
Web-scraping for Social Science Research: A Case Study

Dr. Diarmuid McDonnell
UK Data Service



27 March 2020

UK Data Service



UK Data Service – New Forms of Data

Upcoming webinars:

- [Getting Data from the Internet](#)
- [Web-scraping for Social Science Research: Websites as a Source of Data](#)
- [Web-scraping for Social Science Research: APIs as a Source of Data](#)

Past webinars:

- [Introduction to agent-based modelling for social scientists](#)
- [Adding real world GIS and census data to agent-based modelling for social scientists](#)
- [Conducting experiments, recording output and analysing results of agent-based modelling for social scientists](#)

Case Study



Research Article

Incentivizing Regulatory Participation: Effectiveness of a Fundraising Levy

Alasdair C. Rutherford , Diarmuid McDonnell, Eddy Hogg

First published: 18 March 2020 | <https://doi.org/10.1111/puar.13176>

<https://doi.org/10.1111/puar.13176>

<https://github.com/a1asdair/paper-IncentivizingRegulatoryParticipation>

Table of Contents

1. What is web-scraping?
2. Why is it valuable for social science research?
3. What problem did web-scraping solve in our research?
4. How did we implement web-scraping techniques?
5. What results were we able to produce?
6. General reflections on web-scraping as a social science research method
7. Questions
8. Further learning and resources

What is web-scraping?

It is a computational technique for capturing information stored on a web page.

It is generally implemented using a programming script, although there are software applications that you can use.

It is relatively simple to implement using open-source programming languages e.g., Python, R.

Example



The image shows a browser window with the URL `data.charitycommission.gov.uk`. The page title is "Charity data download" with an "ALPHA" badge. The main heading is "Registered charities in England and Wales", followed by the text "Data download service provided by the [Charity Commission](#)". A section titled "Data extract, March 2020" contains two download options: "Charity register extract (Zip file, 135Mb)" and "Table build scripts (Zip file, 8Kb)". At the bottom, there are links for "Help", "Data definition", and "Built by the Charity Commission", along with a disclaimer: "All content is available under the [Open Government Licence v3.0](#), except where otherwise stated".

data.charitycommission.gov.uk

Charity data download **ALPHA**

Registered charities in England and Wales

Data download service provided by the [Charity Commission](#)

Data extract, March 2020

[Download](#) Charity register extract (Zip file, 135Mb)

[Download](#) Table build scripts (Zip file, 8Kb)

[Help](#) [Data definition](#) Built by the [Charity Commission](#)

All content is available under the [Open Government Licence v3.0](#), except where otherwise stated

Why collect data from the web?

Web pages can be an important source of publicly available information on social phenomena of interest.

Web pages can store a range of different data types including files, text, photos, videos, lists etc, all of which may be collected and marshalled for research purposes.

Once collected, data can be reshaped into a familiar format (tabular) and linked to other sources of social science data.

Research problem

The Fundraising Regulator for England and Wales is the statutory regulator of fundraising activities (e.g., charity shops, direct debit donations) by charities.

The regulator is partly funded by the charities and other organisations it oversees.

Organisations are not compelled to contribute to the cost of regulation, however the regulator expects certain types of organisations to pay a fee (£100k+ spend on fundraising).

Logic and skills

We begin with a web page that contains information we are interested in collecting. We need to **know** the following:

1. The location (i.e., URL or web address) where the web page can be accessed e.g., <https://www.fundraisingregulator.org.uk/directory>).
2. The location of the information we are interested in within the structure of the web page.

Then we need to **do** the following:

3. Request the web page using the URL.
4. Parse the structure of the web page so your programming language can work with its contents.
5. Extract the information we are interested in.
6. Write this information to a file for future use.

Implementing a web-scraping solution

We needed a programming script that will do the following:

- Iterate through a list of web pages (URLs) and extract charity number (unique id) and levy status.
- Perform this task on a routine basis (e.g., monthly).

We settled on a script written in Python, which is a general purpose, easy-to-learn, open-source programming language.

We saved and shared all of this work on a publicly available repository.

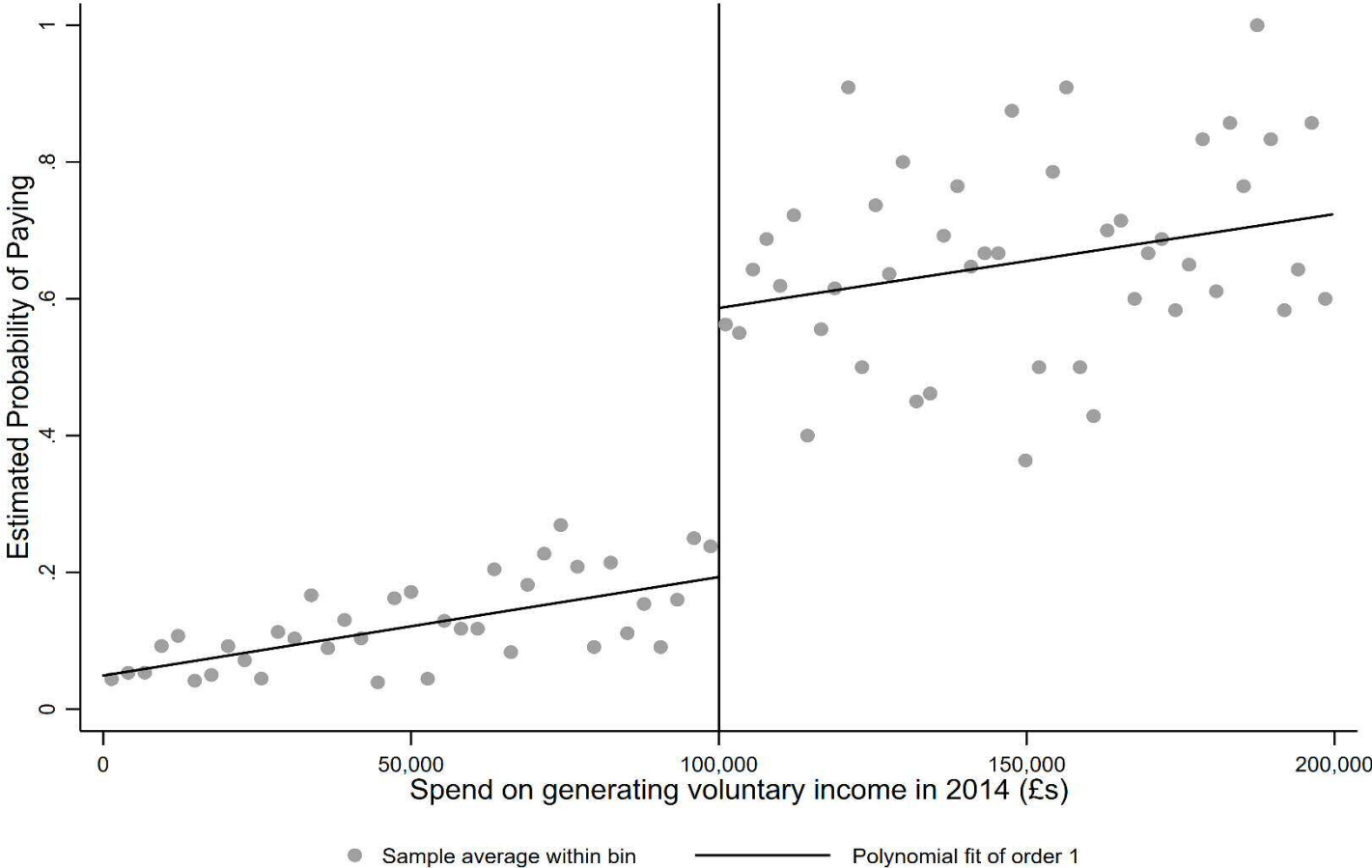
Results

The scraped data were linked to financial data to produce a sample of 4,147 charities.

We exploited the sharp threshold by fundraising expenditure (£100k) in order to make a causal estimate of the effect of the fundraising levy - **RDD**.

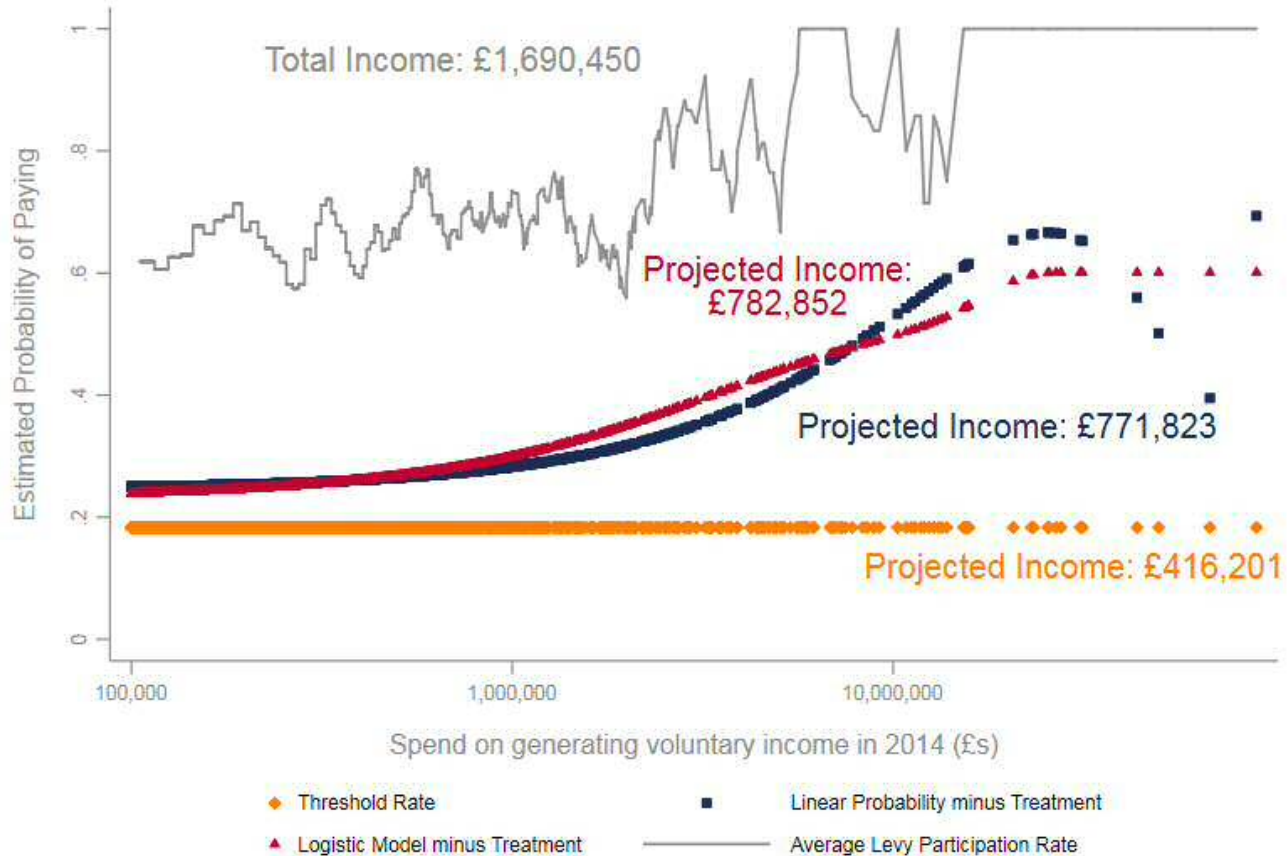
We conducted the analysis in Stata using the `rdrobust` package (Calonico et al. 2017).

Results



Source: Data from Charity Commission & Fundraising Regulator Produced: 25 Jul 2018

Results



Source: Data from Charity Commission & Fundraising Regulator Produced: 16 Nov 2017

Summary

Pros:

- Skill that's relatively easy to learn
- Easy to routinise/automate (important when data are continuously updated)
- Lots of public/charitable bodies in particular share data through websites (e.g. annual reports, statistics and figures)
- Can be easily formatted to permit data linkage

Cons:

- Ethical issues esp. around personal data
- Web page updates can break the script
- Blacklisted from requesting web page
- Requires good internet connectivity for extended periods of time

Questions

Dr. Diarmuid McDonnell

diarmuid.mcdonnell@manchester.ac.uk

 @DiarmuidMc



Further resources and help

Repository: <https://github.com/UKDataServiceOpen>

Youtube: <https://www.youtube.com/user/UKDATASERVICE>

Help: ukdataservice.ac.uk/help/

Subscribe to UK Data Service news at
<https://www.jiscmail.ac.uk>

 @UKDataService

 UKDataService

