THE ECONOMIC EFFECT OF GAINING A NEW QUALIFICATION IN LATER LIFE*

Finn Lattimore[1†], Daniel Steinberg[2] and Anna Zhu[3]

[1]Reserve Bank of Australia

[2]Gradient Institute

[3]RMIT University, IZA

August 23, 2022

Abstract

Pursuing educational qualifications later in life is an increasingly common phenomenon among OECD countries. Despite this, the value-add of gaining a qualification in later-life is unclear. One reason for this is what leads (or enables) mature-age learners to pursue and complete a qualification may itself be a precursor to later-life success. Another reason may be that the benefits are unevenly distributed: some mature-age learners are more likely to benefit than others because of their socio-economic background, degree-type or subject area. Why degree completion may be associated with positive returns and who may benefit are complex questions and not in themselves obvious to the researcher. In this paper we take a data-driven approach to shed light on these unknowns. With the aim of causal inference in mind, we tailor Machine Learning (ML) models to estimate the size of the effect and to identify which mature-age learners tend to reap more benefits than others. We use extremely rich and nationally representative longitudinal data from the Household Income and Labour Dynamics Australia survey. Our ultimate aim is to better inform individuals, employers, and governments about the returns to educational investment, and thus improve the allocation of resources.

JEL: J12, J18, H53

*Corresponding author: Anna Zhu, RMIT University. Email: anna.zhu@rmit.edu.au.

We thank Tim Robinson and numerous seminar and conference participants for helpful comments. The authors would like to thank Tessa Loriggio for her excellent research assistance.

Zhu acknowledges the support of the Australian Research Council Linkage Project (LP170100472). This paper uses unit record data from the Household, Income and Labour Dynamics in Australia (HILDA) Survey. The HILDA Project was initiated and is funded by the Australian Government Department of Social Services (DSS), and is managed by the Melbourne Institute of Applied Economic and Social Research (Melbourne Institute). The findings and views reported in this paper, however, are those of the authors and should not be attributed to either DSS or the Melbourne Institute.

†This work was performed while the author was working at the Gradient Institute. Views expressed in this paper are those of the author and do not reflect those of the Reserve Bank of Australia.

Keywords: Machine Learning, education, mature-age learners, causal impacts

1 Introduction

Pursuing educational qualifications later in life is an increasingly common phenomenon among OECD countries (OECD, 2016). Technological change and automation continues to drive the evolution of skills needed in many professions. This is particularly true for middle-income workers performing routine tasks (Autor et al., 2008; Acemoglu and Autor, 2011). Yet also at the lower end of the income-distribution, such as among welfare recipients, governments are increasingly trying to promote the idea of life-long learning.

We aim to estimate the causal impact of obtaining an educational qualification in later-life on earnings. We also aim to understand which groups of students tend to benefit more from this pursuit.

Much of the existing literature focuses on the returns to education for younger students and/or for the first degree out of secondary school. Among the authors who look at mature-age learners, the analysis often focuses on one type of setting. For example, at community or training colleges (Jacobson et al. (2005), Zeidenburg et al. (2015), Polidano and Ryan (2016), Xi and Trimble (2016), Belfield and Bailey (2017), Dynarski et al. (2016, 2018)) and/or on low-productivity workers such as those who enrol in a government-run training program (Ashenfelter (1978), Ashenfelter and Card (1985), Bloom (1990), Leigh (1990), Raaum and Torp (2002), Jacobson et al. (2005)).

Consequently, the results of such studies may not be generalisable to the entire mature-age education market. Also, they may yield attenuated estimates of the effects as they exclude students who seek different degree types or who study at different institutions.

Moreover, prior studies commonly analyse only measures of total earnings. This ignores the likelihood that further education may be sought in order to allow for more flexibility on the job such as in part-time work. This means that the value-add of further education on hourly wages may be obscured by a total earnings measure.

We fill the gap in the literature by estimating the impact of further education on earnings (both total and hourly) across all categories of a further degree (post-graduate degrees, training certificates, diplomas etc), and spanning all subjects and all institutions at which the study took place. This means we analyse the effects for a group of students with a larger span of demographic and socio-economic background characteristics.

We also add to the literature that estimates the returns across a wider range of qualifications (Angrist and Krueger (1991), Harmon and Walker (1995), Blanden et al. (2012), Perales and Chesters (2017), and Bockerman et al. (2019)). We do this by being the first to use Machine Learning methods to tackle the issue of selection bias and mis-

specification error. Below we describe why the new methods may provide further insights into the causal question.

The first benefit of a data-driven approach is that it can generate a counterfactual group when we use observational data. A key challenge in estimating the causal returns to later-life education is that: what leads (or enables) mature-age learners to pursue and complete a qualification may itself be a precursor to later-life success. Moreover, the drivers of degree completion may be numerous and related to other variables in complex, unknown ways. A data-driven approach can systematically shed light on these unknowns. As a result, ML can help to reduce bias from selection into treatment. This is especially useful when appropriate instrumental variables are unavailable.

A second benefit is that an ML approach provides an a-theoretic way to identify the sub-populations for which the treatment effects are the most different. Traditionally, economists have relied on theory or past experience to identify these groups. Yet this approach risks overlooking important sub-populations. We use a data-driven (permutation importance) procedure. This exercise can elucidate the potential mechanisms as well as inform different policy targeting and/or compensation strategies.

We argue that ML is well placed to achieve these benefits, especially since we apply it to extremely rich, nationally representative, longitudinal data. We use data from the Household Income and Labour Dynamics Australia survey. Two dimensions of these data are important. The first is that they contain a wealth of information about each respondent. For example, we begin with more than 3,400 variables per observation, including information about the respondents' demographic and socio-economic background, and on their attitudes and preferences. Using this broad range of information as control variables we can reduce selection bias issues, and potentially proxy for unobservable differences between those who do and do not obtain a new qualification. Secondly, this dataset contains many variables that are highly correlated and we require a systematic approach to reduce such information redundancy.

An ML approach can achieve these benefits because it can detect patterns and relationships, such as functional forms and control variables, that are unknown to the researcher. This assists with variable selection and with identifying meaningful sub-populations for which the effects differ. Furthermore, as ML models systematically choose control variables to minimise the amount of information redundancy between them, we can improve causal estimation parameters by reducing their variance.

We adapt ML models for the purpose of estimating causal effects. Standard off-the-shelf ML models are better suited to predictive purposes. When obtaining a prediction,

off-the-shelf ML models can find generalisable patterns and minimise overfitting issues because the true outcomes are observed. This means that we can optimize a goodness-of-fit criterion. Causal parameters, however, are not observed in the data, which means we cannot directly train and evaluate our models.

In this paper, we take the difference between the two optimal outcome models, which can achieve the optimum bias-variance trade-off point for the conditional average treatment effect. Specifically, we model the response surfaces for two conditional mean equations - one using the treatment observations and another using the control observations. We estimate these equations with ML methods such as the T-learner and Doubly Robust. Here, we employ both linear (LASSO and Ridge) and non-linear (Gradient Boosting Regression) model classes, which we evaluate using nested cross-validation. We then test the statistical significance of our causal parameters by examining the distribution of the estimates through bootstrapping. Last, we estimate Bayesian ML models to obtain efficiency gains from sharing information between both treatment and control response surfaces during posterior inference (model fitting).

2 Literature review

Much of the causal research on the returns to education for younger cohorts was conducted in the 1990s, some of which relies on data collected in earlier decades. The results show positive and significant wage premiums for those with more education, ranging between 5 and 13% (Machin, 2006). While these studies employ a range of techniques to isolate the impact of education on earnings, endogeneity issues have been identified in many of them i.e., twin studies do not necessarily control for ability, and instrumental variables tend to have weak exclusion restrictions (Harmon, Oosterbeek and Walker, 2003). The more recent literature focuses on improving the estimation methods, but still relies on data from the previous decade or earlier. Despite the large body of work on this topic, the causal evidence on the return to education in the current era is lacking. Drawing on the available research, two studies stand out. Angrist and Krueger (1991) use variation in compulsory school age across US states between 1960 and 1980 to show that those with more schooling, by virtue of reaching the dropout age later in the year, earn more. Using IV estimation to exploit this natural experiment, they place the return to education at 6-11%. Their findings are based on Census data from 1980 where their sample of men were aged between 30 and 50. Harmon and Walker (1995) use the same instrument for education in the UK, which also saw changes in the compulsory school leaving age laws during the 1950s and 1970s. Using data on employed men aged between 18 and 64

over the years 1978 and 1986, they estimate the return to education to be more than 15%. Both studies focus on young cohorts, with the variation in the minimum school age ranging from 14 to 18 years. Our paper expands on this literature by looking at older cohorts.

In comparison, the evidence on the returns to post-secondary education has grown enormously since the 1990s, spurred on by the availability of linked administrative data on student records and earnings. The transition from cross-sectional data to longitudinal data helped to resolve some of the selection bias associated with comparing those with a qualification to those without, as concerns over motivation, ability and other unobserved time-invariant characteristics can be controlled with an individual fixed-effects model. However, some selection bias remains as fixed-effects models cannot fully account for the fact that individuals who expect their prospects to improve by gaining a new qualification select into education.

Vocational and community college degrees have received particular attention in the literature since they tend to attract returning students who have employment earnings prior to attaining a new qualification. This has allowed researchers to compare individuals earnings before and after a qualification, and to compare the size of this change between students who receive a qualification and those who leave without one. In addition, since students enrolled in sub-bachelor programs are disproportionately low-income and low-performing students, the absence of individual characteristics in cross-sectional samples can bias the estimated returns downward (Xi and Trimble, 2016). Studies that compare individual fixed-effects methods with cross-sectional methods using the same sample find opposing effects. Fixed-effects methods show positive returns while cross-sectional methods show negative returns, suggesting that individual fixed-effects models can capture biases resulting from omitted time-invariant characteristics (Xi and Trimble, 2016). In the absence of prior earnings histories, researchers have exploited the student records and earnings of non-completers, as they are a more valid control group for those with qualifications, as opposed to secondary school graduates who never attend TAFE/college. The use of student records has also allowed researchers to disaggregate the effects by length, field of study and type of qualification. Associate degrees in nursing and health tend to yield the highest returns (Zeidenburg et al., 2015; Xi and Trimble, 2016). While these methods may still suffer from some selection bias related to potential unobservable differences between those who choose to leave versus stay in their degree the results have proved to be highly robust and minimally biased (Zeidenburg et al., 2015). Based on these new data and methods, the evidence on the labour market returns to vocational

and community college education is strong and positive, particularly for female students (Belfield and Bailey, 2017; Zeidenburg et al., 2015; Perales and Chesters, 2017).

Dynarski et al. (2018) and Jacobson et al. (2005) extend on these methods by adding individual time-trends, so as to account for time-varying unobserved heterogeneity between students. By doing this they test the common-trends assumption that underlies the individual fixed-effects estimation strategy. Dynarski et al. (2018, 2016) show that accounting for time-trends increases the return to associate degrees by as much as 25% because it can identify that workers who earn awards tend to have flatter earnings growth prior to enrolment. Dynarski et al. (2018) apply this to US community college records in 2003-2004 and earnings data in 2011. They estimate earnings gains of $10,000 per year for associate degree holders and $3,200 per year for certificate holders, placing their estimates at the high end of the range in the literature.

Australian and British research shows that significant earnings gains can be achieved when returning students acquire qualifications above those held previously, and when they are acquired at younger ages (Polidano and Ryan, 2016; Dorsett, Lui and Weale, 2016; Perales and Chesters, 2017). Lower or equivalent level qualifications, or qualifications earned closer to age 40, are associated with smaller or insignificant returns.

Perales and Chesters (2017) use Australian longitudinal data between 2001 and 2015 to estimate a fixed-effects model on individuals who obtained a higher-order qualification between ages 25 and 63. The largest wage gains were associated with acquiring an undergraduate degree, either from a secondary school diploma (Year 12) or post-secondary certificate or diploma, of between 10-23% for men and 15-18% for women. Similarly substantial wage gains were associated with moves from undergraduate to post-graduate degrees of around 15% for men and 10% for women. Only modest gains were associated with transitions from less than Year 12 to a certificate or diploma, for both men (5.9%) and women (3.9%). While the differences between men and women were not found to be statistically significant, there was weak evidence that only men benefit from the transition from Year 12 to a certificate or diploma. This shows that not all qualification advancements benefit Australian workers in the same way.

Polidando and Ryan (2016) explore vocational education in more depth using the same data and methods as Perales and Chesters (2017). Instead of aggregating sub-bachelor qualifications together, Polidano and Ryan (2016) assess each level individually. They found that women who acquire any certificate (level 1-4) after Year 12 showed the largest increase in wages, of about 10%. Men, on the other hand, experienced similar wage gains by acquiring a certificate 3, 4 or diploma without Year 12. In short, men do not increase

their earnings by completing Year 12 first. In general, women showed larger improvements in labour market outcomes following the completion of a vocational qualification. However, qualifications that were not of higher order did not generate benefits. The effects remained 5 years after course completion and suggest strong long-term benefits to vocational education. By comparison, when Polidano and Ryan (2016) analyse the data as pooled cross-sections they found the returns to be around half the size of the fixed effects estimates. However, men showed substantially smaller returns when gaining a university degree in the fixed effects model while women showed similar returns when gaining sub-bachelor qualifications. In general, though, the cross-sectional results doubled the estimated earnings effects, but they also switched which gender benefited most. The fixed effects estimates showed larger qualification effects for women, supporting the general finding in the literature.

Zeidenberg et al. (2015) link community college non-completers with completers according to their field of study to identify differences in their labour market returns. Based on the finding by Jacobson et al. (2005) that non-completers can achieve earnings gains from as little as one semester of credits, Zeidenberg et al. (2015) question whether all credits are equally rewarding i.e., do those who almost complete a course have higher returns than those who progressed less. Matching US community college transcripts between 2002 and 2004 with earnings data in 2011 they find that among non-completers, students who progressed further in their course earned less. Specifically, students who progressed twice as far as the average non-completer earned 4-5% less. The penalty for progression was similar for females and males. Jacobson et al. (2005) found a similar pattern among male displaced workers: those who completed 12-18 months of community college earned more than those who studied beyond this. However, they estimated the opposite effect for female displaced workers. Zeidenberg et al. (2015) posit that after taking the relatively valuable foundation courses, students may lack the motivation, time, or funds for the upper-level courses, which are no more valuable in the labor market but are required for graduation. Grouping all students together, Zeidenberg et al. (2015) find that non-completers earn 32% less than those who graduate with a diploma and 37% less than those who complete an associate degree, revealing large returns for those who obtain a qualification. Certificates holders and non-completers were found to earn about the same, possibly because they acquire a similar number of credits. Xu and Trimbles (2016) analysis illustrates that many adult learners use short-term certificates to switch to a new industry and may explain why certificate holders do not experience an immediate boost in earnings. By subject, earnings differences were not substantial except for nursing, where earnings gains were disproportionately large at 104%. In general, earnings appear to differ more by qualification than by subject (Zeidenberg et al., 2015).

These findings are consolidated in Belfield and Baileys (2017) extensive review of community college returns in the US between 1996 and 2013, spanning 8 states. The studies follow the individual fixed effects approach established in Jacobson et al. (2005). The estimated average earnings were 26% (18%) higher than non-completers for females (males) with an associate degree, amounting to approximately \$7,000[1] (\$5,000) per year, with the gains persisting overtime. Returns to certificate holders were more mixed but in general show positive, albeit modest, returns relative to non-completers. Again, females outperformed males in their returns and on average earned around \$3,000 per year more than non-completers while males earn around \$2,000 per year more than non-completers. In general, certificates that required more credits had higher earnings gains. Credits were also associated with earnings gains, even for those who didnt complete an award. The accumulation of credits with no award, however, does not appear to be as valuable as the award itself, although in some states the difference in returns between an award and an equivalent number of credits was negligible.

Given that students need time to realise the full benefits of training, accounting for this substantially increases estimates of long-term earnings gains. Jacobson et al. (2005) showed that displaced workers were worse off immediately after leaving college relative to what they would have been without returning to school, but that earnings tended to rise quickly in the following year, before leveling out to a new, higher level. This highlights the need for long follow-up periods in the data when assessing the returns to education.

3 Context: Higher education and Vocational study in Australia

Mature-age education in Australia is among the highest in the world. In 2014, Australias participation in vocational education by those aged 25-64 was the highest among OECD countries. The tertiary education rate for those aged 30-64 was the second highest (Perales and Chesters, 2017). Mature-age Australians are increasingly enrolling in university or college to change employers, change careers, gain extra skills, improve their promotion prospects and earning capability or search for better work/life balance. Redundancy and unemployment have also been driving forces for individuals to return to education later in life (Coelli, Tabasso and Zakirova, 2012).

The increase in mature-age learners accessing higher education has in part been driven by government policy. In 2009, the Australian government adopted a national target of

[1] US dollars in 2014.

at least 40% of 25-34-year-olds having attained a qualification at bachelor level or above by 2025 (O'Shea, 2014). In 2017, 39% of 25-34-year-olds had a bachelors degree or higher (Caruso, 2018).

The transition to a demand-driven system in Australia in 2009 (Universities Australia, 2020) also had a large effect on access to higher education, as it removed the cap on the number of university student places. As a result, the number of Commonwealth Supported Places (CSPs) increased by 150% between 2008 and 2017[2] (Universities Australia, 2020) as more students could access university at a subsidised rate. The introduction of the Higher Education Participation and Partnerships Program (HEPPP) in 2010 helped to improve access for disadvantaged students as higher education providers were granted funding according to their share of domestic undergraduate students from low SES backgrounds (DESE, 2021). In 2021, the definition of disadvantage was expanded to include the share of students from regional and remote areas and Indigenous backgrounds. The Higher Education Loan Program (HELP), introduced in 1989 as HECS and originally designed to support CSP (undergraduate) students by offering income-contingent loans, was expanded in 2005 to support full-fee paying domestic students, and again in 2007 to support students undertaking higher education VET courses (Parliament of Australia, 2017). This broadened access to loans to students undertaking courses other than a university bachelors degree. Since most CSPs are awarded for undergraduate degrees, the majority of CSP students are under 25 years old (78% in 2018) (Universities Australia, 2020). Students pursuing post-graduate or vocational courses generally do not qualify for a CSP place. Instead, they can defer their tuition fees through HELP loans, allowing the mature-age cohort to access higher education while juggling other financial commitments.

While the initial uptake of university places in the demand-driven system was strong, especially among mature-age students[3] (Universities Australia, 2019), growth in undergraduate enrolments slowed since 2012. In 2018, mature-age enrolments even dropped below the previous year. The 40+ age group showed the worst growth, receding by 10%, while the 25-29s and 30-39s showed growth of around -4% (Universities Australia, 2020). The decline of enrolments coincided with the freezing of the Commonwealth Grant Scheme (CGS) which capped funding at 2017 levels, effectively ending the demand-driven system (Universities Australia, 2020). Access to Commonwealth Supported Places (CSPs) have since been limited to 2017 levels, with cap raises from 2020 subject to performance measures (Universities Australia). As a proportion of the working age population, mature-age students also participated less in vocational education and training (VET) over the same

[2]From approximately 4,000 places to 12,000 places.

[3]Between 2010 and 2012, growth in mature-age enrolments in undergraduate courses doubled for the 30-39 age group and tripled for the 40+ age group.

period. It appears the introduction of the demand-driven system also increased VET participation between 2010 and 2012, before continuing its decline (Atkinson and Stanwick, 2016). Total VET enrolments since 2018 stabilised, with 2019 and 2020 enrolments slightly above 2018 levels[4] (NCVER, 2021). The impact of COVID-19 on 2021 enrolments is yet to be fully determined. So far, VET enrolments for the first half of 2021 are well above the previous 4 years across all age groups, with ∼1 million enrolments in 2021 compared to ∼870 thousand enrolments in 2017[5] (NCVER, 2021).

The demand-driven system almost doubled the number of undergraduate enrolments from underrepresented groups between 2008 and 2018, including students from low socio-economic backgrounds, regional and remote areas, Indigenous backgrounds, and students with a disability (Universities Australia, 2020). Despite the large increase in enrolments, attrition and completion rates have not changed dramatically over the same period and show mixed results. Low SES students show lower rates of attrition at TAFE[6] in both postgraduate (-14pp) and bachelor courses (-5pp) (DESE, 2019a). While low SES students also improved their attrition at university[7] at post-graduate level (-1pp), this group had higher attrition at the bachelor level (2pp). Completion rates followed a similar trend (in the opposite direction) (DESE, 2019b). The pattern is very similar for remote and regional students, with attrition rates falling in all categories except for a slight increase for regional bachelor students. Post-graduate Indigenous students show large drops in attrition at both the university and TAFE level, by between 7 and 11pp, while bachelor students at both institution types show a slight increase in attrition (0-1pp). Mature-age cohorts, across all students and not just disadvantaged groups, reveal the same trend, with attrition and completion rates improving across all categories except university bachelors degrees. However, compared with 20-24-year-olds, mature-age students have lower completion rates in general, regardless of level or institution, but this has not changed overtime. Similar to the 20-24 age group, mature-age students have twice as much success completing postgraduate degrees than bachelors degrees (∼ 30% vs 60%). The same is true for advantaged groups[8] , with bachelor students at university also showing slightly higher attrition (∼2pp) over the same period. Since all groups of students are less likely to complete university-level bachelor courses, it appears that broadening the student cohort to disadvantaged groups has not reduced the success rates in higher education, and

[4]Total VET enrolments 2016-2020.

[5]Government funded program enrolments Jan-June 2017-2021.

[6]Table C and NUHEI (non-university higher education institutions) providers, includes private universities, colleges, and TAFEs.

[7]Table A and B providers, includes public universities and self-accredited private universities.

[8]Medium & High SES, Metropolitan, and Non-indigenous students.

in fact, disadvantaged groups have improved their rates of completion and attrition in the last 10 years since 2008 in all other courses.

The Australian government introduced an additional scheme in 2019 to support workers aged 45-70 looking to reskill or upskill in order to remain employed (DESE, 2022a). Workers must be at risk of entering the income support system or recently unemployed to be eligible for support. The Skills and Training Incentive program subsidises accredited and non-accredited courses as long as the training is linked to their current job, a future job opportunity or an industry/skill in national shortage. Eligible courses are identified when individuals complete a Skills Checkpoint assessment and can be subsidised up to $2200, regardless of the number of courses taken. Within this limit, up to 75% of course fees can be subsidised when they relate to skills in national shortage, while all others can receive up to 50% of course fees. The number of incentives available per year are capped at increasing levels up to 2024 (between 1,800 and 7,500) when the scheme is set to end (DESE, 2022b).

The cost of a bachelors degree for domestic students in Australia is the sixth highest among OECD countries (Universities Australia, 2020). In 2018, the average annual cost of a bachelors degree was around $5,000 in Australia, about half of the top 2 most expensive countries where it costs around $9,000 in the US and $12,000 in the UK[9]. VET and TAFE courses in Australia cost a minimum of $4,000 per year on average while postgraduate courses cost a minimum of $20,000 per year on average[10] (Studies in Australia, 2018).

To cover the cost of higher education, Australian domestic students have two main ways to receive support from the government. The Commonwealth Supported Place (CSP) scheme, mainly offered to undergraduate students, subsidises tuition fees for those studying at public universities and some private higher education providers. Students who meet the eligibility criteria (domestic students at an approved education provider) are automatically assigned a CSP. Most CSPs are for undergraduate study but some providers offer CSPs at the postgraduate level. From 2022, CSP places are capped at 7 years of full-time study. Once this allotment has been used, an additional 3 years of full-time subsidised study will be available 10 years from start date of the last course (StudyAssist, 2022a).

The second, and broader, source of financial support is the Higher Education Loan Program (HELP) which provides income-contingent loans to students at university or higher education providers. This allows students to defer their tuition fees until their earnings

[9]Values are in US dollars.
[10]Values are in Australian dollars.

reach the compulsory repayment threshold, upon which repayments are deducted from their pay throughout the year at a set rate. Voluntary repayments can also be made. In 2022, HELP loans were limited to around $110,000 and the repayment threshold was around $47,000 [11] (StudyAssist, 2022b). To qualify for HELP, students must be studying at approved HELP providers. Once accepted into their course, students can apply for HELP via their institution by submitting their tax file number and a HELP form.

HECS-HELP is the HELP scheme available to CSP students while FEE-HELP is the HELP scheme available to full-fee paying students who dont qualify for a CSP i.e., post-graduate students. VET Students Loans (formerly VET FEE-HELP) are also part of the HELP scheme and are available to students undertaking vocational education and training (VET) courses outside of higher education (Universities Australia, 2020). CSPs and HELP loans are withdrawn from students who fail half of their subjects, assessed on a yearly or half-yearly basis depending on the level of study[12].

4 Data

We use data from the Household Income and Labour Dynamics Australia survey. These data are rich, and we exploit the full set of background information on individuals (beginning with more than 3,400 variables per observation).

HILDA covers a long time span of nearly 20 years, starting in 2001. We use the 2019 release. This means we observe respondents annually from 2001 to 2019.

Sample exclusions

Our analysis sample contains respondents who were 25 years or above in 2001. This allows us to focus on individuals who obtain a further education beyond that acquired in their previous degree. To ensure we analyse those who pursue further study, we delete any individuals who was currently studying in 2001. This also ensures that our covariates, which are defined in 2001 are not contaminated by the impacts of studying but clearly precede the study spell of interest. These sample exclusions result in 7,359 respondents being dropped because they are below the age of 25 in 2001 and a further 1,387 respondents being dropped because they were studying in 2001.

We restrict the sample to those who are present in both 2001 and 2019. This ensures that we observe base characteristics and outcomes for every person in our analysis sample.

[11] Values are in Australian dollars.

[12] Yearly at bachelor level and per trimester for courses lower than bachelor level.

This results in a further 5,727 respondents being dropped from the sample. Our final analysis sample is 5,441 observations.

Outcomes

We measure outcomes in two ways. The first is to compare outcomes in 2019 across the groups of individuals who did and did not get re-educated. The second is to compare within-person changes in outcomes between 2001 and 2019.

We analyse the outcomes related to the labour market such as earnings, employment, changes in earnings, changes in occupation, industry, and jobs. We also analyse well-being and mental health measures to assess the potential broader impacts of getting re-educated.

Treatment

We define further education as an individual who obtains a further degree in a formal, structured educational program. These programs must be delivered by a certified training, teaching or research institution. Thus, we do not analyse informal on-line degrees (such as Coursera degrees). We also do not consider on-the-job training as obtaining further education.

Our treatment variable is a binary variable that takes the value of 1 if an individual has obtained an additional degree anytime between wave 2 (2002) and wave 17 (2017). As we analyse outcomes in 2019, this means we calculate the average returns between 2 years and up to 17 after course completion. We delete any respondent who obtained a qualification after wave 17. This allows us to analyse outcomes at least two years after course completion.

HILDA documents formal degree attainment in two ways. The first is to ask respondents, in every, wave what is their highest level of education. The second way is to ask respondents, in every wave, if they have acquired an additional educational degree since the last time they were interviewed.

We utilise both these questions to construct our measure of further education. Using the first question, we compare if the highest level of education in 2019 differs from that in 2001. If there has been an upgrade in educational qualification between these two years, we set the treatment indicator to be one and zero otherwise. This question, however, only captures upgrades in education; it fails to capture additional qualifications that are at the same level or below as the degree acquired previously by the respondent. We rely on the second survey question to fill this gap.

These two survey questions thus capture any additional qualification obtained from 2002 to 2017, inclusive. Additional qualifications refer to the following types of degrees: Trade certificates or apprenticeships; Teaching or nursing qualifications, Certificate I to IV, Associate degrees, Diplomas (2-year and 3-year fulltime), Graduate Certificates, Bachelor, Honours, Masters and Doctorate degrees.

Covariates

We define our covariates using 2001 as the base year. Since we delete any respondents who were currently studying in 2001, we ensure that all of the covariates were defined before a respondent begins further study.

A unique approach to our feature selection strategy is that we use all the information available to us from the HILDA survey in 2001. This means that we have more than 3,400 raw variables per observation. Before using the covariates in a ML model, we delete any features that are identifiers or otherwise deemed irrelevant for explaining the outcome.

In order to reduce redundancy in this vast amount of information, we first apply a supervised Machine learning model to predict outcomes 5 years ahead of 2001 i.e., in 2006. We then select the top 100 variables that are most predictive of the outcome in 2006.[13] These variables are listed in Table 1.

Variables selected by the ML model but not the baseline model

As a descriptive exercise, Table 2 presents the features that were missed by the baseline model. In the baseline model, we included covariates such as age, gender, state of residence, household weekly earnings, highest level of education attained, and current work schedule. This collection of variables have been informed by theory or empirical experience. For brevity, we call this the theoretical model.

The data-driven model that we estimate includes far more variables compared to the baseline model. We ascertain which of these variables explain the residual variation in the outcome: we regress the residuals from the theoretical models (without the treatment included) on the features included in the data-driven model and train a LASSO model to highlight the salient variables that were missed. The variables that are chosen are listed in Table 2. We also document how these variables are correlated to the outcome and to

[13]Confounders are features that both have an impact on the outcome and on the treatment. Chernozhukov et al. 2018 suggest including the union of features kept in the two structural equations (outcome on features and treatment on features). Here, we only include the features that predict the outcome equation because including features that are only predictive of the treatment can erroneously pick up instrumental variables (see paper for a discussion of this issue).

the treatment in order to give us a sense of the direction of the bias their omission may induce.

Variables excluded from the theoretical model include doctorate qualifications; employment conditions such as work schedule, casual employment, firm size, tenure or years unemployed; financial measures such as weekly wage, investment income and mortgage debt; health measures such as limited vigorous activity and tobacco expenses; and work-life preferences related to working hours and child care. Most of the omitted variables bias the OLS estimates upwards, except for doctorate qualifications, casual employment, years unemployed, parental child care and dividend and business income.

It is important to highlight that this exercise is a quick, descriptive one. As previously mentioned, the ML algorithm randomly selects variables that are highly correlated thus we may have missed out on reporting the label of important variables omitted from the theoretical model.

5 Descriptive Figures and Tables

We calculate the average returns to degree completion for mature-age students who completed degrees between 2002 and 2017. The window in which study and degree-completion took place is noticeably large. However, sample size limitations with our survey data mean that it is not feasible to run an ML analysis, disaggregated by the timing-of-completion.

In order to obtain some insights into the potential heterogeneity over time, we present a series of descriptive graphs in this section. Here, our aim is not to present any causal analysis but to describe which groups studied earlier in the time period (and thus had more time to accumulate returns). These graphs can also point to the potential different factors driving study across the time period, and different effects on earnings depending on how much time has elapsed since completion.

Figure 3 presents the distribution of degree completion over time. There is a steep decline in degree-completion proportions over time. This is likely to reflect the aging profile of HILDA survey respondents and that further study is disproportionately higher among the younger cohorts (25-44 year olds) (See Figure 8).

Over time, Figure 4 shows that the composition of degrees completed has shifted. Among those who completed a degree in later years, compared to those who completed a degree in the earlier period, a higher percentage completed a Certificate III or IV, Diploma or Advanced Diploma as opposed to a lower-level degree (Certificate I or II or below). In all

years, the most frequently completed degrees are Cert 3 or 4, Associate degrees, Diplomas and Advanced Diplomas.

Figure 5 shows that the gender composition for degree completions between 2002 and 2017, inclusive. Females tend to be more likely to complete degrees than males and this gender disparity has slightly increased over time. However, there is a stark difference in the types of degrees completed across gender.

Figure 6 shows the distribution across degree type over time, and further broken down by gender. Males are more likely to complete a Certificate III or IV, Diploma or Advanced Diploma, than females. By contrast, females are more likely to complete a Certificate I or II or a degree below this level. Females, however, as also more likely to complete a degree at the level of Bachelor or higher, than males. This trend that has increased across time with a sharp discontinuity occurring in 2012.

Figure 9 shows an increase in both average earnings and employment overtime between 2002 and 2017. Despite the upward trajectory, these outcomes show more volatility following 2008. Figure 10 breaks these outcomes down by gender. Among females, the proportion employed increased up to 2012 to rise above the proportion of males employed. Following 2012, female employment trended down while male employment trended up. While female average earnings were below male average earnings through most of the period, the catch up to male earnings subsided in 2012 also.

6 Method

We use a range of ML-based techniques.

6.1 T-Learner model

The first is the T-learner approach. We aim to measure the amount by which the response Y_i would differ between hypothetical worlds in which the treatment was set to $T = 1$ versus $T = 0$, and to estimate this for average across subpopulations defined by attributes X.

Under the potential outcomes framework of Imbens and Rubin 2015, we use $Y(0)$ and $Y(1)$ to denote the outcomes we would have observed if treatment were set to zero or one, respectively. In reality, we only observe the potential outcome that corresponds to the realised treatment:

$$Y_i = T(Y(1) + (1 - T)Y(0) \tag{1}$$

Here, we denote the binary treatment indicator $T \in \{0; 1\}$.

Identification assumptions

To interpret the estimated parameter as a causal relationship, the following assumptions are needed:

1. Conditional independence (or conditional ignorability/exogeneity or conditional unconfoundedness) Rubin (1980): $Y(0)$ and $Y(1)$ are orthogonal to T conditional on X.

   This assumption requires that the treatment assignment is independent of the two potential outcomes. Practically, this amounts to assuming that components of the observable characteristics available in our data, or flexible combinations of them, can proxy for unobservable characteristics. Otherwise, unobservable confounding bias remains.

   A benefit of using all the features the HILDA dataset has to offer is that we may minimise unobserved confounding effects. Specifically, we rely on the 3,400 covariates and complex interactions between them as well as flexible functional forms to proxy for components of this unobserved heterogeneity. For example, while we do not observe ability or aptitude directly, we may capture components of it with other measures that are observed in HILDA such as past educational attainment or the long list of income and other sources of income variables (see Table 1 for a list of the covariates).

   The reader is likely to conceptualise other dimensions of unobserved heterogeneity that may not be captured in Table 1. There are two likely scenarios in this case. First, HILDA may not be exhaustive enough, even with its existing richness, to capture all dimensions of unobserved heterogeneity. As a result, our estimates may be biased.

   Another potential scenario is that the source of unobserved heterogeneity in question (or some components of it) is still captured but modelled under the guise of another variable label. Variables that are highly correlated with each other are unlikely to be simultaneously included in the model. This is because the ML algo-

rithm, in attempting to reduce the amount of information redundancy, may have randomly dropped one or more of those correlated variables.

2. Stable Unit Treatment Value Assumption (SUTVA) (or counterfactual consistency): $Y_i = Y0_i + T_i(Y1_i - Y0_i)$. Assumption 2 ensures that there is no interference, no spill-over effects, and no hidden variation between treated and non-treated observations. SUTVA may be violated if individuals who complete further education influence the labour market outcomes of those who do not complete further education. For example, if the former group absorb resources that would otherwise be channelled to the latter group. Alterntiavely, the former group may be more competitive in the labour market and reduce the probabiltty of promotions or job-finding for the latter group. As those who complete further education are a relatively small group, it is unlikely that these general equilibrium effects would occur.

3. Overlap Assumption (or common support or positivity). Assumption 3 states that no subpopulation defined by $X = x$ is entirely located in the treatment or control group, hence the treatment probability needs to be bounded away from zero and one.

   The overlap is an important assumption because extrapolation using the two predictive equations, $\mu_1(x)$ and $\mu_0(x)$ models, is likely to perform best for subpopulations defined by $X = x$ entirely located in the treatment or control group.

   As these models are trained and evaluated only in the regions of the X-covariates or features for which there are treated observations and control observations, respectively, non-overlap means there is no way for us to validate whether the function performs well.

   This means the optimum bias-variance trade-off point for the conditional average treatment effect may not align with the optimum bias-variance trade-off point for the separate $\mu_1(x)$ and $\mu_0(x)$ models. Since, ultimately we are interested in the CATEs (as opposed to the predictive accuracy of the individual conditional mean functions), this can mean that we have biased CATEs.

4. Exogeneity of covariates. This assumption means that the covariates included in the conditioning set are not affected by the treatment. To ensure this, we define all of our covariates at a time point before any individual started studying. Specifically, we use the first wave of HILDA (in 2001) to define our covariates. We only look at those individuals who completed further education in 2002 onwards. Furthermore, we delete any individuals who were currently studying in 2001 to ensure the covariates cannot reflect downstream effects of current study.

With the strong ignorability and overlap assumptions in place, treatment effect estimation reduces to estimating two response surfaces one for treatment and one for control.

The T-learner is a two-step approach where the conditional mean functions $\mu_1(x) = E[Y_1, X_i = x]$ and $\mu_0(x) = E[Y_0, X_i = x]$ are estimated separately with any generic machine learning algorithm.

Machine learning methods are well suited to find generalizable predictive patterns, and we employ a range of model classes including linear (LASSO and Ridge), non-linear (Gradient Boosting Regression) and Bayesian ML models. Once we obtain the two conditional mean functions, for each observation, we can predict the outcome under treatment and control by plugging each observation into both functions. Taking the difference between the two outcomes results in the Conditional Average Treatment Effect (CATE).

To show this, we define our parameter of interest, the CATE ($\tau(x)$), which is formally defined as:

$$\tau(x) = E[Y1_i - Y0_i | X_i = x] \tag{2}$$

which, with the assumptions above, is equivalent to taking the difference between two conditional mean functions $\mu_1(x) - \mu_0(x)$:

$$\begin{aligned} \tau(x) &= \mu_1(x) - \mu_0(x) && (3)\\ &= E[Y_i | D_i = 1, X_i = x] - E[Y_i | D_i = 0, X_i = x] && (4)\\ &= E[Y1_i - Y0_i | X_i = x] && (5) \end{aligned}$$

In this estimation, we are not interested in the coefficients from regressing Y on X. What we want instead is to have a good approximation of the function and hence good estimates from e.g., $\mu_1(x)$ and $\mu_0(x)$. This is why ML methods are well suited for the job.

A benefit of our set-up is that when we take the difference between the two conditional mean functions, we coincidently find the optimum bias-variance trade-off point for the conditional average treatment effect. This means that we have an indirect way to obtain the best prediction of the CATE through two predictive equations, where we observe the true outcomes (and thus are able to regularise).

In practice, however, this indirect way of minimising the mean squared error for each separate function to proxy for the minimum mean squared error of the treatment effect can be problematic. See, for example, Knzel et al. (2019); Kennedy (2020) for settings when the T-learner is not the optimal choice.

One potential estimation problem arises when there are fewer treated individuals than control individuals and the individual regression functions are complicated (very non-smooth). This means they can be difficult to estimate well on their own. The T-learner can be inefficient because it does not exploit the shared information between treatment and control observations: for example, if X relates to Y in the same fashion for treated and control observations. As a result, the estimate μ_1 tends to over smooth the function; in contrast, the estimate μ_0 regularises to a lesser degree because there are more control observations. This means a naive plug-in estimator of the CATE that simply takes the difference between $\mu_1 - \mu_0$ will be a poor and overly complex estimator of the true difference. It will tend to overstate the presence of heterogeneous treatment effects.

6.2 Doubly Robust model

The second approach is Doubly Robust. A benefit here is that we use additional information from the propensity score (we employ machine learning models to gain a better understanding of the treatment assignment process, the students background, and the nature and complexity of their situation that may have led them to pursue further education). Thus, the doubly robust approach can improve upon the T-learner approach because it can reduce misspecification error either through a correctly specified propensity score model or through correctly specified outcome equations. Another feature of the Doubly Robust approach is that it places a higher weight on observations in the area where the relative count of treatment and control observations is more balanced (i.e. the area of overlap). A benefit of this is that it can also provide better extrapolations of the predicted outcomes.

$$\hat{ATE} = \frac{1}{n}\sum_{i=1}^{n}[\frac{T_i(Y_i - \hat{\mu_1}(X_i))}{\hat{p}(X_i)} + \hat{\mu_1}(X_i)] - \frac{1}{n}\sum_{i=1}^{n}[\frac{(1-T_i)(Y_i - \hat{\mu_0}(X_i))}{1-\hat{p}(X_i)} + \hat{\mu_0}(X_i)] \quad (6)$$

where:

$\hat{p}(X_i)$ is an estimation of the propensity score (using logistic regression)

$\hat{\mu_1}(X_i)$ is an estimation of $E[Y|X, T = 1]$ (using any ML model)

$\hat{\mu_0}(X_i)$ is an estimation of $E[Y|X, T = 0]$ (using any ML model)

Previously, with the T-learner, we were just estimating μ_0 and μ_1. With the DR estimator, we augment μ_0 and μ_1. For example, for the treated observations, we augment μ_1 by multiplying the prediction error by the Inverse Propensity Weight (or 1 divide by the Propensity Score). This upweights those who get treated but who are statistically similar to the control observations. We then apply this same augmentation to the μ_0 for the control observations.

6.3 Bayesian Linear Models

A third approach is to use a Bayesian linear model. The Bayesian approach models the outcome surface as,

$$\begin{aligned} y_i &\sim N(\mu(x_i, t_i), \sigma^2) \\ \mu(x_i, t_i) &= w_0 + w_t t_i + w_x^\top x_i + t_i w_{tx}^\top x_i \end{aligned}$$

Where we have used the following priors,

$$\begin{aligned} \{\lambda_0, \lambda_t, \lambda_x, \lambda_{tx}\} &\sim Gamma(1, 1) \\ \sigma^2 &\sim Gamma(1, 1) \\ w_0 &\sim N(0, \lambda_0) \\ w_t &\sim N(0, \lambda_t) \\ \mathbf{w}_x &\sim N(0, \lambda_x \mathbf{I}_d) \\ \mathbf{w}_{tx} &\sim N(0, \lambda_{tx} \mathbf{I}_d) \end{aligned}$$

The advantage of this model is that it is free to share information between both treatment and control response surfaces during posterior inference (model fitting). This is unlike a T-learner, which uses entirely separate models and data to estimate the control and response surfaces. We believe this information sharing property of the Bayesian model helps reduce [error/variance?] in effect estimation from model extrapolation in circumstances where there is little overlap in the support of $P(x|t = treatment)$ and $P(x|t = control)$, which is an issue with the T-learner approach. This information sharing comes from the fact that the λ^* hyper-parameters are all from a shared distribution, which incorporates

information from t_i, x_i and the interaction $t_i \cdot x_i$ to update the posterior over these parameters. For model inference, we use the no U-turn MCMC sampler [Hoffman et. al. 2014] in the pyro software package [Bingham et. al. 2019]. The choice of a Gamma(1, 1) prior is partially motivated by the fact that we standardise the inputs, x, to the model to have zero mean and unit variance and $t \in \{0, 1\}$. We also tried a sparsity inducing Laplacian prior over weights, w, but we did not find a substantial difference from using Normal priors.

6.4 Model selection and model evaluation

We separate the evaluation of the model class and estimation of the ATE and CATE parameters in two procedures. We evaluate the predictive capacity of each model class using nested cross-validation. The procedure is represented in Figure 1. Here, our aim is to compare the predictive performance of three model classes: LASSO, Ridge and Gradient Boosting Regression (GBR). Our second procedure is to estimate the ATE and CATE parameters. The procedure is represented in Figure 2. We use bootstrap sampling (with replacement) to generate uncertainty estimates for the parameters, which we obtain over several draws of the same model class.

Focusing on the first procedure, we apply nested cross-validation to evaluate which model class performs best. In a first step, as Figure 1 shows, we pre-process the full dataset (containing 3,400 variables) to generate a dataset with a smaller set of highly predictive features (containing 100 variables). We apply a supervised machine learning approach with a LASSO model to select our top 100 predictors of the outcome of interest (as measured in 2006). Note that in our later estimations of the treatment effect, the outcome is measured in 2019. We implement this intermediary step in order to reduce the correlation between variables and eliminate redundant information.

We assume that the top 100 features that are most predictive of the outcome in 2006 correlate with the features that would be most predictive of the outcome in 2019. By choosing to apply this pseudo-supervised ML approach on the same outcome variable (but measured at a different time point) means that we obtain a good indication of the features that are useful for a model to perform well. Improved model performance here will also mean that the selected features are likely to represent the important confounders. We have chosen 2006 to ensure there is no overlap with 2019 outcomes, and thus to avoid overfitting issues.

Using the top 100 predictors, we apply nested CV approach in order tell us which model class is best. Here we measure the performance of the model class on unseen data. First,

we split the data into train and test folds (80-20 split). Within the 80 percent train fold we perform 5-fold cross-validation in order to train and evaluate the performance of each configuration of parameters. We do this separately for the outcome surface using the treated observations and the outcome surface using the control observations. From this, we select the models with the best mean scores. We then evaluate the selected model on the holdout test set.

We repeat this process ten times (10-outer scores) for each model class. This allows us to evaluate the performance based on the mean and standard deviation of these scores. Note that thus far, we have not evaluated any particular configuration of the model, rather the performance of the model class.

Table 3 shows that the GBR is the best performing model class. It is the model that yields the highest out-of-sample R-squared or the lowest (in absolute value) MSE. This is true for both the outcome surfaces.

Inference via bootstrapping

Once we have selected the best performing model class, we turn to the estimation of the parameters and parameter uncertainty. We use bootstrapped validation for generating uncertainty estimates. This captures the uncertainty arising from model configuration (or the selection of hyperparameters) in addition to that which stems from estimating parameters of a fixed model from noisy, finite data. Here, we estimate the same model class on different realisations of the data.

A common approach to inference in the causal machine learning literature is to use sample splitting (Athey and Wager 2019). Sample splitting ensures that the standard errors on the estimators are not underestimated because it avoids using the same data point to both select the configuration (hyperparameter selection) of the model and to estimate the parameters. When the same data point is used to perform both tasks then the standard errors would not reflect both the uncertainty stemming from model selection and that which stems from estimating parameters on noisy, finite data. The result of this is that our standard errors suffer from pre-test bias: i.e. selecting the model with the same data you use to estimate the model, which can lead to overfitting.

However, sample splitting is appropriate when the sample size is large. An issue with studies that rely on survey-based data is that sample sizes are not large enough to perform sample splitting. For example, there is not enough data to split the dataset into train and test datasets such that each of these splits would cover all the common and uncommon values of the X-covariates that are observed in the full sample. If we were to use a training dataset that was insufficiently sized or non-representative, it would be difficult for the

ML models to effectively map the X-covariates to the outcome surfaces, 1(x) and 0(x). There would also not be enough data in the test set to effectively estimate the parameters of the model configuration chosen in the train set. As a result, our estimate treatment effects are likely to be very imprecisely estimated.

A suitable alternate procedure is to use bootstrapping. Resampling gives us a decent estimate on how the point estimates might vary. In this way, we side-step the need to rely on the assumption of asymptotical normality (and we do not need to utilise sample splitting to generate the standard errors). In our bootstrapping procedure, we ensure that the standard errors reflect the sources of uncertainty stemming from both the selection of the model and the estimation of the model. As a result, we generate standard errors that avoid any potential pre-test issues.

As a first step, as shown in Figure 2, we obtain the 100 top predictors from the initial pre-processing of the full dataset. i.e. we train a supervised machine learning LASSO model to extract which features best predict earnings in 2006.

As a second step, we train our models using the 100 top predictors on the first bootstrapped sample to select the best models for $\mu_1(x)$ and $\mu_0(x)$. Within this bootstrap sample, we divide the dataset into five folds and perform cross-validation to select the best model configuration. Similar to the cross-validation description above, our model configuration is trained on subsets of the data, and then evaluated on holdout samples (where we compare the predictions made from the models for the outcome surfaces for the treated and control observations to the real values). We employ group-5-fold cross validation such that groups are the bootstrap indices of the original data. This ensures that training data does not simultaneously appear in the validation set. We perform this model selection step within the bootstrapping procedure to capture the uncertainty coming from the selection of hyperparameters. If we simply re-estimated the same model (with a given set of hyperparameters) in each bootstrap model then the uncertainty is only over the model parameters, and not the model choice (e.g. the GBR tree depth).

This process allows us to select the best model configurations for the outcome surfaces: $\mu_1(x)$ and $\mu_0(x)$. Third, and once we have these predicted outcome equations, we are able to calculate the individual treatment effect, $\tau(x)$ for each person in the original sample (note not the individuals from the bootstrap sample). We substitute the values of their features or independent variables into the LASSO/ Ridge equations for $\mu_1(x)$ and $\mu_0(x)$ or tree-based models.

At this point, each individual in the original sample has one value of $\tau(x)$. We can obtain a mean $\tau(x)$ by averaging over all the individuals. This is the mean $\tau(x)$ for the first

bootstrap sample. We repeat this procedure over 100 bootstrap samples. This provides an empirical distribution of the mean $\tau(x)$'s. The final mean over the bootstrap samples (the mean of the mean $\tau(x)$'s), which should be the same as the $\tau(x)$ from the original sample when the number of bootstrap samples is large enough, is how we calculate the ATE and CATE estimators. The standard deviation of this series through bootstrapping represents the standard error estimate for the ATE or CATE measure.

Note we want to generate the value of ATE or CATE parameters, keeping the original sample size constant, thus all our bootstrap samples have the same sample size as that of the original sample.[14]

The main goal of repeating over the bootstrap samples is to capture uncertainty coming from the model selection procedure (i.e. the choice of hyperparameters) as well as the uncertainty over the model parameters.

Each bootstrap sample gives us an estimate of the C(ATE). Across 100 bootstrap samples, we have a distribution of C(ATE) estimates. We take the standard deviation of these 100 C(ATE) values to obtain the standard error.

To obtain the confidence intervals, we undertake the following procedure:

1. Multiply the C(ATE) by two
2. Take the 5th and 95th percentile of the C(ATE) distribution
3. Calculate the bounds of the confidence interval
   (a) Upper bound = Step 1 - 5th percentile
   (b) Lower bound = Step 1 - 95th percentile

To do this we approximate the critical values of the confidence interval as $\delta^* = \bar{x}^* - \bar{x}$

such that the confidence interval

$P(\delta_{0.95} \leq \bar{x} - \mu \leq \delta_{0.05}|\mu) \Leftrightarrow P(\bar{x} - \delta_{0.95} \geq \mu \geq \bar{x} - \delta_{0.05}|\mu) = 0.90$

can be reduced to $[2\bar{x} - \bar{x}^*_{0.05}, 2\bar{x} - \bar{x}^*_{0.95}]$.

[14]The bootstrap resample is the same size as the original sample because the variation of the ATE depends on the size of the sample. Thus, to approximate this variation we need to use resamples of the same size.

7 Results

There are clear economic benefits to gaining an additional qualification in later life (25 years or older). Table 4 displays a gain of approximately $88-110 per week in gross earnings. In 2019, this is roughly 7-8 percent of the average gross weekly earnings of $1256.20 for all Australian employees (ABS, 2019; 6345.0 Wage Price Index, Australia).

The effect sizes from the GBR model are smaller than that of the two linear models. GBR is better capture non-linearities. For example, age is likely to exhibit a highly non-linear relationship with earnings in 2019. Those who were aged 46 or above in 2001 will be aged 65 or above in 2019. This means they are more likely to have retired by 2019 compared to those who were aged below 46 in 2001. As a result, we may expect a shift down in earnings at age 46.

To illustrate how GBR adequately captures non-linearities we re-estimated our results focusing on those who were aged 25-45 in 2001. This is the same as interacting a binary variable (for age 25-45) with every other covariate in the model. In Appendix Figure 29, we see that the results across the models are now more similar than when we use the full sample.

Turning to other measures of earnings, the value-add in earnings (taking the growth in earnings between 2001 and 2019 - all expressed in 2019 terms) is also higher for those who gained an additional qualification, compared to those who did not advance in their education, by approximately $60-80 per week. As a proportion of the growth in earnings for all Australian employees between 2001 and 2019, this represents 18 percent of the overall growth in earnings.

The similarity in the results between the value-add increase in earnings and the level increase in 2019 earnings suggests that, once we control for background characteristics, we have accounted for inherent or base level differences between those who do and do not obtain an additional qualification. In other words, the difference in 2019 earnings reflects the causal effect of advancing education in later-life.

Proportionate changes in earnings can be measured by taking the log of the earnings measures. In Appendix Figure 30, we see that the proportionate change in earnings was large at 50 percent. This is likely to be because many people entered the labour market as a result of the new qualification.

Our ML models estimate returns that are smaller than the returns estimated in a Fixed Effects (FE) or cross-sectional models (OLS with and without controls) where features have been selected based on theory. Figure 18 compares the returns from five different

approaches: an ordinary least squares model with no covariates, with covariates selected using theory, a FE model and ML models.

The effects remain strong up to a decade-and-a-half after course completion. The largest earnings gains are associated with acquiring an undergraduate degree or above (including Graduate certificates and Graduate diplomas). As displayed in Figure 19, only modest gains are associated with secondary or post-secondary certificates or diplomas. More generally, significant earnings gains are achieved when returning students acquire qualifications above those previously held. In fact, the earnings gains are much higher if students advance their qualification status rather than gaining an additional degree at the same or lower level.

The magnitude of the returns to education also differs depending on the type of subject studied. Obtaining a degree in a technical subject is likely to yield a return that is twice as high as obtaining a degree in a nontechnical degree. The returns are statistically significant for the former but estimated with a high variance for the latter. Technical subjects are STEM subjects, medicine, and health-related fields whereas nontechnical subjects encompass those in the creative arts, arts, humanities, and social science disciplines. These results are also displayed in Figure 19.

Qualification advancements do not benefit Australian workers in the same way: women appear to benefit more from later-life study than men (see Figure 20). The gains to further education for women are likely to stem from the process of completing the degree. By contrast, for men, the difference in earnings (in 2019) between those who do and do not gain an additional qualification appear to emerge due to selection into adult learning: men who completed further education had higher earnings prior to enrolment compared to men who did not acquire an additional degree. Comparing the earnings growth, as opposed to the level differences in 2019 earnings, between these two groups of men shows that the value-add of further study is much smaller and now statistically insignificant.

Larger earnings gains are also achieved when qualifications are acquired at older ages. The value-add of an additional degree for older adult learners is more than double that of learners who acquire additional degrees at younger ages. This contrasts with findings from previous studies, which show higher returns when degrees are acquired at younger ages (Polidano and Ryan, 2016; Dorsett, Lui and Weale, 2016; Perales and Chesters, 2017). A possible reason for the difference is that we analyse older Australians and compared to some of these studies, we consider a broader range of degree types.

On the face of it, the returns to education appear larger for those with lower starting individual earnings in the main job compared to those who start out (in 2001) with a

higher weekly salary. Specifically, Figure 21 shows that individuals who earn less than the median value (of $425 per week) benefit twice as much from an additional qualification compared to those who start out with a salary below the median rate. However, this is largely driven by selection effects. Among workers with starting salaries below the median rate, those who completed further education had higher earnings prior to enrolment compared to those who did not acquire an additional degree. This positive selection into further education may arise because the base group of those who do not pursue further education includes the long-term unemployed. In other words, the value-add of an additional qualification for those who start out with lower earnings is smaller once we account for selection effects.

Among workers who begin with higher levels of individual earnings i.e., their weekly gross earnings in the main job are larger than the median rate, the value-add of gaining an additional degree is in fact, larger than the comparison of level earnings in 2019. One potential explanation for this is that among this group, workers with relatively lower earnings than their higher earning counterparts may seek further education in order to improve their earnings capacity. By acquiring an additional degree, they successfully lift their earnings relative to those who did not obtain further education.

Acquiring an additional qualification may increase earnings through a number of potential mechanisms. We find evidence that, in Figure 22 for example, it increases the chance that individuals move from being unemployed or out of the labour force to being employed. The increase in employment is approximately 7 percentage points and is statistically significant. We also find evidence pointing to workers switching occupations or industries. This suggests that further education in later life can support the economic goals of a larger workforce as well as a more mobile one.

8 Sub-group analysis

In this section we analyse if there is heterogeneity in the treatment impacts.

To begin, we identify the important variables for which we expect to see the largest changes in the treatment effects. We use a Permutation Importance procedure.

8.1 Permutation importance feature selection method

We use a permutation importance selection method to evaluate the relative importance of individual features. Our aim here is to understand where the heterogenous treatment

effects are most pronounced. In other words, we aim to identify the sub-groups for which the treatment effects differ most significantly. In selecting the important features, our objective is to understand how to partition the data by the treatment effects as opposed to predicting the outcomes themselves.

The permutation importance method involves testing the performance of a model after removing each individual feature and replacing that feature with random noise (while keeping the underlying distribution of that feature in-tact). The model performance we are interested in, as previously mentioned, is the one that maps the features to the individual treatment effects.

Specifically, we take each bootstrap sample and train two outcome surfaces - one for $\hat{Y}_0$ and one for $\hat{Y}_1$ - and compute the individual treatment effects (τ's in Equation 2) for every person. Note that we train the model on the bootstrapped sample but estimate the individual treatment effects using the feature values for individuals from the original sample. Thus, for every individual we have 100 values of their individual treatment effects.

After obtaining the individual treatment effects for each individual, we train a model that maps the features to the individual treatment effects. We use cross-validation to select our hyperparameters and obtain the optimal model.

Using the original data, we take a single column amongst the features and permute the order of the data and calculate a new set of individual treatment effects. We compare the new and original individual treatment effects (based on the permuted data and those from the non-permuted data) and calculate the Mean Squared Errors (MSE).

We repeat this for all the features, permuting them individually and evaluating how they change the prediction of the individual treatment effect target. Features that yield the largest MSEs are likely to be more important than those features with lower MSEs. This is because permuting them changed the prediction of the individual treatment effect the most, which suggests that the individual treatment effect prediction relies heavily on those features.

We then repeat the above steps across all the bootstrap samples. Note that a different bootstrap sample will change the value of the individual treatment effects (since we train different outcome surfaces for $\hat{Y}_0$ and $\hat{Y}_1$ for each bootstrap sample).

We embed the permutation importance selection method in a bootstrapping procedure in order to capture hyperparameter uncertainty. For example, a different tree depth could be chosen between different bootstrap samples. This would affect the type of non-

linear/interaction relationships that would be captured by the models, which in turn would affect which features turn out to be important.

Finally, we obtain an average MSE for each feature, averaged across all bootstrap samples. This average value allows us to rank the features by their importance. Again, those with the largest average MSE values are the most important. We can also evaluate the uncertainty of this estimate since we obtain a distribution of MSE values across the different bootstrap samples.

Figure 11 displays the top ten features (based on the permutation importance procedure described above) and a residual category for all the other features. The features that are most important are: weekly gross wages on the main job and income- or wage-related variables. Together, this class of income/wage variables accounts for 36% of the importance of all variables. Figure 11 reiterates the importance of features related to employment, including occupational status, employment expectations, and employment history. The demographic background of the individual, namely their age, is also important. Figure 11 shows the importance breakdown for the GBR model.

The LASSO and RIDGE models show a similar story to the results from the permutation importance procedure using the GBR model. Overall, as Appendix Figure 31 and Figure 32 show, income and employment-related variables are the most salient in explaining treatment effect heterogeneity. However, the LASSO and RIDGE models also demonstrate the importance of other demographic characteristics such as the number of and presence of young children in the household, as well as economic-related variables such as previous educational attainment, amount remaining on the home loan and work aspirations.

GBR selects different features from the LASSO and RIDGE models because it is a non-linear model.

Figure 13 displays the distribution of the MSE values across the bootstrap samples for the GBR model. It displays the distributions for the top 3 features. The feature with the highest importance score: weekly gross wage in the main job has a markedly flatter distribution than the other two features. This suggests that in some of the bootstrap samples, where the MSE is larger, the individual treatment effects from the permuted data differ greatly from the original individual treatment effects.

Focusing on the results from the GBR models for both the T-learner and Doubly Robust models, Figure 23 shows that there is heterogeneity in the treatment impacts in the GBR model. We have identified the features that were considered most important according to the permutation procedure. For each feature, we divide the sample into two groups.

For continuous variables, we take the median value and divide the sample into those who are above and below this median value.

Both the T-Learner and Doubly Robust models show that weekly wage has a large impact on the effect size. Those with below median wage in 2001 derive more benefits than those with above median wage, possibly because high income earners hit an earnings ceiling. Younger people in 2001 also derive more returns, as they may have had more time to accumulate returns. Females also demonstrate higher returns compared to males. The T-Learner and Doubly Robust models show opposing results with regards to resident children aged 0-4. While the T-Learner model reveals greater returns among those with non-resident children and those above median household annual wage, the Doubly Robust model shows little difference in effects by these characteristics.

9 Conclusions

With the aim of causal inference in mind, we tailor ML models to examine the degree to which gaining a new qualification provides mature-age learners with economic (earnings and employment) and well-being (mental health and subjective well-being) benefits. We also examine which mature-age learners tend to reap more benefits than others.

We use data from the Household Income and Labour Dynamics Australia survey, which combined with ML models, allow us to exploit the full set of background information on individuals (beginning with more than 3,400 variables).

The key benefits of our automated feature selection process include the potential to minimise confounding bias while minimising overfitting issues. It can do this by identifying features (and functional forms) that may have been previously overlooked or unanticipated.

We find that our automated feature selection method selects a set of controls/ features/ variables that include those that have theoretical foundations and/or align with those chosen in past empirical studies. However, we also choose features that have been traditionally overlooked. These include variables such as household debt, wealth, housing, and geographic mobility variables. Other important predictors include the ages of both resident and non-resident children: non-resident children aged 15 or above matter and resident children aged 0-4 are important.

ML models randomly choose between variables that are highly correlated, which means that our selected features do not suggest any causal relationship between the features

and the outcome. However, as the features act as nuisance parameters, helping up to obtain the best prediction of the outcomes (or the response surfaces for treatment and control groups), this random selection of variables does not compromise our primary aim of obtaining the ATE and CATE estimates. We use cross-validation for feature selection.

Another innovation of our approach is inference of causal parameters in ML models. We use bootstrapping (with replacement), generating 100 samples to produce a distribution of ATE and CATE values. This allows us to derive standard errors for our causal parameters.

There are clear economic benefits to gaining an additional qualification in later life. Analysing those who obtain an additional qualification anytime between 2002 and 2017, we estimate a gain of approximately $100 per week in gross earnings. This represents roughly 8 percent of the Average Weekly Gross Earning for the average worker in Australia.

The effects remain strong up to a decade-and-a-half after course completion. The largest earnings gains are associated with acquiring an undergraduate degree or above. Only modest gains are associated with postsecondary certificates or diplomas. Furthermore, the subject-area of study is important: undertaking study in technical subjects such as the STEM or medicine and health-related disciplines yield significantly higher returns that study in the fields of Arts, Humanities or the Social Sciences.

Qualification advancements do not benefit Australian workers in the same way: women appear to benefit more from later-life study than men and older Australians benefit more than their younger counterparts.

Acquiring an additional qualification may increase earnings through a number of potential mechanisms. We find evidence that it increases the chance that individuals move from being unemployed or out of the labour force to being employed. We also find evidence pointing to workers switching occupations or industries. This suggests that further education in later-life can support the economic goals of a larger workforce as well as a more mobile one.

10 Tables and Figures

Table 1: Summary Statistics

Variable label	Variable name	Mean	SD
Outcomes			
Annual Earnings individual in 2019	y_wscei	614.730	1044.717
Imputed wages			
Change in annual earnings between 2001 and 2019	y_dwscei	129.029	980.754
Treatment Indicators			
Highest level of educ changed between 2001 and 2017	reduhl	0.097	0.296
Extra degree attained in 2002 to 2017	redufl	0.257	0.437
Extra degree Bachelor and/or above	bachab	0.072	0.259
Below bachelor	bbach	0.209	0.406
Technical degree	techdeg	0.151	0.358
Qualitative degree*	qualdeg	0.080	0.272
Covariates			
Demographics			
Sex	hgsex	1.536	0.499
Section of State	hhsos	0.690	1.046
Age	hgage1	46.025	12.832
Age of youngest person in HH	hhyng	27.115	21.886

Continued on next page

Continued from previous page

Variable label	Variable name	Mean	SD
No. persons aged 0-4 years in HH	hh0_4	0.257	0.589
No. persons aged 10-14 years in HH	hh10_14	0.274	0.606
Age when first left home	fmagelh	21.502	11.230
Living circumstances	hgms	1.997	1.708
English fluency	hgeab	1.604	0.262
Unemployment rate in region	hhura	6.884	1.075
Education			
Highest year of school completed/attending	edhists	2.383	1.439
Bachelor degree (without honours) obtained	edqobd	0.211	0.330
Masters degree obtained	edqoms	0.041	0.160
Doctorate obtained	edqodc	0.011	0.085
No. qualifications unknown	edqunk	0.078	0.403
Employment			
Occupation	jbmo61	3.772	1.825
Years in paid work	ehtjbyr	21.963	11.907
Tenure with current employer	jbempt	8.505	7.369
Type of work schedule	jbmday	3.785	2.612
Current work schedule	jbmsch	2.255	1.819
Casual worker	jbcasab	1.797	0.291
Hours/week worked at home	jbmhrh	12.372	7.174
Hours/week travelling to and from work	lshrcom	3.052	3.716

Continued on next page

Continued from previous page

Variable label	Variable name	Mean	SD
Satisfaction with employment opportunities	losateo	6.693	2.557
Occupational status - current main job	jbmo6s	50.177	19.199
No. persons employed at place of work	jbmwpsz	3.746	1.961
Age intends to retire	rtiage1	345.709	230.208
Age retired/intends to retire	rtage	113.904	130.211
Prob. of losing job in next 12 months	jbmploj	15.196	35.018
Prob. of accepting similar/better job	jbmpgj	59.585	26.196
Looked for work in last 4 weeks	jsl4wk	1.272	0.411
Years unemployed and looking for work	ehtujyr	0.464	1.647
Hours per week worked in last job	ujljhru	34.990	6.922
Industry of last job	ujljin1	9.373	1.822
Work preferences			
Total hours per week would choose to work	jbprhr	34.378	6.407
Importance of work situation to your life	loimpew	6.854	2.908
Childcare			
Child looks after self	chu_sf	0.128	0.144
Uses child care while at work	cpno	1.257	0.139
Parent provides child care	cpu_me	0.434	0.151
Work-family balance			
Do fair share of looking after children	pashare	2.411	0.671
Miss out on home/family activities	pawkmfh	3.904	1.069

Continued on next page

Continued from previous page

Variable label	Variable name	Mean	SD
Working makes me a better parent	pawkbp	4.038	0.979
Family			
No. dependent children aged 5-9	hhd5_9	0.261	0.584
No. dependent children aged 10-14	hhd1014	0.269	0.604
No. non-resident children	tcnr	0.993	1.373
Sex of non-resident child	ncsex1	1.509	0.320
Likely to have a child in the future	icprob	1.188	0.374
Finances			
Owned a home previously	hspown	1.368	0.424
Amount outstanding on home loans	hsmgowe	96803.720	43547.610
Time until home loan paid off	hsmgfin	2011.858	4.157
Food expenses outside the home	xposml	36.982	42.522
SEIFA (level of economic resources)	hhec10	5.463	2.897
Taxes on total income	txtottp	7476.727	14035.510
Change in total gross income since 1 year ago	wslya	2231.465	1950.065
Had an incorporated business	bifinc	1.715	0.199
Had a non-LLC or unincorporated business	bifuinc	1.259	0.193
Income			
HH current weekly gross wages - all jobs	hiwscei	992.666	918.261
Current weekly gross wages - main job	wscme	468.062	556.185
HH financial year gross wages	hiwsfei	52472.490	49458.180

Continued on next page

Continued from previous page

Variable label	Variable name	Mean	SD
Financial year gross wages	wsfe	25463.770	30265.630
Financial year regular market income	tifmktp	30734.790	33618.860
Financial year disposable total income	tifditp	27477.160	22701.270
Imputation flag: current weekly gross wages - all jobs	wscef	0.070	0.256
Imputation flag: current weekly gross wages - other jobs	wscoef	0.044	0.205
Imputation flag: financial year gross wages	wsfef	0.071	0.256
Other sources of income			
Receive superannuation/annuity payments	oifsup	0.059	0.232
Receive redundancy and severance payments	oifrsv	0.002	0.038
Receive other irregular payment	oifirr	0.001	0.027
Receive government pensions or allowances	bncyth	0.004	0.027
Receive Disability Support Pension	bnfdsp	0.151	0.181
Receive other regular public payments	oifpub	0.000	0.019
Financial year regular private income	tifprin	77.299	1409.625
Financial year investments	oifinvp	1951.052	10569.050
Financial year dividends	oidvry	744.263	4651.593
Financial year interest	oiint	666.116	3448.494
Financial year regular private pensions	oifpp	967.101	5055.004
Financial year business income (loss)	bifn	185.652	3274.511
Financial year business income (profit)	bifip	2597.792	13649.410
Financial year irregular transfers from non-resident parents	oifnpt	35.067	1305.812

Continued on next page

Continued from previous page

Variable label	Variable name	Mean	SD
Financial year public transfers	bnfapt	2865.540	4717.042
Financial year government non-income support payments	bnfnis	1025.031	2237.987
HH financial year public transfers	hifapti	5542.675	7937.136
HH financial year business income	hibifip	4880.589	18393.360
Health			
Imputation flag: current weekly public transfers	bncapuf	0.044	0.204
Imputation flag: financial year investments	oifinf	0.124	0.330
Imputation flag: financial year dividends	oidvryf	0.079	0.270
Imputation flag: financial year rental income	oirntf	0.071	0.257
Imputation flag: financial year business income	biff	0.071	0.258
Health limits vigorous activities	gh3a	2.108	0.718
How much pain interfered with normal work	gh8	1.704	0.971
Health condition/disability developed last 12 months	helthyr	1.870	0.151
Tobacco expense in average week	lstbca	37.771	10.690
Housing			
Years at current address	hsyrcad	9.541	10.226
External condition of dwelling	docond	1.970	0.870
No dwelling security	dosecno	0.552	0.497
No. homes lived in last 10 years	mhn10yr	3.456	1.107
Moved to be near place of work	mhreawp	0.084	0.111
Moved because I was travelling	mhrearo	0.009	0.038

Continued on next page

Continued from previous page

Variable label	Variable name	Mean	SD
Attitudes			
Importance of religion	loimprl	4.612	3.483
Working mothers care more about work success	atwkwms	3.729	1.807
Mothers who don't need money shouldn't work	atwkmsw	3.951	1.982
Identifiers			
Family number person 02	hhfam02	NA	NA
Relationship to person 03	rg03	NA	NA
ID of other responder for HH Questionnaire	hhp2	NA	NA

*Definition of technical and qualitative degree: Technical: STEM, Architecture, Agriculture and Environment, Medicine, Other Health-related Studies and Nursing, Management and Commerce and Law. Non-technical: Education, Society and Culture (includes economics!), Creative Arts, and Food, Hospitality and Personal Services.

Table 2: ML variables omitted by OLS theory model

Variable label	Variable name	Relationship with re-education (redufl)	Relationship with outcome (y_wscei)	Bias direction in OLS models
Education				
Doctorate obtained	edqodc	-	+	-
Employment				
Tenure with current employer	jbempt	-	-	+
Current work schedule	jbmsch	-	-	+
Casual worker	jbcasab	-	+	-
Occupational status - current main job	jbmo6s	+	+	+
No. persons employed at place of work	jbmwpsz	+	+	+
Prob. of accepting similar/better job	jbmpgj	+	+	+
Years unemployed and looking for work	ehtujyr	+	-	-
Work-life balance				
Total hours per week would choose to work	jbprhr	+	+	+
Parent provides child care	cpu_me			-
Do fair share of looking after children	pashare	-	+	-
Miss out on home/family activities	pawkmfh	+	+	+
Income				
Current weekly gross wages - main job	wscme	+	+	+
Imputation flag: current weekly gross wages - all jobs	wscef	+	+	+

Continued on next page

Continued from previous page

Variable label	Variable name	Relationship with re-education (redufl)	Relationship with outcome (y_wscei)	Bias direction in OLS models
Change in total gross income since 1 year ago	wslya	+	+	+
Financial year investments	oifinvp	-	-	+
Financial year business income (profit)	bifip	-	-	+
Amount outstanding on home loans	hsmgowe	+	+	+
Imputation flag: financial year dividends	oidvryf	+	-	-
Imputation flag: financial year rental income	oirntf	+	+	+
Imputation flag: financial year business income	biff	+	-	-
Health				
Health limits vigorous activities	gh3a	+	+	+
Tobacco expense in average week	lstbca	-	-	+
Identifiers				
ID of other responder for HH Questionnaire	hhp2	-	-	+

*Notes can be entered here

Table 3: Nested CV Holdout Sample: Level Earnings

Model	Outcome surface	Negative MSE	NMSE Std	R-squared	R-squared Std	ATE	ATE_std
GBR	Treated	-886515	452077	0.22	0.06	68.2	28.4
	Control	-659056	107251	0.36	0.07		
LASSO	Treated	-955958	361911	0.15	0.09	94.1	14.5
	Control	-710521	178030	0.32	0.05		
Ridge	Treated	-966849	434518	0.16	0.08	97.8	14.5
	Control	-712374	174033	0.32	0.04		

Notes: 5 fold CV performed on 80% train sample. All statistics presented in this table are based on the 20% holdout sample. Ten outer folds are used. See Figure 1 for more details .

Table 4: Average Treatment Effects: comparison across models

Outcome	Model	N	ATE	S.E (ATE)
Level earnings	OLS	5441	64.41	28.70
	T-learner (GBR)	5441	88.38	35.05
	T-learner (LASSO)	5441	110.08	50.89
	T-learner (Ridge)	5441	108.95	39.61
	Doubly Robust (GBR)	5441	87.73	8.54
	Doubly Robust (LASSO)	5441	87.32	5.83
	Doubly Robust (Ridge)	5441	89.78	3.34
	Bayesian Ridge	5441	45.23	22.77
	Bayesian Ridge (cf. s)	5441	46.15	25.54
	Gaussian Process	5441	0.19	6.41
	Hierarchical BLM	5441	0.28	

Notes: Sample of 25 or older respondents who had completed a degree at any point between 2002 and 2017. Total completions: 1,383.

Figure 1: Selecting and Evaluating Model Class

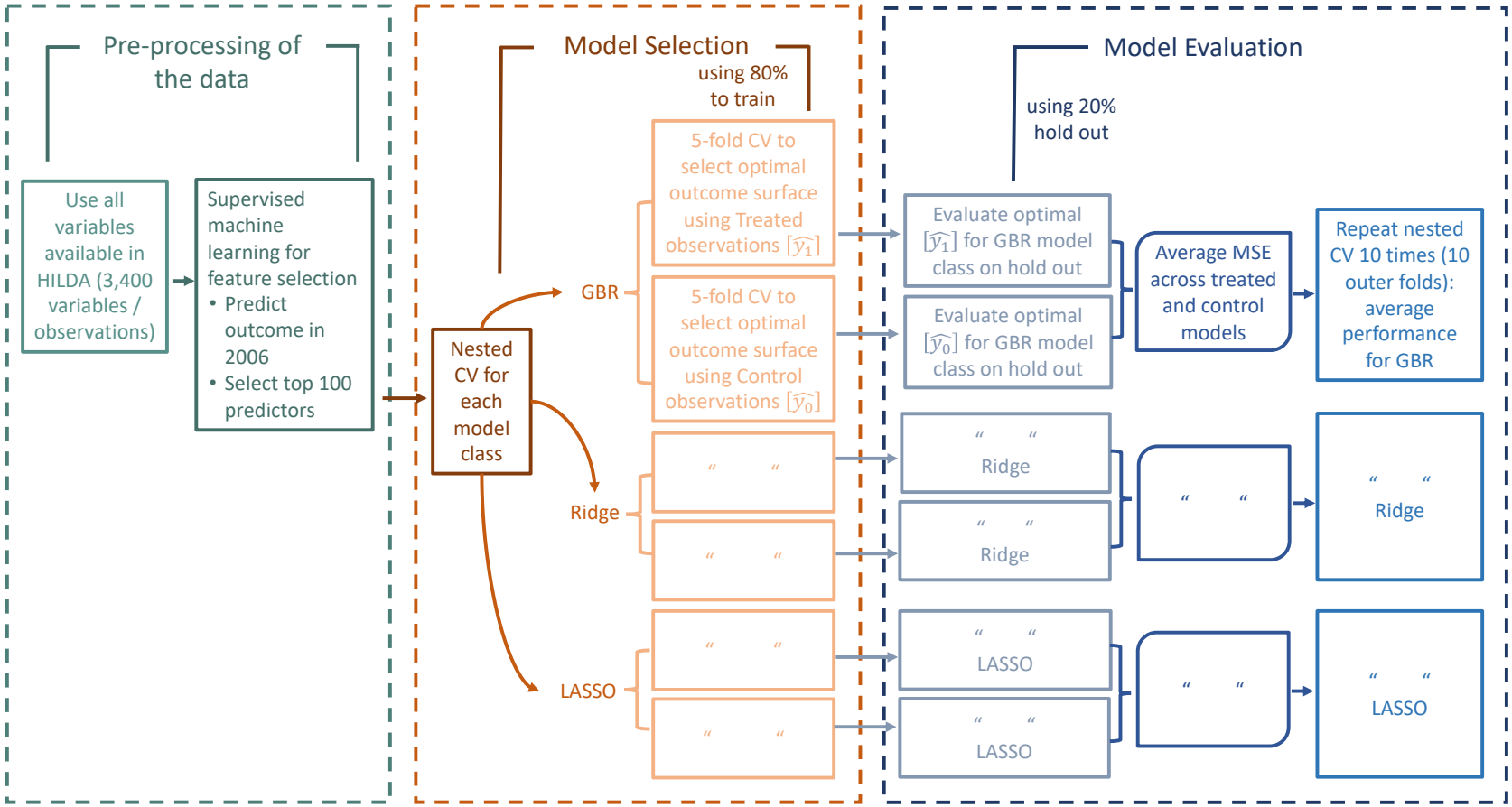


Figure 2: Generating Uncertainty Parameters

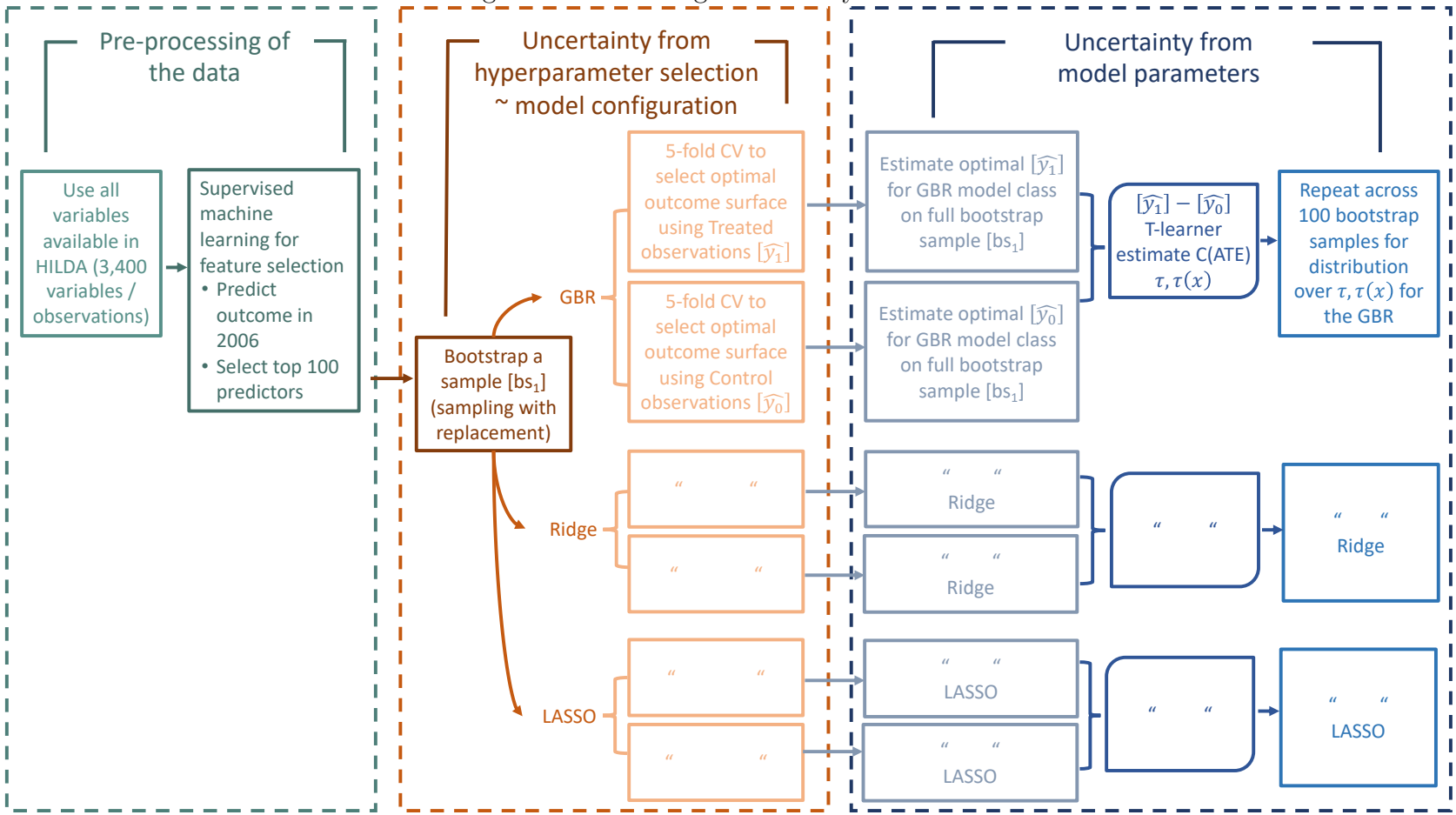
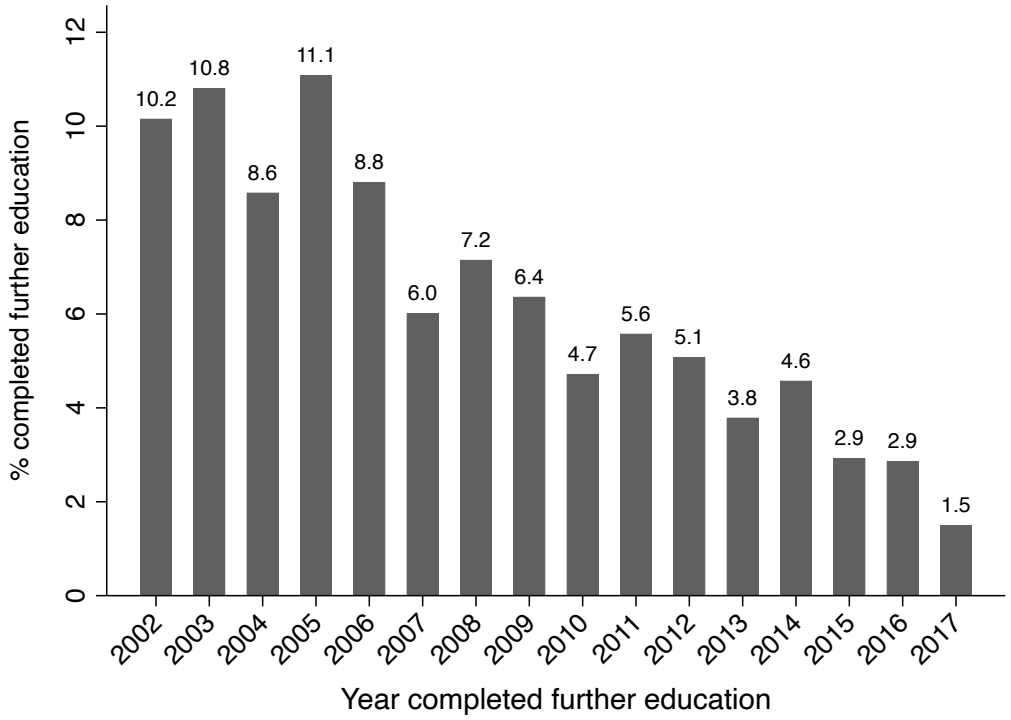
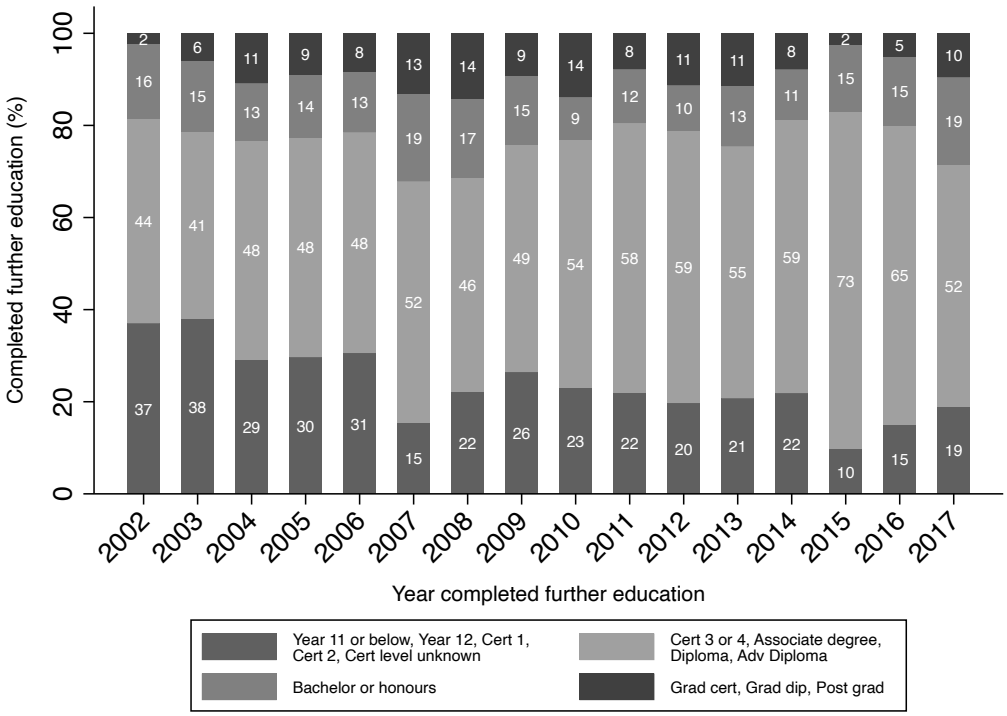


Figure 3: Timing of Completion



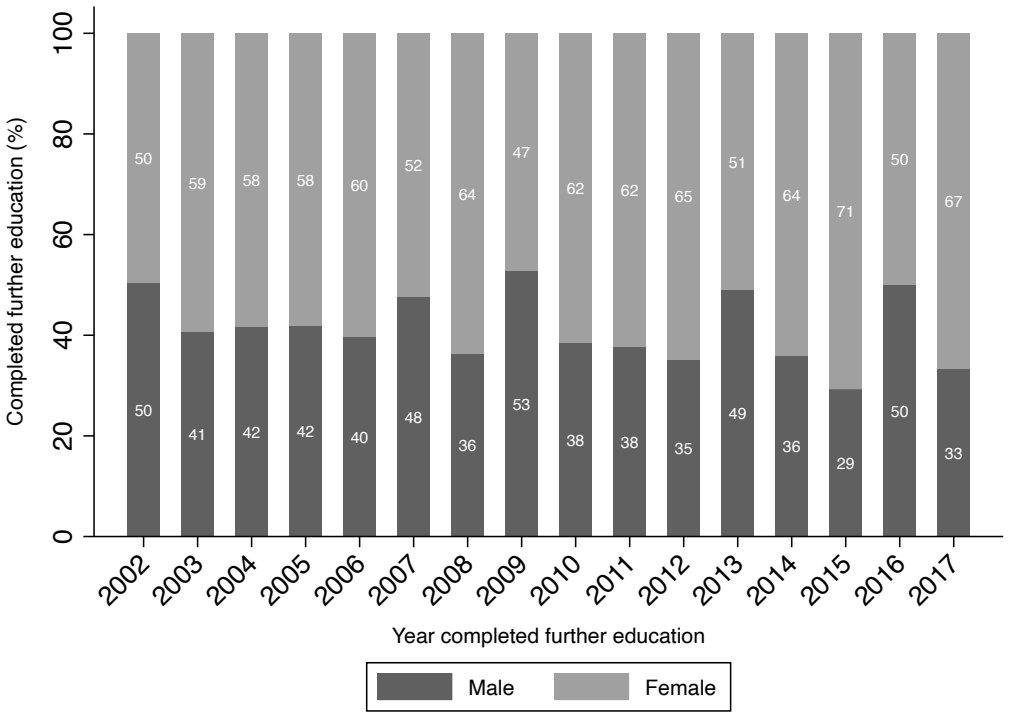
Notes: Sample of 25 or older respondents who had completed a degree at any point between 2002 and 2017. Total completions: 1,383.

Figure 4: Timing of Completion by Type of Degree



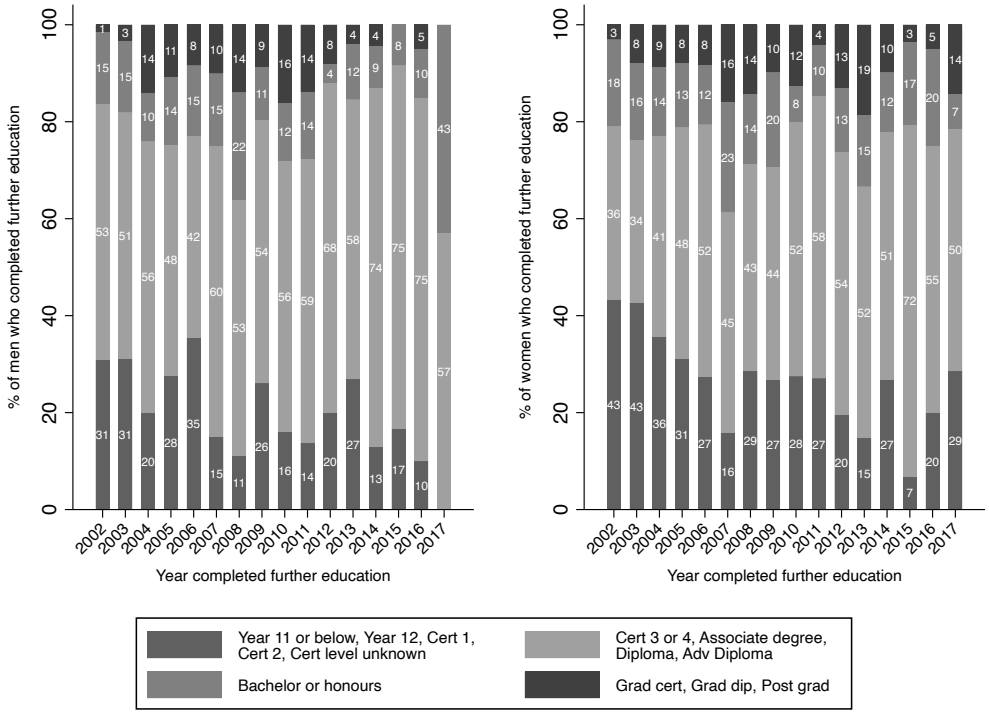
Notes: Sample of 25 or older respondents who had completed a degree at any point between 2002 and 2017. Total completions: 1,383.

Figure 5: Timing of Completion by Gender



Notes: Sample of 25 or older respondents who had completed a degree at any point between 2002 and 2017. Total completions: 1,383.

Figure 6: Timing of Completion by Type of Degree and Gender



Notes: Sample of 25 or older respondents who had completed a degree at any point between 2002 and 2017. Total completions: 1,383.

Figure 7: Degree completions by sex

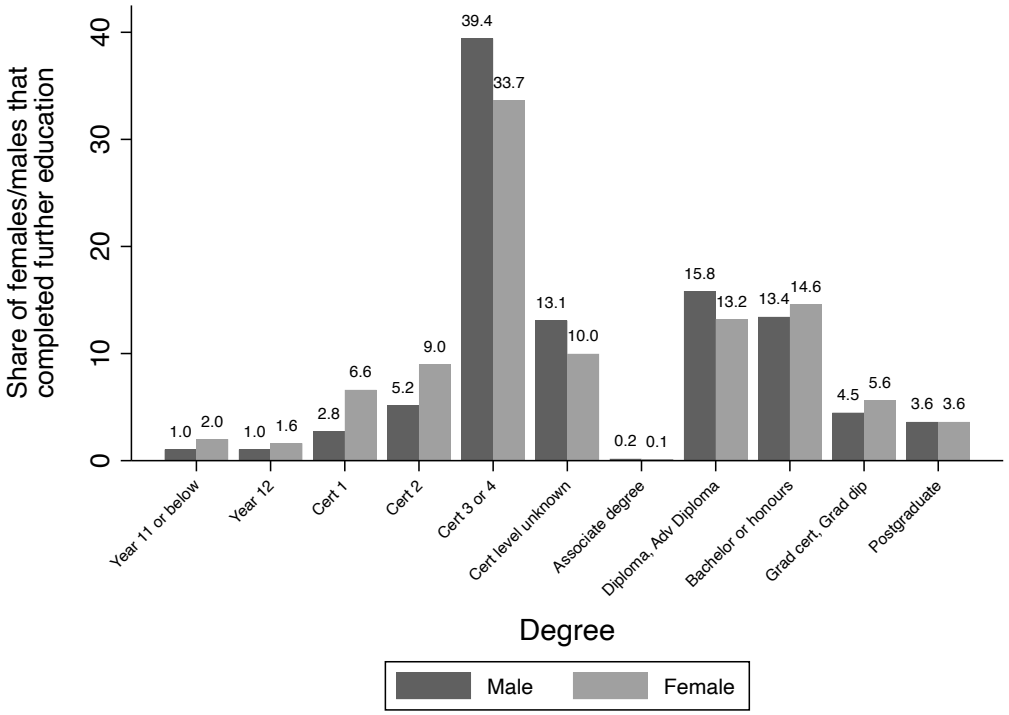


Figure 8: Degree completions by age

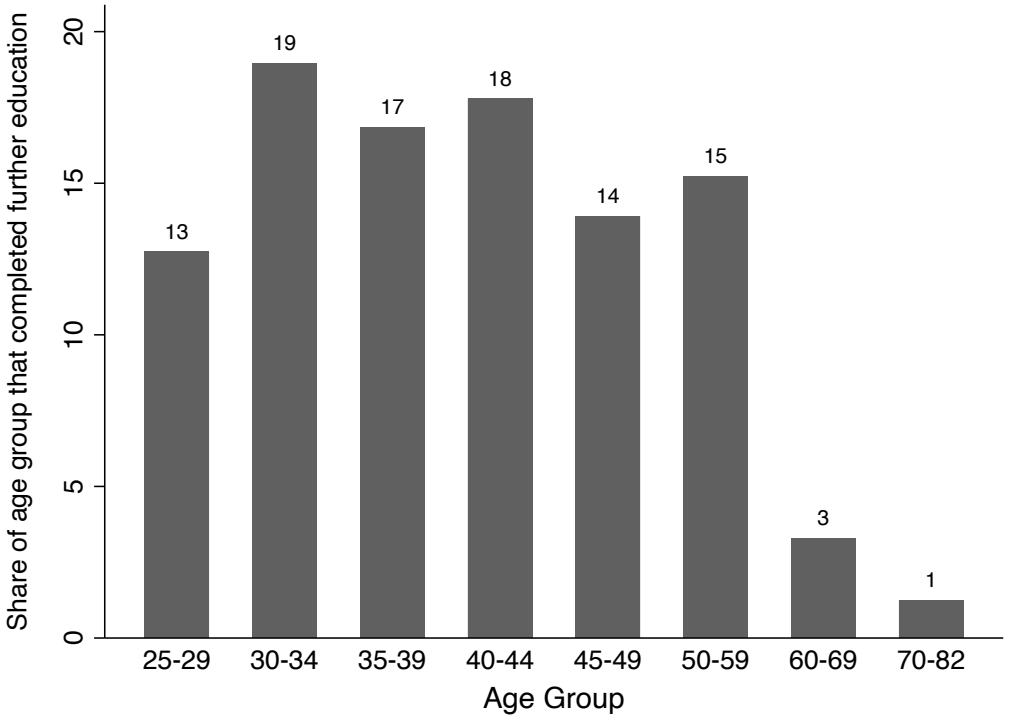


Figure 9: Earnings and Employment by year

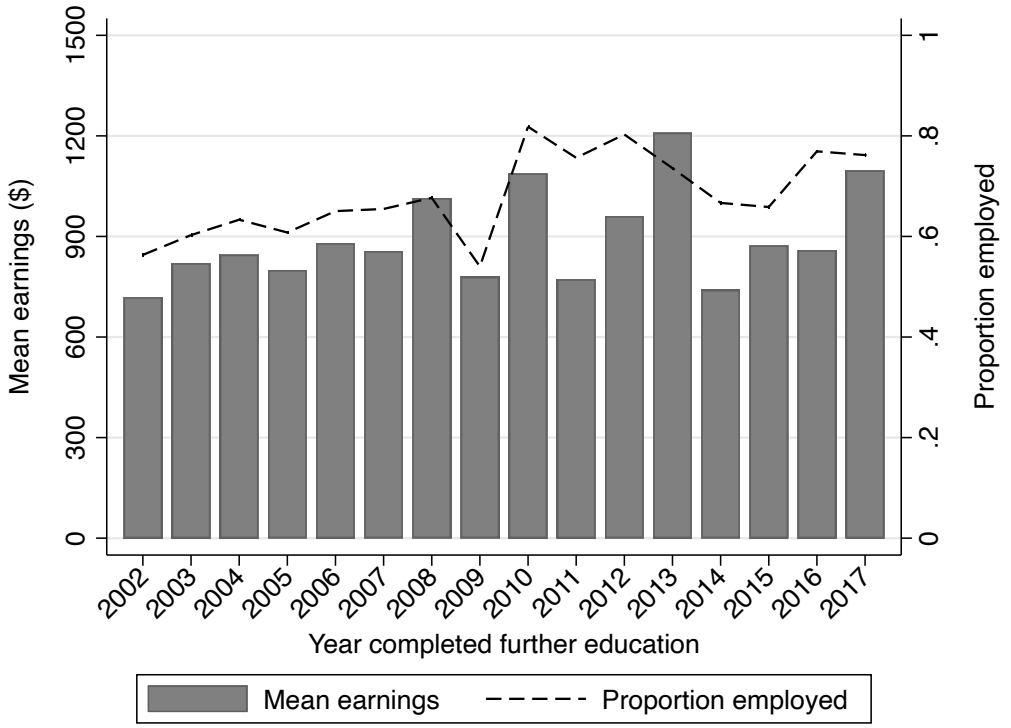


Figure 10: Earnings and Employment by year and sex

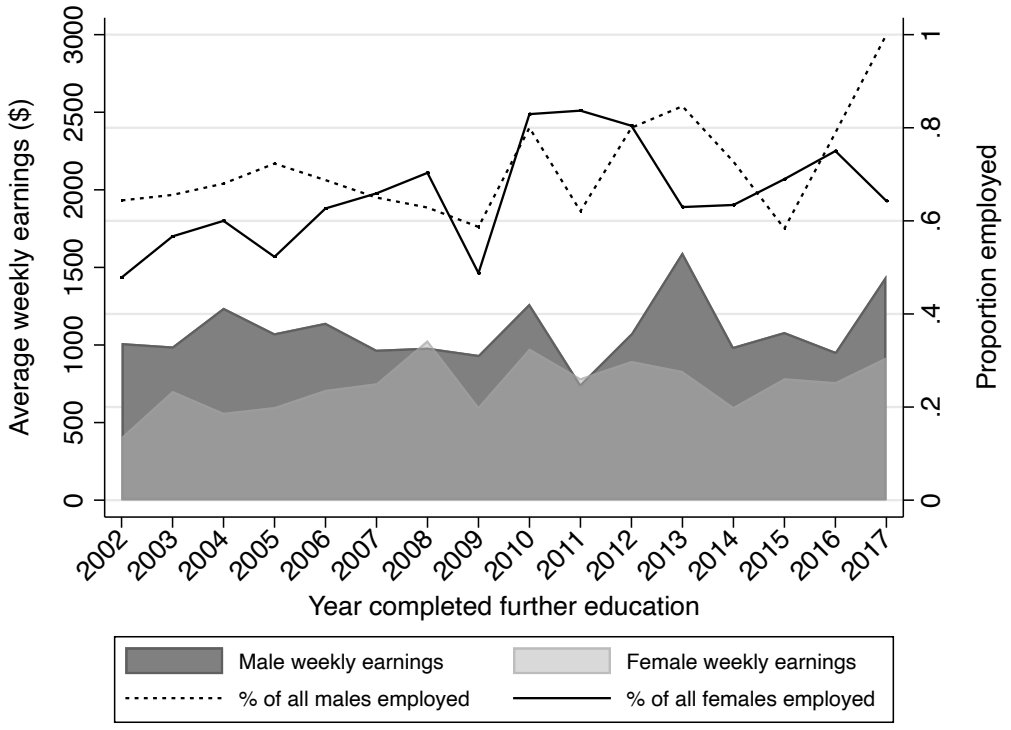
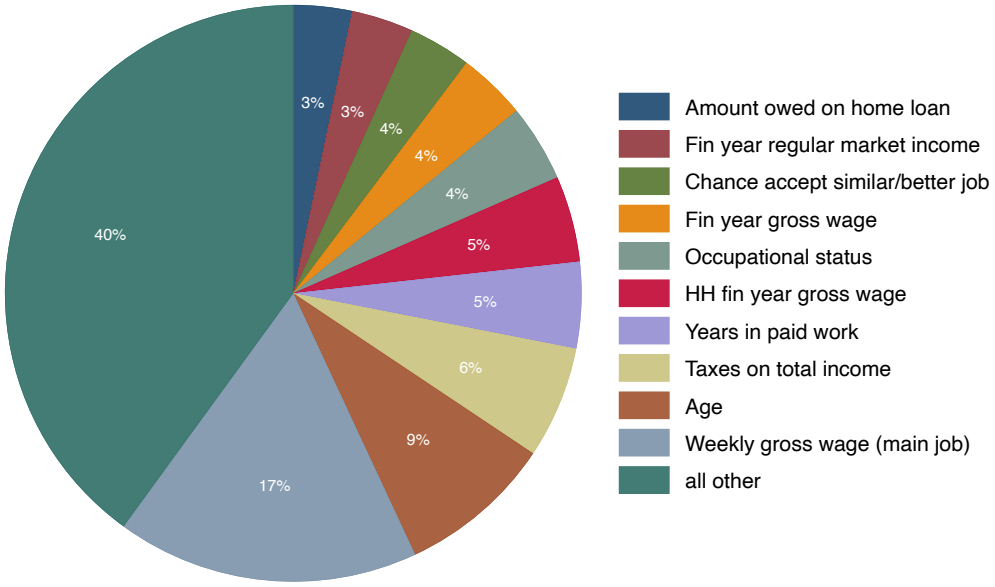
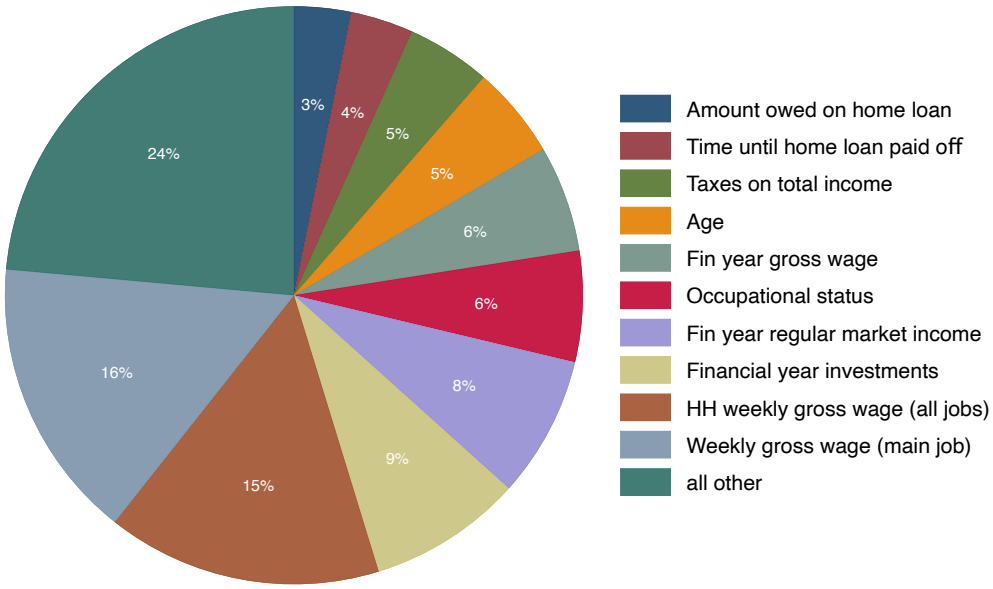


Figure 11: Importance Features in Heterogeneous Treatment Effects Estimation using T-Learner: Level Earnings



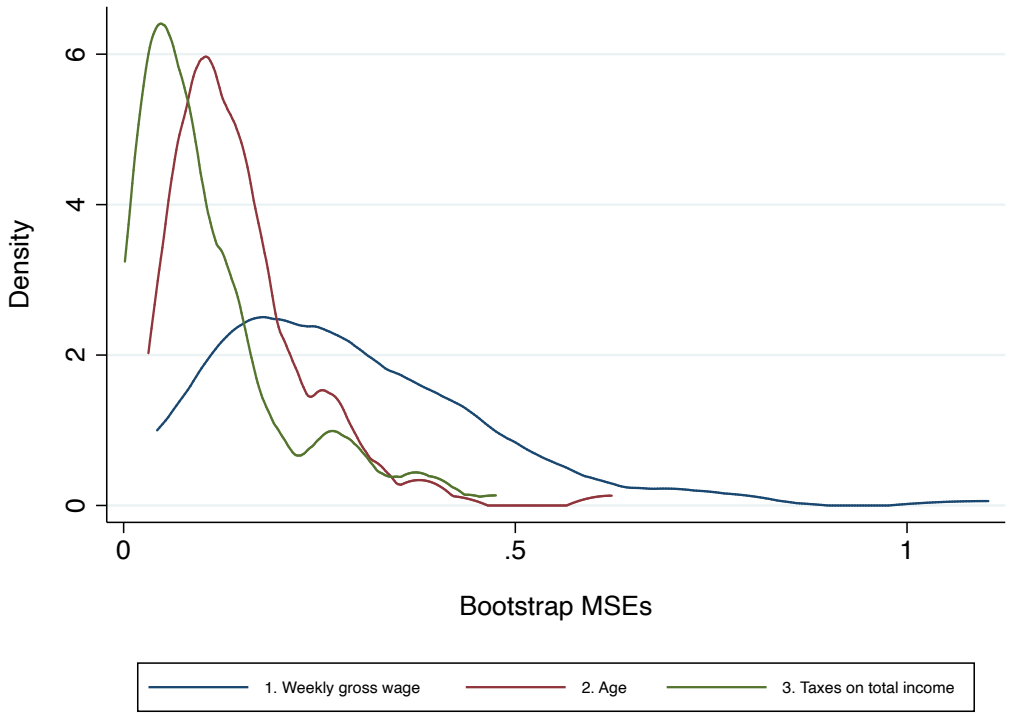
Notes:

Figure 12: Importance Features in Heterogeneous Treatment Effects Estimation using DR: Level Earnings



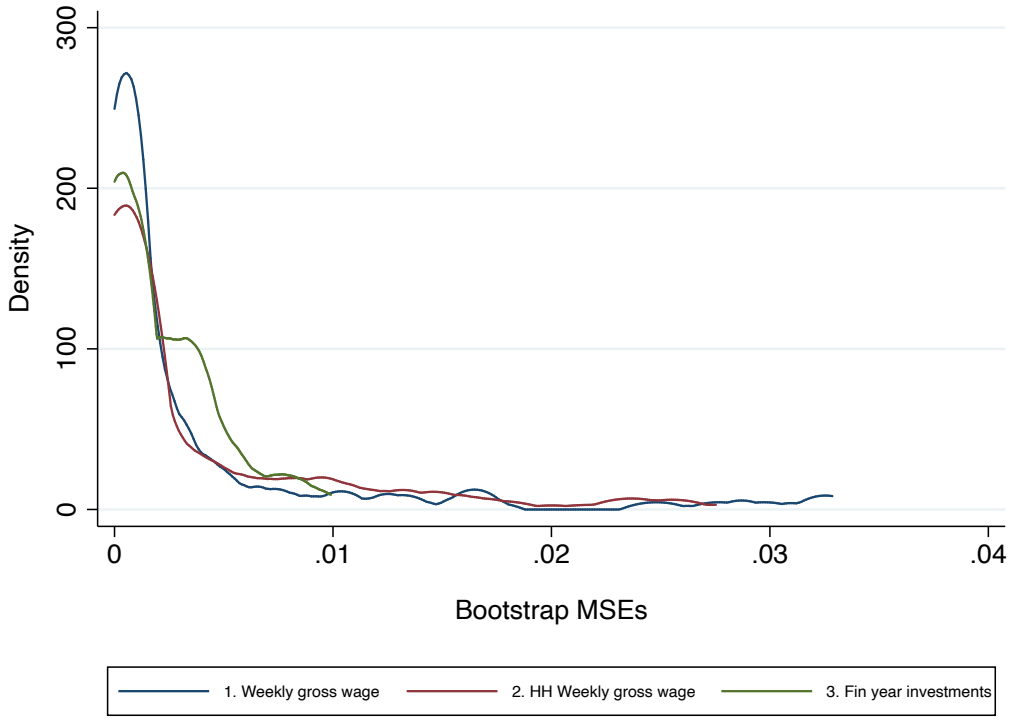
Notes:

Figure 13: Top 3 Features Distribution of Importance using T-Learner: Level Earnings



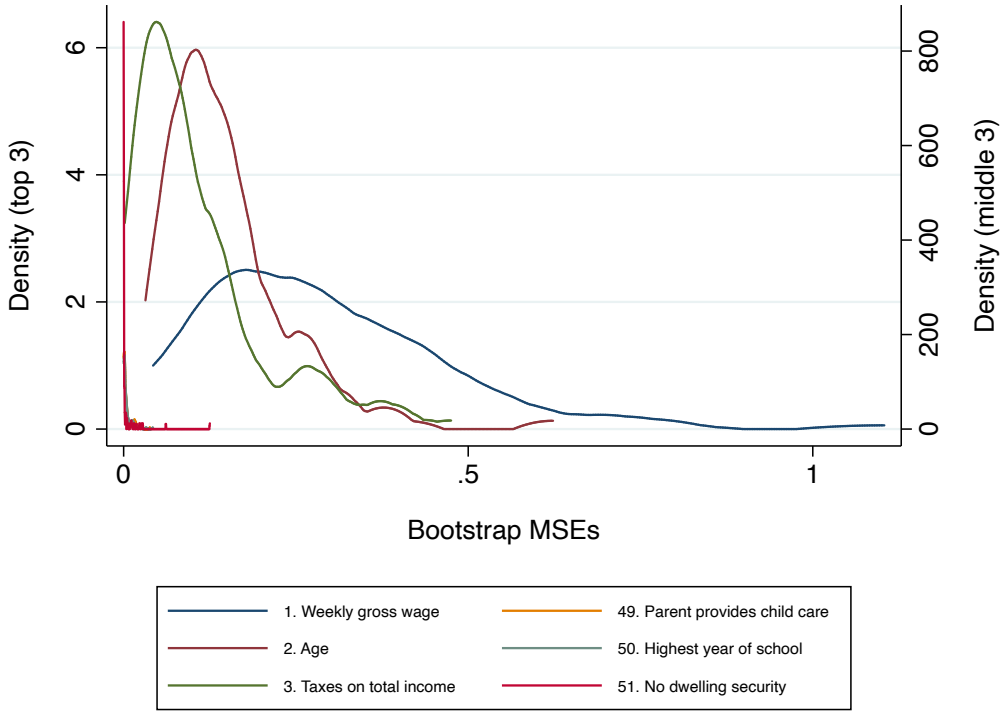
Notes:

Figure 14: Top 3 Features Distribution of Importance using DR: Level Earnings



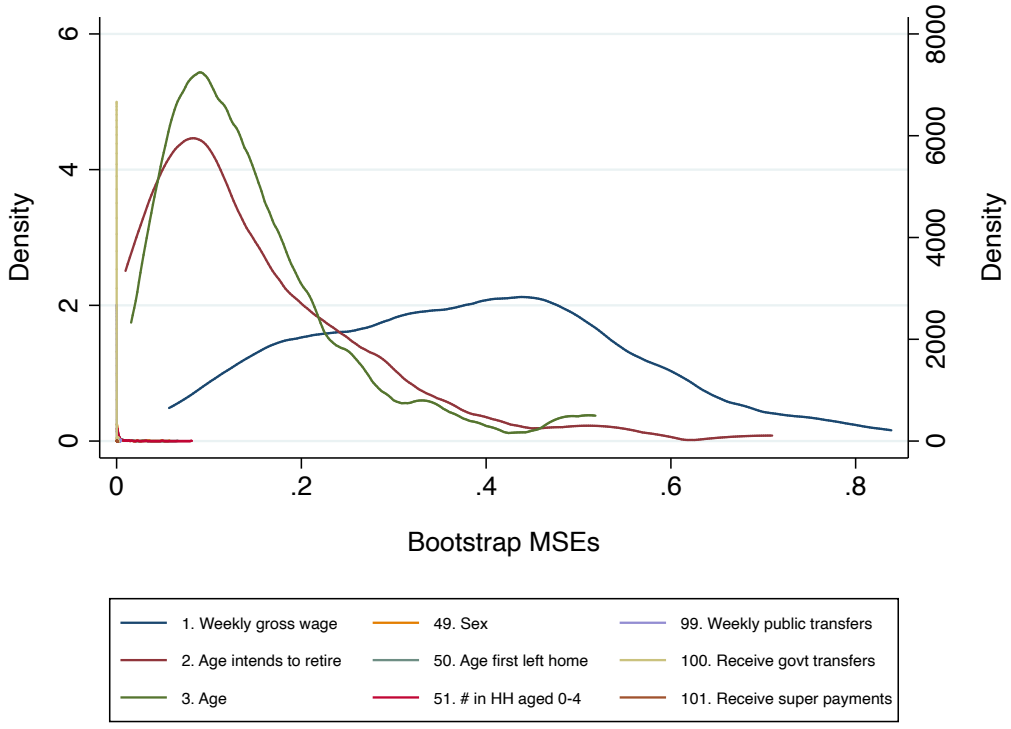
Notes:

Figure 15: Top and Middle 3 Features Distribution of Importance using T-Learner: Level Earnings



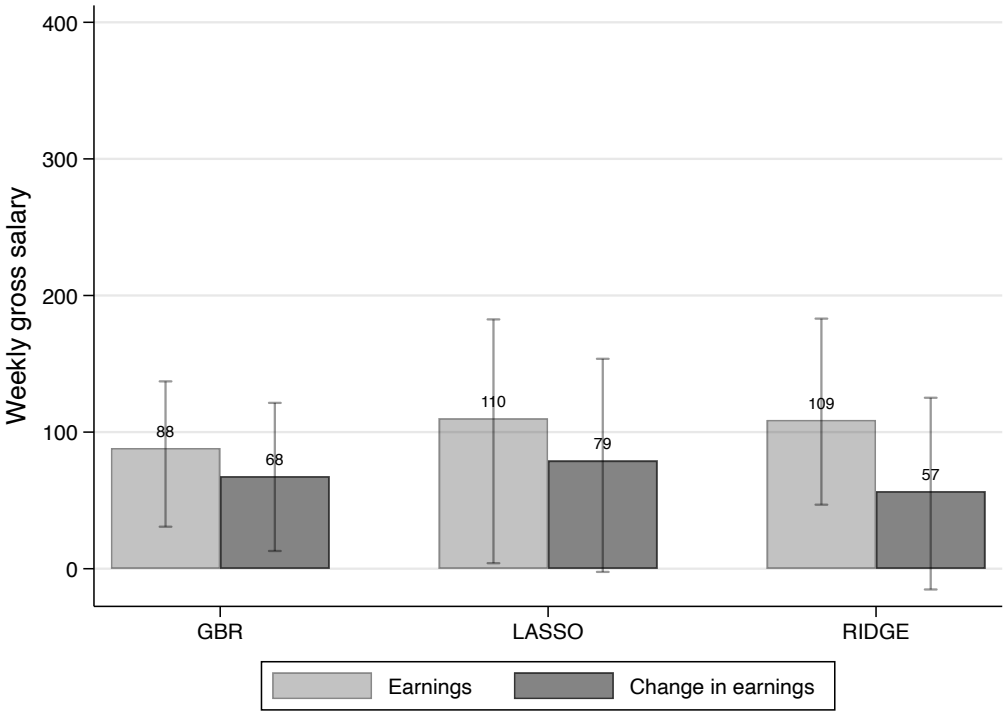
Notes:

Figure 16: Top 3 Features Distribution of Importance using T-Learner: Change in Earnings



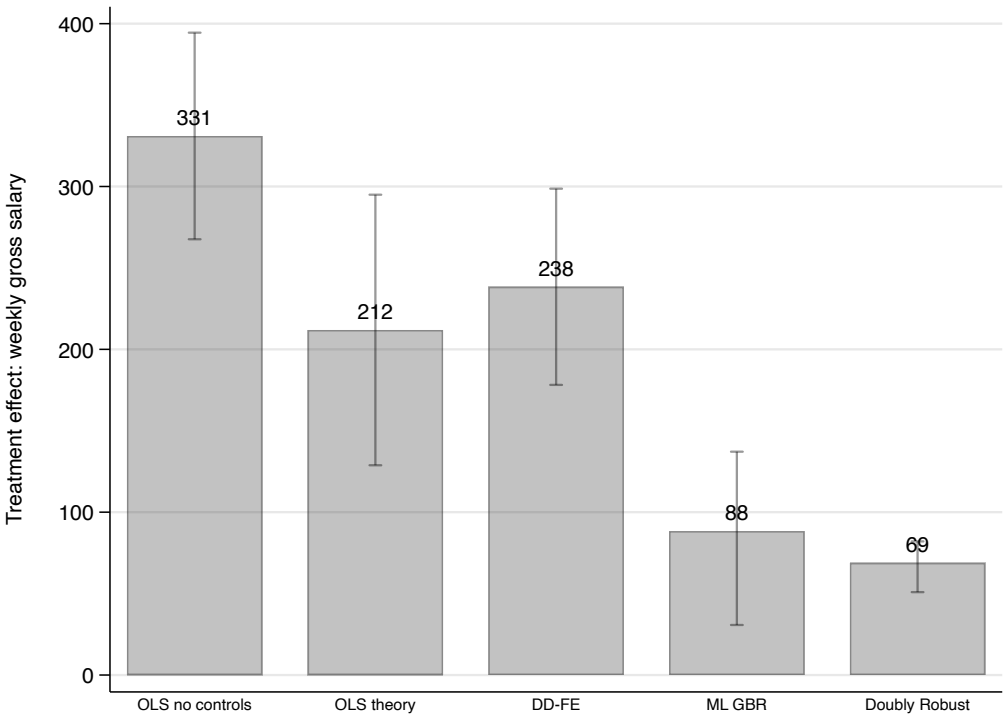
Notes:

Figure 17: Value-add in earnings



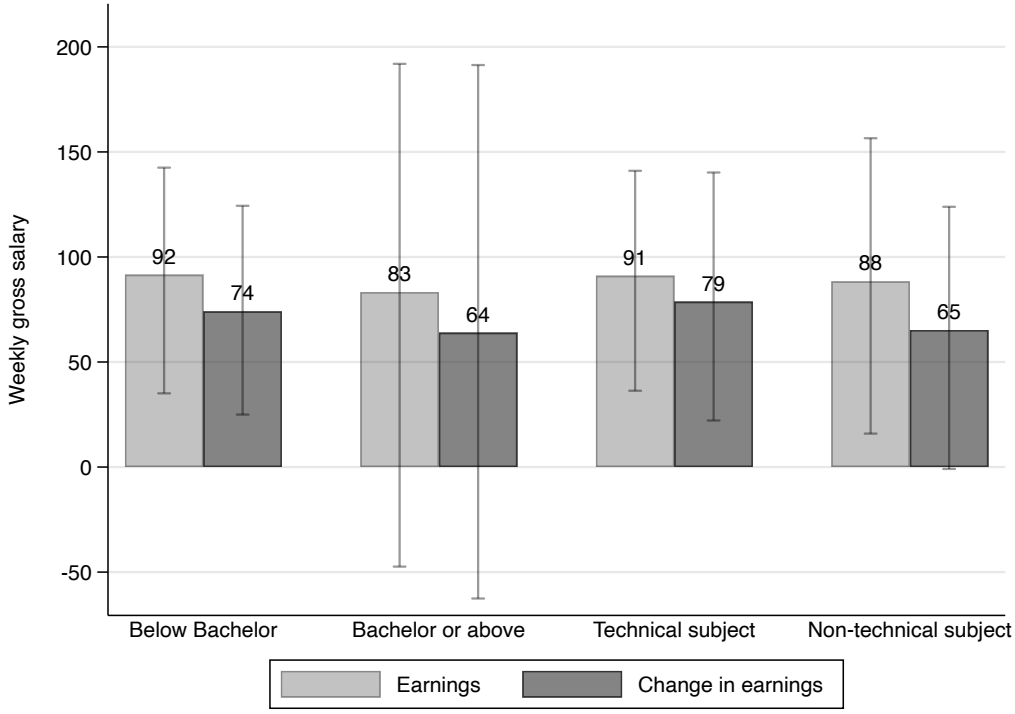
Notes:

Figure 18: Methodology



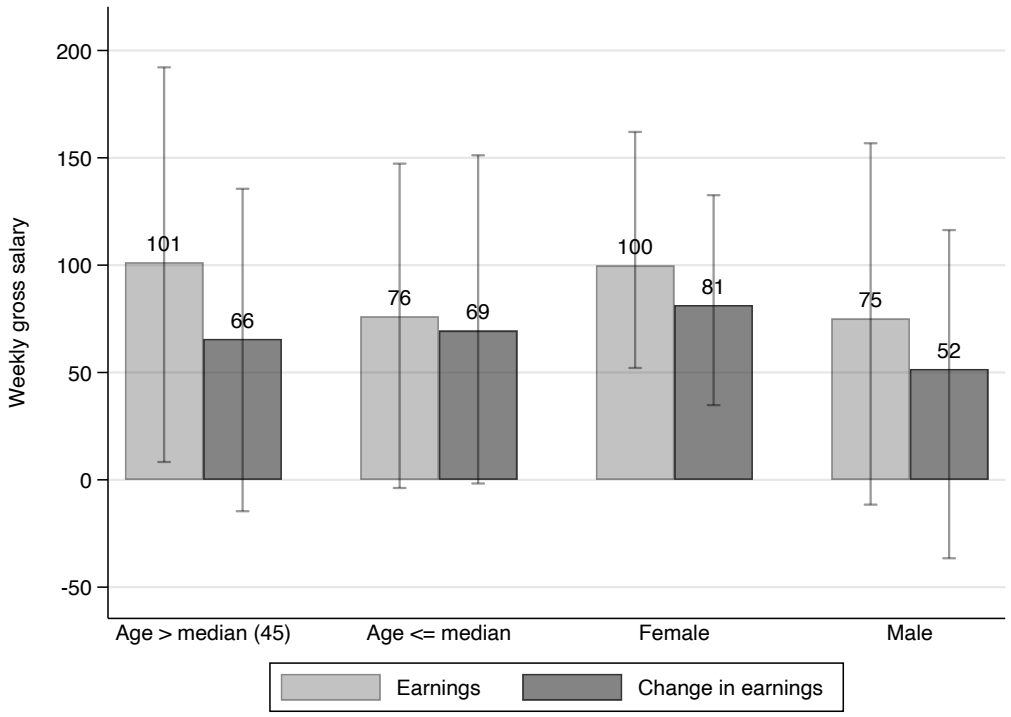
Notes:

Figure 19: Type of degree and subject area



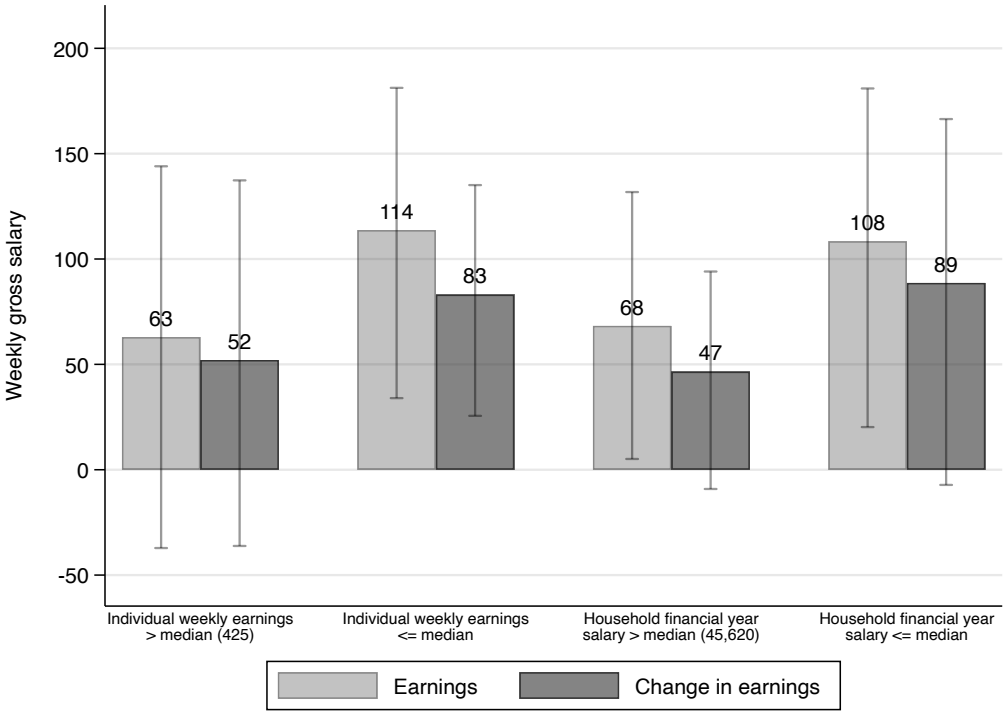
Notes:

Figure 20: Heterogenous Treatment Effects: Demographics



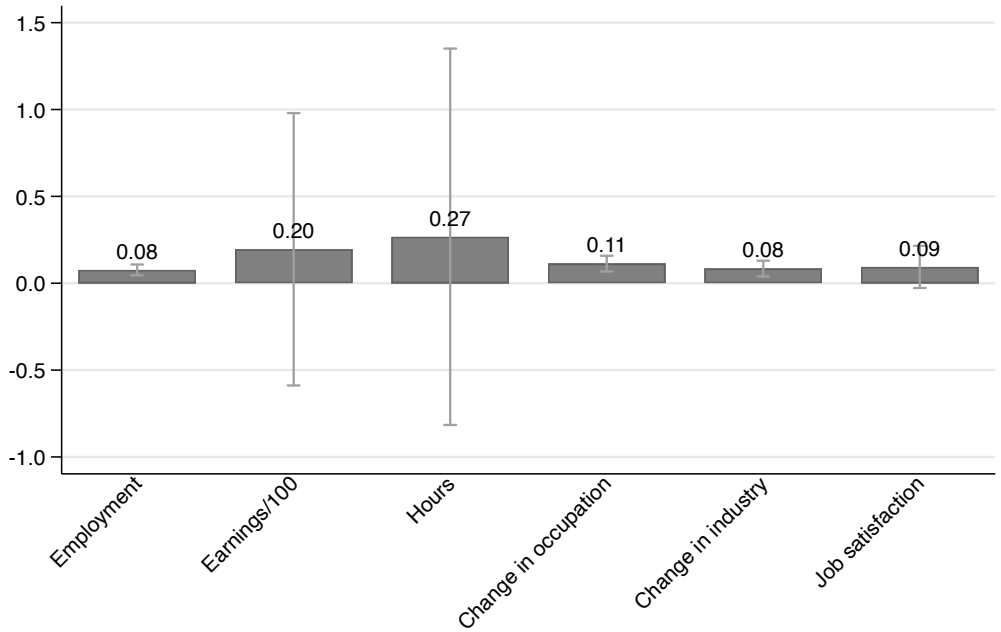
Notes:

Figure 21: Heterogenous Treatment Effects: SES background



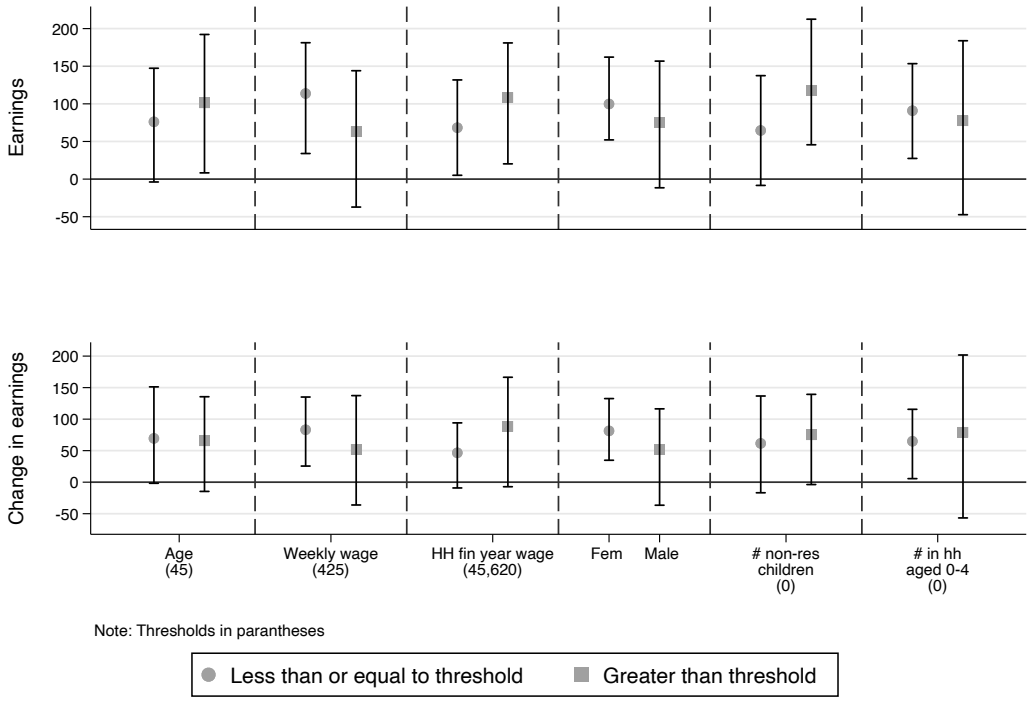
Notes:

Figure 22: Employment Transitions



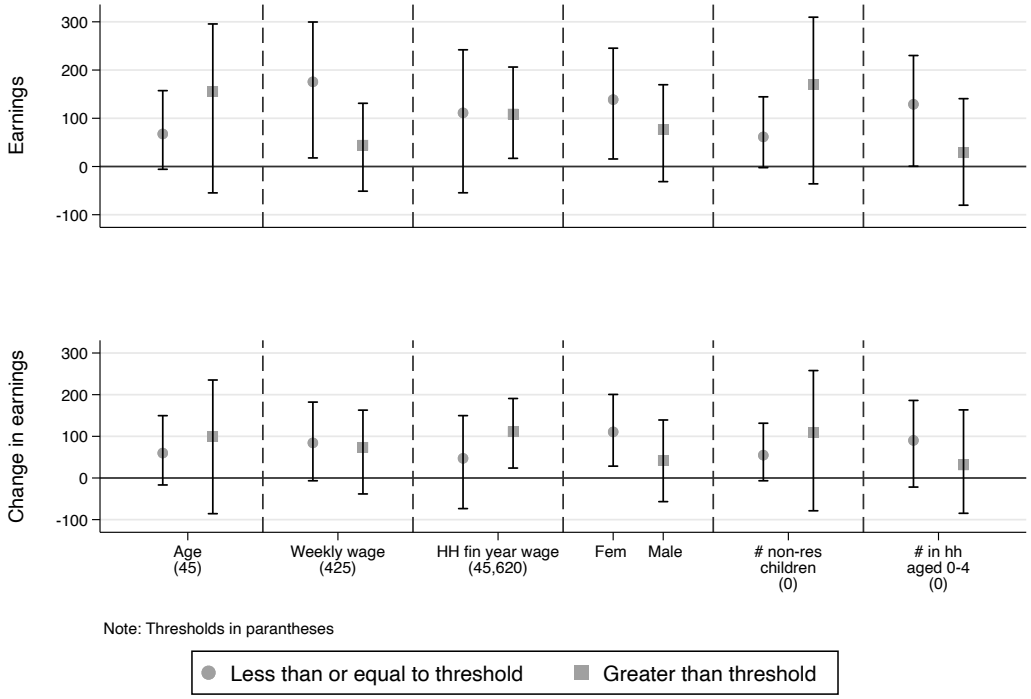
Notes:

Figure 23: Earnings HTEs: GBR



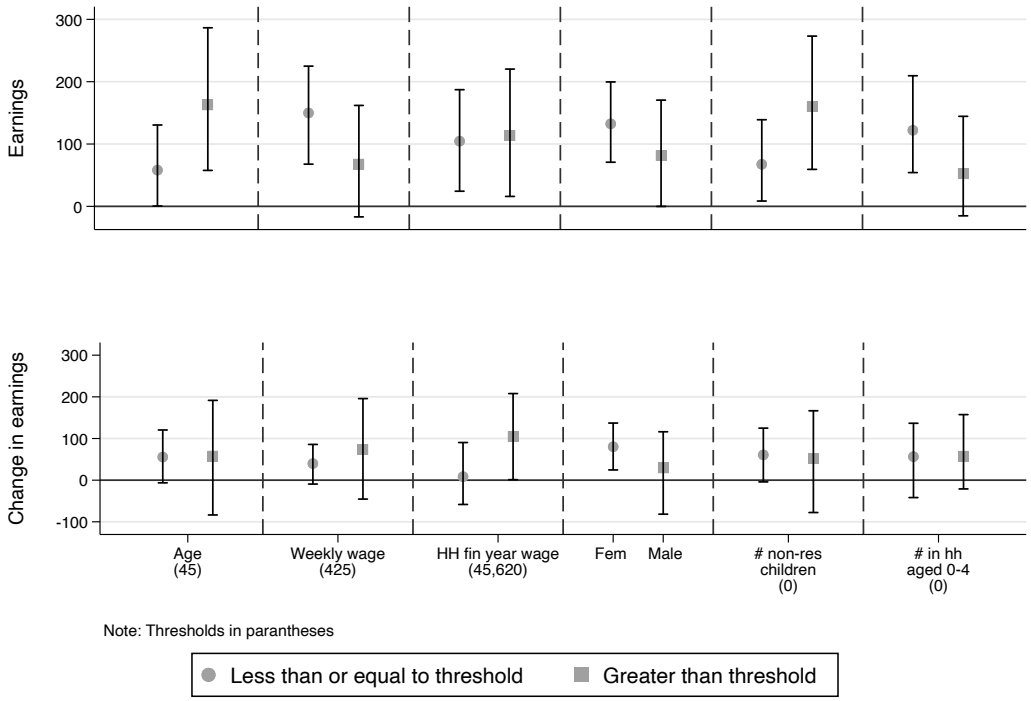
Notes:

Figure 24: Earnings HTEs: LASSO



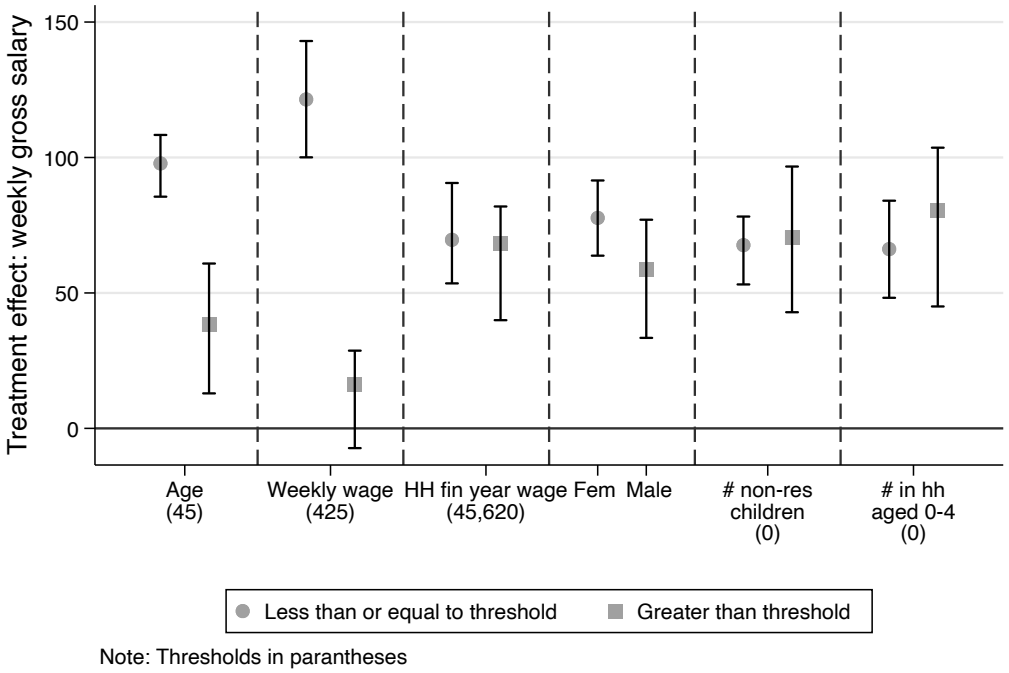
Notes:

Figure 25: Earnings HTEs: Ridge



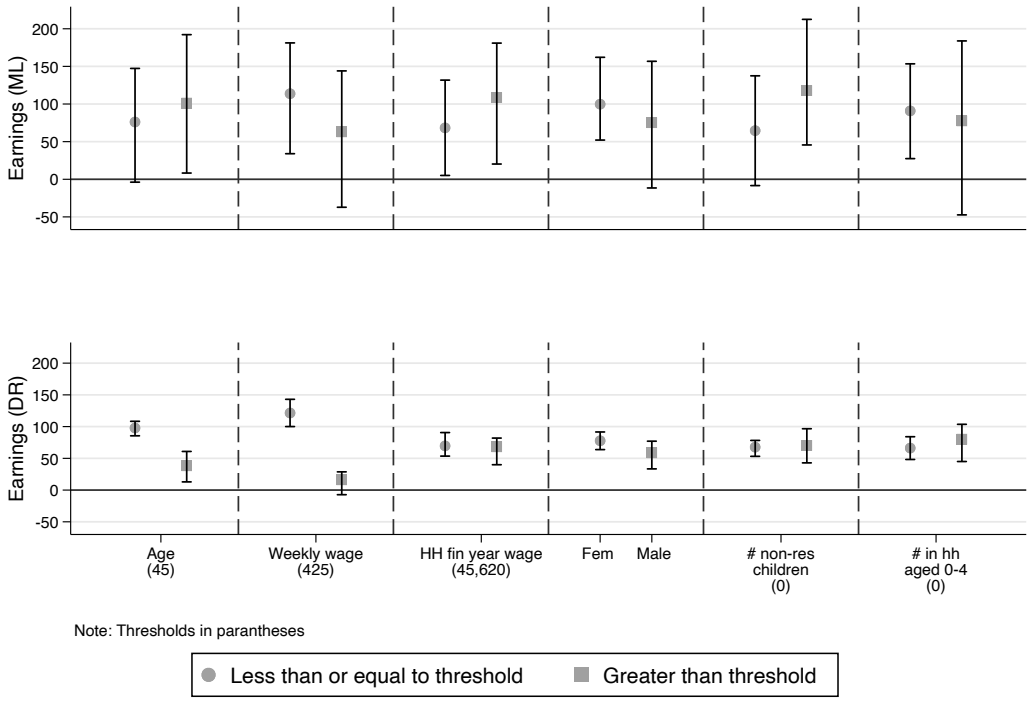
Notes:

Figure 26: Earnings HTEs: DR



Notes:

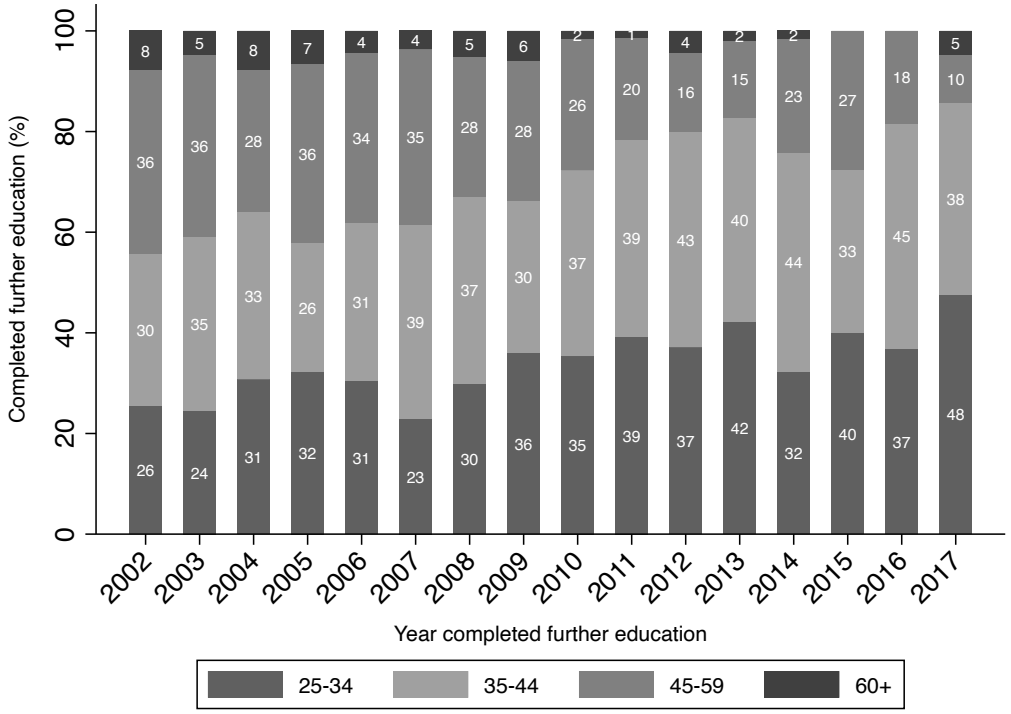
Figure 27: Earnings HTEs: ML and DR



Notes:

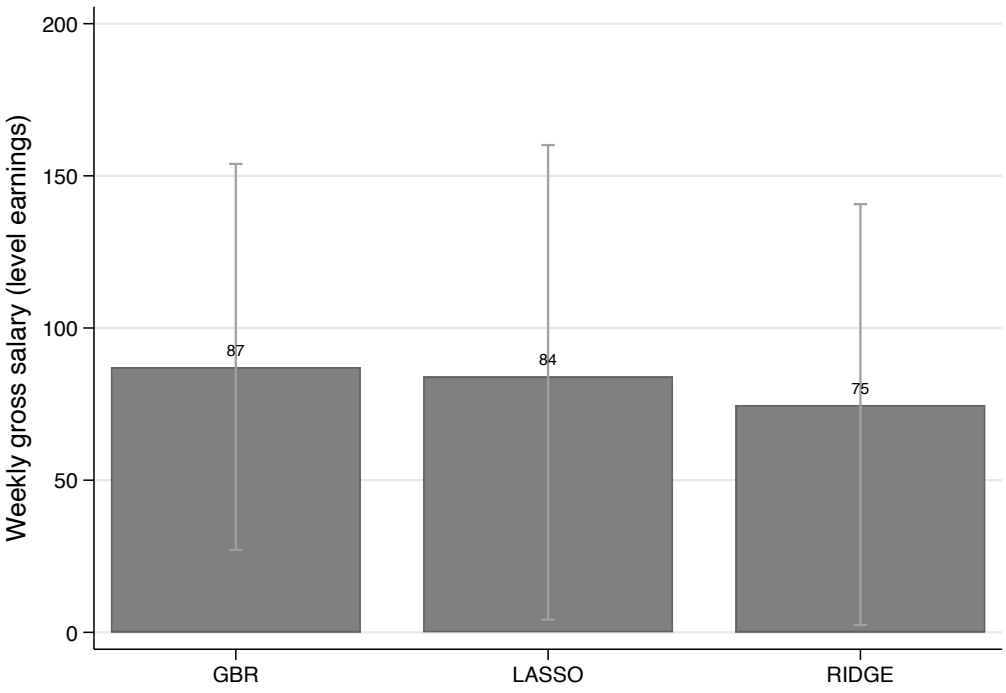
11 Appendix

Figure 28: Timing of Completion by Age



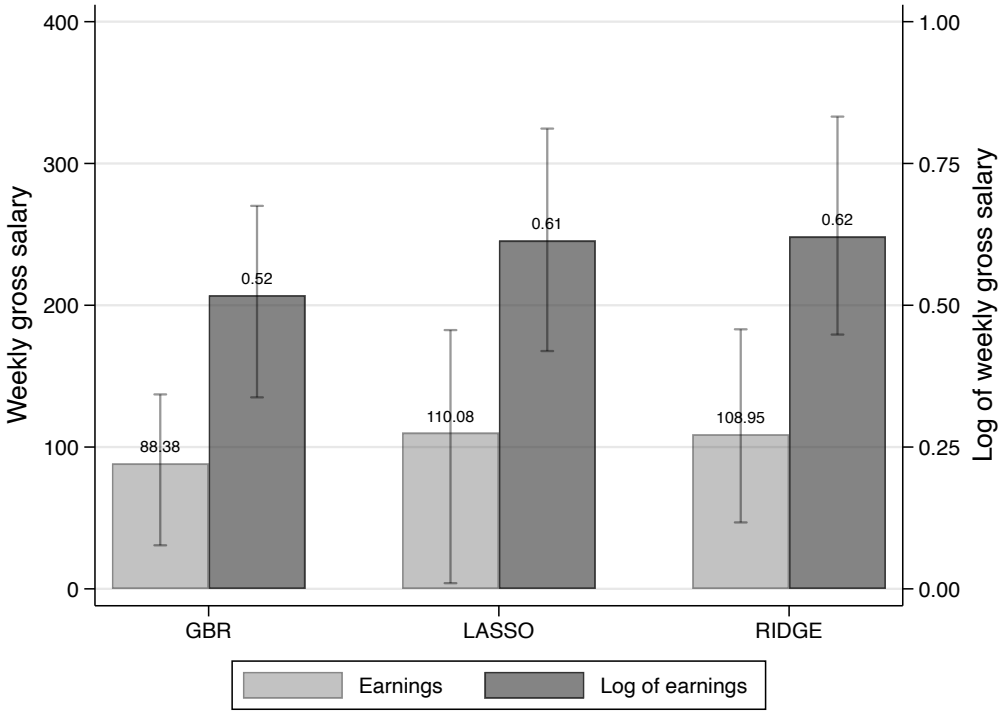
Notes: Sample of 25 or older respondents who had completed a degree at any point between 2002 and 2017. Total completions: 1,383.

Figure 29: Value-add in earnings: 25-45 year-old sample



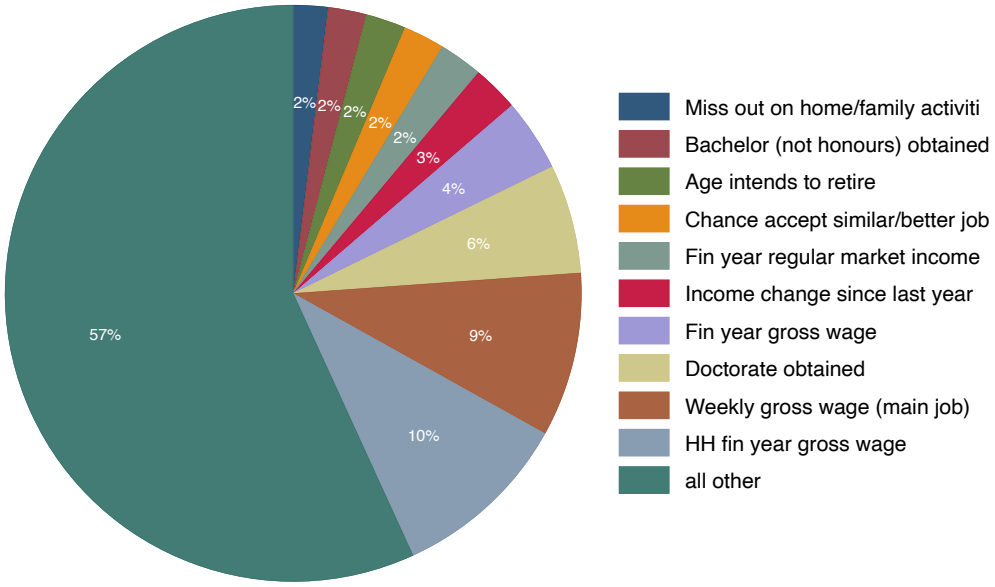
Notes:

Figure 30: Value-add in log earnings



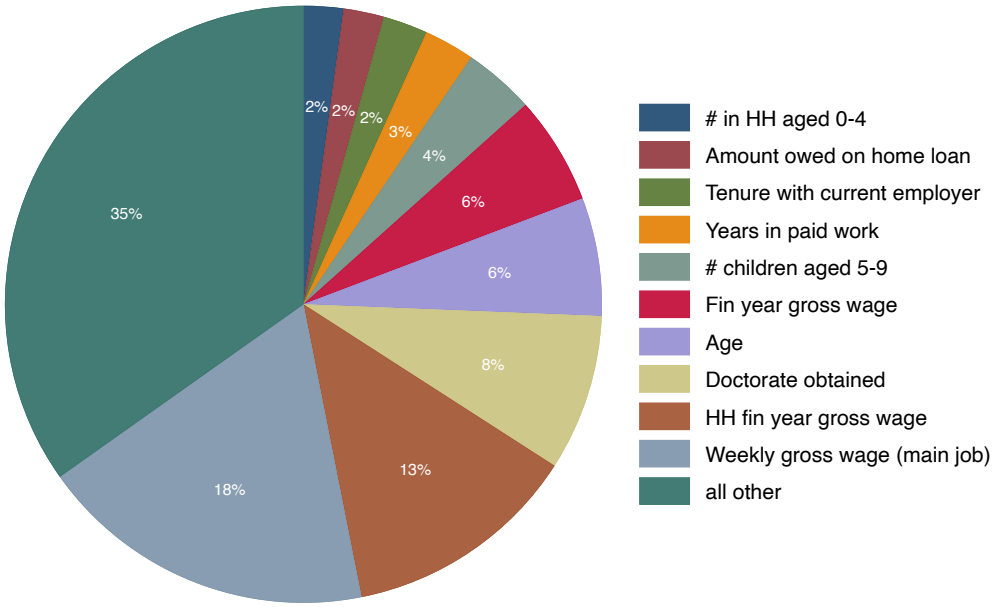
Notes:

Figure 31: Importance Features in Heterogeneous Treatment Effects Estimation using Ridge: Level Earnings



Notes:

Figure 32: Importance Features in Heterogeneous Treatment Effects Estimation using LASSO: Level Earnings



Notes:

Table 5: Nested CV Holdout Sample: Change in Earnings

Model	Outcome surface	Negative MSE	NMSE Std	R-squared	R-squared Std	ATE	ATE_std
GBR	Treated	-1013520	389355	0.13	0.11	53.2	20.4
	Control	-664787	113211	0.25	0.06		
LASSO	Treated	-981927	364876	0.15	0.10	73.1	19.7
	Control	-703529	178858	0.20	0.07		
Ridge	Treated	-1004753	446164	0.15	0.09	43.7	11.0
	Control	-708521	174065	0.20	0.05		

Notes: 5 fold CV performed on 80% train sample. All statistics presented in this table are based on the 20% holdout sample. Ten outer folds are used. See Figure 1 for more details

Table 6: Average Treatment Effects: comparisons across models

Outcome	Model	N	ATE	S.E (ATE)
Change in earnings	OLS	5441	57.33	28.61
	T-learner (GBR)	5441	67.62	32.85
	T-learner (LASSO)	5441	79.10	46.08
	T-learner (Ridge)	5441	56.67	42.86
	Doubly Robust (GBR)	5441	60.44	10.89
	Doubly Robust (LASSO)	5441	65.20	5.33
	Doubly Robust (Ridge)	5441	63.69	3.48
	Bayesian Ridge	5441	44.17	24.11
	Bayesian Ridge (cf. s)	5441	41.08	22.70
	Gaussian Process	5441	0.52	7.06
	Hierarchical BLM	5441	0.20	

Notes: Sample of 25 or older respondents who had completed a degree at any point between 2002 and 2017. Total completions: 1,383. See Figure 2 for more details.

Reference List

Acemoglu, D. and D. Autor (2011). Skills, tasks and technologies: Implications for employment and earnings, Volume 4 of Handbook of Labor Economics, pp. 10431171. Elsevier.

Angrist, Joshua D., and Alan B. Krueger. "Does compulsory school attendance affect schooling and earnings?." The Quarterly Journal of Economics 106.4 (1991): 979-1014.

Ashenfelter, Orley, and David J. Zimmerman. "Estimates of the returns to schooling from sibling data: Fathers, sons, and brothers." Review of Economics and Statistics 79.1 (1997): 1-9.

Ashenfelter, Orley. Estimating the Effect of Training Programs on Earnings. The Review of Economics and Statistics 60.1 (1978): 4757.

Ashenfelter, Orley, and David Card (1985). "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs", Review of Economics and Statistics, 67, 648-660.

Athey, Susan and Stefan Wager. 2019. Estimating treatment effects with causal forests: An application. arXiv preprint arXiv:1902.07409 .

Atkinson, Georgina, and John Stanwick. Trends in VET: policy and participation, NCVER, Adelaide (2016). Australian Government Department of Education, Skills and Employment (DESE). Higher Education Participation and Partnerships Program (HEPPP). 2021, https://www.dese.gov.au/heppp.

Australian Government Department of Education, Skills and Employment (DESE). Selected Higher Education Statistics 2019 Student Data, Attrition, retention, and success rates, 2019a, https://app.powerbi.com/view?r=eyJrIjoiOTY3ZmQ2OTUtODdlYi00YTMxLWIwNzAtM2RiZWMwNjk0Mzc1IiwidCI6ImRkMGNmZDE1LTQ1NTgtNGIxMi04YmFkLWVhMjY5ODRmYzQxNyJ9.

Australian Government Department of Education, Skills and Employment (DESE). Selected Higher Education Statistics 2019 Student Data, Cohort analysis completion rates, 2019b, https://app.powerbi.com/view?r=eyJrIjoiOTY3ZmQ2OTUtODdlYi00YTMxLWIwNzAtM2RiZWMwNjk0Mzc1IiwidCI6ImRkMGNmZDE1LTQ1NTgtNGIxMi04YmFkLWVhMjY5ODRmYzQxNyJ9.

Australian Government Department of Education, Skills and Employment (DESE). Skills and Training Incentive. 2022a, https://www.dese.gov.au/skills-and-training-incentive.

Australian Government Department of Education, Skills and Employment (DESE). Skills and Training Incentive Guidelines. 2022b, https://www.dese.gov.au/skills-and-training-incentive/resources/skills-and-training-incentiveguidelines.

Autor, D., L. Katz, and M. Kearney (2008). Trends in U.S. wage inequality: Re-assessing the revisionists. Review of Economics and Statistics 90(2), 300323.

Belfield, C., and Bailey, T. (2017). The Labor Market Returns to Sub-Baccalaureate College: A Review. A CAPSEE Working Paper. Center for Analysis of Postsecondary Education and Employment.

Bingham, E., Chen, J.P., Jankowiak, M., Obermeyer, F., Pradhan, N., Karaletsos, T., Singh, R., Szerlip, P., Horsfall, P. and Goodman, N.D., (2019). Pyro: Deep universal probabilistic programming. The Journal of Machine Learning Research, 20(1), 973-978.

Blanden, J., Buscha, F., Sturgis, P., and Urwin, P. (2012). Measuring the earnings returns to lifelong learning in the UK. Economics of Education Review, 31(4), 501-514.

Bloom, Howard S. Back to Work. Testing Reemployment Services for Displaced Workers. WE Upjohn Institute for Employment Research, 1990.

Bckerman, Petri, Mika Haapanen, and Christopher Jepsen. "Back to school: Labor-market returns to higher vocational schooling." Labour economics 61 (2019): 101758.

Card, David. The causal effect of education on earnings. In Handbook of labor economics, Vol. 3A, edited by Orley Ashenfelter and David Card. New York and Oxford: Elsevier Science, North-Holland, pp. 180163, 1999.

Card, David. Using geographic variation in college proximity to estimate the return to schooling. In Aspects of labour market behaviour: Essays in honour of John Vanderkamp, edited by Louis Christofides, E. Kenneth Grant, and Robert Swidinsky. Toronto: University of Toronto Press, pp. 20122, 1995.

Caruso, Stephanie. The Changing Face of a Student: Returning to Education at a Mature Age in Australia. Short courses, 2018, https://www.shortcourses.com.au/ed/studying-as-a-mature-age-student/.

Coelli, Michael, Domenico Tabasso, and Rezida Zakirova. Studying beyond Age 25: Who Does It and What Do They Gain? Research Report. National Centre for Vocational Education Research Ltd. PO Box 8288, Stational Arcade, Adelaide, SA 5000, Australia, 2012.

Dorsett, Richard, Silvia Lui, and Martin Weale. "The effect of lifelong learning on mens wages." Empirical Economics 51.2 (2016): 737-762.

Duchini, Emma. "Is college remedial education a worthy investment? New evidence from a sharp regression discontinuity design." Economics of Education Review 60 (2017): 36-53.

Dynarski, Susan, Brian Jacob, and Daniel Kreisman. "How important are fixed effects and time trends in estimating returns to schooling? Evidence from a replication of Jacobson, Lalonde, and Sullivan, 2005." Journal of Applied Econometrics 33.7 (2018): 1098-1108.

Dynarski, Susan, Brian Jacob, and Daniel Kreisman. "The Fixed-Effects Model in Returns to Schooling and Its Application to Community Colleges: A Methodological Note." Centre for Analysis of Postsecondary Education and Employment (2016).

Edward H. Kennedy. Optimal doubly robust estimation of heterogeneous causal effects. arXiv:2004.14497 (2020).

Harmon, Colm, and Ian Walker. "The marginal and average returns to schooling in the UK." European Economic Review 43.4-6 (1999): 879-887.

Harmon, Colm, and Ian Walker. Estimates of the Economic Return to Schooling for the United Kingdom. American Economic Review, 85: (1995), 12781286.

Harmon, Colm, Hessel Oosterbeek, and Ian Walker. "The returns to education: Microeconomics." Journal of economic surveys 17.2 (2003): 115-156.

Heckman, James J., Robert J. LaLonde, and Jeffrey A. Smith. "The economics and econometrics of active labor market programs." Handbook of labor economics. 3 (1999): 1865-2097.

Hoffman, M.D. and Gelman, A., 2014. The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. J. Mach. Learn. Res., 15(1), pp.1593-1623.

Imbens, Guido W., and Donald B. Rubin. Causal inference in statistics, social, and biomedical sciences. Cambridge University Press, 2015.

Jacob, Brian A., and Lars Lefgren. "Remedial education and student achievement: A regression-discontinuity analysis." Review of economics and statistics 86.1 (2004): 226-244.

Jacobson, Louis, Robert LaLonde, and Daniel G. Sullivan. "Estimating the returns to community college schooling for displaced workers." Journal of Econometrics 125.1-2 (2005): 271-304.

Kristoffersen, Ingebjrg. "Great expectations: Education and subjective wellbeing." Journal of Economic Psychology 66 (2018): 64-78.

Knzel, Sren R., Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. "Metalearners for estimating heterogeneous treatment effects using machine learning." Proceedings of the national academy of sciences 116.10 (2019): 4156-4165.

Leigh, Duane E. Does Training Work for Displaced Workers? A Survey of Existing Evidence. WE Upjohn Institute for Employment Research, 300 South Westnedge Avenue, Kalamazoo, MI 49007, 1990.

Machin, Stephen. Social Disadvantage and Education Experiences. No. 32. OECD Publishing, 2006.

Melguizo, Tatiana, Fabio Sanchez, and Tatiana Velasco. "Credit for low-income students and access to and academic performance in higher education in Colombia: A regression discontinuity approach." World development 80 (2016): 61-77.

Mincer, Jacob. Schooling, Experience and Earnings, National Bureau of Economic Research, New York, 1974. NCVER. Total VET students and courses 2020: program enrolments DataBuilder, Total by Year (2021).

O'Shea, Sarah Elizabeth, Josephine May, and Cathy Stone. Breaking the barriers: supporting and engaging mature age first-in family university learners and their families. 17th International FYHE Conference. Brisbane, Australia: QUT Publications, 2014.

Organisation for Economic Co-operation and Development. Indicator C1: Who participates in education? Education at a glance 2016: OECD indicators. OCED Publishing, 2016.

Parliament of Australia. Higher Education Loan Program (HELP) and other student loans: a quick guide. 2017, https://www.aph.gov.au/About_Parliament/Parliamentary_Departments/Parliamentary_Library/pubs/rp/rp1617/Quic k_Guides/HELP.

Perales, Francisco, and Jenny Chesters. "The returns to mature-age education in Australia." International Journal of Educational Research 85 (2017): 87-98.

Polidano, Cain, and Christopher Ryan. Long-Term Outcomes from Australian Vocational Education. Melbourne Institute of Applied Economic and Social Research, The University of Melbourne, 2016.

Powdthavee, Nattavudh, Warn N. Lekfuangfu, and Mark Wooden. What's the good of education on our overall quality of life? A simultaneous equation model of education and

life satisfaction for Australia. Journal of Behavioral and Experimental Economics, 54 (2015): 1021.

Raaum, Oddbjrn, and Hege Torp. "Labour market training in Norwayeffect on earnings." Labour economics 9.2 (2002): 207-247.

Rubin, Donald B. "Randomization analysis of experimental data: The Fisher randomization test comment." Journal of the American statistical association 75.371 (1980): 591-593. RMIT Classification: Trusted

Studies in Australia. Study Costs. 2018, https://www.studiesinaustralia.com/studying-in-australia/how-to-study-inaustralia/study-costs.

Study Assist. Student learning entitlement. 2022a, https://www.studyassist.gov.au/help-loans-commonwealthsupported-places-csps/student-learning-entitlement.

Study Assist. Student learning entitlement. 2022b, https://www.studyassist.gov.au/help-loans/combined-help-loanlimit.

Universities Australia. "2020 Higher Education Facts and Figures." Universities (2020): Australia.

Universities Australia. "The Demand Driven System." https://www.universitiesaustralia.edu.au/policysubmissions/diversity-equity/the-demand-driven-system/.

Wang, Weidong, Yongqing Dong, Xiaohong Liu, Linxiu Zhang, Yunli Bai, and Spencer Hagist. "The more educated, the healthier: evidence from rural China." International journal of environmental research and public health 15.12 (2018): 2848.

Xu, Di, and Madeline Trimble. "What about certificates? Evidence on the labor market returns to nondegree community college awards in two states." Educational Evaluation and Policy Analysis 38.2 (2016): 272-292.

Zeidenberg, Matthew, Marc Scott, and Clive Belfield. "What about the non-completers? The labor market returns to progress in community college." Economics of Education Review 49 (2015): 142-156.