UK Data Service

Web-scraping for Social Science Research: A Case Study

The webinar will begin at 1pm

You now have a menu in the top right corner of your screen.

The red button with a white arrow allows you to expand and contract the webinar menu, in which you can write questions/comments.

Feel free to type questions as we go, we will answer as many as we can at the end

We can't hear you.

Web-scraping for Social Science Research: A Case Study

Dr. Diarmuid McDonnell
UK Data Service

27 March 2020

UK Data Service

UK Data Service

Can you hear us?

UK Data Service

Can you hear us?

If not:

- Check your volume, and that your speaker/headset is plugged in
- Click on audio to change to listening via phone.
- We are recording this webinar – we plan to put it on our website

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

UK Data Service

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

UK Data Service

Table of Contents

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

UK Data Service

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**.

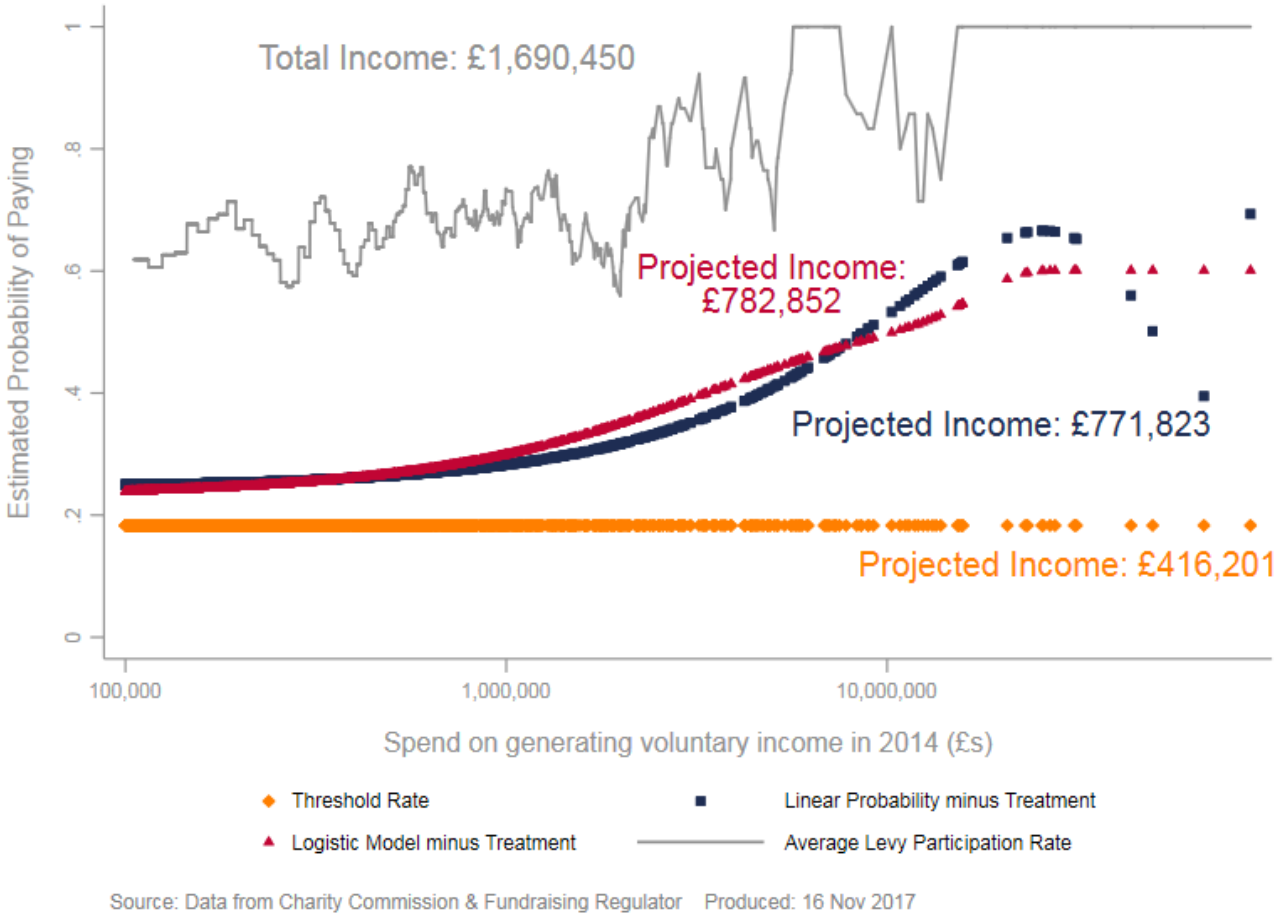
We conducted the analysis in Stata using the `rdrobust` package (Calonico et al. 2017).

Results

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

UK Data Service

Results



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

UK Data Service

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

UK Data Service