The Stata Journal

The *Stata Journal* publishes reviewed papers together with shorter notes or comments, regular columns, book reviews, and other material of interest to Stata users. Examples of the types of papers include 1) expository papers that link the use of Stata commands or programs to associated principles, such as those that will serve as tutorials for users first encountering a new field of statistics or a major new technique; 2) papers that go "beyond the Stata manual" in explaining key features or uses of Stata that are of interest to intermediate or advanced users of Stata; 3) papers that discuss new commands or Stata programs of interest either to a wide spectrum of users (e.g., in data management or graphics) or to some large segment of Stata users (e.g., in survey statistics, survival analysis, panel analysis, or limited dependent variable modeling); 4) papers analyzing the statistical properties of new or existing estimators and tests in Stata; 5) papers that could be of interest or usefulness to researchers, especially in fields that are of practical importance but are not often included in texts or other journals, such as the use of Stata in managing datasets, especially large datasets, with advice from hard-won experience; and 6) papers of interest to those who teach, including Stata with topics such as extended examples of techniques and interpretation of results, simulations of statistical concepts, and overviews of subject areas.

The *Stata Journal* is indexed and abstracted by *CompuMath Citation Index*, *Current Contents/Social and Behavioral Sciences*, *RePEc: Research Papers in Economics*, *Science Citation Index Expanded* (also known as *SciSearch*, *Scopus*, and *Social Sciences Citation Index*.

For more information on the *Stata Journal*, including information for authors, see the webpage

http://www.stata-journal.com

Subscriptions are available from StataCorp, 4905 Lakeway Drive, College Station, Texas 77845, telephone 979-696-4600 or 800-STATA-PC, fax 979-696-4601, or online at

http://www.stata.com/bookstore/sj.html

Subscription rates listed below include both a printed and an electronic copy unless otherwise mentioned.

U.S. and Canada		Elsewhere	
1-year subscription	$ 79	1-year subscription	$115
2-year subscription	$155	2-year subscription	$225
3-year subscription	$225	3-year subscription	$329
3-year subscription (electronic only)	$210	3-year subscription (electronic only)	$210
1-year student subscription	$ 48	1-year student subscription	$ 79
1-year university library subscription	$ 99	1-year university library subscription	$135
2-year university library subscription	$195	2-year university library subscription	$265
3-year university library subscription	$289	3-year university library subscription	$395
1-year institutional subscription	$225	1-year institutional subscription	$259
2-year institutional subscription	$445	2-year institutional subscription	$510
3-year institutional subscription	$650	3-year institutional subscription	$750

Back issues of the *Stata Journal* may be ordered online at

http://www.stata.com/bookstore/sjj.html

Individual articles three or more years old may be accessed online without charge. More recent articles may be ordered online.

http://www.stata-journal.com/archives.html

The *Stata Journal* is published quarterly by the Stata Press, College Station, Texas, USA.

Address changes should be sent to the *Stata Journal*, StataCorp, 4905 Lakeway Drive, College Station, TX 77845, USA, or emailed to sj@stata.com.

The *Stata Journal*, electronic version (ISSN 1536-8734) is a publication of Stata Press. Stata, Stata, Stata Press, Mata, Mata, and NetCourse are registered trademarks of StataCorp LP.

The Stata Journal (2012)
12, Number 3, pp. 454–460

Importing presidential approval poll results

Mehmet F. Dicle
Loyola University New Orleans
New Orleans, LA
mfdicle@loyno.edu

Betul Dicle
New Orleans, LA
betuldicle@hotmail.com

Abstract. The American Presidency Project (http://www.presidency.ucsb.edu) provides presidential job approval poll results. These data are available for each U.S. president since President Franklin D. Roosevelt and for all the job approval polls conducted since his presidency. In this article, we propose the Stata command `approval`, which downloads these approval poll results in their original format, an HTML table. `approval` then parses the HTML table and prepares the data as a usable Stata dataset.

Keywords: dm0064, approval, presidential job approval, presidential popularity, U.S. presidents, parse HTML

1 Introduction

The American Presidency Project provides a wide range of valuable data related to the U.S. presidents. Among these publicly available data are presidential job approval poll results, compiled by Gerhard Peters using the Gallup Poll. These data are available for each U.S. president since President Franklin D. Roosevelt and for all the job approval polls conducted since his presidency. You can access the data in HTML format through The American Presidency Project website (http://www.presidency.ucsb.edu). You can copy and paste the poll results into a text editor for further editing before the data are used by Stata.

We propose the Stata command `approval` to automate the process of accessing and parsing the approval data, which are available for each president separately. With `approval`, poll results are accessed, downloaded as HTML, and parsed. The result is a dataset usable in Stata. With `approval`, poll results may be processed either for one U.S. president or for multiple presidents. If multiple presidents are preferred, then the data are appended and the presidency number may be used as the panel variable.

2 The approval command

The presidential job approval poll results are provided through the following website: http://www.presidency.ucsb.edu/data/popularity.php?pres=44. The number that you enter in the URL corresponds to a particular U.S. president, in this case, President Barack Obama. Through the URL, a list of HTML tables is provided. Only one of these tables is related to the presidential job approval poll results.

As its first step, `approval` fetches the URL as a string variable. After the web content is retrieved, each table within the table HTML tags (<table>) is parsed into a string vector. Because the table with the first column and first row content that is equal to "President" belongs to the presidential job approval poll results, the corresponding vector cell is kept and the others are discarded. The vector cell that contains the data is then assigned to a string, and all end-of-row HTML tags (</tr>) are replaced with carriage return [char(13)].

Up to this point, `approval` uses Mata code. The resulting string variable is tokenized by carriage returns, transposed, and transferred to Stata as a string variable. The final processing with Stata splits each observation (each table row of the data) using the end-of-column HTML tags (</td>). In the resulting data, columns of the original table are the variables, and rows of the original table are the observations. Two additional variables are generated: `president`, which contains the name of the president, and `president2`, which contains the presidency number of the president. All variables are formatted to their original formats: string for `president`, float for `president2`, byte for `approving/disapproving/unsure`, and float for `startdate/enddate`.

2.1 Important Mata functions used in the approval code

Table 1 provides the Mata functions used in the `approval` code. These functions are for general parsing purposes and can be used in creating other Stata commands that parse HTML code.

Table 1. Important Mata functions used in the `approval` code

Task	Code or function

Get HTML source code from WWW	`string file_get_contents (string scalar raw)`
	`{`
	`  fh = fopen(raw, "r")`
	`  raw=""`
	`  while ((line=fget(fh))!=J(0,0,"")) {`
	`  raw=raw+line`
	`  }`
	`  fclose(fh)`
	`  return (raw)`
	`}`

Strip common HTML tags	`string strip_tags (string scalar raw)`
	`{`
	`  tags = ("tr", "TR", "td", "TD", "strong", "STRONG",`
	`  "/strong", "/STRONG", "span", "SPAN", "/span",`
	`  "/SPAN", "img", "IMG", "/img", "/IMG", "br",`
	`  "BR", "!-", "table", "TABLE", "/table", "/TABLE")`
	`  for (j=1; j<=cols(tags); j++) {`
	`  tag = tags[j]`
	`  while (strpos(raw, "<" + tag)) {`
	`  bas_pos = strpos(raw, "<" + tag)`
	`  bas_txt = substr (raw, 1, bas_pos - 1)`
	`  son_txt = substr (raw, bas_pos, .)`
	`  bas_pos2 = strpos(son_txt, ">")`
	`  son_txt = substr (son_txt, bas_pos2 + 1, .)`
	`  raw = bas_txt + son_txt`
	`  }`
	`  }`
	`  return (raw)`
	`}`

Strip specific HTML tags	`string remove_tags (string scalar raw, string scalar tag)`
	`{`
	`  while (strpos(strlower(raw), "<" + tag)) {`
	`  bas_pos = strpos(strlower(raw), "<" + tag)`
	`  bas_txt = substr (raw, 1, bas_pos - 1)`
	`  son_txt = substr (raw, bas_pos, .)`
	`  bas_pos2 = strpos(strlower(son_txt), "</" + tag +`
	`  ">") + 3 + strlen(tag)`
	`  son_txt = substr (son_txt, bas_pos2 + 1, .)`
	`  raw = bas_txt + son_txt`
	`  }`
	`  return (raw)`
	`}`

Remove unnecessary white space	`string remove_space (string scalar raw)`
	`{`
	`  while (strpos(raw, "  ")) {`
	`  raw = subinstr(raw, "  ", " ")`
	`  }`
	`  return (raw)`
	`}`

2.2 Syntax

`approval, president(`*numlist*`)` [`save(`*filename*`) timeseries`]

2.3 Options

`president(`*numlist*`)` is the list of U.S. presidents' presidency numbers. The list may contain only one president or multiple presidents. `president()` is required. The name of the president will become the content of the variable `president`, which will be based on the presidency number provided. The presidency number will become the content of the variable `president2`. *numlist* must be greater than 31. Presidential numbers are as follows:

Franklin D. Roosevelt is the 32nd president
Harry S. Truman is the 33rd president
Dwight D. Eisenhower is the 34th president
John F. Kennedy is the 35th president
Lyndon B. Johnson is the 36th president
Richard Nixon is the 37th president
Gerald R. Ford is the 38th president
Jimmy Carter is the 39th president
Ronald Reagan is the 40th president
George Bush is the 41st president
William J. Clinton is the 42nd president
George W. Bush is the 43rd president
Barack Obama is the 44th president

`save(`*filename*`)` is the output filename. A Stata data file is created in the current working directory.

`timeseries` converts the data into a time series. If a poll starts on April 23 and ends on April 27, then the days between April 23 and April 27 are filled in with `tsfill` (not `tsfill, force` across presidents though). Thus, for instance, April 24 will have the same approval rating as April 23, etc.

3 Using approval to import presidential job approval poll results

▷ Example

A single U.S. president's job approval poll results: In this example, job approval poll results for President Barack Obama, the 44th U.S. president, are downloaded and parsed.

```
. approval, president(44)
Poll results for President Barack Obama is downloaded and parsed.
. summarize
```

Variable	Obs	Mean	Std. Dev.	Min	Max
president	0				
president2	1084	44	0	44	44
approving	1084	49.32657	6.818881	38	69
disapproving	1084	42.72325	7.515142	12	55
unsure	1084	7.941882	1.836668	0	21
startdate	1084	18482.63	327.8573	17918	19052
enddate	1084	18484.71	327.8727	17920	19054

```
. list in 1/3
```

	president	presid~2	approv~g	disapp~g	unsure	startdate
1.	Barack Obama	44	68	12	21	21jan2009
	enddate					
	23jan2009					
2.	Barack Obama	44	69	13	18	22jan2009
	enddate					
	24jan2009					
3.	Barack Obama	44	67	14	19	23jan2009
	enddate					
	25jan2009					

◁

▷ Example

All U.S. presidents' job approval poll results: In this example, job approval poll results for all U.S. presidents since President Franklin D. Roosevelt, the 32nd U.S. president, are downloaded, parsed, and appended.

```
. approval, president(32/44)
Poll results for President Franklin D. Roosevelt is downloaded and parsed.
Poll results for President Harry S. Truman is downloaded and parsed.
Poll results for President Dwight D. Eisenhower is downloaded and parsed.
Poll results for President John F. Kennedy is downloaded and parsed.
Poll results for President Lyndon B. Johnson is downloaded and parsed.
Poll results for President Richard Nixon is downloaded and parsed.
Poll results for President Gerald R. Ford is downloaded and parsed.
Poll results for President Jimmy Carter is downloaded and parsed.
Poll results for President Ronald Reagan is downloaded and parsed.
Poll results for President George Bush is downloaded and parsed.
Poll results for President William J. Clinton is downloaded and parsed.
Poll results for President George W. Bush is downloaded and parsed.
Poll results for President Barack Obama is downloaded and parsed.
. summarize
```

Variable	Obs	Mean	Std. Dev.	Min	Max
president	0				
president2	2431	41.43439	3.382555	32	44
approving	2431	51.96791	11.46093	22	91
disapproving	2431	38.32086	12.68241	2	71
unsure	2431	8.940354	4.24062	0	43
startdate	2431	13024.94	6886.804	-6737	19052
enddate	2431	13027.31	6886.686	-6737	19054

```
. list in 1/3
```

1.

president	presid~2	approv~g	disapp~g	unsure
Franklin D. Roosevelt	32	69	24	6

startdate	enddate
22jul1941	22jul1941

2.

president	presid~2	approv~g	disapp~g	unsure
Franklin D. Roosevelt	32	65	25	8

startdate	enddate
29jul1941	29jul1941

3.

president	presid~2	approv~g	disapp~g	unsure
Franklin D. Roosevelt	32	68	23	7

startdate	enddate
05aug1941	05aug1941

```
. list in 1030/1032

      +----------------------------------------------------------------+
1030. |          president | presid~2 | approv~g | disapp~g | unsure |
      | William J. Clinton |       42 |       56 |       39 |      4 |
      |----------------------------------------------------------------|
      |          startdate           |            enddate              |
      |          21may1999           |          23may1999              |
      +----------------------------------------------------------------+

      +----------------------------------------------------------------+
1031. |          president | presid~2 | approv~g | disapp~g | unsure |
      | William J. Clinton |       42 |       59 |       35 |      4 |
      |----------------------------------------------------------------|
      |          startdate           |            enddate              |
      |          04jun1999           |          05jun1999              |
      +----------------------------------------------------------------+

      +----------------------------------------------------------------+
1032. |          president | presid~2 | approv~g | disapp~g | unsure |
      | William J. Clinton |       42 |       61 |       34 |      3 |
      |----------------------------------------------------------------|
      |          startdate           |            enddate              |
      |          11jun1999           |          13jun1999              |
      +----------------------------------------------------------------+
```

◁

4 Conclusion

In this article, we showed how to use `approval` to download, parse, and save presidential job approval poll results provided by The American Presidency Project. Although the data are available as an HTML webpage for public use, the proposed Stata command `approval` converts the HTML data into a usable Stata dataset.

About the authors

Mehmet F. Dicle is an assistant professor of finance at Loyola University New Orleans.

Betul Dicle is a PhD in political science.