Australian Government
Department of Education, Skills and Employment

Factors Affecting Higher Education Completions

Methodology

- *This project utilised the Australian Bureau of Statistics Multi-Agency Data Integration Project[1] (MADIP), which contains highly detailed de-identified unit level data. This factsheet explains the three main processes that supported the analysis of factors affecting undergraduate bachelor's degree completions.*
- *Data was linked from various administrative datasets across government.*
- *Data cleaning and generation of analytical datasets.*
- *Methodologies and statistical techniques that draw associative and causal inferences between the variables of interest in our analysis.*

Introduction

The Data Integration Partnership for Australia[2] (DIPA) was a three-year partnership across 20 Commonwealth agencies which commenced on 01 July 2017 and ended 30 June 2020. DIPA supported this project to better understand the factors that affect bachelor's degree student completion in higher education. This paper details the data and methodology we employed to deliver the findings in our factsheets.[3]

- The Data Linkage and Assembly section explains how the Australian Bureau of Statistics (ABS) facilitated the migration and linkage of the administrative datasets used in this project.
- The Construction of the Analytical Dataset section details the steps taken to integrate the datasets into a form suitable for analysis, and the steps for deriving variables of interest.
- Methodologies and Statistical Techniques is split into two sections: Descriptive and predictive modelling, and Causal inference. The Descriptive and predictive modelling section reveals how we were able to use machine learning to identify key variables that affect undergraduate completion rates as seen in the Introduction factsheet[4]. The Causal inference section provides an explanation of how we were able to estimate average treatment effects for variables such as: gap years[5], living with a disability[6], student payments[7] and more.
- The endnotes section contains further readings on the subjects discussed in this paper.

[1] https://www.abs.gov.au/websitedbs/d3310114.nsf/home/statistical+data+integration+-+madip
[2] https://www.pmc.gov.au/public-data/data-integration-partnership-australia
[3] https://www.education.gov.au/transition-education-work
[4] https://www.education.gov.au/transition-introduction
[5] https://www.education.gov.au/transition-gap-year
[6] https://www.education.gov.au/transition-disability
[7] https://www.education.gov.au/transition-study-assistance

Contents

Introduction **1**

Data Linkage and Assembly **3**

Methodologies and statistical techniques **3**

Descriptive and predictive modelling 3

Logistic regression 4

Random forests 4

Causal inference modelling 4

Directed Acyclic Graphs (DAGs) 4

Coarsened Exact Matching (CEM) 4

Relative risk estimation 5

Causal Forests 5

Construction of the analytical dataset **6**

Variables prior to commencement 6

Variables during study 8

Appendix A: Income rules **10**

References **14**

Data Linkage and Assembly

The Multi-Agency Data Integration Project (MADIP) is a partnership among Australian Government agencies to develop a secure and enduring approach for combining information on healthcare, education, government payments, personal income tax, and population demographics (including the Census) to create a comprehensive picture of Australia over time.

This project uses a customised MADIP extract which contains data from the following administrative collections:

- Social Security and Related Informations (SSRI)
- Medicare Benefits Schedule (MBS)
- Pharmaceutical Benefits Schedule (PBS)
- Census of Population and Housing 2016
- Higher Education Management System (HEIMS)
- Personal Income Tax and Pay As You Go summaries (PIT/PAYG)
- Registries Death Data

The ABS provided the infrastructure and support to link these datasets and provide us access to confidential unit record files. Linkage rates were above ninety percent and the overall data linkage quality was high.

Methodologies and statistical techniques

To investigate the factors affecting bachelor's degree completions we used several types of statistical inference techniques: descriptive and predictive modelling, and causal inference modelling.

The differences between these techniques can be understood by the types of questions they are trying to answer.

Descriptive modelling seeks to understand how variables may be related to one another, e.g. *Is there a trend between commencement age and bachelor's completion rates*?

Predictive modelling is used to achieve the best possible accuracy when trying to predict an outcome based on covariates, e.g. *How well can we predict student completion rates*?

Causal inference modelling answers the '*what if*?' questions, e.g. *what if we gave all undergraduate bachelor's students youth allowance*? *How would this affect their completion rates*?[i]

Descriptive and predictive modelling

Our team used R[ii], Python[iii], SAS[iv] and Stata[v] to generate our results. RStudio[vi] was used for R programming. Spyder[vii] and Jupyter Notebook[viii] were used for Python programming. Packages from these programs are also referenced throughout this document. Two main models were used to assess the relative importance and relative effect sizes of our covariates: random forests and logistic regression.

Logistic regression

Logistic regression was used as a descriptive tool to model the completion rates of students against all the variables. This gave an overview of the effect sizes for each variable and how they compare. An application of this technique is in the Introduction.[4] This analysis was performed with SAS.

Random forests

A random forest model[ix] was used to complement the logistic regression results, however taking the non-parametric approach allowed for greater predictive accuracy. The random forest was primarily used to retrieve a variable importance list. While the logistic regression may tell us the effect sizes of our variables, the variable importance tells us which have a greater impact on predicting completions. *Relative importance* was calculated using Gini impurity[x] which calculates the decrease in node purity if a variable was to be removed from the model. An application of this technique in the Introduction.[4] This analysis was performed with Python, using the Scikit-learn package.[xi]

Random forests were also used to generate lists of the most confounding variables to control for in causal inference modelling. A random forest was fit to predict completion rates and our treatment variable of interest, i.e. confoundedness.

The Synthetic Minority Oversampling Technique (SMOTE)

Some sample sizes for our confoundedness random forests had very low numbers (i.e. imbalanced data). This is a problem for the random forest algorithm which optimises global accuracy. To combat this, we used the SMOTE[xii] (Synthetic Minority Oversampling Technique) algorithm to oversample small sub-populations of interest.

Causal inference modelling

For causal inference estimation we developed a doubly robust methodology using coarsened exact matching plus relative risk estimation. When we had poor exchangeability[8] we used a causal forest[xiii] with overlap weighting.[xiv]

Directed Acyclic Graphs (DAGs)

To estimate the causal effect of a treatment a DAG must first be drawn to identify the variables (confounders) which must be controlled for and those that must be ignored (i.e. colliders or mediators)[xv]. The remaining variables that do not fit into either of these groups can be selected for matching using the confoundedness random forest variable importance list.

Coarsened Exact Matching (CEM)

Matching is used in causal inference studies to group observations with similar covariates together and emulate randomised control trials.[xvi] The concept is to 'hold all else equal' and observe the average effect of a treatment over our sample and infer to the population. Traditionally propensity scores[9] have been used for matching,[xvii] however, a treated unit may have a very similar propensity score to an untreated unit but may not have many characteristics exactly in common. Propensity score matching is a simple way of reducing the dimensionality of a unit's covariates into a single

[8] Exchangeability is an assumption in causal inference. It requires the treated and untreated populations to have some shared covariate representation so that we can emulate a randomised controlled trial.

[9] A propensity score is the probability of a student being treated (i.e. receiving student payments) given their background characteristics

dimension, however exact matching allows us to retain more dimensions. Exact matching works by grouping treated and untreated individuals into strata based on their covariate levels, e.g. women of age 18 that come from a regional background with and without student payments may be grouped together. Any units that do not have a counterfactual (any treated or untreated unit that does not have a statistical pair) gets discarded, so we used coarsened exact matching to allow our covariate groups to be more flexible. An example of this is grouping women of ages 18 – 20 that come from regional or remote backgrounds. This increases our sample retention while still approximating a randomised controlled trial. Within each matching stratum the treated are weighted such that their representation in each stratum is 50 per cent (or a 1:1 relationship between treated and untreated). To maintain the unweighted post-matching sample size these groups are then down-weighted appropriately.[xviii]

Relative risk estimation

Relative risk estimation was used on the matched sub-samples to estimate the causal effect of the treatments. Logistic regression (which fits odds-ratios) was used in our larger model to understand the effect sizes of our predictors on completion rates, however, for causal inference relative risk estimation was preferred due to the collapsibility property which allows for inference from the matching stratum to the population.[xix] Relative risk was estimated using a Poisson generalised linear model with Stata.[xx] Calculating the margins post-regression allows for an easily interpretable treatment effect as probability of completion rate increases.

Causal Forests

A causal forest is a doubly robust technique which uses random forests to group observations into categories and estimate their propensity scores. Causal forests perform inverse propensity score weighting (IPSW)[xxi] in the leaves of the forests averaged over all the trees generated, resulting in an average treatment effect for a variable of interest, randomising all other variables. For the work and study factsheet, we were interested in the non-financial effect of jobs on completion rates. Matching plus regression was not possible due to almost perfect prediction of job status by controlling for income, resulting in extreme propensity scores.[10] The causal forest procedure was paired with overlap weights, a technique that estimates how likely an observation is to be in the opposite group (e.g. if someone doesn't have a job, what is the probability of them having a job based on other covariates). These overlap weights are robust to extreme scores and were used instead of the IPSW. Causal forest analysis was performed using the R package grf.[xxii]

[10] Extreme propensity scores are an issue in inverse propensity score weighting as dividing by a very small number causes a few observations being the majority representation of the whole sample.

Construction of the analytical dataset

The majority of the MADIP data covers the years 2011 – 2016 so we decided to look at a 6 year completion rate for first time bachelor's students commencing in 2011. If they did not complete their degree by semester 2 2016, we considered them incomplete.[11]

Variables prior to commencement

Most static demographic variables were taken from the 2016 Census. These include:

- Aboriginal and Torres Strait Islander (Indigenous) status
- Gender
- English speaking country of birth
- Language spoken at home
- Age at commencement of study

If there were non-binary genders, those students were excluded from the analyses due to low sample. When either **Gender** or **Indigenous status** was missing from the census, we used the statistical mode from all our other datasets to supplement.

When **English speaking country of birth** was missing from census, we used the HEIMS enrolment file to supplement.

As **Language spoken at home** is a potentially dynamic indicator, we took the modal language spoken at home during a student's 2011 enrolment from the HEIMS enrolment file. When enrolment data was missing, we took the language spoken at home from census 2016 data.

English speaking country of birth was aggregated into three categories:

1. Australia
2. English speaking
3. Non-English speaking

Similarly, **Language spoken at home** was aggregated into:

1. English
2. Non-English

Commencement age was simply derived by using a student's age reported at census and subtracting the number of months relative to their 2011 commencing semester. Note that age was also supplemented with other MADIP datasets[12] and the modal age was used.

An Index of Relative Socio-economic Advantage and Disadvantage (**IRSAD**) was constructed at the statistical area 1 level and grouped into deciles.[xxiii] It was derived using an address history file collated from all possible addresses across the MADIP datasets. The most common address at any given point in time was used.

The same methodology was also used to derive a **Remoteness** value, categorised into six levels:

- Major cities

[11] Some 2016 semester 2 completions may not be recorded until 2017 semester 1. This has likely introduced negative bias to the completion rates.
[12] MBS/PBS, PIT, SSRI, HEIMS, Census

- Inner regional
- Outer regional
- Remote
- Very remote
- Missing or unknown

Remoteness was further aggregated into Major cities (Metro) and regional/remote for specific analyses.

Tertiary entrance score (TES) and **Parental educational attainment** were derived from the HEIMS enrolment file.

TES was numeric from 30 to 100 for those that had an ATAR[13]. Those that did not have an ATAR were grouped into:

- Prior VET[14] completions
- Prior VET study, but not completed
- Missing TES

Those with ATARs were further binned into the following groupings of roughly equal sample size:

- ≤ 60
- 61 – 67
- 68 – 73
- 74 – 78
- 79 – 82
- 83 – 86
- 87 – 90
- 91 – 93
- 94 – 97
- 98 – 100

Parental educational attainment was grouped into 4 categories based off the number of parents a student reported as having a tertiary qualification:

- No parents with tertiary education
- 1 parent with tertiary education
- 2 parents with tertiary education
- Missing or unknown

We created two **gap year** variables. The first was specifically made for the gap year factsheet and strictly focussed on a school year cohort for school leavers 2008 to 2014. The second variable was a simplified flag used as a covariate for the 2011 six-year completion cohort.

For the gap year factsheet, the **gap year** variable and cohort were derived from HEIMS course, enrol and load files, plus data from PIT and census. We defined students as gap year takers when a 1-year break was taken between completing high school and starting a bachelor's degree. We defined non-gap year takers as those who directly enter university the year following graduating high school. For

[13] ATAR is the Australian Tertiary Entrance Rank, these were rounded to integers.
[14] Vocational Education and Training

the purposes of income support analysis, we also analysed those who took a 2-year break before starting a bachelor's degree.

Following advice from university statistics team, we applied the following filters:

- Unit status: 2, 3 (pass/fails only, removes instances of withdrawals);
- Course type: 09, 10 (bachelor's degree courses);
- New Admission: 33 (university criterion for acceptance was secondary school completion);
- Highest Participation: 7X (Highest education was secondary school completion);
- Commencement Indicator: 1 (newly commencing student);
- Tertiary entrance score: 0 to 100 (Award courses only);
- First year of enrolment 2009 to 2016;
- Age left school between 15 and 21;
- Combined Year left school 2008 to 2014.

For the 2011 cohort analyses, we derived a simpler derivation of gap year to be used as a covariate. This simple gap year definition provided better coverage of our overall cohort. It was defined as having a 1- or 2-year break between year left school and first year of bachelor's degree enrolment based on the enrol file. All other students were considered non-gap year takers.

Variables during study

We created a semester definition which splits the whole calendar year in half to account for students that may be taking courses outside the standard two semesters per year. We defined a unit as being in the first or last half of the year using the census date for that unit contained in the HEIMS load file.

Study variables taken from HEIMS were aggregated across time for analyses, these include equivalent full-time student load[xxiv] (**EFTSL**), field of education (**FOE**), **attendance mode**, and **institution** of study. As students regularly change their degrees throughout their study, we assigned **FOE** and **institution** based off their completion entry.[15] For those that did not complete their studies, we took their statistical modes. **EFTSL** was calculated as a per-semester-studied value as opposed to a per-semester or per-year. This way students who took breaks from studying were not misrepresented as having low study loads overall. **Attendance mode** was also categorised using the statistical mode over the duration of the student's enrolment. **FOE** was used at the 2-digit level[xxv] and **institution** only contained Table A and B providers[xxvi] and we aggregated all non-university providers into a single grouping.

Mental health, **disability**, and **chronic health** services flags were based on advice from the Department of Health and were derived using a combination of items from MBS/PBS as well as self-reporting on enrolment for **disability** and access to **non-student support payments** through SSRI. The flags were used as indicators of present conditions during study. Health data was provided in six-monthly volumes per student from MBS and PBS[16]. Health flags were derived based on guidance from Department of Health.

The **chronic health flag** was based on MBS six-monthly volume data. A student was flagged as having a chronic health condition if they accessed any services for MBS chronic disease management (CDM)

[15] Some honours graduates are reported as finishing the same degree twice in consecutive years
[16] Only data from 2013 – 2016 was available

items during study. CDM items included chronic disease and other allied health services. The 'during study' time period is defined as between a student's initial commencing 2011 semester and the last semester of a student's course of completion or modal course of study inclusive.

The **mental health flag** was derived in a similar way to the **chronic health flag**[17]. A student was flagged as having a known mental health condition if they accessed any of the following mental health services items during study:

- Psychiatric services;
- General Practitioner mental health services, such as a mental health plan;
- Clinical psychology services;
- Other psychology services; and
- Other mental health professional services.

The 'during study' time period is defined as between a student's initial commencing 2011 semester and the last semester of a student's course of completion or modal course of study inclusive.

Within the mental health factsheet, we additionally explored any mental health condition between 2011 and 2016 regardless of study period. We also analysed mental health after study, the 'after study' time period was defined as after the last semester of a student's course of completion or modal course of study.

Gross income[xxvii] while studying was calculated using a combination of PIT, PAYG and SSRI. **Gross income** was averaged over semesters studied. To ensure that we did not include income from semesters where students weren't studying, we only include semesters where there was full coverage over the financial year. PAYG was more flexible as sometimes the payment summaries contained detailed start and end dates. **Gross income** is a summation of:

- **Employment income**
- **Business income**
- **Investment income**
- **Superannuation**
- **Other income**
- **Government pensions**

For the work and study factsheet, **employment income** was used to investigate the labour earnings of students. For the income derivation rules please see Appendix A.

Student support and **non-student support payments** were identified using the SSRI six monthly dataset. If a student had accessed Youth Allowance, Austudy or ABSTUDY at any point in time during their study we gave them a student payment flag and if they had received any other support payment (such as disability payments or parental payments etc.) they received a separate flag.[xxviii]

[17] Mental and chronic health flags share Group E services (Other allied health providers).

Appendix A: Income rules

Income type	Version	Operator	Variable	Description	Notes
Employee/Exertion	PIT	=	Grs_Pmt_Totl_Calcd_Amt	Salary or wages	
Employee/Exertion	PIT	+	AlwncErngs_TipsDrctrsFees_Amt	Allowances earnings tips directors' fees etc	
Employee/Exertion	PIT	+	ETPs_OthrThn_ExcsvCmpnt_Amt	Employment termination payments taxable component	
Employee/Exertion	PIT	+	Prsnl_Srvcs_Atrbd_Incm_Amt	Attributed personal services income	
Employee/Exertion	PIT	+	RFBs_Totl_Amt	Total reportable fringe benefits amount	
Employee/Exertion	PIT	+	RprtblEmplr_Spntn_Cntrbtns_Amt	Reportable employer superannuation contributions	
Employee/Exertion	PIT	+	Assbl_FSI_Amt	Foreign source income assessable foreign source income	
Employee/Exertion	PIT	+	TAX_FREE_AMT	Tax-free component	EMP termination payments from PAYG
Employee/Exertion	PIT	+	Lump_A	Lump sum A	From PAYG
Employee/Exertion	PIT	+	Lump_B	Lump sum B	From PAYG
Employee/Exertion	PIT	+	Lump_D	Lump sum D	From PAYG
Employee/Exertion	PIT	+	Exmt_Forgn_Emplt_Incm_Amt	Foreign source income exempt foreign employment income	
Employee/Exertion	PAYG	=	Lump_A	Lump sum A	
Employee/Exertion	PAYG	+	Lump_B	Lump sum B	
Employee/Exertion	PAYG	+	Lump_D	Lump sum D	

Employee/Exertion	PAYG	+	Lump_E	Lump sum E	
Employee/Exertion	PAYG	+	TOTL_ALWNC_AMT	Total allowances	
Employee/Exertion	PAYG	+	GRS_AMT	Gross Payment Amount	
Employee/Exertion	PAYG	+	EXMPT_FORGN_EMPLT_INCM_AMT	Exempt foreign employment income	
Business	PIT	=	Net_Incm_or_Lss_PP_Amt	Net income or loss from business - primary production	
Business	PIT	+	Net_Incm_or_Lss_NPP_Amt	Net income or loss from business - non-primary production	
Business	PIT	+	Trsts_PP_Dstbn_Amt	Distribution from trusts primary production	
Business	PIT	+	PSI_Net_Amt	Net PSI	
Business	PIT	+	Pshps_NPP_Less_Forgn_Incm_Amt	Distribution from partnerships less foreign income non primary production	
Business	PIT	+	Pshps_PP_Dstbn_Amt	Distribution from partnerships primary production	
Investment	PIT	=	Grs_Intst_Amt	Gross interest	
Investment	PIT	+	Divs_Unfrnkd_Amt	Dividends unfranked	
Investment	PIT	+	Divs_Frnkd_Amt	Dividends franked	
Investment	PIT	+	Divs_FCR_Amt	Dividends franking credit	
Investment	PIT	+	Trsts_Npp_less_CGForgnIncm_amt	Share of net income from trusts less net capital gains and foreign income non primary production	
Investment	PIT	+	NonPPFrnkdDstbnsFrmTrstsAmt	Franked distributions from trusts non primary production	
Investment	PIT	+	Ausn_FCRs_Frm_NZC_Amt	Foreign source income Australian franking credits from a NZ company	
Investment	PIT	+	Net_Forgn_Rnt_Amt	Foreign source income net foreign rent	
Investment	PIT	+	Rntl_Net_Rnt_Amt	Rent net rent	
Superannuation	PIT	=	AASIS_Txbl_Cmpnt_Txd_Elmnt_Amt	Australian annuities and superannuation income streams taxable component taxed element	
Superannuation	PIT	+	AASIS_TxblCmpnt_UtaxdElmnt_Amt	Australian annuities and superannuation income streams taxable component untaxed element	
Superannuation	PIT	+	AASISLSInArrsTaxPrtTxdElmntAmt	Australian annuities and superannuation income streams lump sum in arrears taxable component taxed element	

Superannuation	PIT	+	AASISLSArrsTaxPrtUtaxdElmntAmt	Australian annuities and superannuation income streams lump sum in arrears taxable component untaxed element	
Superannuation	PIT	+	SLS_Txd_Elmnt_Amt	Australian superannuation lump sum payments taxed element	Missing from our MADIP extract
Superannuation	PIT	+	SLS_Utaxd_Elmnt_Amt	Australian superannuation lump sum payments untaxed element	Missing from our MADIP extract
Superannuation	PIT	+	LfInsrncOrScts_Bonus_Amt	Bonuses from life insurance companies and friendly societies	
Other	PIT	=	NetForgnPnsnAnnty_Wtht_UPP_Amt	Foreign source income net foreign pension or annuity without UPP	
Other	PIT	+	NetForgnPnsnAnnty_With_UPP_Amt	Foreign source income net foreign pension or annuity with UPP	
Other	PIT	+	CFC_Incm_Amt	Foreign entities controlled foreign company income	
Other	PIT	+	Tfrr_Trst_Incm	Foreign entities Transferor trust income	
Other	PIT	+	Othr_Net_FSI_Amt	Foreign source income other net foreign source income	
Other	PIT	+	Totl_Othr_Incm_Ctgry_1_Amt	Other income category 1	
Other	PIT	+	Totl_Othr_Incm_Ctgry_2_Amt	Other income category 2	
Government Pensions	PIT	=	AUSNGOVTPNSNS_AND_ALWNC_AMT	Australian Government pensions and allowances	
Government Pensions	SSRI	=	tot_welfare_pay	Total welfare paid	All payment types considered
Tax	PIT	=	net_tax_amt	Net tax	
Tax	PIT	+	Bsc_ML_Calcd_Amt	Medicare levy	
Tax	PIT	+	MLS_Calcd_Amt	Medicare levy surcharge	
Tax	PIT	-	Mdcl_Expnss_Tax_Ofst_Amt	Medical expenses offset available	
Tax	PAYG	=	TAX_WHELD_AMT	Tax Withheld Amount	
Total	PIT	=	Employee/Exertion	Derived	
Total	PIT	+	Business	Derived	
Total	PIT	+	Investment	Derived	
Total	PIT	+	Superannuation	Derived	
Total	PIT	+	Other	Derived	
Total	PIT	+	Government Pensions	Derived	

Total	PAYG	=	Employee/Exertion	Derived	
Total	PAYG	+	RPRTBL_FBT_AMT	Reportable fringe benefits amount	
Total	PAYG	+	RPRTBL_EMPLYR_SUPER_CNTRBN_AMT	Reportable employer superannuation contributions	
Total	PAYG	+	TAX_FREE_AMT	Tax-free component	
Total	PAYG	+	TOTL_TXBL_AMT	Taxable component	
Total	PAYG	+	tot_welfare_pay	Total welfare paid	From SSRI
Disposable	PIT	=	Total	Derived	
Disposable	PIT	-	DDCTNS_TOTL_AMT	Total deductions	
Disposable	PIT	-	Tax	Derived	
Disposable	PAYG	=	Total	Derived	
Disposable	PAYG	-	Tax	Derived	
Net business inc	PIT	=	Net_Incm_or_Lss_NPP_Amt	Net income or loss from business primary production	
Net business inc	PIT	+	Net_Incm_or_Lss_PP_Amt	Net income or loss from business non primary production	
Net business inc	PIT	-	bus_net_tax_amt	Business net tax	

References

[i] Hernán MA, Hsu J, and Healy B (2019) 'A second chance to get causal inference right: a classification of data science tasks'. *Chance*, *32*(1), 42-49.
[ii] R Core Team (2019) *R: A language and environment for statistical computing. R Foundation for Statistical Computing*, Vienna, Austria.
[iii] Van Rossum G, and Drake FL (2011) *The python language reference manual*. Network Theory Ltd.
[iv] SAS Enterprise Guide 7.15 (2017) SAS Institute Inc., Cary, NC.
[v] StataCorp (2019) *Stata Statistical Software: Release 16*. College Station, TX: StataCorp LLC.
[vi] RStudio Team (2020) *RStudio: Integrated Development for R*,RStudio, PBC, Boston, MA.
[vii]Córdoba C (2020) *Spyder: The Scientific Python Development Environment*, Spyder.
[viii] Kluyver T, Ragan-Kelley B, Pérez F, Granger BE, Bussonnier M, Frederic J, Kelley K, Hamrick JB, Grout J, Corlay S, Ivanov P, Avila D, Abdalla S, Willing C (2016, May) 'Jupyter Notebooks-a publishing format for reproducible computational workflows', In *ELPUB* (pp. 87-90).
[ix] Breiman L (2001) Random forests. *Machine learning*, *45*(1), 5-32.
[x] Louppe G, Wehenkel L, Sutera A, and Geurts P (2013) 'Understanding variable importances in forests of randomized trees', *Advances in neural information processing systems* (pp. 431-439).
[xi] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M and Duchesnay E (2011) 'Scikit-learn: Machine Learning in Python', *Journal of Machine Learning Research*, 12, pp. 2825-2830.
[xii] Chawla NV, Bowyer KW, Hall LO, and Kegelmeyer WP (2002) 'SMOTE: synthetic minority over-sampling technique', *Journal of artificial intelligence research*, *16*, 321-357.
[xiii] Wager S, and Athey S (2018) 'Estimation and inference of heterogeneous treatment effects using random forests', *Journal of the American Statistical Association*, *113*(523), 1228-1242.
[xiv] Li F, Thomas LE, and Li F (2019) 'Addressing extreme propensity scores via the overlap weights', *American journal of epidemiology*, *188*(1), 250-257.
[xv] Pearl J (2009) *Causality*, Cambridge university press.
[xvi] Stuart EA (2010) 'Matching methods for causal inference: A review and a look forward', *Statistical science: a review journal of the Institute of Mathematical Statistics,* 25(1), 1.
[xvii] Rubin DB (1974) 'Estimating causal effects of treatments in randomized and nonrandomized studies', *Journal of educational Psychology*, 66(5), 688.
[xviii] Iacus SM, King G and Porro G (2012) 'Causal inference without balance checking: Coarsened exact matching', *Political analysis*, *20*(1), 1-24.
[xix] Hernán MA, Robins JM (2020) *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC.
[xx] Cummings, P (2009) 'Methods for estimating adjusted risk ratios', *The Stata Journal*, *9*(2), 175-196.
[xxi] Xu S, Ross C, Raebel MA, Shetterly S, Blanchette C, and Smith D (2010) 'Use of stabilized inverse propensity scores as weights to directly estimate relative risk and its confidence intervals', *Value in Health*, *13*(2), 273-277.
[xxii] Tibshirani J, Athey S and Wager S (2020*) grf: Generalized Random Forests. R package version 1.2.0.*
[xxiii] ABS (Australian Bureau of Statistics) (2014) *Statistics*, ABS, Accessed 20 March 2020.
[xxiv] Australian Government Department of Education, Skills and Employment (2020) *Equivalent Full-Time Student Load (EFTSL)*, Department of Education, Skills and Employment, Accessed 20 March 2020.
[xxv] Australian Government Department of Education, Skills and Employment (2020) *Field of education code*, Department of Education, Skills and Employment, Accessed 20 March 2020.
[xxvi] Australian Government Department of Education, Skills and Employment (2020) *Determine your provider type*, Department of education, Skills and Employment, Accessed 20 March 2020.
[xxvii] ABS (Australian Bureau of Statistics) (2018) *Government Benefits, Taxes and Household Income*, ABS, Accessed 20 March 2020.
[xxviii] Services Australia (2020) *Payments for students and trainees*, Services Australia, Accessed 28 April 2020.