase to 19% (no mobility) and -4% (mobility). If there is perfect positive correlation then the taxpayer subsidy estimate decreases to 10% (no mobility) and -11% (mobility).

Figure 15: Proportion of ICL Loan Repaid by Decile of Lifetime Earnings: Dynamic vs No dynamic earning simulations



Note: Overall graduate contribution shown by horizonal lines and the subsidy can be calculated by taking this amount away from 100%.

As shown in Barr et. al. (2018) and Britton et. al (2018) this subsidy can be reduced or increased very easily by varying the ICL parameters including introducing a surcharge on the loan, changing the interest rate and/or changing other ICL parameters such as repayment thresholds and repayment rates. But given the income mobility of BA graduates in the US, it is clear that a well designed ICL will work (see Barr et. al. (2018) for more details) and would have considerable advantages over the current Stafford Loan system. Additional work simulating the earning dynamics of 2-year college graduates and drop-outs suggest a similarly designed ICL could work beyond BA graduates. Of course tight regulation of loans would need

to be implemented, particularly with the for-profit sector, as is the case with the current US loan system.

6. Conclusions

This paper reviewed the empirical approaches that are needed to both evaluate and design student loans systems. A particular innovation of the paper, is that it has suggested relatively straight forward methods for improving income and earnings simulation when data is poor (e.g. the data has banded income or there is not good panel data available in the country). Another innovation is that the method proposed extends work that is already routinely done in countries evaluating their student loan system.

The paper shows that for RB analysis it is generally better to use raw percentile estimates of income or earnings by age and gender and age smoothing rather than use UQR methods. Having banded income data (as is the case in countries like Japan) does not appear to be a significant problem for all but the highest earners and RB analysis (or indeed student loan design) is not affected by grouping of income or earnings data.

The paper shows how income and earnings dynamics can be easily introduced even with short panels which have a minimum of two observations for the same individual. This involves using copula functions which better capture the complex dependence between income or earnings over one year. With traditional dynamic panel data methods this is only possible to do in a reliable way with longer panels (longer T).

Finally the paper highlights the importance of including dynamics in both assessing the RBs associated with current loans systems as well as designing ICLs. Ignoring dynamics will firstly underestimate the proportion of individuals facing repayment hardship with a TBRL and secondly will result in over-estimating the taxpayer costs of an ICL.

References

Barr, N. Chapman, B., Dearden, L. and Dynarski, S. (2018), 'Reflections on the US College Loans System: Lessons from Australia and England', submitted to this issue.

Britton, J., Higgins, T. and van der Erve, L. (2018), Income contingent student loan design: Lessons from around the world, submitted to this issue.

Bonhomme, S. and Robin, J-M, (2009). 'Assessing the Equalizing Force of Mobility Using Short Panels: France, 1990-2000,' *Review of Economic Studies*, Oxford University Press, vol. 76(1), 63-92.

Borah and Basu (2013). 'Highlighting differences between conditional and unconditional quantile regression approaches through an application to assess medication adherence', *Health Economics*, 22(9):1052-70

Chapman, B and Dearden, L (2017), 'Conceptual and Empirical Issues for Alternative Student Loan Designs: The Significance of Loan Repayment Burdens for the US', *Annals*, Volume: 671 (1), 249-268.

Chapman and Doris, (2017), 'Modelling Higher Education Financing Reform for Ireland', this issue.

Chapman, B. and Lounkaew, K (2015). 'An analysis of Stafford loans repayment burdens', *Economics of Education Review* 45 (3): 89–102.

Crawford, C. Crawford, R. and Jin, W (2014), Estimating the Public Cost of Student Loans, IFS Report R94, 2014, http:www.ifs.org.uk/comms/r94.pdf

Dearden, L. Fitzsimons, E. Goodman, A. and Kaplan, G. (2008), "Higher Education Funding Policy", *Economic Journal*, vol. 118, no.526, F100-F125.

Dearden, L. and Nagase, N (2017), 'Getting higher education finance right in Japan: problems and possible solutions', submitted to this issue.

Firpo, Fortin and Lemieux (2009). 'Unconditional quantile regressions', *Econometrica* 77(3): 953–973.

Higgins, T. and Sinning, M. (2013), 'Modeling income dynamics for public policy design: An application to income contingent student loans', *Economics of Education Review*, 37: 273-285

Sklar (1959), Fonctions de Répartition à n Dimensions et Leurs Marges. Vol. 8, Institut Statistique de l'Université de Paris, Paris, 229-231.

Stewart, M.B (1983). 'On Least Squares Estimation when the Dependent Variable is Grouped', Review of Economic Studies, vol. 50(4), 737—753.

Appendix A

Table A1: Distribution of Grouped Income Variable BA Graduates 2014-2017 CPS

Income lower bound ($US per year)	Income upper bound ($US per year)	Proportion Males	Proportion Females
0	0	2.95	6.65
0	5000	2.49	8.19
5000	10000	1.79	3.77
10000	15000	2.24	4.35
15000	20000	2.56	4.11
20000	25000	3.14	4.43
25000	30000	3.14	4.35
30000	35000	3.66	5.17
35000	40000	3.74	5.21
40000	45000	4.29	5.92
45000	50000	4.23	5.32
50000	55000	4.80	5.56
55000	60000	3.61	4.17
60000	65000	4.35	4.22
65000	70000	3.36	3.11
70000	80000	7.38	6.06
80000	90000	5.94	4.30
90000	100000	4.73	3.14
100000	125000	10.80	5.52
125000	150000	5.63	2.29
150000		15.17	4.14
Sample Size		64,376	78,009

Appendix B: Kendall Tau (τ)

In the paper all measures of rank correlation/dependence use Kendall tau (τ). This is a measure of association based on the number of concordant, discordant and tied paired the cumulative distribution of income at age *t* (u_t) and *t+1* (u_{ti}) in the CPS panel. A pair of cdfs { (u_{ti}, u_{t+1i}), (u_{tj}, u_{t+ji})}, are:

- concordant if $u_{ti} < u_{tj}$ and $u_{t+1i} < u_{t+1j}$ or $u_{ti} > u_{tj}$ and $u_{t+1i} > u_{t+1j}$ and the number of concordant pairs is denoted by n_c
- discordant if $u_{ti} > u_{tj}$ and $u_{t+1i} < u_{t+1j}$ or $u_{ti} < u_{tj}$ and $u_{t+1i} > u_{t+1j}$ and the number of discordant pairs is denoted by n_d
- tied if $u_{ti} = u_{tj}$ or $u_{t+1i} = u_{t+1j}$ and the number of tied pairs are denoted by n_t and n_{t+1} respectively.

Kendall's tau-b rank correlation (τ) which is used in the paper is given by

$$\tau = \frac{n_c - n_d}{\sqrt{(n_c + n_d + n_t)(n_c + n_d + n_{t+1})}}$$

and the total number of pairs that can be constructed and compared for a sample of size T is:

- n= ½ T*(T-1).