1: Introduktion

Videregående kvantitative metoder i studiet af politisk adfærd

Frederik Hjorth
adjunkt, Ph.D.
fh@ifs.ku.dk
fghjorth.github.io
@fghjorth

Institut for Statskundskab
Københavns Universitet

6. september 2018

Mig:

- cand.scient.pol., ph.d. i statskundskab
- adjunkt ved IfS
- interesser: politisk psykologi, holdningsdannelse, indvandring, anvendt metode

Jer:

- navn
- ønsker til faget
- god tekst jeg har læst (under studiet eller iø.)

- "Videregående Kvantitative Metoder I Studiet Af Politisk Adfærd" (VKM)
- Seminar m. 28 timers holdundervisning
- Hjemmeside: `github.com/fghjorth/vkme18`

Fagets opbygning

Gang	Tema	Litteratur	Case
1	Introduktion til R	Leeper (2016)	
2	R workshop I + tidy data	Wickham (2014), Zhang (2017)	
3	Data fra online-kilder	MRMN kap 9+14	Hjorth (2016)
4	Tekst som data	Grimmer & Stewart (2013), Benoit & Nulty (2016)	Baturo & Mikhaylov (2013)
5	Regression I: OLS brush-up	AP kap 3	Mutz (2018)
6	Regression II: Paneldata	AGS kap 4	Mutz (2018)
Efterårsferie			

Fagets opbygning

7	R workshop II	tba	
8	Introduktion til kausal inferens	Angrist & Pischke (2010), Samii (2016)	Carroll (2018)
9	Eksperimenter I	AP kap 1+2, GG kap 1+2	Gerber, Green & Larimer (2008)
10	Eksperimenter II	GG kap 3+4+5	Gerber & Green (2000)
11	Instrumentvariable	AP kap 4	Colantone & Stanig (2018)
12	Difference-in-differences	AP kap 5	
13	Regressionsdiskontinuitetsdesigns	AP kap 6	Eggers & Hainmueller (2009)
14	'Big data' og maskinlæring	Varian (2014), Montgomery & Olivella (2017)	Theocharis et al. (2016)

Pensum

- AGS: Andreß, H. J., Golsch, K., & Schmidt, A. W. (2013). Applied panel data analysis for economic and social surveys. Springer Science & Business Media.
- GG: Gerber, A. S., & Green, D. P. (2012). Field experiments: Design, analysis, and interpretation. WW Norton.
- AP: AP: Angrist, J. D., & Pischke, J. S. (2014). Mastering 'metrics: The path from cause to effect. Princeton University Press.
- MRMN: Munzert, S., Rubba, C., Meißner, P., & Nyhuis, D. (2014). Automated data collection with R: A practical guide to web scraping and text mining. John Wiley & Sons.

+ hertil artikler og case-artikler

Målbeskrivelse

Seminarets målsætning er at sætte den studerende i stand til efter endt undervisning at kunne (ILO's):

- Identificere relevante designs og teknikker for at løse politologiske problemstillinger.
- Bearbejde data i strukturerede og ustrukturerede formater mhp. senere analyse.
- Analysere empiriske politologiske problemstillinger med udgangspunkt i kvantitative data.
- Reflektere over fordele og ulemper ved forskellige designs og teknikker fra kursets pensum og samt i andres og eget arbejde.

Eksamen

Formelle krav:

- fri opgave
- afleveringsfrist 9. januar 2019
- øvrige rammer fremgår af studieordningen

Koncepter for opgaven:

1. Fri opgave med anvendelse af fagets metoder
2. Replikationsstudie
3. Specialeforstudie

3 'ben' i faget:

1. Logik: styrker og svagheder ved forskellige undersøgelsesdesigns
2. Teknik: den underliggende økonometri/statistik
3. Implementering: hvordan man faktisk gennemfører analysen i R

Fokus i VKM på 1+3

Typisk struktur for holdtime: iht. “Particular General Particular” princippet

- Motiverende eksempel på metode
- Præsentation af principper
- Gennemgang af implementering i R

Bærende motivation for faget: 2 revolutioner har drevet kolossal vækst i kvantitativ samfundsvidenskab

1. 'data revolution'
2. 'computational revolution'

(OBS: i uge 8 møder vi en tredje, den såkaldte 'credibility revolution')

Konsekvens: fortrolighed med velstrukturerede data er utilstrækkeligt

»It is simply not sufficient to achieve 'statistical literacy' by learning about common statistical concepts and methods. Instead, all students in the social sciences should acquire basic data analysis skills so that they can exploit ample opportunities to learn from data (...)« (Kosuke Imai: *Quantitative Social Science: An Introduction*, p. 3)

Hvad er R?

- et program til statistisk programmering
- et programmeringssprog (som C++, Python, Perl, etc.)
- fungerer generelt *objekt-orienteret* (ctr. fx. Stata)
- open source
- opfindere: Ross Ihaka & Robert Gentleman
- videreudvikling af *S*

»In 1992, Gentleman - then a professor at the University of Waterloo in Canada - travelled 8600 miles to the University of Auckland to lecture for three months. One day, he found himself needing a manual for a particularly tricky piece of software and Ihaka - still a professor of statistics in those days - was the only one in the department who had a copy. In time, both realised they shared an interest in what Ihaka calls “playing academic fun and games” with statistical computing languages. (...)«

Kilde: Thieme, N. (2018), R generation. *Significance*, 15: 14-19. doi:10.1111/j.1740-9713.2018.01169.x

»Around that time, the University of Auckland needed a programming language to use in its undergraduate statistics courses as the school's current tool had reached the end of its useful life. There was one major caveat: the program needed to run on Macintosh. According to Gentleman, the Department of Statistics took inventory and decided "that thing Ross and Robert are working on", which happened to run on Macintosh, was better than their current language. The professors called it R, as both a nod to S and in reference to their forenames.«

Kilde: Thieme, N. (2018), R generation. *Significance*, 15: 14-19. doi:10.1111/j.1740-9713.2018.01169.x

Første spadestik til R for ca. 20 år siden:

Peter Dalgaard
@pdalgd

Follow

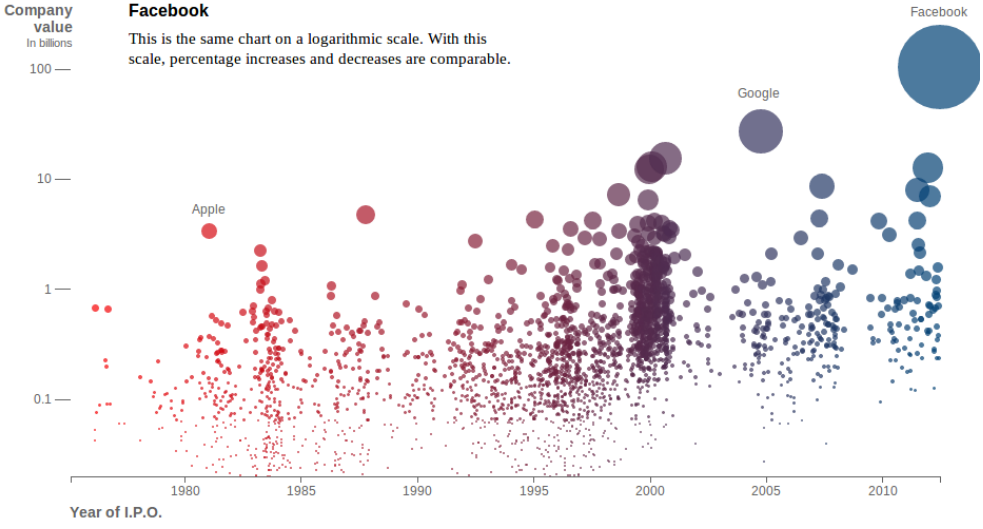
It was twenty years ago today, Ross Ihaka got the band to play....
#rstats

```
Date: Sat, 16 Aug 1997 09:15:45 +1200 (NZST)
From: Ross Ihaka <ihaka@stat.auckland.ac.nz>
To: Kurt.Hornik@ci.tuwien.ac.at, p.dalgaard@kubism.ku.dk,
        thomas@biostat.washington.edu
Subject: Invitation ...
Cc: maechler@stat.math.ethz.ch, rgentlem@stat1.stat.auckland.ac.nz

We have had a bit of a discussion on enlarging the R "core team" At
present this seems to consist of Robert, Martin and myself although the
following people also have commit privileges in the CVS tree:

        Luke Tierney            no introduction needed
        Heiner Schwarte         developer of dyn.load etc
        Paul Murrell            my PhD student
[...]
As major contributors (and apparently sane people) we would like to
invite you to be part of the R "core team".
[...]
We can't promise you anything much in return, except a free copy of R
:-) and perhaps a publication on "distributed development of statistical
software".  Since you are clearly hopeless software junkies, perhaps you
don't need any more incentive.
[...]
```

1:21 AM - 16 Aug 2017

Hvorfor R?

- næsten uendelige anvendelser
- reproducerbart workflow
- den nye analytiske standard
- free as in free speech and free beer
  - sml. m. pris for 1 Stata single user business license: $1.195
- absolut bedst til datavisualisering


Company value
In billions
100
10
1
0.1
Facebook
This is the same chart on a logarithmic scale. With this scale, percentage increases and decreases are comparable.
Facebook
Google
Apple
1980
1985
1990
1995
2000
2005
2010
Year of I.P.O.

The New York Times

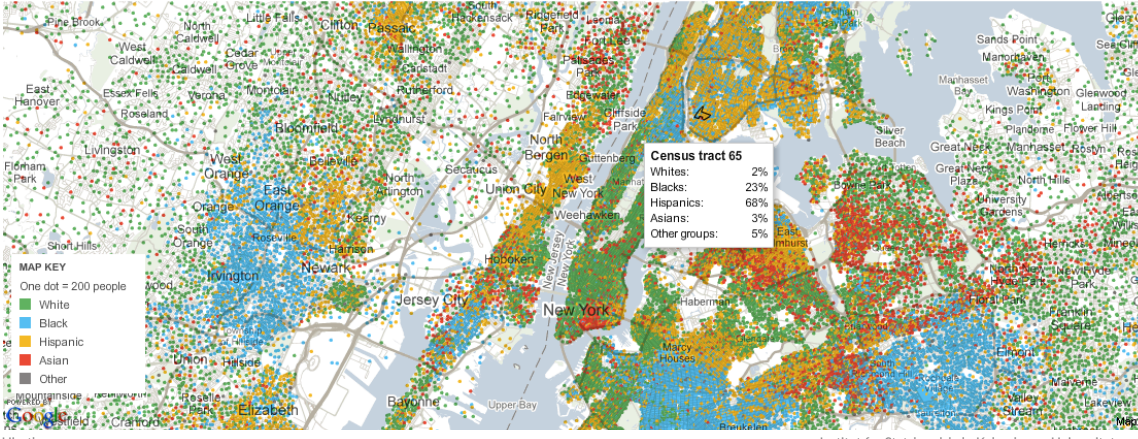
Mapping America: Every City, Every Block

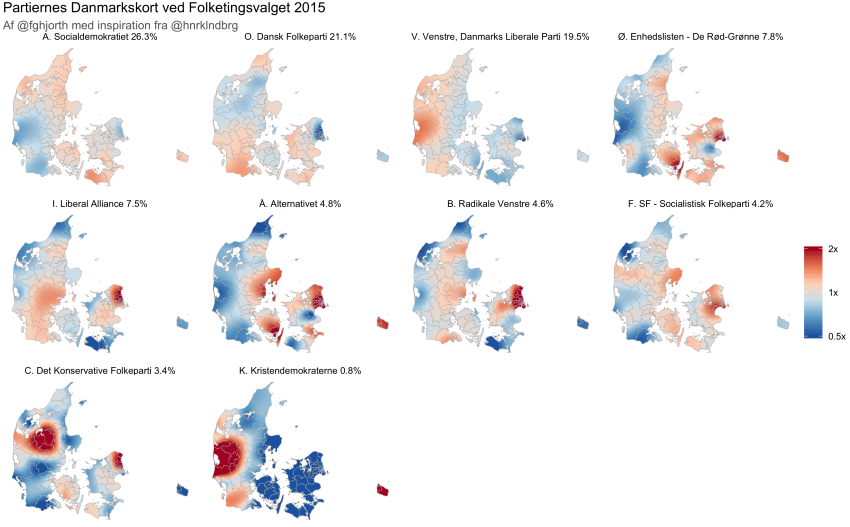
Find something interesting? Share this view

Browse local data from the Census Bureau's American Community Survey, based on samples from 2005 to 2009.

Distribution of racial and ethnic groups

View More Maps | Address, ZIP code or city Go

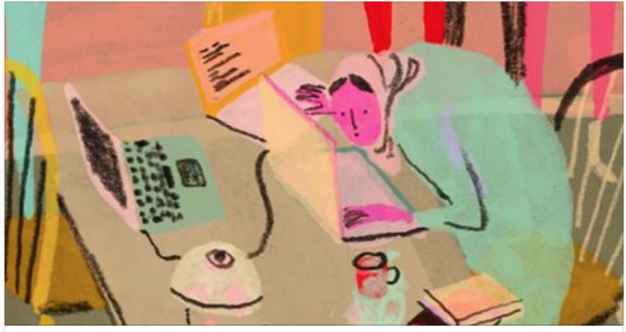



Partiernes Danmarkskort ved Folketingsvalget 2015
Af @fghjorth med inspiration fra @hnrklndbrg
A. Socialdemokratiet 26.3%
O. Dansk Folkeparti 21.1%
V. Venstre, Danmarks Liberale Parti 19.5%
Ø. Enhedslisten - De Rød-Grønne 7.8%
I. Liberal Alliance 7.5%
Å. Alternativet 4.8%
B. Radikale Venstre 4.6%
F. SF - Socialistisk Folkeparti 4.2%
C. Det Konservative Folkeparti 3.4%
K. Kristendemokraterne 0.8%
2x
1x
0.5x

Hvorfor *ikke* R?

- ingen grafisk brugerflade (GUI)
- mere 'ordrigt' end fx. Stata
- konstant import af ekstrapakker e
- meget følsom over for fejl
- kryptiske fejlmeddelelser

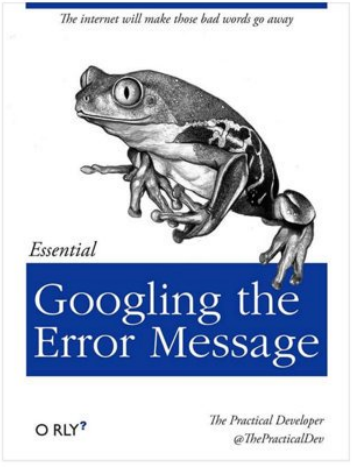
Men:

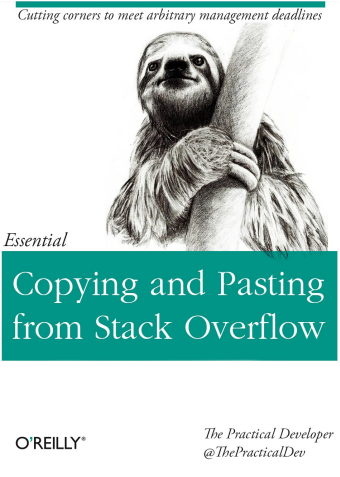


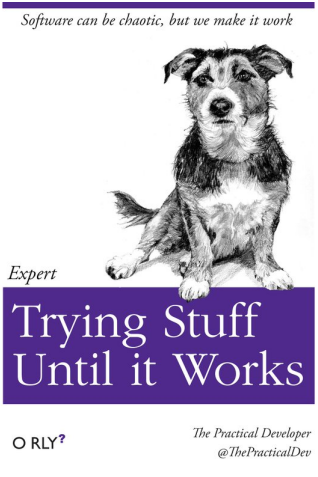
I Was Intimidated by Coding Until I Learned This Secret Strategy: Googling

"You don't need to go to grad school. Save your money. I'll teach you how to code." Seven years ago, in a bar near downtown Los Angeles, I was sharing ...

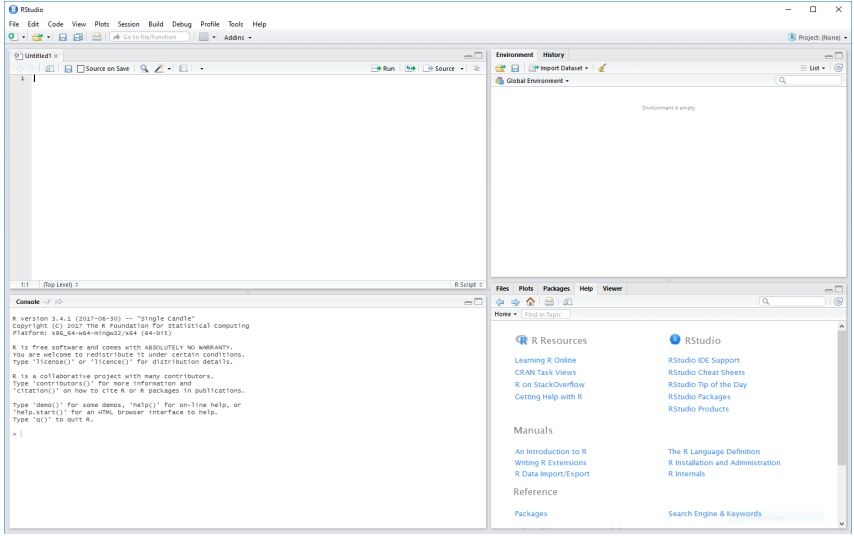
SLATE.COM


The internet will make those bad words go away
Essential
Googling the Error Message
O RLY?
The Practical Developer
@ThePracticalDev


Cutting corners to meet arbitrary management deadlines
Essential
Copying and Pasting from Stack Overflow
O'REILLY
The Practical Developer
@ThePracticalDev


Software can be chaotic, but we make it work
Expert
Trying Stuff Until it Works
O RLY?
The Practical Developer
@ThePracticalDev

Det bedste IDE til R: RStudio



Nyttige tastaturgenveje i RStudio:

- Assignment operator: Alt + -
- Autocomplete: ⇤⇥
- Comment and uncomment: Ctrl + ⇧ + C
- Run code in script: Ctrl + ↵
- Run code in console: ↵
- Toogle through executed code: ↑ and ↓
- Clear screen: Ctrl + L
- Run to current line: Ctrl + B

Leeper, *Really Introductory Introduction*:

- Getting started
  - brug af R som regnemaskine: fx. `(2+4)/7`
  - parsing errors ctr. syntax errors
  - nye vektorer: fx. `dice <- c(2,2,3,4)`
  - ekstrahering fra vektorer: fx. `dice[1:3]`
  - ny data frame: fx. `df <- data.frame(dice,number=1:4)`
  - data framens struktur: `str(df)`
  - centrale tendenser: `summary(df)`

- Real data
  - installer pakker: `install.packages()`
  - indlæs pakker: `library()`
  - importér data: `import()` fra `rio`-pakken

- Randomness
  - sample fra en vektor: `sample()`
- Plots
  - pakke: `ggplot2`
  - fx. `ggplot(iris,aes(x=Sepal.Length)) + geom_histogram`

- Basic programming tools
  - funktioner: fx. `ftoc <- function(f){ c<-((f-35)*5)/9 ; print(c) }`
  - for loops: fx. `for (i in 1:10) print(i*i)`

Næste gang:

- intro til 'tidy data': Wickham
- databehandling med tidyverse: Zhang
- lektie:
  - 'Introduction to the Tidyverse' på Datacamp
  - kapitel 1+3 om tidy verber
  - optionally 2+4 om visualisering

Tak for i dag!