Causal Inference Project:
Impact of Scholarships on Student Success

Anushka Mukherjee, Lucie Marimar, Jort Koks, Jakob Sarrazin

April 04th, 2025

Machine Learning for Econometrics
ENSAE, IP Paris
Bruno Crépon, Matthieu Doutreligne

1. Motivation

Motivation I

Retention and Completion: A Core Challenge for Universities

- → **High dropout rates** are a persistent issue in higher education, especially during the first years of study.
- → **Timely graduation** is crucial for both students (career entry) and universities (funding, reputation
- ⇒ **Financial constraints** are a major barrier to academic success — especially for socio-economically disadvantaged students.

Motivation II

Scholarships as a Tool to Improve Student Retention and Graduation

- → **Scholarship programs** are widely used as an intervention, but:
  - Their **causal effect** on student outcomes is difficult to measure
  - Many studies show correlations, but few rigorously identify causality.
- → This study uses a **causal machine learning framework (DML)** to estimate the **true effect of scholarships**, adjusting for observed confounders.
- → Findings can inform **policy decisions** on financial aid allocation and **targeting of support** for at-risk students.

2. PICO & Research Question

PICO Formulation

Population, Intervention, Comparison, Outcome

P - Undergraduate students at a Portuguese university (N = 4,424), with data on demographics, socio-economic background, and prior academic performance.

I - Receiving a scholarship during university studies.

C - Students without scholarships, adjusted for observed confounders (grades, family background, gender, etc.).

O - Two binary outcomes observed 3 years after enrollment:

1. Dropout vs. Enrolled/Graduated
2. Graduated vs. Dropout/Enrolled

Research Question

RQ1

Does receiving a scholarship **reduce** the likelihood of **dropping out** within 3 years?

RQ2

Does receiving a scholarship **increase** the likelihood of **graduating** within 3 years?

3. Data Overview and Exploratory Analysis

The Data

Source

– UCI Machine Learning Repository – Predict Students Dropout and Academic Success

Scope

– Administrative records from a Portuguese university $\longrightarrow$ 4,424 undergraduate students across various degree programs

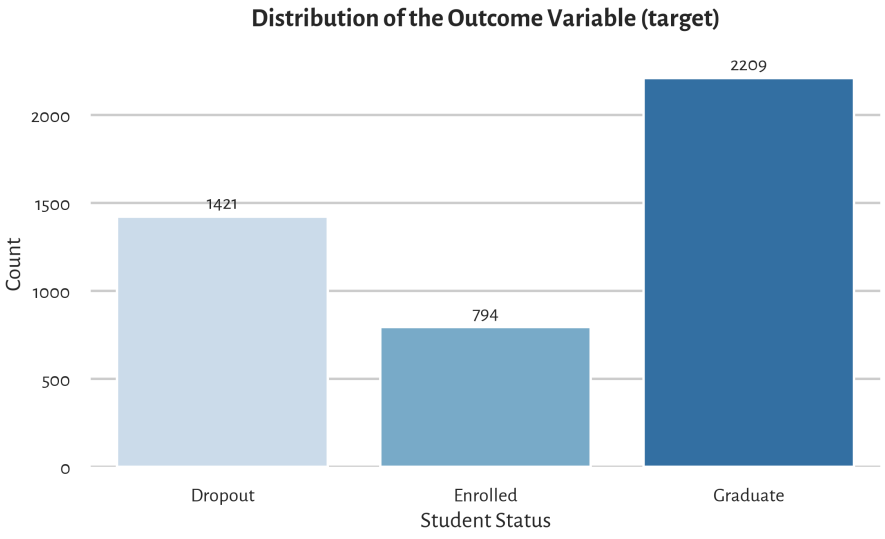
Observation Period

– Students tracked for 3 years after enrollment.

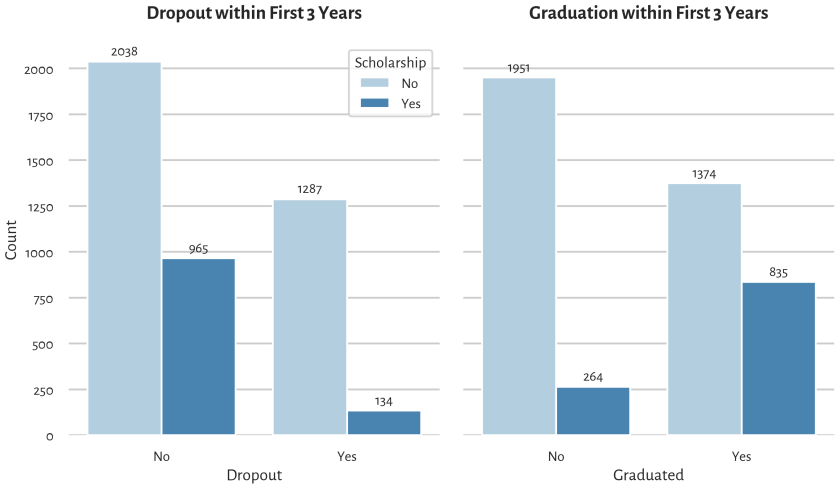
Variables

Outcome Variable	**Treatment Variable**	**Covariates** (Pre Treatment)
Student status after 3 years: – Dropout – Still enrolled – Graduated → *Re-coded into two binary variables for RQ1 & RQ2*	Received scholarship or not (*Binary variable*)	– Academic performance before university – Family background – Economic context – Demographics

Outcome Variable

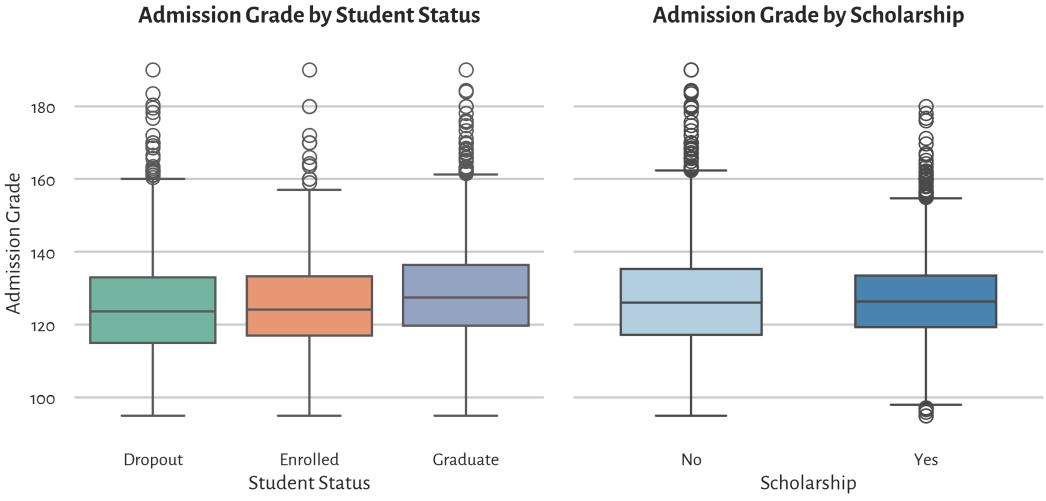


Treatment vs Outcome Variable

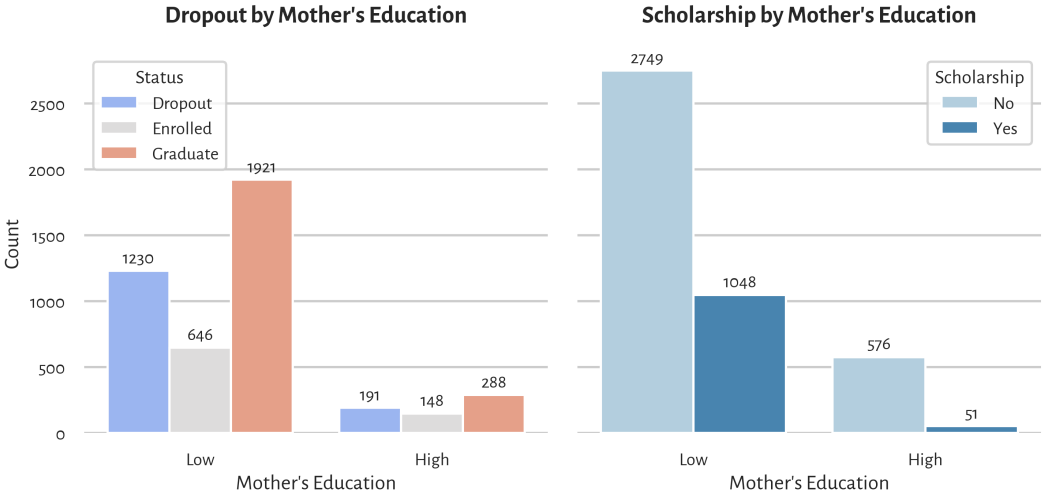


Dropout (Graduation) probability decreases (increases) by 68.50% (83.86%) for scholarship holders.

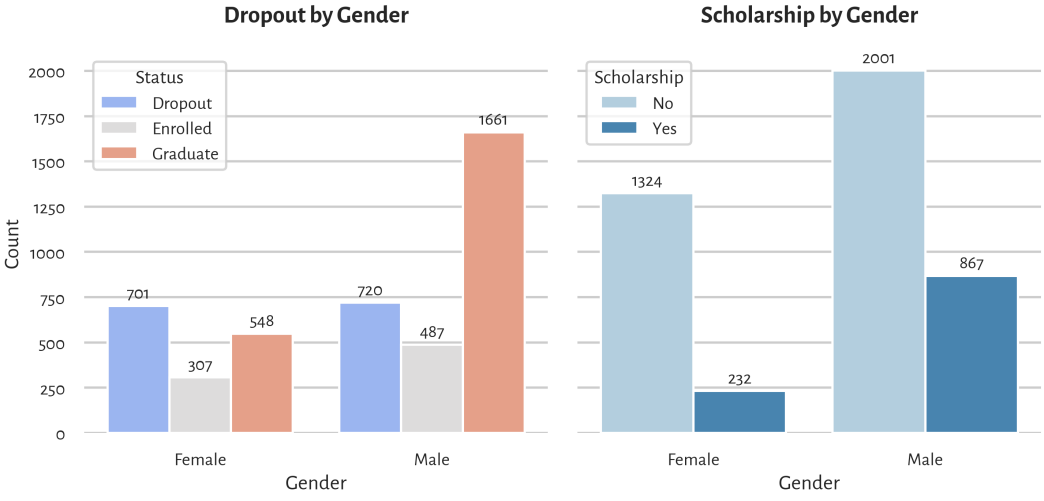
Covariates: Academic Preperation



Covariates: Family Background

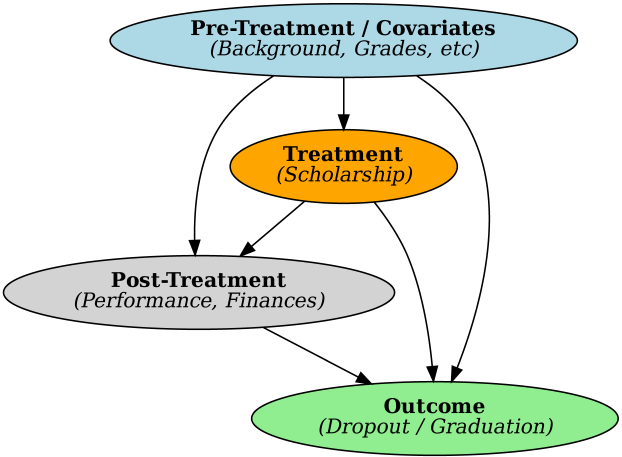


Covariates: Gender



4. Causal Graph and Covariate Selection

Simplified causal graph



Relevant paths

Three relevant paths between Treatment (incoming) and Outcome

1. Treatment (incoming) $\leftarrow$ Covariates $\rightarrow$ Outcome
2. Treatment (incoming) $\leftarrow$ Covariates $\rightarrow$ Post-Treatment $\rightarrow$ Outcome
3. Treatment (incoming) $\leftarrow$ Covariates $\rightarrow$ Post-Treatment $\leftarrow$ Treatment (outgoing) $\rightarrow$ Outcome

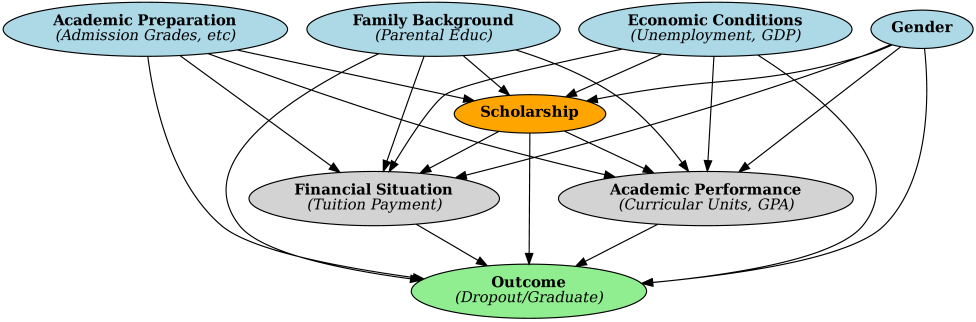
All paths are blocked simultaneously by conditioning on the covariates, but NOT conditioning on the post-treatment effects.

Covariate selection

37 features in the dataset, including one treatment and one target

Pre-treatment (14)	Post-treatment (14)	Unsuitable (7)
– Academic performance before university – Family background – Economic context – Individual characteristics (age, gender, etc.)	– Academic performance during university – Financial sitatation during university	– Marital status – Course selection – Parental occupation – Nationality – Etc.

Full causal graph



Assumptions

1. **Ignorability (Conditional Independence)**: no unmeasured factors that both influence scholarship status and dropout rate.
2. **Positivity**: There should be overlap in characteristics – for any combination of covariates, there are both scholarship and non-scholarship students.
3. **No interference**: One student's scholarship doesn't directly affect another's outcome.

5. Causal Effect Estimation Using Double Post Lasso

6. Causal Effect Estimation Using Double Machine Learning

Steps (using all suitable covariates)

1. Normalize numerical features
2. Random forest classifier to model the treatment (who receives scholarships). It learns $P(T = 1|X)$, the propensity score.
3. Lasso regression with cross-validation, used to model the outcome $E[Y|X]$, i.e., the expected dropout probability given covariates.
4. DoubleMLPLR (Partially Linear Regression) estimator for the causal effect of treatment on outcome, controlling for covariates.

Results research question 1

RQ1: Does receiving a scholarship reduce the likelihood of dropping out within 3 years?

Estimated Treatment Effect: -0.1701 (17% decrease)

Standard Error: 0.0137

95% Confidence Interval: [-0.1969, -0.1432]

T-statistic: -12.4159

P-value: 0.0000

The treatment effect is statistically significant at the 5% level.

Results research question 2

Does receiving a scholarship increase the likelihood of graduating within 3 years?

Estimated Treatment Effect: 0.2348 (23% increase)

Standard Error: 0.0162

95% Confidence Interval: [0.2031, 0.2666]

T-statistic: 14.4859

P-value: 0.0000

The treatment effect is statistically significant at the 5% level.

Heterogeneous treatment effects

RQ1: Does receiving a scholarship reduce the likelihood of dropping out within 3 years?

Treatment Effect (Males): -0.1620 $\pm$ 0.0153

Treatment Effect (Females): -0.2167 $\pm$ 0.0297

Does receiving a scholarship increase the likelihood of graduating within 3 years?

Treatment Effect (Males): 0.2378 $\pm$ 0.0187

Treatment Effect (Females): 0.2499 $\pm$ 0.0338

8. Robustness and Sensitivity Analysis

Sensitivity Analysis: Covariate Sets

How robust are our results to different model specifications?

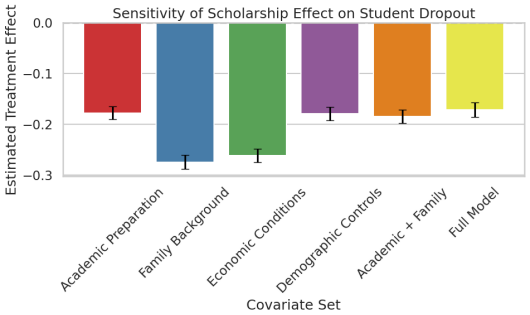
We estimate the treatment effect using DoubleML while varying the covariates included in the model:

- Academic preparation
- Family background
- Economic context
- Demographic controls (age, gender, etc.)
- Combined and full models

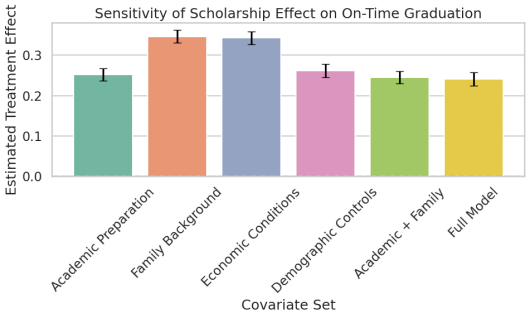
Result: The direction and significance of the effect remains stable, though magnitude varies. Strongest effects seen when economic and family background variables are included.

Sensitivity Analysis: Visual

Treatment effect estimates across different covariate sets.



RQ1: Dropout



RQ2: On-Time Graduation

Placebo Test (RQ1 & RQ2)

What if scholarship assignment were random?

We randomly permute the treatment variable (scholarship).

Re-run DML using this randomized "placebo" treatment.

Expectation: no causal effect should be detected.

Result: Estimated placebo effects for both dropout and graduation are close to zero and statistically insignificant.

$\Rightarrow$ Supports that original results are unlikely to be due to spurious correlation or overfitting.

Robustness Summary

Treatment effects are statistically significant and stable across covariate sets.

Stronger effects seen for students from lower socio-economic backgrounds.

No significant treatment effect detected in placebo test $\rightarrow$ supports causal interpretation.

Conclusion

Our results are robust to model specification and randomization checks, suggesting that scholarships have a genuine causal impact on student success.

X. Conclusion