NOVA IMS
Information Management School

“THE NEW NORMAL”

Isabel Dias 20191215 | Joana Sousa 20191205 | Luís Fernandes 20221649 | Tiago Santos 20191263

What is the effect of online studying on student performance?

PROBLEM 01

This project aims to understand if online learning has any impact on students' performance on state level exams. For that, a dataset with emulated data was used, with variables deemed important for a similar analysis [1][2]. It was retrieved from *Kaggle - “COVID-19 Effect on Grades”* [3]. The dataset contains panel data with 6 semesters – with the first three taking place before COVID-19 lockdowns and the final three coming after lockdowns. The variables were originally manipulated to meet real world trends, and demographic patterns for Portland Oregon.

Dataset info:

18 Variables
8400 Observations

Variables:

Personal Information (12)

School Performance information (3)

State Performance Information (3)

METHODOLOGY 02

Performed data analysis and feature engineering on the dataframe, and visualized relationships with “ggplot”. The target variable was created by summing the variables regarding grades of state level exams, and the individual state level scores were removed.
The modelling was done in two phases:

- ✓ The first phase aimed to test whether different time periods had a significant effect on students' grades, with a panel data structure.
- ✓ In the second phase, two datasets were created (as cross-sectional data), representing the online and presential phase. Then, a Robust Chow test [4] was used to check if the variables had a different effect on grades for the different time periods. This also had the advantage of testing variables that were removed in the Fixed Effect model of the first phase.

RESULTS 03

First Phase

Two estimations were made, one with Random Effects (RE) and another one with Fixed Effects (FE). Using Hausman test, a p-value of 0.0001279 was obtained. So, for a 5% significance, there was statistical evidence that only FE was consistent.
Moreover, assumption RE.3 was verified by applying White Special test, which evidenced heteroskedasticity in the model.
Functional form misspecification was tested using RESET (Regression Specification Error Test) with two different specifications. In both, there was statistical evidence of misspecification, but the better model was chosen.
In that model, the *timeperiod* variables were jointly significant, and *mothereduc* was not. However, there was not a significant difference within each phase (online or presential).

Variables	Estimate	P-value
fathereduc	1.702896	0.1396
readingscore	0.589417	$< 2x10^{-16}$
writingscore	0.588991	$< 2x10^{-16}$
mathscore	0.604135	$< 2x10^{-16}$
timeperiod_1	-0.297754	0.6232
timeperiod_2	0.485978	0.4292
timeperiod_3	-9.753899	$< 2x10^{-16}$
timeperiod_4	-9.866280	$< 2x10^{-16}$
timeperiod_5	-9.817424	$< 2x10^{-16}$
R-Squared	**Adjusted R-Squared**	**Global Significance**
0.54879	0.45792	$< 2.22x10^{-16}$

Table 1: Summary Statistics for Fixed Effects

Second Phase

In both models of online data and presential data, the White Special test was performed with the conclusion of heteroskedasticity in the models.
Then, a RESET test was performed before and after introducing interactions between the grade scores, and even though there was always evidence of functional form misspecification, the interactions improved the model.
The variables *gradelevel* and *familysize* were removed, as they were not statistically significant. The *numcomputers* was statistically significant in one of the models, so it was kept.
Finally, after performing the Robust Chow Test, it was concluded that the effect of the variables when students learnt online was different than presentially.

Variables	Estimate (Presential)	P-value (Presential)	Estimate (Online)	P-value (Online)
(Intercept)	$-2.292x10^{2}$	$< 2x10^{-16}$	$-1.722x10^{2}$	$< 2x10^{-16}$
school	-10.29	$8.83x10^{-10}$	-6.368	0.000178
gender	13.94	$< 2x10^{-16}$	11.02	$< 2x10^{-16}$
covidPos	-4.330	$4.47x10^{-16}$	-5.538	$1.36x10^{-8}$
householdincome	$6.755x10^{-5}$	0.00124	$7.352x10^{-5}$	0.000560
freelunch	-19.93	$< 2x10^{-16}$	-19.75	$< 2x10^{-16}$
numcomputers	$4.051x10^{-1}$	0.16365	$7.367x10^{-1}$	0.013787
I(mathscore * writingscore)	$-1.712x10^{-2}$	$1.68x10^{-11}$	$-1.524x10^{-2}$	$1.39x10^{-8}$
I(reading * writingscore)	$-1.96x10^{-2}$	$9.12x10^{-15}$	$-1.958x10^{-2}$	$9.37x10^{-14}$
I(readingscore * mathscore)	$-1.590x10^{-2}$	$8.23x10^{-10}$	$-1.569x10^{-2}$	$1.87x10^{-8}$
fathereduc	2.051	$1.30x10^{-5}$	2.640	$5.95x10^{-6}$
mothereduc	3.085	$5.63x10^{-8}$	2.323	$4.02x10^{-6}$
readingscore	3.426	$< 2x10^{-16}$	3.108	$< 2x10^{-16}$
writingscore	3.401	$< 2x10^{-16}$	3.054	$< 2x10^{-16}$
mathscore	3.193	$< 2x10^{-16}$	2.765	$< 2x10^{-16}$

R^2 (Presential)	Adj. R^2 (Presential)	Global Significance (Presential)	R^2 (Online)	Adj. R^2 (Online)	Global Significance (Online)
0.702	0.699	$< 2.2x10^{-16}$	0.6661	0.6627	$< 2.2x10^{-16}$

Table 2: Summary Statistics for Cross-Sectional data models

CONCLUSIONS 04

In the first phase, it was concluded that *timeperiod* variables were jointly significant (even though there was not a significant difference between the different moments within online and presential classes). For the online *timeperiod*, it is expected that the student's grades decrease almost 10 points, meaning that the online classes have a negative impact on student's learning.
In the second phase, it was concluded that the parameters were significantly different between presential and online periods. In the online period, *school* and *gender* were less important to determine the grades, while the *householdincome* became more important. The variable *numcomputers* (number of computers at home) was not significant in presential periods, however, it became significant in online ones, which meets the expectations.
The fact that the dataset is not real can be considered the main limitation for future utility of this project, however the same strategies of this project can be directly applied to a new dataset with real data.

REFERENCES

[1] Pandey, Mrinal & Sharma, Vivek. (2013). A Decision Tree Algorithm Pertaining to the Student Performance Analysis and Prediction. International Journal of Computer Applications. 61. 1-5. 10.5120/9985-4822.

[2] Castillo-Merino, D. and Serradell-López, E. (2014) “An analysis of the determinants of students' performance in e-learning,” Computers in Human Behavior, 30, pp. 476–484. Available at: https://doi.org/10.1016/j.chb.2013.06.020.

[3] COVID-19 Effect on Grades. (2021, April 23). Kaggle. https://www.kaggle.com/datasets/dylanbollard/covid19-effect-on-grades-constructed-dataset?select=COVID-19-Constructed-Dataset.xlsx

[4] Toyoda, T. (1974). Use of the Chow Test under Heteroscedasticity. Econometrica, 42(3), 601–608. https://doi.org/10.2307/1911796