Berkeley-Haas Data Task

David Scolari

April 2022

Description of Data-Cleaning Approach

The goal of my data cleaning approach is to make it a balanced panel of book level loan counts between 2003 and 2011, such that there are a total of 88,006 observations (one per book) in each calendar year.

The main challenge that I found in doing this is that there is not a loan even for every unique book in each year. So to obtain the desired balanced panel, I need to implement "empty" loan events into each year to account for books that are not loaned in said year.

The main pipeline that I use for this task involve Stata's "collapse" and "reshape commands". This is perhaps not the best way to go about this, mainly because, as I will describe in detail, I am interested in the byproduct results of these commands and not the primary results. However, I was unable to find a more direct method of creating this result, and since this method does produce the desired panel, I proceed with it. In a research setting, I would consult colleagues and spend more time looking for a more direct methods of making this panel.

Collapsing the data by book will result in a column of summary statistics (the primary result) and a column of 88,006 unique book ids. The latter column is the vector we are after, and to it, I add a column for each year, which after reshaping, expand the unique cross-sectional book units into a panel that spans 9 years, 2003-2011. Each book is reflected once in each year. There are 792,054 observations in this matrix.

After creating this "blank" panel, I merge the raw data back onto it. The result will be the same as the original raw data with the addition of "empty" loan evens so that every unique book is considered in each year, even if not loaned. I then collapse again to sum loan events by book and year, making sure to preserve location and scanned variables. The result of this collapse is nearly the finished product, however, location and scanned variables do not carry over to the books that have 0 loans for a given year. I have a loop method that is described in detail in the comments of my code, however, I do want to mention that this loop relies on the variables yearscanned and location being invariant within cross-sectional units. This is true by construction for yearscanned—a book cannot be scanned again after it is already in the Google Books system. However, it seems possible for a physical book to be moved from Harvard Library location to another. I do some data exploration to convince myself that location does not vary within books. Given more time, I would prefer to double check this and perhaps check this with colleagues.

Table 5 Replication

The Table 5 replication results from estimating the following regression. In the first column, the outcome variable is zero-inflated log-loans. In the second column, the outcome is a binary indicating whether or not book i is loaned in year t.

$$y_{ti} = \beta_0 + \beta_1 PostScanned_{ti} + \sum_{l=1}^{k} \sum_{t=1}^{9} YearLocation_{lti} + \gamma_i \tag{1}$$

Table 1: Table 5 Replication

	(1) log-OLS	(2) LPM	(3) est3	(4) est4
Post-Scanned	-0.0511***	-0.0811***	-0.0613***	-0.0968***
	(0.00152)	(0.00119)	(0.00170)	(0.00133)
Book FE	Yes	Yes	Yes	Yes
Year-Location FE	Yes	No	Yes	No
N	792054	792054	792054	792054

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 2: TWFE vs. CS Estimators

V1	TWFE	CS	significance	sign_change
Estimated ATT	-0.051	-0.067	Less	No
V1	(0.001)	(0.005)		

Standard errors in parenthesis

Additional Results

In Table 2, I provide additional results using borrowstatus definitions. For the models summarized by Table 2, I use the same model specification used in the Table 5 replication, adding indicator variables for subgroups of interest. I choose to use the log-OLS model for this table. The difference in observations between models result from the number of subgroups of borrows I consider with each model.

Column 1 divides borrows into faculty and non-faculty. The partial effect of the scanned variable is positive for faculty and negative for non-faculty, perhaps indicating that improved search functionality of google books dominates the substitution effects for faculty borrowers. This might be due to faculty members, relative to non-faculty members, being highly adept at using the library search functions which existed before Google Books, making the improved search effect weak and leaving the substitution effect to dominate.

Column 2 compares master, doctorate, and undergraduate students. Here, the effect of scanning is found to be slightly positive for undergraduate students and negative for masters and doctoral students. Dynamics similar to those governing the faculty and non-faculty comparison may be responsible for this.

Column 3 compares in-building loans to inter-library loans. I expected to see a small effect negative of scanning for in-building loans as people who are browsing books in-building may not be using the search functions at all and positve inter-library library effects as these loans likely involve a search tool, so improved search functionality may benefit these types of loans a lot. However, the model finds small effects for both. This may be due to the fact that inter-library loans are made by borrowers adept at using library search functions in the pre-scanning period.

Table 3: Additional Results

	(1) log-OLS	(2) log-OLS	(3) log-OLS
Post-Scanned	-0.0293***	0.00935***	0.00207***
	(0.000951)	(0.00141)	(0.000112)
Faculty	-0.0624***		
	(0.000489)		
Faculty × Scanned	0.0377***		
	(0.00148)		
Doctorate Student		-0.00344***	
		(0.000150)	
Masters Student		-0.00393***	
		(0.000150)	
Doctorate × Scanned		-0.0333***	
		(0.00191)	
Masters × Scanned		-0.0149***	
		(0.00216)	
In-Building			0.0242***
			(0.000168)
In-Building × Scanned			0.000884*
			(0.000394)
Book FE	Yes	Yes	Yes
Year-Location FE	Yes	Yes	Yes
N	1584108	2376162	1584108

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$