Le Van

Coding Task

July 2024

1 Introduction

I'd like to start by thanking the Data Innovation Lab for giving me the opportunity to perform this task as part of the selection process for the predoc position. I'd love to be selected for the position as it will allow me to contribute to original economic research, be mentored by renowned professors, improve my coding and reasoning abilities and get closer to my Ph.D. dream. Here is the link to view this overleaf file.

2 My work

2.1 Table 5 replication

For this task, I tried my best to follow the regression model in the paper and the instructions provided. I, unfortunately, did not get the same coefficients though I've checked my work several times. My guess is there might be an issue with the 'collapse' command as I only selected variables that I thought are needed for the regression analysis.

2.2 Data retrieval

For this task, I tried using a few websites/API such as Google Books API which successfully produced some good results but it has a limit of 1000 queries per day so it ran into lots of error (429 specifically) shortly. I also experimented using either oclc numbers or book titles. It turned out that each has its own pros and cons. Oclc numbers are numerical so they are less likely prone to spelling errors than book titles; however, the pattern of oclc numbers need to be fixed. Eventually, I ran into a github repository that listed all sorts of free APIs including Gutendex - JSON web API for Project Gutenberg ebook metadata. I did some experiments and thankfully, the openai lending key and model were compatible. Initially, I did not limit the length for answer so I got a long answer for each question about genre such as: "The book titled "The Minor Mathers: A List of Their Works" likely falls within the genre of non-fiction, specifically

in the realms of literary criticism or bibliography. It seems to provide a catalog or examination of the works of the Minor Mathers." This is a detailed answer but since we are interested in the specific genre, it would require a lot of manual work to extract the genre from that answer (though manual work might also mean more accurate.) Thus, I made some changes to the code and now it only extracted the key word for genre.

2.3 Tables and Figures

	Main effect	
	log-OLS	LPM
Post_Scanned	-0.0086***	-0.0094***
	(0.0007975)	(0.0007433)
Book FE	Yes	Yes
Year-Location FE	Yes	Yes
N	792054	792054

Note: This table reports effects of digitization on loans at Harvard's libraries. Columns (1) and (2) provide baseline estimates from a zero-inflated Log-OLS estimation (1) and a linear probability model (2). All models include book and year-location fixed effects. Standard errors are in parentheses, clustered at the book level. * p < 0.10, ** p < 0.05, *** p < 0.01

Table 1: Effects on Loans (Replicated)

	Main effect	
	log-OLS	LPM
Post_Scanned	-0.0088***	-0.0095***
	(0.000803)	(0.000749)
Book FE	Yes	Yes
Year-Location FE	Yes	Yes
Genre FE	Yes	Yes
N	783612	783612

Note: This table reports effects of digitization on loans at Harvard's libraries. Columns (1) and (2) provide baseline estimates from a zero-inflated Log-OLS estimation (1) and a linear probability model (2). All models include book, year-location and genre fixed effects. Standard errors are in parentheses, clustered at the book level. * p < 0.10, ** p < 0.05, *** p < 0.01

Table 2: Effects on Loans with Genre Fixed Effect

The first two columns of the regression imply that digitization decreases loans on average by about 0.9 percent, and increases the probability that a book is checked out at Harvard at all in a year by 1 percentage points. We

see that the coefficient changes very little after adding fixed effects on genres, which indicates that the omitted variable bias was minimal. In other words, the unobserved factors captured by the genre fixed effects do not substantially correlate with the variable of interest - number of loans after digitization.