Research Methods in Political Science I

5. Collecting Data

Yuki Yanai

November 4, 2015

KOBE UNIVERSITY

Today's Menu

1. Data Sets
   - Introduction
   - What kinds of data sets do we need?
   - Where to get data?
2. Getting Data form the Web
   - Method 1: Copy and Paste
   - Method 2: Using OutWit Hub
3. Web Scraping
   - Introduction
   - Web Scraping with Python (optional)

Data

- Data analyses: Not doable without data!

- **What kinds** of data do we need ?
- **How** should we obtain data?

Rectangular Data

- We usually analyze rectangular data sets
- Each row represents an observed unit
- E.g., each candidate in a row in the figure
- Each column represents a variable
- Each cell has a value (numeric or character)

	A	B	C	D	E	F
1	year	ku	kun	party	name	age
2	1996	aichi	1	1000	KAWAMURA, TAKASHI	
3	1996	aichi	1	800	IMAEDA, NORIO	
4	1996	aichi	1	1001	SATO, TAISUKE	
5	1996	aichi	1	305	IWANAKA, MIHOKO	
6	1996	aichi	1	1014	ITO, MASAKO	
7	1996	aichi	1	1038	YAMADA, HIROSHIB	
8	1996	aichi	1	1	ASANO, KOSETSU	
9	1996	aichi	2	1000	AOKI, HIROYUKI	
10	1996	aichi	2	800	TANABE, HIROO	
11	1996	aichi	2	1001	FURUKAWA, MOTOHISA	
12	1996	aichi	2	305	ISHIYAMA, JYUNICHI	
13	1996	aichi	2	1003	FUJIWARA, MICHIKO	
14	1996	aichi	2	1014	ISHIKAWA, KAZUMI	
15	1996	aichi	2	1	MURAMATSU, YOICHI	
16	1996	aichi	2	1038	YAMAZAKI, YOSHIAKI	
17	1996	aichi	3	1000	YOSHIDA, YUKIHIRO	
18	1996	aichi	3	800	KATAOKA, TAKESHI	
19	1996	aichi	3	1001	KONDO, SHOICHIA	
20	1996	aichi	3	305	YANAGIDA, SAEKO	
21	1996	aichi	3	1038	NAKANO, YOKO	
22	1996	aichi	3	1014	OGAWA, OSAMU	
23	1996	aichi	3	1	ATOJI, MASAO	
24	1996	aichi	4	1000	MISAWA, JUN	
25	1996	aichi	4	800	TSUKAMOTO, SABURO	
26	1996	aichi	4	305	SEKO, YUKIKO	
27	1996	aichi	4	1001	TAKAGI, HIROSHI	
28	1996	aichi	4	1038	ITO, TAKAYOSHI	
29	1996	aichi	4	1014	SHIOKAWA, CHIKANAO	

Figure: hr96-09.csv

CSV Files

CSV: Comma Separated Values

- Text file
- Versatile
  - Can be edited by spread sheet applications such as Calc or Excel
  - Any data-analysis package can read CSV
- **Always keep your data sets in CSV!**
  - **Reproducibility**: for others and for future use

Example of CSV: hr96-09.csv（1）

```
gyear,ku,kun,party,name,age,status,nocand,wl,rank,previous,vote,voteshare,eligible,turnout,exp
1996,aichi,1,1000,"KAWAMURA, TAKASHI",47,2,7,1,1,2,66876,40,346774,49.22,9828097
1996,aichi,1,800,"IMAEDA, NORIO",72,3,7,0,2,3,42969,25.7,346774,49.22,9311555
1996,aichi,1,1001,"SATO, TAISUKE",53,2,7,0,3,2,33503,20.1,346774,49.22,9231284
1996,aichi,1,305,"IWANAKA, MIHOKO",43,1,7,0,4,0,22209,13.3,346774,49.22,2177203
1996,aichi,1,1014,"ITO, MASAKO",51,1,7,0,5,0,616,0.4,346774,49.22,.
1996,aichi,1,1038,"YAMADA, HIROSHIB",51,1,7,0,6,0,566,0.3,346774,49.22,.
1996,aichi,1,1,"ASANO, KOSETSU",45,1,7,0,7,0,312,0.2,346774,49.22,.
1996,aichi,2,1000,"AOKI, HIROYUKI",51,2,8,1,1,2,56101,32.9,338310,51.79,12940178
1996,aichi,2,800,"TANABE, HIROO",71,3,8,0,2,1,44938,26.4,338310,51.79,16512426
1996,aichi,2,1001,"FURUKAWA, MOTOHISA",30,1,8,2,3,1,43804,25.7,338310,51.79,11435567
1996,aichi,2,305,"ISHIYAMA, JYUNICHI",31,1,8,0,4,0,21337,12.5,338310,51.79,2128510
1996,aichi,2,1003,"FUJIWARA, MICHIKO",44,1,8,0,5,0,2670,1.6,338310,51.79,3270533
1996,aichi,2,1014,"ISHIKAWA, KAZUMI",61,1,8,0,6,0,701,0.4,338310,51.79,.
1996,aichi,2,1,"MURAMATSU, YOICHI",47,1,8,0,7,0,418,0.2,338310,51.79,.
1996,aichi,2,1038,"YAMAZAKI, YOSHIAKI",43,1,8,0,8,0,348,0.2,338310,51.79,.
1996,aichi,3,1000,"YOSHIDA, YUKIHIRO",35,1,7,1,1,1,52478,32.3,331808,50.38,11245219
1996,aichi,3,800,"KATAOKA, TAKESHI",46,2,7,0,2,3,43884,27,331808,50.38,5365436
1996,aichi,3,1001,"KONDO, SHOICHIA",38,1,7,2,3,1,38351,23.6,331808,50.38,11767342
1996,aichi,3,305,"YANAGIDA, SAEKO",50,1,7,0,4,0,26225,16.1,331808,50.38,2110540
1996,aichi,3,1038,"NAKANO, YOKO",54,1,7,0,5,0,773,0.5,331808,50.38,.
1996,aichi,3,1014,"OGAWA, OSAMU",35,1,7,0,6,0,722,0.4,331808,50.38,.
```

Figure: A CSV file opened in a text editor

Example of CSV: hr96-09.csv（2）

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	
1	year	ku	kun	party	name	age	status	nocand	wl	rank	previous	vote	voteshare	eligible	turnout	e
2	1996	aichi	1	1000	KAWAMURA, TAKASHI	47	2	7	1	1	2	66876	40	346774	49.22	
3	1996	aichi	1	800	IMAEDA, NORIO	72	3	7	0	2	3	42969	25.7	346774	49.22	
4	1996	aichi	1	1001	SATO, TAISUKE	53	2	7	0	3	2	33503	20.1	346774	49.22	
5	1996	aichi	1	305	IWANAKA, MIHOKO	43	1	7	0	4	0	22209	13.3	346774	49.22	
6	1996	aichi	1	1014	ITO, MASAKO	51	1	7	0	5	0	616	0.4	346774	49.22	.
7	1996	aichi	1	1038	YAMADA, HIROSHIB	51	1	7	0	6	0	566	0.3	346774	49.22	.
8	1996	aichi	1	1	ASANO, KOSETSU	45	1	7	0	7	0	312	0.2	346774	49.22	.
9	1996	aichi	2	1000	AOKI, HIROYUKI	51	2	8	1	1	2	56101	32.9	338310	51.79	1
10	1996	aichi	2	800	TANABE, HIROO	71	3	8	0	2	1	44938	26.4	338310	51.79	1
11	1996	aichi	2	1001	FURUKAWA, MOTOHISA	30	1	8	2	3	1	43804	25.7	338310	51.79	1
12	1996	aichi	2	305	ISHIYAMA, JYUNICHI	31	1	8	0	4	0	21337	12.5	338310	51.79	
13	1996	aichi	2	1003	FUJIWARA, MICHIKO	44	1	8	0	5	0	2670	1.6	338310	51.79	
14	1996	aichi	2	1014	ISHIKAWA, KAZUMI	61	1	8	0	6	0	701	0.4	338310	51.79	.
15	1996	aichi	2	1	MURAMATSU, YOICHI	47	1	8	0	7	0	418	0.2	338310	51.79	.
16	1996	aichi	2	1038	YAMAZAKI, YOSHIAKI	43	1	8	0	8	0	348	0.2	338310	51.79	.
17	1996	aichi	3	1000	YOSHIDA, YUKIHIRO	35	1	7	1	1	1	52478	32.3	331808	50.38	1
18	1996	aichi	3	800	KATAOKA, TAKESHI	46	2	7	0	2	3	43884	27	331808	50.38	
19	1996	aichi	3	1001	KONDO, SHOICHIA	38	1	7	2	3	1	38351	23.6	331808	50.38	1
20	1996	aichi	3	305	YANAGIDA, SAEKO	50	1	7	0	4	0	26225	16.1	331808	50.38	
21	1996	aichi	3	1038	NAKANO, YOKO	54	1	7	0	5	0	773	0.5	331808	50.38	.
22	1996	aichi	3	1014	OGAWA, OSAMU	35	1	7	0	6	0	722	0.4	331808	50.38	.
23	1996	aichi	3	1	ATOJI, MASAO	43	1	7	0	7	0	246	0.2	331808	50.38	.
24	1996	aichi	4	1000	MISAWA, JUN	44	1	6	1	1	1	57361	35.7	315704	51.95	1
25	1996	aichi	4	800	TSUKAMOTO, SABURO	69	3	6	0	2	10	48209	30	315704	51.95	1
26	1996	aichi	4	305	SEKO, YUKIKO	49	1	6	2	3	1	30976	19.3	315704	51.95	
27	1996	aichi	4	1001	TAKAGI, HIROSHI	43	1	6	0	4	0	23411	14.6	315704	51.95	
28	1996	aichi	4	1038	ITO, TAKAYOSHI	61	1	6	0	5	0	348	0.2	315704	51.95	.
29	1996	aichi	4	1014	SHIOKAWA, CHIKANAO	40	1	6	0	6	0	243	0.2	315704	51.95	.
30	1996	aichi	5	1001	AKAMATSU, HIROTAKA	48	2	7	1	1	3	48648	30.9	319846	50.27	1
31	1996	aichi	5	800	KIMURA, TAKAHIDE	41	1	7	2	2	1	46485	29.5	319846	50.27	

Figure: A CSV file opened in a spread sheet

Internet（1）

When data sets are available

- Public institutions' websites
  - World Bank
  - OECD
  - Statistics Japan
  - etc.
- Websites of researchers and universities
  - Polity IV Project
  - Global Election Database (by Dawn Brancati)
  - etc.
- Open data archives
  - Dataverse
  - ICPSR
  - SSJ
  - etc.

Internet (2)

Data are available, but not as rectangular data sets

- Manually type numbers
- Copy the content and past it onto a spreadsheet
- Use OutWit Hub
- **Web scrauping** by R (or Python)

Visit Libraries!

- Electronic data kept in (out-of-data) media such as CD-ROM
- Access to online data bases
- Data printed in books
  - Manually type the data
  - Scan $\rightarrow$ OCR $\rightarrow$ Scrape! (R or Python)

Purchasing Data

- Some data sets are sold
- They are usually expensive: not a practical choice for a student

↓

- Check if the library owns the data set
- If not, ask the library to buy it

Building Your Original Data Sets

- Collect data by surveys, observations, or experiments
- Gather information by reading data sources
- Note: Data collecting process must be reproducible too
  - Record everything including data sources (Sensitive information can be masked when you publish the data set. You must manage such information carefully, though.)
  - Set the coding rulues **a priori**, and **write them down in a document**

When copy-and-paste is acceptable

When the information we need is provided in a table in a single web page

- Copy the table (Cmd + c or Ctrl + c)
- Paste it on to a spreadsheet (Cmd + v or Ctrl + v)
- If you get the rectangular data set, you're good to go
- Might need some tweaks

When copy-and-paste doesn't work (1)

Some tables are not properly copied

- The table contains a lot of irrelevant information
- Many variables are crammed into a single column when you paste the table into a spreadsheet
- Copy doesn't work in the first place
- **What should we do?**
- Use OutWit Hub (Free!!!)

When copy-and-paste doesn't work (2)

The information we need is scattered across multiple webpages

- Visit each page and use c-&-p or OutWit Hub (lame...)
- What if you have to visit 100 pages?
  - Use OutWit Hub Pro (not free)
  - **Scrape the web** (recommended)

What Is Web Scraping?

Web scraping: method to extract information from the webpages

1. Find a website that contains the information you need
2. Find the pages that have the info within the website
3. Specify where in the pages the information exists by HTML tags
4. Extract the information by R or Python
5. Pull the gathered information together and make a data set

How far you automate the process depends on your skill and purpose

See the course website for some examples

What Is Python?

Programming language

- Script language
- Accept object-oriented programming, imperative programming, functional programming, procedural programming, etc.
- Handle Japanese (and other non-Western) characters (Unicode)
- Run on Linux, Mac, and Windows

Installing Python

- Go to http://www.python.org/
- Click “Download” on top and download Python 3.5.x
- Choose an appropriate installer for your environment
- Follow the instructions

Note: if you use homebrew, you should install Python by homebrew. Google for more information.

Setting the PATH

Set the PATH so that you can run Python in any directory

- On Windows: http://docs.python-guide.org/en/latest/starting/install/win/
- On Mac: http://docs.python-guide.org/en/latest/starting/install/osx/

Installing ActiveTCL 8.5.16.0

This is (probably) necessary for Mac only

- Visit ActiveState's website http://www.activestate.com/activetcl/downloads
- From "Download Tcl", choose 8.5.16.0
- Follow the instructions to install the package

Installing pip

- Visit https://pip.pypa.io/en/latest/installing.html
- From “Install pip”, download `get-pip.py`
- In Terminal (or Command Prompt on Windows), type:

```
python get-pip.py
```

(add path to get-pip.py if necessary)

Installing Beautiful Soup

Beautiful Soup:

- Python library for web scraping
- HTML Parser
- To install, on Terminal (or Command Prompt), type:
  `pip install beautifulsoup4`

You can use (`pip install`) to install other python libraries

Next Class

Linear Regression（1）

- Review OLS
- Calculate OLSE with R
- Presenting Regression Results