Reproduction of Erin Hengel 2017

Chia-Wei, Chen [*]

April 24, 2022

Contents

Introduction

This project is to reproduce Erin Hengel (2017) on ”Publishing While Female”. The main work is to get all articles with abstracts [1] from five well-known economics journals and match each articles with a gender score.

Therefore scaping articles and authors’ genders are a crucial task of this job.

[*] r10323045@ntu.edu.tw

[1] Some are not explicitly abstracts. See further chapter

1 Current Progress

I mainly finish the following tasks.

1. Scraping abstract from AER, ECA, JPE, QJE, and RES [2].
2. For articles in JPE, some abstracts are provided explicitly, not in html text form, but in the first-page image. I extracted them via image processing.
3. Extracted Authors' genders from Erin Hengel's Website.
4. Construct a python file to convert all results into regression-ready format.

I will then describe in detail what I have done and what I am not able to complete. Note that the work is somehow messy and might be hard to tract. I will do my best to guide you through my processes.

1.1 Scraping

Table 1 shows the source of abstract for each journal:

Journal	Years	Source	Scraped by previous person
AER	After 1999	AEA web	✓
	1980 to 1999	JSTOR	
ECA	1950 to 2022	JSTOR	
JPE	1960 to 2022	Chicago Press Journals	
QJE	1980 to 2022	Oxford Academic	✓
REStud	1980 to 2022	Oxford Academic	✓

Table 1: Source and Years of scraping for each journal.

Several results are already provided hence I worked on and scraped only AER, ECA, JPE. Technical details are provided in future chapters. I simply scraped the "Abstract" element on the `html`. However, there are discrepancies with the results provided by Hengel (2017), see figure 1

The result I scraped is shown in table 2 in page 4. Table 2 shows several discrepancies that have to be checked manually, and I haven't tackle, and have no clue how, yet.

Non-technical Description of the Scraping Process

For each journal, I do the scraping in two steps.

1. Extract the list of articles from each issues within the time period. Save all articles into a single file, named `~_all_issue.csv`
2. Scrape the abstracts according to the source url.

This allows me to easily separate the process into several segments.[3]

[2] These are respectively, *The American Economic Review, Econometrica, Journal of Political Economy, Quarterly Journal of Economics*, and *The Review of Economic Studies*

[3] The website usually arranges its articles in the following hierarchy: 1. Issues within decades 2. Articles withing issues. 3. Abstract within the webpage for each articles. Therefore it is a good idea to first extract all articles and its urls from all issues, then scrape the articles from the list of urls.

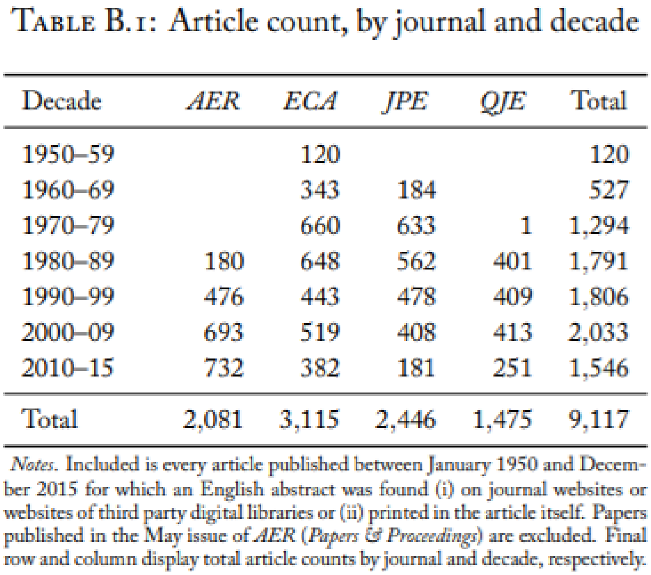

TABLE B.1: Article count, by journal and decade

Decade	*AER*	*ECA*	*JPE*	*QJE*	Total
1950–59		120			120
1960–69		343	184		527
1970–79		660	633	1	1,294
1980–89	180	648	562	401	1,791
1990–99	476	443	478	409	1,806
2000–09	693	519	408	413	2,033
2010–15	732	382	181	251	1,546
Total	2,081	3,115	2,446	1,475	9,117

Notes. Included is every article published between January 1950 and December 2015 for which an English abstract was found (i) on journal websites or websites of third party digital libraries or (ii) printed in the article itself. Papers published in the May issue of *AER* (*Papers & Proceedings*) are excluded. Final row and column display total article counts by journal and decade, respectively.

Figure 1: Numbers of articles with abstract provided By Hengel

AER

- Articles after 1999 are already scraped when I started the project; the original result is in the following folder:

  ```
  Already Scraped/American Economic Review
  ```

- I scraped those before 1999 because it is not provided in the official website of AER. I accessed those from JSTOR, and saved my raw results in

  ```
  AER/AER.csv
  ```

- Following Hengel (2017), articles in May and ”Comments”, ”Reply” etc., are neglected, and were not scraped in the first place.

- Discrepancies are shown in table2.

ECA

- Results are stored in

  ```
  ECA\ECA.csv
  ```

- To speed up scraping, I loaded all articles in JSTOR but skipped the ones with *Report of,Report on, Annual Reports,Criticism Invited.* I kept the orginal list of articles (before filtering) in

  ```
  ECA\ECA_all_issue.csv
  ```

  and kept those I skipped in

Table 2: Scraped and image processed result compared with Hengel (2017)

Decade	Scrape	Hengel	Note
		AER	
1980 - 1989	181	180	
1990 - 1999	512	476	
2000 - 2009	682	693	
2010 - 2015	733	732	
		ECA	
1950 - 1959	130	120	Non-english articles
1960 - 1969	344	343	
1970 - 1979	661	660	
1980 - 1989	648	648	
1990 - 1999	443	443	
2000 - 2009	520	520	
2010 - 2015	384	382	
		JPE	
1960 - 1969	0	184	No official abstract
1970 - 1979	448	633	OCR results added
1980 - 1989	559	562	OCR results added
1990 - 1999	478	478	
2000 - 2009	408	408	
2010 - 2015	181	181	
		QJE	
1980 - 1989	393	401	
1990 - 1999	409	409	
2000 - 2009	413	413	
2010 - 2015	251	251	

* For non-english articles, I excluded them when counting, but the article is still included in the combined raw data.

`ECA\ECA_skipped.csv`

- Ignore the other `csv` files. Those are for temporary use.
- I scraped all of them from JSTOR.
- Some are not in English. The discrepancies reduces to one or two after the filtering.
- The non-English articles are still included in the `combined/all_combined.csv`, but it is easy to filter out.
- Discrepancies are shown in table 2. Which does not differ much.

JPE

- Results with abstracts are stored in

`JPE/JPE.csv`

- Original list of all articles is saved in

`JPE/JPE_all_issue.csv`

- The source is Chicago Press Journal. Some articles have abstracts in the front-page image, but the text is not provided as web element that is available for copying.
- I handle some articles during 1970 - 1989 by image processing, which automatically extracted the abstract block and save them to the result file.
- Images and its result are stored in

`JPE\first_pages`

- It seems that Hengel did not only consider explicitly the abstract, but also some fist paragraphs of articles, when abstracts are not found.
- Manual work must be done to match the results of Hengel.

QJE

- Already scraped by previous work.
- I cleaned the format to match those I scraped, making it easier to merge.

REStud

- Already scraped by previous work.
- Hengel (2017) didn't provide her results.
- I also cleaned the format.

1.2 Gender

Gender detail is extracted from Hengel's website. The original data can be found in

`Authors/.`

Each folder in the `Authors/` directory is a list of readability scores of Economists, categorized by the first letter of its last name, as shown in figure 2

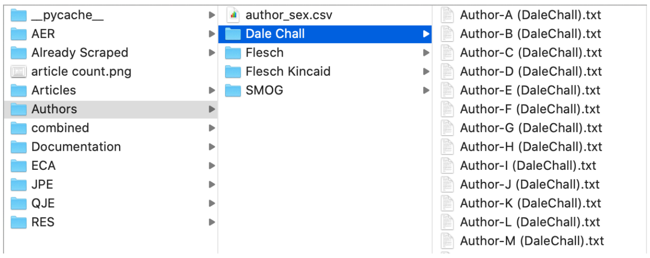

Figure 2: Where lists of authors and its information can be found

Author-Gender Detail

The first-hand gender information that was directly extracted from the website is stored in

`Authors\author_sex.csv`

Directly using his data has several difficulties when matching it to the authors scraped.

- First name is abbreviated
- Middle names are neglected
- Special characters in name is interchanged with some English letters, such as ¨ø" to "o"

To overcome this, for each author, I created several equivalent "names". Transformation rules are coded in the first section of `combined/transform_reg_data.ipynb`. The final result is saved in

`combined/sex.csv`

This file matched all authors appearing in the `all_combined.csv` and match its gender.

Note that there are still several exceptions that are not captured by this rule, and is left as blank. Also, authors that only appears in REStud wasn't provided in Hengel's website[4]. I excluded REStud atricles and save it in

`combined/sex_noRes.csv`

The descriptive statistics are summarized in table 3

Authors from	Male	Female	# of Missing Gender
Including REStud	6241	1971	1173
Without REStud	6101	1104	318

Table 3: Statistics of gender for journals including and excluding REStud

I recommend doing further corrections in gender manually instead of looking for other patterns.

1.3 Regression-ready data

I also came up with the program that turns the combined articles and the authors; information into a new data frame that is regression-ready. It is written in the second section[5] of the file

`combined/transform_reg_data.ipynb`

Named *Get the data to be ready for regression*

The procedure is as following

1. Load the combined data (From configuration)
2. Load the gender information. It should be a `json` format so that is can be turned into a `dict` type.
3. Run all the function definitions and apply them.

[4] Her website only show results for AER, ECA, JPE, QJE for demonstrations.

[5] Note that the first section is the author-gender transformation part, described in the previous section.

References

Hengel, E. (2017) “Publishing while Female. Are women held to higher standards? Evidence from peer review,” Cambridge Working Papers in Economics 1753, Faculty of Economics, University of Cambridge.