政治学方法論 I

第 5 回：データの効率的な集め方

矢内 勇生

神戸大学　法学部/法学研究科

2014 年 10 月 29 日

今日の内容

1. データセット
   - イントロダクション
   - どのようなデータセットが必要か
   - どこからデータを手に入れるか
2. ウェブサイトからのデータ入手法
   - 方法 1：コピペする
   - 方法 2：OutWit Hub を使う
3. Python によるウェブスクレイピング
   - 導入

データ

- データ分析：データがないとできない！

- どのようなデータを用意すべき？
- データはどうやって手に入れる？

長方形データ (rectangular data)

- 分析に使うデータセットは基本的に長方形
- 各行は、分析単位（右の例では候補者）
- ただし、第 1 行には変数名を書くのが基本
- 各列は、変数
- 各セルに情報（数値、文字列）を入力する

	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 ファイル

CSV: Comma Separated Values

- テキストファイル
- 汎用性が高い
  - スプレッドシートソフト（Excel 等）で編集可能
  - どんな統計分析ソフトでも開ける
- データは常に CSV 形式で保存しておくべき
  - 再現性の確保：他人の使用、将来の使用に備える

CSV ファイルの例：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: テキストエディタで開いたとき

CSV ファイルの例：hr96-09.csv（2）

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

Figure: Excel で開いたとき

インターネットで手に入れる（1）

データセットとして手に入る場合

- 公的機関のウェブサイト
  - World Bank
  - OECD
  - 総務省統計局
  - etc.
- 研究者・研究機関のウェブサイト
  - Polity IV Project
  - Global Election Database (by Dawn Brancati)
  - etc.
- データアーカイブ
  - Dataverse
  - ICPSR
  - SSJ データアーカイブ
  - etc.

インターネットで手に入れる（2）

データはあるが、そのままでは使えない場合

- 手入力
- スプレッドシートにコピペ
- OutWit Hub をつかって CSV に保存
- Python でウェブスクレイピング

図書館で手に入れる

- CD-ROM 等の電子資料
- データベースへのアクセス
- 紙媒体
  - 手入力
  - スキャン → OCR → Python

データを買う

- 新聞社などが売っているデータを買うこともできる
- 高額なものが多い：一般の学生にとってはあまり現実的ではない

↓

- 図書館が購入していないか調べる
- なければ図書館に購入依頼を出す

自分で作る

- 調査、観察などによって自分でデータを集める
- 新聞記事などを自分でデータ化する
- 注意：再現性の確保に努める
  - 情報源はすべて記録する（個人情報などでも公開するか否かは後で考えればよい。ただし厳重な管理が必要）
  - コーディングルールを事前に決め、それを文書として記録しておく

コピペできるとき

欲しい情報がひとつのウェブページに表として掲載されているとき

- 表をコピー (Cmd + c or Ctrl + c) する
- 新規スプレッドシート (Excel 等）にペースト (Cmd + v or Ctrl + v) する
- 長方形データセットができれば OK
- 多少の微調整は必要かも

コピペできないときは・・・

欲しい情報がひとつのウェブページに表として掲載されていても、コピペできないときがある

- 欲しいデータ以外の余計な情報が（大量に）コピーされる
- スプレッドシートにペーストすると、複数の変数がひとつの行にまとめて入れられてしまう
- そもそもうまくコピーできない
- どうする？

→ ソフト (OutWit Hub [無料!!!]）を使う

コピペできないときは・・・

欲しい情報がひとつのウェブページにまとまっていないとき

- 複数のページをひとつひとつ訪問し、コピペまたは OutWit Hub で解決する
- 訪問すべきページがたくさんあったらどうする？
  - 有料版 (OutWit Hub Pro) を使う
  - Python を使う! （推奨）

ウェブスクレイピング

Web scraping：ウェブサイトから情報を抽出する方法

1. 必要な情報が掲載されているウェブサイトを見つける
2. ウェブサイトにアクセスし、情報がある「ページ」を見つける
3. HTML タグなどを手掛かりに、必要な情報がある場所を特定する
4. 必要な情報を抜き取る
5. 抜き取った情報を分析可能なデータセット形式に整形する

どこからどこまで自動化するかは対処する問題と技術次第：

比較的簡単な Python スクリプトで、ある程度自動化できる

Python とは何か

プログラミング言語

- スクリプト言語
- オブジェクト指向、命令型、関数型、手続き型などに対応
- 標準で日本語 (Unicode) が使用できる
- Mac, Linux, Windows 等で利用可能

Python のインストール

- http://www.python.jp/ のサイトに移動
- 左側にあるメニューから、「Pyhon 2.7.8, ダウンロード」を選択する
- 自分の環境に合ったインストーラをダウンロード
- 指示に従ってインストールする

注：homebrew を使えるなら、そちらでインストールしたほうがよい。

参考：http://qiita.com/tetsuya/items/f9a01d6bdea9639aff26

PATH の設定

どのフォルダからも Python を実行できるよう、PATH を設定したほうがよい

- Windows の場合：
  http://www.pythonweb.jp/install/setup/index1.html を参照
- Mac の場合：
  http://www.pythonweb.jp/install/setup/index1.html を参考に、パスに /usr/bin を追加する

ActiveTCL 8.5.16.0 のインストール

Mac の場合（Windows は必要なし？）

- ActiveState のサイト http://www.activestate.com/activetcl/downloads に移動
- “Download Tcl” のメニューから 8.5.16.0 を選ぶ
- 指示に従ってインストールする

pip のインストール

- pip のサイト https://pip.pypa.io/en/latest/installing.html に移動
- “Install pip” の `get-pip.py` をダウンロード
- ターミナル（コマンドプロンプト）で

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

  と、打つ（get-pip.py にパスが必要なときは足す）

Beautiful Soup のインストール

Beautiful Soup:

- スクレイピングに役立つ Python ライブラリ
- HTML のタグを利用した情報の選り分けが可能
- インストール：ターミナル（コマンドプロンプト）で

```
pip install beautifulsoup4
```

と、打つ

他の Python ライブラリをインストールするときも、同様の方法 (`pip install`) が使える

来週の内容

線形回帰分析（1）

- OLS の復習
- R による OLSE の計算
- R で得た分析結果の提示法