A Causal Inference Study on the Effects of First Year Workload on the Dropout Rate of Undergraduates

Kaoruko Kawamoto

Keio University

July 3, 2023

Thesis Information

- Title:A Causal Inference Study on the Effects of First Year Workload on the Dropout Rate of Undergraduates
- Author : Marzieh Karimi-Haghighi, Carlos Castillo, and Davinia Hernández-Leo
- Date: Accepted July 27, 2022
- Journal: LNCS,volume 13355

Abstract

- The paper evaluate the risk of early dropout in undergraduate studies using causal inference methods.
- It is determined that taking a relatively lighter workload in the first year have the effect on dropout risk by using PSM, IPW, AIPW, and DROrthoForest.
- The results show that a reduction in workload reduces dropout risk.

Contents

1.1 Introduction

Among students who discontinue their studies, there is some subgroups have specific features.

- Older students at point of entry (over 21 years) are more likely to drop out after the first year compared to younger students who enter university directly from high school.

- Graduation rates among ethnic minority university students are lower than among White students.

2 Related Works

Research has been conducted to predict the dropout risk of students in advance and to seek effective factors to prevent dropouts.

ex1

A model that predicts realtime dropout risk for each student during an online course using the Student Information and Course Management System.

Some studies look at the features driving dropout.

ex2

By using propensity score matching (PSM), it is investigated whether university dropout in the first year is affected by participation in Facebook groups created by students.

3.1 Dataset

The anonymized dataset was used.

- It was provided by Universitat Pompeu Fabra.(2009 - 2018)
- It consisted of 23096 undergraduate students' cases after the dataset was cleaned.

Students were admitted to university through four access types.

- type 1 : students who took a standard admission test (0.81)
- type 2 : students who moved from incomplete studies in another university or were older than 25 (0.10)
- type 3 : students who completed vocational training before (0.07)
- type 4 : students completed a different university degree before (0.02)

3.2 Dataset

Center	N	Dropout rate	Underperf. rate	National %	Male %	Avg. age	Avg. credits	Avg. grade	Access type I
ENG	2,444	41%	56%	89%	79%	19.4	63.4	4.6	65%
HUM	1,749	22%	33%	90%	32%	20.3	63.1	5.9	76%
TRA	2,292	16%	28%	88%	18%	19.3	62.9	6.3	83%
POL	1,683	14%	27%	94%	55%	18.8	63.1	6.2	87%
HEA	1,206	14%	16%	93%	25%	19.0	60.2	7.2	82%
LAW	5,479	12%	32%	92%	33%	19.3	62.5	6.0	79%
ECO	5,707	9%	26%	93%	47%	18.5	62.9	6.3	88%
COM	2,536	7%	7%	96%	27%	18.8	61.7	7.5	84%
All	**23,096**	**15%**	**29%**	**92%**	**40%**	**19.1**	**62.6**	**6.2**	**81%**

- Engineering and Faculty of Humanities have the highest dropout and underperformance rates and the type1 students are lower.
- The average age in the two centers is higher compared to the other faculties.
- The Faculty of Communication, which has the lowest dropout and underperformance rates, there are more national students compared to other schools.

4.1 Methodology

Most important features for predicting dropout risk

Feature extraction based CART was conducted.

- the number of credits in the first year (workload)
- admission grade
- age
- study access type (type1, 2, 3, 4)

Based on the four feature values, three scenarios to be verified by causal inference was set up.

4.2 Methodology

3 scenarios of the treat groups and the control groups

1. The first-year students who are older than the mean
   Treat group : Credits are less than median (60)
   Control group : Credits are more than median (60)

2. All the students
   Treat group : Credits are less than median (60)
   Control group : Credits are more than median (60)

3. Students from access types 3 and 4
   Treat group : Credits are less than median (60)
   Control group : Credits are more than median (60)

The difference in dropout risk between the treatment group and the control group is calculated.

4.3 Methodology

Methods used for causal inference

- Inverse-Propensity score Weighting : IPW assigns weights to each individual based on the inverse of their propensity scores. This method gives more weight to individuals who are less likely to receive the treatment they actually received.

$$TE_i = \frac{W_i Y_i}{p_i} - \frac{(1 - W_i) Y_i}{1 - p_i}$$

- Augmented Inverse-Propensity Weighted : AIPW combines the IPW approach. It estimates the causal effect by weighting the data using inverse propensity scores and then incorporating additional adjustments.

$$TE_i = \frac{W_i Y_i - (W_i - p_i)\widehat{Y}_i}{p_i} - \frac{(1 - W_i) Y_i - (W_i - p_i)\widehat{Y}_i}{1 - p_i}$$

4.4 Methodology

Methods used for causal inference

- This method uses Doubly Robust Orthogonal Forests (DROrthoForest) which are a combination of causal forests and double machine learning to non-parametrically estimate the treatment effect for each individual.

- a negative TE_i shows a reduced dropout risk
- a positive TE_i indicates an increased dropout risk

5.1 Results

In general, the results suggest that in high propensity to treatment conditions (students who are already likely to take less workload) there is a substantial reduction of the probability of dropout, particularly in scenarios 2 and 3.

Table 5. ATE obtained using Propensity Score Matching with five buckets.

Propensity	1. Low	2. Med-low	3. Med	4. Med-high	5. High
Scenario 1	0.18	0.04	-0.05	0.02	-0.08
Scenario 2	-0.04	0.03	0.00	0.02	-0.42
Scenario 3	0.04	-0.08	-0.17	0.30	-0.22

5.2 Results

- In the case of IPW and AIPW, we can see that the 0.95 confidence intervals contain the value zero
  → This means that the uncertainty in these methods is large

- The results with the DROrthoForest method are all negative with confidence intervals that do not contain the zero
  → They show a reduction of the probability of dropout of about 5 percentage points in all three scenarios because of the treatment.

Table 6. IPW, AIPW, and DROrthoForest results estimating the Average Treatment Effect (ATE) and its 95% confidence interval [lower-ci, upper-ci] in three scenarios.

	IPW			AIPW			DROrthoForest		
Scenario	lower-ci	ATE	upper-ci	lower-ci	ATE	upper-ci	lower-ci	ATE	upper-ci
Scenario 1	-0.06	0.02	0.11	-0.01	0.07	0.15	-0.07	-0.06	-0.05
Scenario 2	-0.03	0.03	0.09	-0.06	0.01	0.08	-0.04	-0.04	-0.03
Scenario 3	-0.12	-0.01	0.10	-0.10	0.01	0.12	-0.07	-0.05	-0.03

6 Discussion, Conclusions, and Future Work

About causal inference

- To avoid bias due to confounding factors, we performed matching based on propensity scores between the control group and the treatment group.
- The results suggest a negative effect, i.e., a reduction of risk of dropout, following a lower number of credits taken on the first year.

- $\rightarrow$ we should ask students at risk (in this study, types III and IV) to consider taking a reduced workload, or to ask educational policy makers to consider revising the regulations that establish the minimum number of credits.