Gov 50: 2. Introduction to R and R Markdown

Matthew Blackwell

Harvard University

Fall 2018

1. Today’s agenda

2. R logistics

3. Measuring turnout

1/ Today’s agenda

Where are we?

- What you've been doing:
    - Reading QSS, Ch 1
    - Creating DataCamp, rstudio.cloud, and Perusall accounts
- DataCamp Assignment 1:
    - On Canvas/DataCamp now.
    - Due Tues, 9/11 at 11:59 ET
    - DC Assignment 2 due on Thurs, 9/13.
    - Get started early!
- Prerequisites.
- Any other questions?

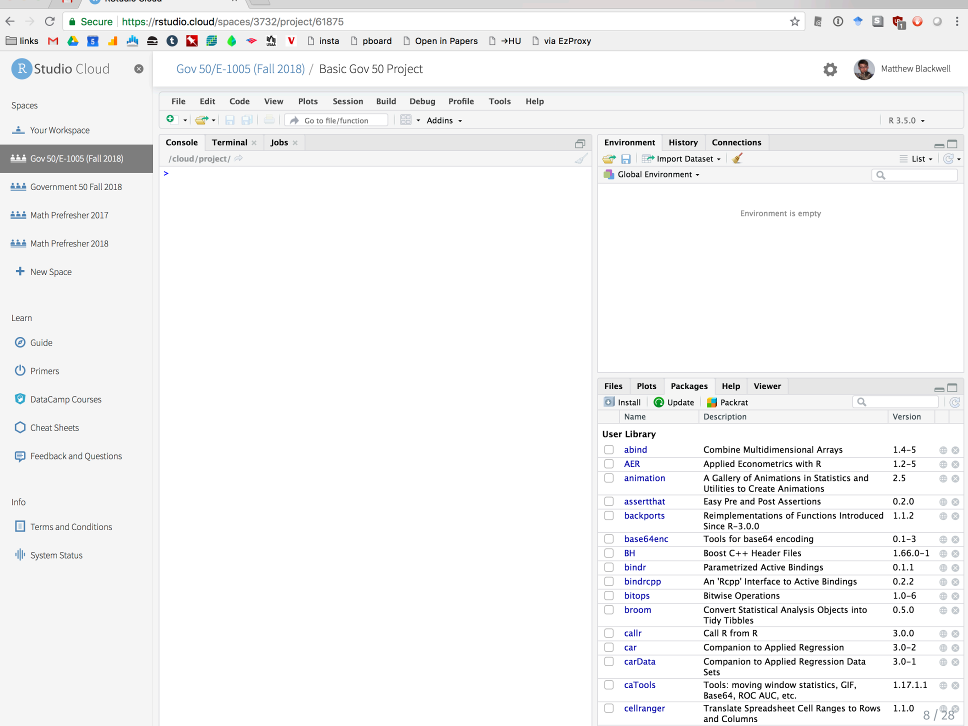
Where are we going?

- Today:
  - Introduction to R, RStudio, and DataCamp
  - Quick exercise on measuring turnout to get familiar with R

2/ R logistics

RStudio Cloud

- rstudio.cloud (we'll set you with accounts):
  - Online version of R pre-loaded with all the goodies.
  - Minimize the headaches of installation/packages/etc.
  - Allows us to distribute HW code/data/templates to you very easily.
- You're free to download RStudio (a program to use R) on your own machine to test it.



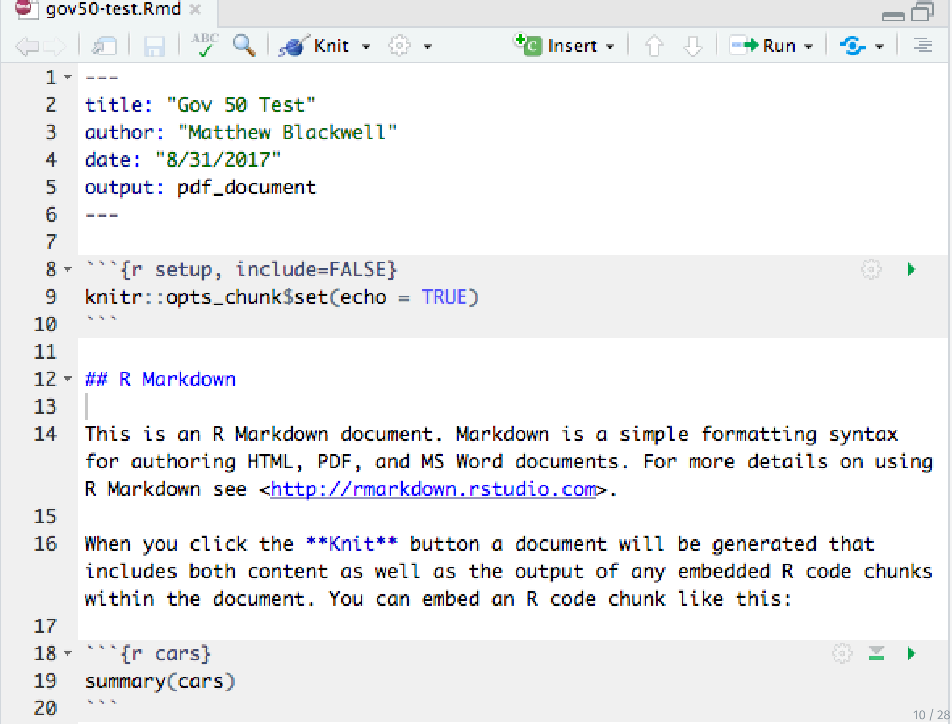
Environment is empty

User Library

	Name	Description	Version
□	abind	Combine Multidimensional Arrays	1.4-5
□	AER	Applied Econometrics with R	1.2-5
□	animation	A Gallery of Animations in Statistics and Utilities to Create Animations	2.5
□	assertthat	Easy Pre and Post Assertions	0.2.0
□	backports	Reimplementations of Functions Introduced Since R-3.0.0	1.1.2
□	base64enc	Tools for base64 encoding	0.1-3
□	BH	Boost C++ Header Files	1.66.0-1
□	bindr	Parametrized Active Bindings	0.1.1
□	bindrcpp	An 'Rcpp' Interface to Active Bindings	0.2.2
□	bitops	Bitwise Operations	1.0-6
□	broom	Convert Statistical Analysis Objects into Tidy Tibbles	0.5.0
□	callr	Call R from R	3.0.0
□	car	Companion to Applied Regression	3.0-2
□	carData	Companion to Applied Regression Data Sets	3.0-1
□	caTools	Tools: moving window statistics, GIF, Base64, ROC AUC, etc.	1.17.1.1
□	cellranger	Translate Spreadsheet Cell Ranges to Rows and Columns	1.1.0

Writing Scripts

- For HWs, we'll have you write up your answers in a file called an "R markdown" file.
- Essentially a mix of text answers and your code to analyze data/produce graph.
- Benefits:
  - Reproducible, automatic report creation, automation.
- Downsides:
  - Might be unfamiliar, but we'll provide resources online and in section!
- I write my slides in R markdown and I'll post the source so you can see what it's like.



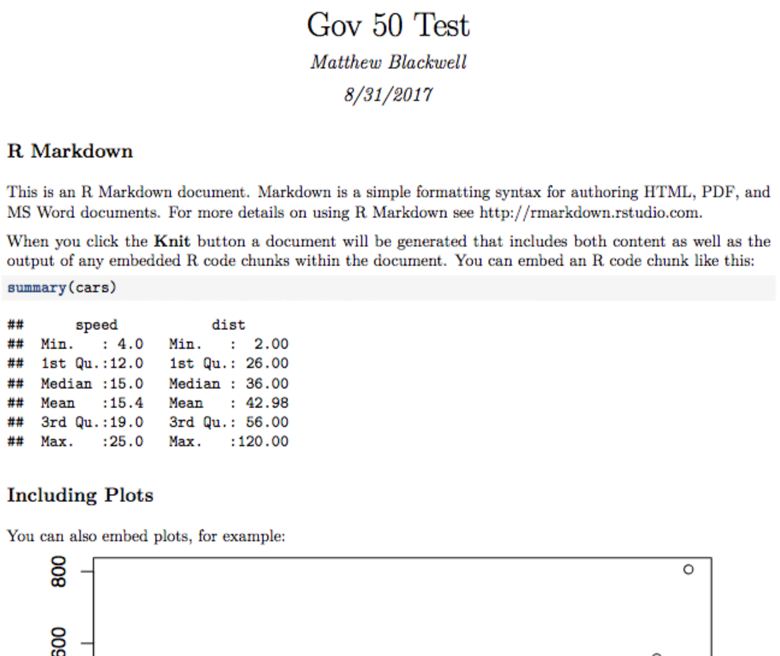
~~~
---
title: "Gov 50 Test"
author: "Matthew Blackwell"
date: "8/31/2017"
output: pdf_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
~~~

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax
for authoring HTML, PDF, and MS Word documents. For more details on using
R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that
includes both content as well as the output of any embedded R code chunks
within the document. You can embed an R code chunk like this:

```{r cars}
summary(cars)
```
```
```



Gov 50 Test

Matthew Blackwell

8/31/2017

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00  
```

Including Plots

You can also embed plots, for example:

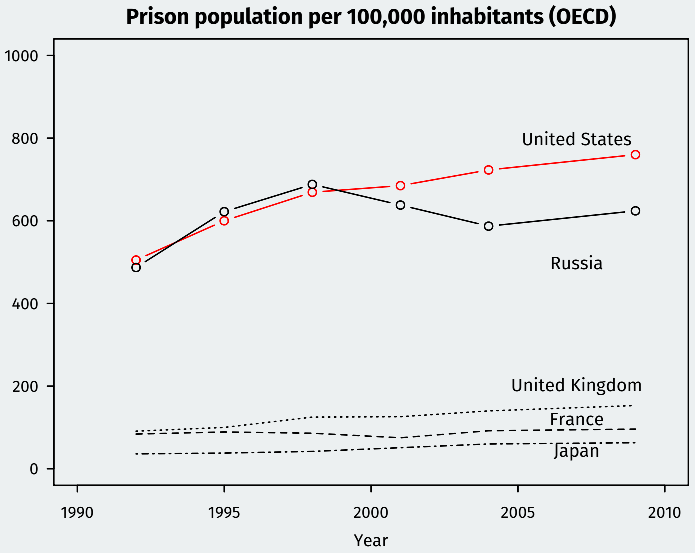
Break to show Rstudio/DataCamp

3/ Measuring turnout

Measuring turnout

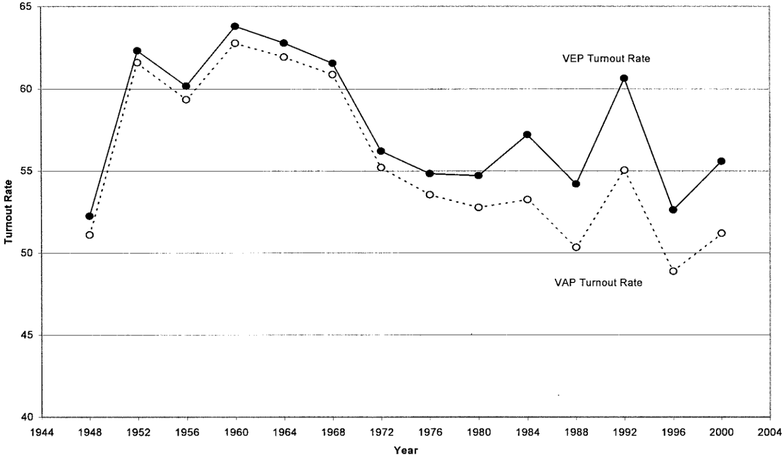
- Question: How do you measure turnout rates?
- Numerator: Total votes cast
- Denominator:
  1. **Registered voters**
  2. **VAP** (voting-age population) form Census
  3. **VEP** (voting-eligible population)
- **VEP** = VAP + overseas voters − ineligible voters
  - overseas voters: military personnel and civilians
  - ineligible voters: non-citizens, disenfranchised felons, those who failed to meet states' residency requirement, etc.

Growing Prison Populations



VAP and VEP are different

FIGURE 1. National VAP and VEP Presidential Turnout Rates, 1948–2000



McDonald and Popkin (2001) *American Political Science Review*

Bias in self-reported turnout

- Measuring individual turnout:
    - **voter file**: registered voters only
    - **survey**: American National Election Study (ANES)
- **Social desirability bias**: “Did you vote?” “....yeah, sure!”
- Data set: `turnout.csv`

Variables	Description
`year`	election year
`ANES`	ANES estimated turnout rate
`VEP`	Voting Eligible Population (in thousands)
`VAP`	Voting Age Population (in thousands)
`total`	total ballots cast for highest office (in thousands)
`felons`	total ineligible felons (in thousands)
`noncitizens`	total non-citizens (in thousands)
`overseas`	total eligible overseas voters (in thousands)
`osvoters`	total ballots counted by overseas voters (in thousands)

- Load the dataset (there is an easy pull-down menu too):

```
turnout <- read.csv(”data/turnout.csv”)
class(turnout)
```

```
## [1] ”data.frame”
```

- Every object in **R** belongs to a class: `character`, `numeric`, etc.
- `data.frame` is like a *matrix* with rows (observations) and columns (variables):

```
dim(turnout)
```

```
## [1] 14  9
```

```
turnout[1:3, c(”year”, ”total”, ”VEP”, ”VAP”, ”felons”)]
```

```
##   year total    VEP    VAP felons
## 1 1980 86515 159635 164445    802
## 2 1982 67616 160467 166028    960
## 3 1984 92653 167702 173995   1165
```

Vectors

- Each column of the `data.frame` is a vector:

```
turnout$year
```

```
##  [1] 1980 1982 1984 1986 1988 1990 1992 1994 1996 1998
## [11] 2000 2002 2004 2008
```

- We can subset the vector using brackets:

```
turnout$year[2]
```

```
## [1] 1982
```

```
turnout$year[2:4]
```

```
## [1] 1982 1984 1986
```

Creating vetors

- Create a vector using `c()` for “concatenate”:

```r
c(2,3,4)
```

```
## [1] 2 3 4
```

- We can save vectors with new names to keep track of things:

```r
eighties <- turnout$year[1:5]
eighties
```

```
## [1] 1980 1982 1984 1986 1988
```

- We can also do basic arithmetic on vectors:

```r
eighties + 10
```

```
## [1] 1990 1992 1994 1996 1998
```

VAP-based turnout

- total votes / (VAP + overseas voters) × 100:

```
VAPtr <- turnout$total /
    (turnout$VAP + turnout$overseas) * 100
VAPtr
```

```
##  [1] 52.0 40.2 52.5 36.1 49.7 35.9 54.0 38.0 47.5 34.8
## [11] 49.3 35.8 54.5 55.7
```

- Add informative labels:

```
names(VAPtr) <- turnout$year
VAPtr
```

```
## 1980 1982 1984 1986 1988 1990 1992 1994 1996 1998 2000
## 52.0 40.2 52.5 36.1 49.7 35.9 54.0 38.0 47.5 34.8 49.3
## 2002 2004 2008
## 35.8 54.5 55.7
```

VEP-based turnout

- total votes / VEP × 100:

```
VEPtr <- turnout$total / turnout$VEP * 100
names(VEPtr) <- turnout$year
```

- Difference between VEP and VAP-based turnout rates:

```
diff <- VEPtr - VAPtr
names(diff) <- turnout$year
diff
```

```
## 1980 1982 1984 1986 1988 1990 1992 1994 1996 1998 2000
## 2.16 1.89 2.71 2.06 3.05 2.48 4.07 3.10 4.12 3.26 4.88
## 2002 2004 2008
## 3.68 5.55 5.88
```

Percent change vs. percentage point change

- *Percentage-point change*:

$$\text{new turnout rate}(\%) - \text{base turnout rate}(\%)$$

- *Percentage change*:

$$\frac{\text{new turnout rate} - \text{base turnout rate}}{\text{base turnout rate}} \times 100$$

```
(VEPtr - VAPtr) / VAPtr * 100
```

```
##  1980  1982  1984  1986  1988  1990  1992  1994  1996
##  4.14  4.70  5.16  5.72  6.13  6.90  7.54  8.14  8.68
##  1998  2000  2002  2004  2008
##  9.36  9.89 10.28 10.18 10.56
```

Self-reported vs VAP & VEP turnout

- Comparison between VAP and ANES:

```
diffVAP <- turnout$ANES - VAPtr
summary(diffVAP)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    11.1    18.2    20.6    20.3    22.4    26.2
```

- Comparison between VEP and ANES:

```
diffVEP <- turnout$ANES - VEPtr
summary(diffVEP)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.58   15.27   16.89   16.84   18.53   22.49
```

Presidential vs. midterm elections

- Elections in the data:

```
turnout$year
```

```
##  [1] 1980 1982 1984 1986 1988 1990 1992 1994 1996 1998
## [11] 2000 2002 2004 2008
```

- Presidential elections: odd entries of vector (1st, 3rd...) plus the last

```
pres <- c(1, 3, 5, 7, 9, 11, 13, 14)
mids <- c(2, 4, 6, 8, 10, 12)
```

```
turnout$year[pres]
```

```
## [1] 1980 1984 1988 1992 1996 2000 2004 2008
```

```
turnout$year[mids]
```

```
## [1] 1982 1986 1990 1994 1998 2002
```

- Presidential elections:

```
pVEPtr <- VEPtr[pres]
names(pVEPtr) <- turnout$year[pres]
pVEPtr
```

```
## 1980 1984 1988 1992 1996 2000 2004 2008
## 54.2 55.2 52.8 58.1 51.7 54.2 60.1 61.6
```

- Midterm elections:

```
mVEPtr <- VEPtr[mids]
names(mVEPtr) <- turnout$year[mids]
mVEPtr
```

```
## 1982 1986 1990 1994 1998 2002
## 42.1 38.1 38.4 41.1 38.1 39.5
```

Sample averages

- **Mean** or **average** of a set of numbers:

$$\text{mean} = \frac{\text{sum of the numbers}}{\text{how many numbers}}$$

$$\text{mean}(x_1, x_2, x_3) = \frac{x_1 + x_2 + x_3}{3}$$

- Average difference between presidential and midterm elections:

```r
mean(pVEPtr) - mean(mVEPtr)
```

```
## [1] 16.4
```

Wrap up

- What to do next?
  - Create accounts!
  - DataCamp assignments!
  - Try loading the data from this lecture and implementing some of the commands.
  - Toy around with Rmd file to see how it works.
- Next week: Causality.