5737

concepts + computation

week 1
TOOLS

computation is key

Who are these people?

An Ice Breaker

Business as Usual? Economic Responses to Political Tensions

Christina L. Davis Princeton University
Sophie Meunier Princeton University

Do political tensions harm economic relations? Theories claim that trade prevents war and political relations motivate trade, but less is known about whether smaller shifts in political relations impact economic exchange. Looking at two major economies, we show that negative events have not hurt U.S. or Japanese trade or investment flows. We then examine specific incidents of tensions in U.S.-French and Sino-Japanese relations over the past decade—two case pairs that allow us to compare varying levels of political tension given high existing economic interdependence and different alliance relations. Aggregate economic flows and high salience sectors like wine and autos are unaffected by the deterioration of political relations. In an era of globalization, actors lack incentives to link political and economic relations. We argue that sunk costs in existing trade and investment make governments, firms, and consumers unlikely to change their behavior in response to political disputes.

Do political tensions have economic consequences? The relationship between economic interdependence and conflict has been a central debate in international relations. Leading scholars contend that "states with good relations should have more trade than states with poor relations" and import decisions of firms will respond to "the climate of friendliness or hostility that exists between the importer and exporter" (Morrow, Siverson, and Tabares 1998, 650; Pollins 1989b, 739). Analysis of trade and conflict in a simultaneous equations model concludes that "political relations are driving commerce, not the other way around" (Keshk, of force, and to war. While most analysis of the interdependence debate focuses on militarized disputes, we analyze the shift at the lower level from normal relations to political tensions. As noted by Pevehouse, "much of the nuance of interdependence theory has been discarded" in recent empirical studies that use dichotomous measures for conflict, and new insights may be gained by returning to the earlier approach in the literature that measured conflict and cooperation with events data (2004, 247). A large range of interactions determines the status of political relations between states. By political tensions, we mean disagreement over policy issues, hostility between

think + write + discuss

Here are some questions:

What is the key causal claim of the paper?

Do they have good evidence for their causal claim?

What are some descriptors for this paper?

What do the authors mean by "political tensions"?

How do they measure political tensions?

Let's take a close look....

How is this going to go?

Teaching Philosophy

Learning results from what the student does and thinks, and only from what the student does and thinks.

—Herbert Simon

It is the one who does the work who does the learning.

—Terry Doyle

Aaron Williams
@awunderground

Follow

This is the best advice from @hadleywickham. True of learning R. True of learning anything. #rstats r-posts.com/advice-to-youn…

> **It's easy when you start out programming to get really frustrated and think, "Oh it's me, I'm really stupid," or, "I'm not made out to program." But, that is absolutely not the case. Everyone gets frustrated.** I still get frustrated occasionally when writing R code. It's just a natural part of programming. So, it happens to everyone and gets less and less over time. Don't blame yourself. Just take a break, do something fun, and then come back and try again later.

11:16 AM - 25 Aug 2018

153 Retweets 426 Likes

8 153 426

Assignments

Weekly

- 14 conceptual homeworks (3% each; 42% total)
- 7 computational homeworks (3% each; 21% total)
- 14 reflections (1% each; 14% total)

End-of-semester

- Data assignment related to FYP (8%)
- Final exam (15%)

Orienting Ourselves

The Ten Commandments

of Success in (My) POS 5737

I

slow your roll


thoughtful empirical analysis
p-values from regression models
first-year grad students

Journal of Economic Literature
Vol. XXXIV (March 1996), pp. 97–114

Significance testing (p-value) discourages thoughtful social science.

The Standard Error of Re

By DEIRDRE N. MCCLOSKEY
and
STEPHEN T. ZILIAK
University of Iowa

Suggestions by two anonymous and patient referees greatly improved the paper. Our thanks also to seminars at Clark, Iowa State, Harvard, Houston, Indiana, and Kansas State universities, at Williams College, and at the universities of Virginia and Iowa. A colleague at Iowa, Calvin Siebert, was materially helpful.

THE IDEA OF statistical significance is old, as old as Cicero writing on forecasts (Cicero, *De Divinatione*, I. xiii. 23). In 1773 Laplace used it to test whether comets came from outside the solar system (Elizabeth Scott 1953, p. 20). The first use of the very word "significance" in a statistical context seems to be John Venn's, in 1888, speaking of differences expressed in units of probable error:

> They inform us which of the differences in the above tables are permanent and signifi-

cant for science or policy and yet be insignificant statistically, ignored by the less thoughtful researchers.

In the 1930s Jerzy Neyman and Egon S. Pearson, and then more explicitly Abraham Wald, argued that actual investigations should depend on substantive not merely statistical significance. In 1933 Neyman and Pearson wrote of type I and type II errors:

> Is it more serious to convict an innocent man or to acquit a guilty? That will depend on the

Regression models are not magic. So don't treat them like they are.

10

STATISTICAL MODELS AND SHOE LEATHER

*David A. Freedman**

Regression models have been used in the social sciences at least since 1899, when Yule published a paper on the causes of pauperism. Regression models are now used to make causal arguments in a wide variety of applications, and it is perhaps time to evaluate the results. No definitive answers can be given, but this paper takes a rather negative view. Snow's work on cholera is presented as a success story for scientific reasoning based on nonexperimental data. Failure stories are also discussed, and comparisons may provide some insight. In particular, this paper suggests that statistical technique can seldom be an adequate substitute for good design, relevant data, and testing predictions against reality in a variety of settings.

the simple tools we discuss in our first few weeks

histogram, avg, SD, scatterplot, simple linear model

ARE POWERFUL

I want you to learn to use them well.

II

master the simple things


GRAD STUDENT
SCATTERPLOT
HETEROSKEDASTIC
ORDERED PROBIT

III

computation is key

IV

change-review-commit-push

V

engage with me where you are

that’s enough

slow your roll

master the simple things

computation is key

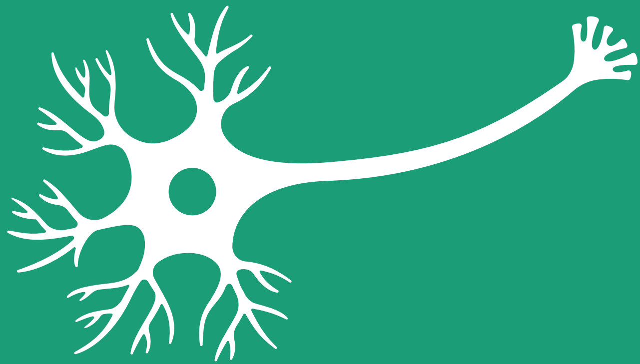
change-review-commit-push

engage with me where you are

What should I take from this class?

Build a Foundation

think + write + discuss



What should we build a foundation for?

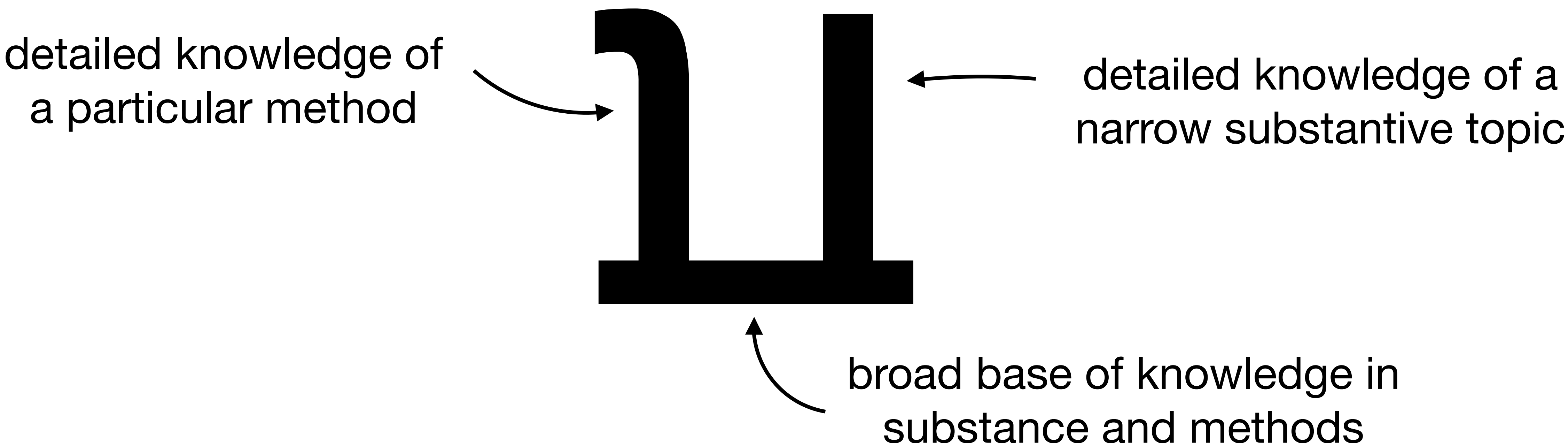
Where do you want to be in 10 years?

What do you need to accomplish in the next 5 years?

What do you need to take from this course to get there?


π

π


detailed knowledge of
a particular method
detailed knowledge of a
narrow substantive topic
broad base of knowledge in
substance and methods

broad base of knowledge in methods

that allows us to produce great research projects

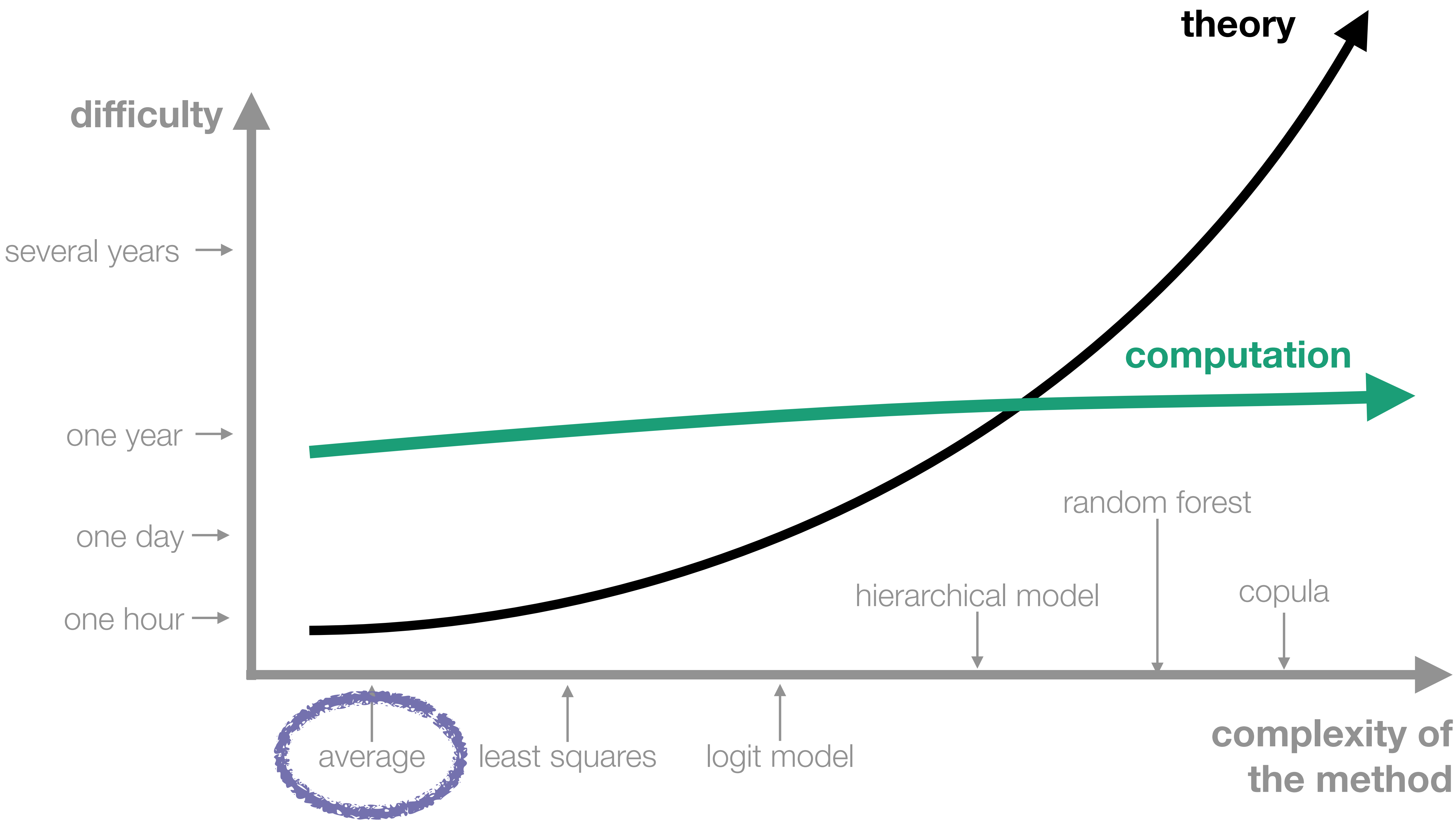
broad base of knowledge in methods

1. **basic statistical tools**, such as a histogram, average, standard deviation, normal approximation, scatterplot, correlation, simple regression, sample surveys
2. **basic concepts in probability theory**, such as conditional probability, the law of averages, the expected value, the standard error.
3. **basic concepts in inference**, such as a point estimate, interval estimate, and hypothesis test.
4. **advanced concepts in probability theory** (that rely on calculus), such as a pmf or pdf, moments, and the central limit theorem.

concepts <u>and computation</u>

Why should I care about computation?


computation


theory
difficulty
several years
computation
one year
random forest
one day
hierarchical model
copula
one hour
average
least squares
logit model
complexity of
the method

Prices of over 50,000 round cut diamonds

Description

A dataset containing the prices and other attributes of almost 54,000 diamonds. The variables are as follows:

Usage

```
diamonds
```

Format

A data frame with 53940 rows and 10 variables:

price

price in US dollars (\$326–\$18,823)

carat

weight of the diamond (0.2–5.01)

cut

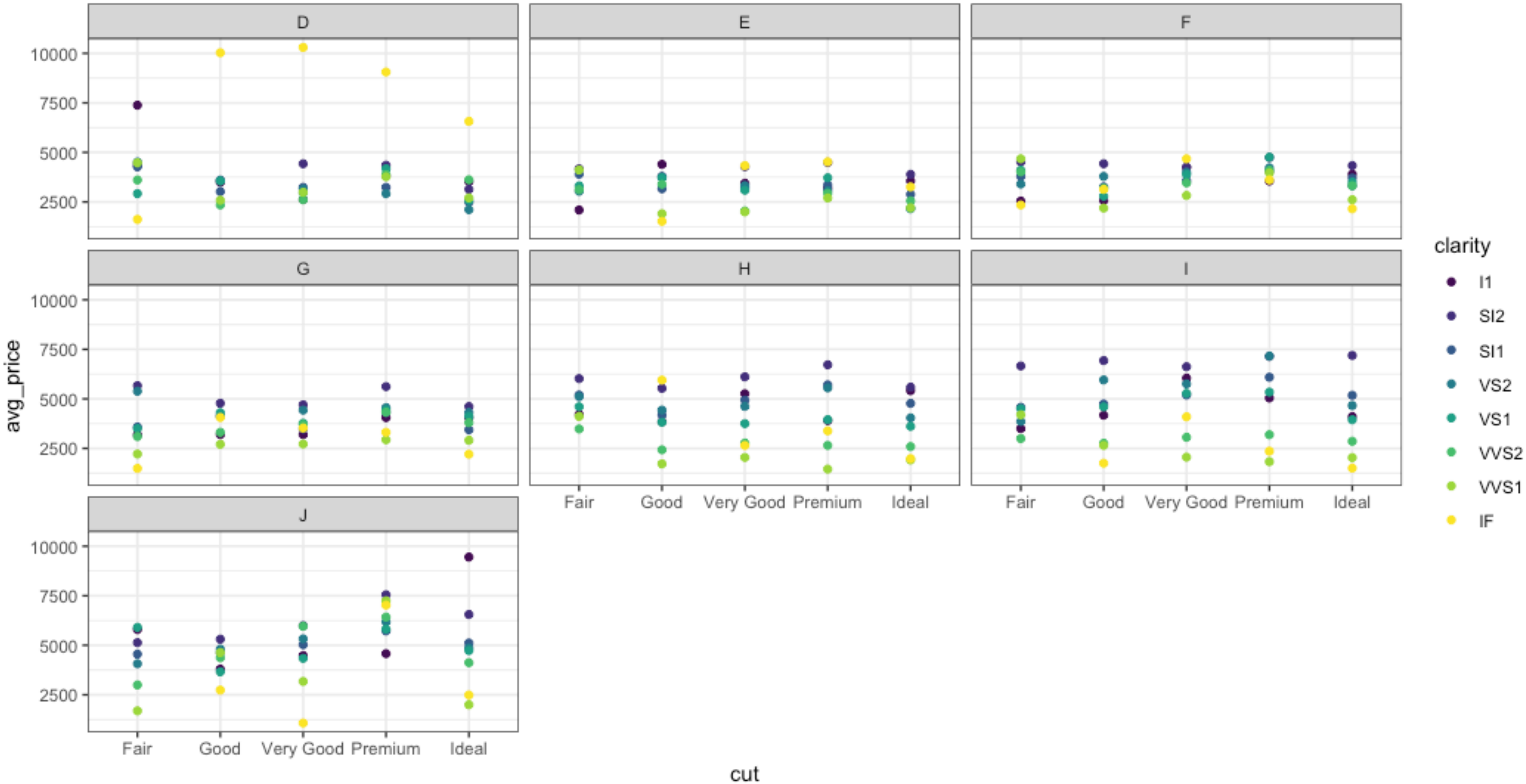
quality of the cut (Fair, Good, Very Good, Premium, Ideal)

color

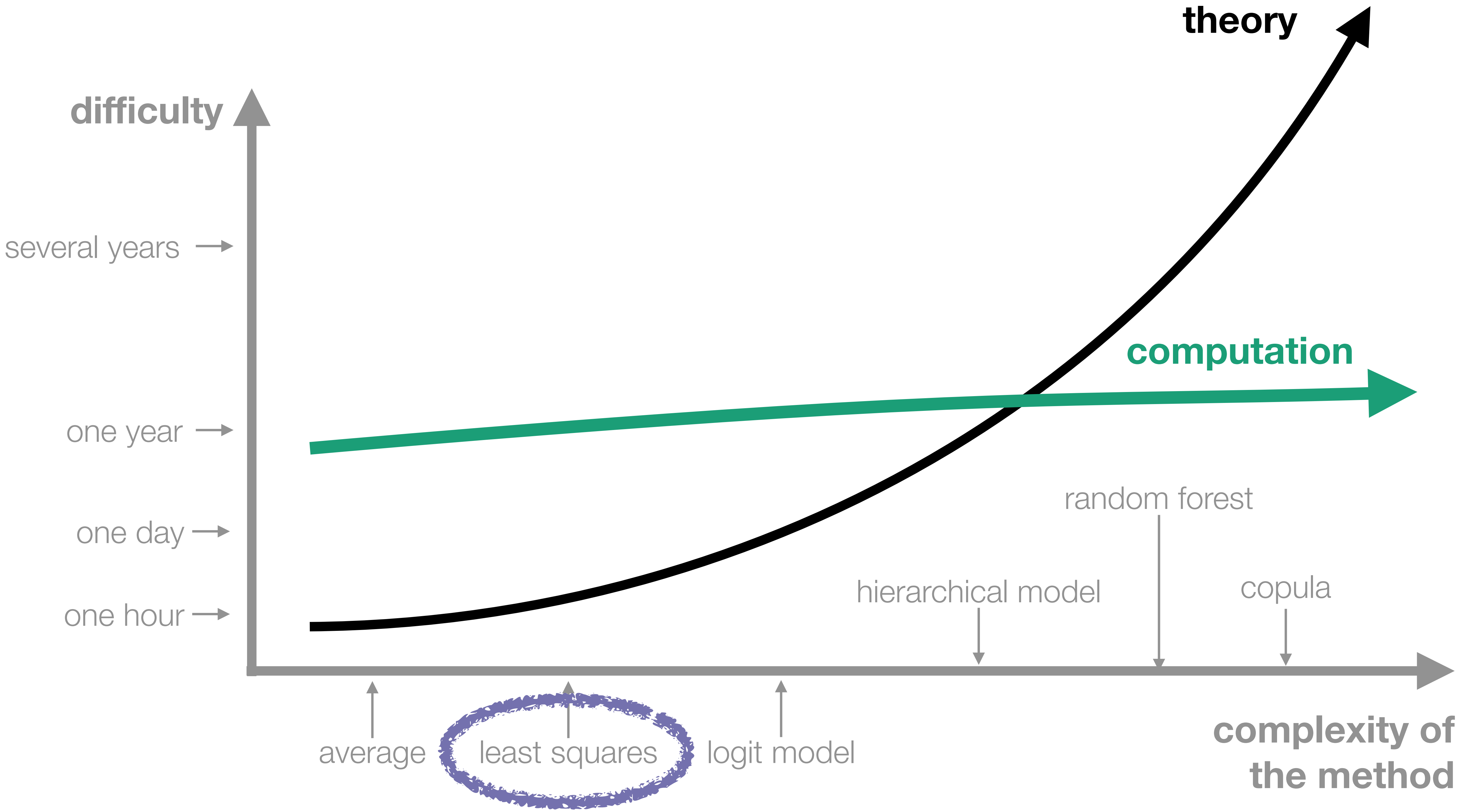
diamond colour, from D (best) to J (worst)

clarity

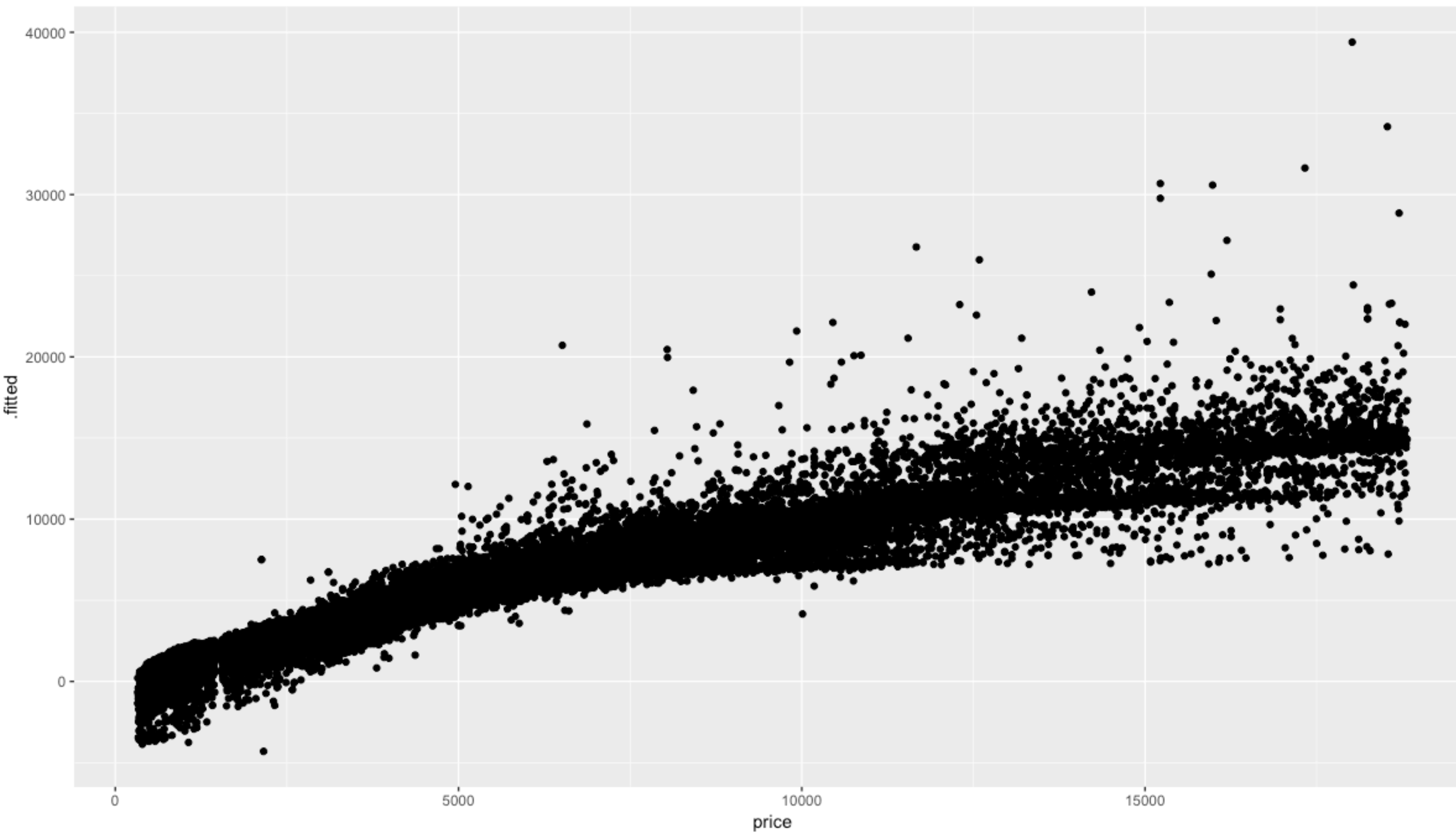
a measurement of how clear the diamond is (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best))


D
E
F
G
H
I
J
10000
7500
5000
2500
avg_price
Fair
Good
Very Good
Premium
Ideal
cut
clarity
I1
SI2
SI1
VS2
VS1
VVS2
VVS1
IF

```r
# load packages
library(tidyverse)

# loads a default data set in R
data(diamonds)

# quick look at the data set
glimpse(diamonds)

# find average for each cut, color, and clarity
sum_df <- diamonds %>%
  group_by(cut, color, clarity) %>%
  summarize(avg_price = mean(price))
sum_df

# plot averages
ggplot(sum_df, aes(x = cut, y = avg_price, color = clarity)) +
  geom_point() +
  facet_wrap( ~ color) +
  theme_bw()

```



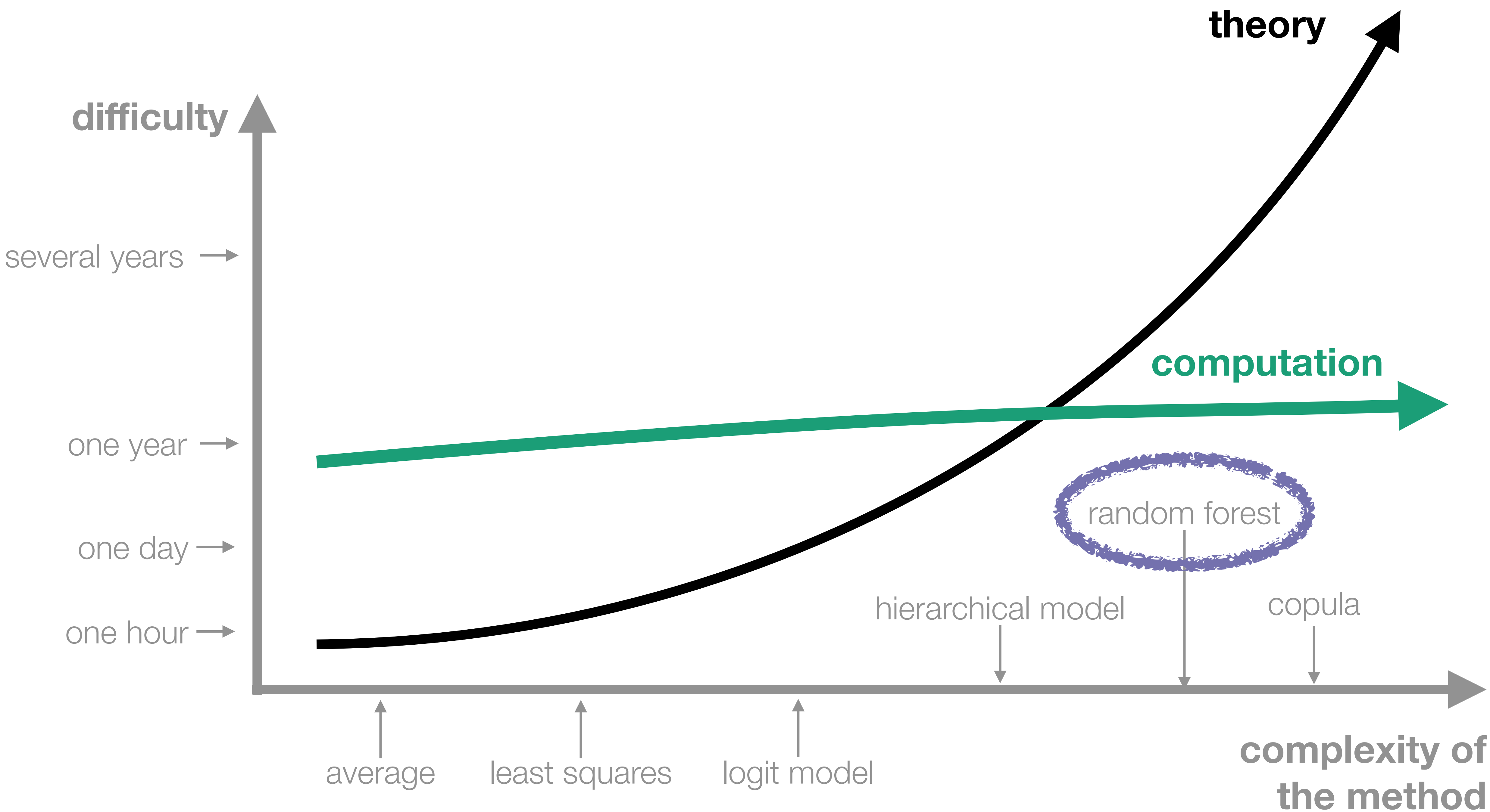
theory
difficulty
several years
computation
one year
random forest
one day
hierarchical model
copula
one hour
average
least squares
logit model
complexity of the method

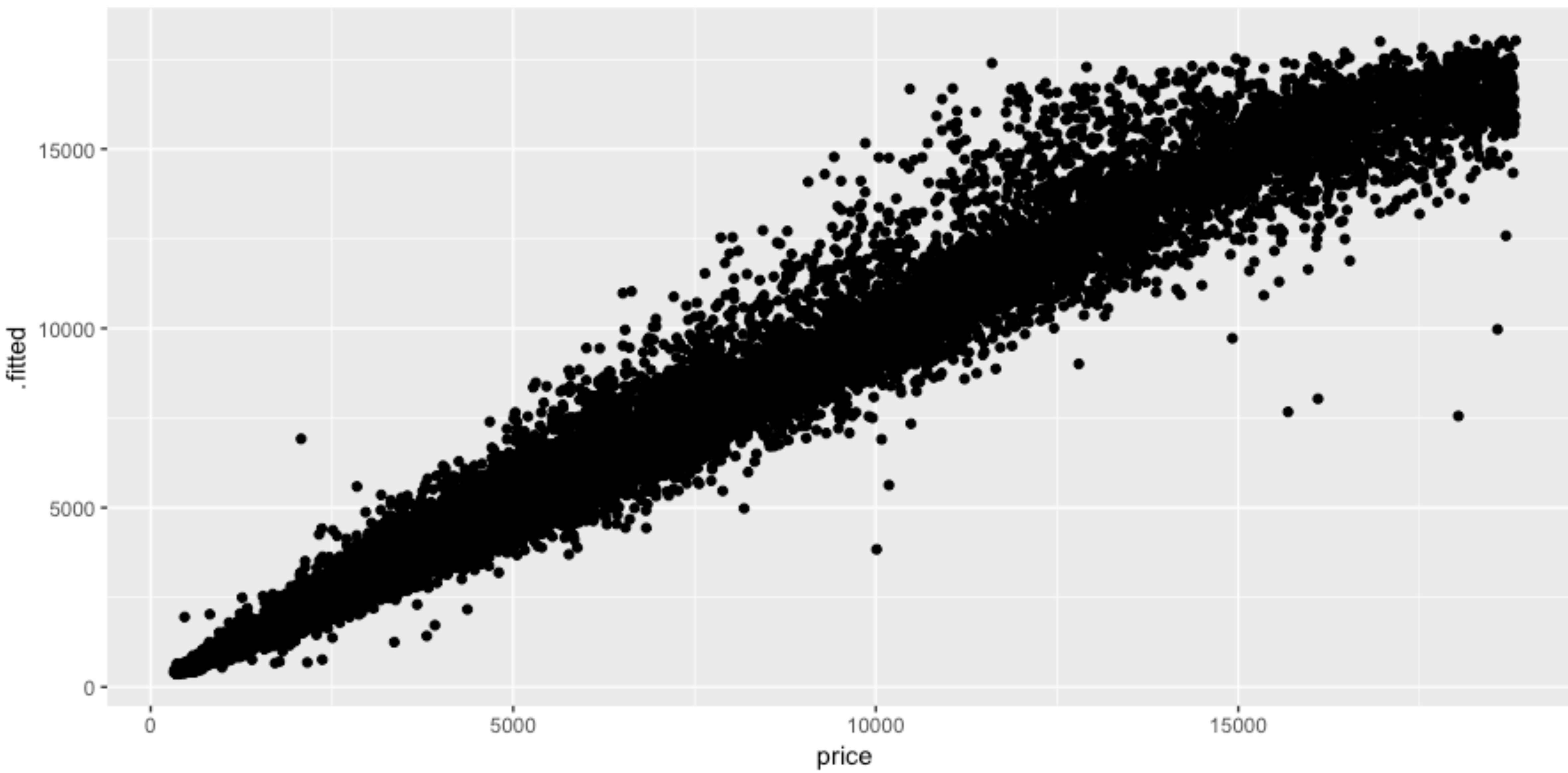


40000
30000
20000
10000
0
.fitted
0
5000
10000
15000
price

```r
# load packages
library(tidyverse)

# loads a default data set in R
data(diamonds)

# quick look at the data set
glimpse(diamonds)

# find average for each cut, color, and clarity
sum_df <- diamonds %>%
  group_by(cut, color, clarity) %>%
  summarize(avg_price = mean(price))
sum_df

# plot averages
ggplot(sum_df, aes(x = cut, y = avg_price, color = clarity)) +
  geom_point() +
  facet_wrap( ~ color) +
  theme_bw()
```

```r
# load packages
library(tidyverse)
library(broom)

# loads a default data set in R
data(diamonds)

# quick look at the data set
glimpse(diamonds)

# fit regression model using all predictors
fit <- lm(price ~ ., data = diamonds)

# tidy fit
diamonds <- augment(fit, diamonds) %>%
  glimpse()

# plot predictions
ggplot(diamonds, aes(x = price, y = .fitted)) +
  geom_point()
```

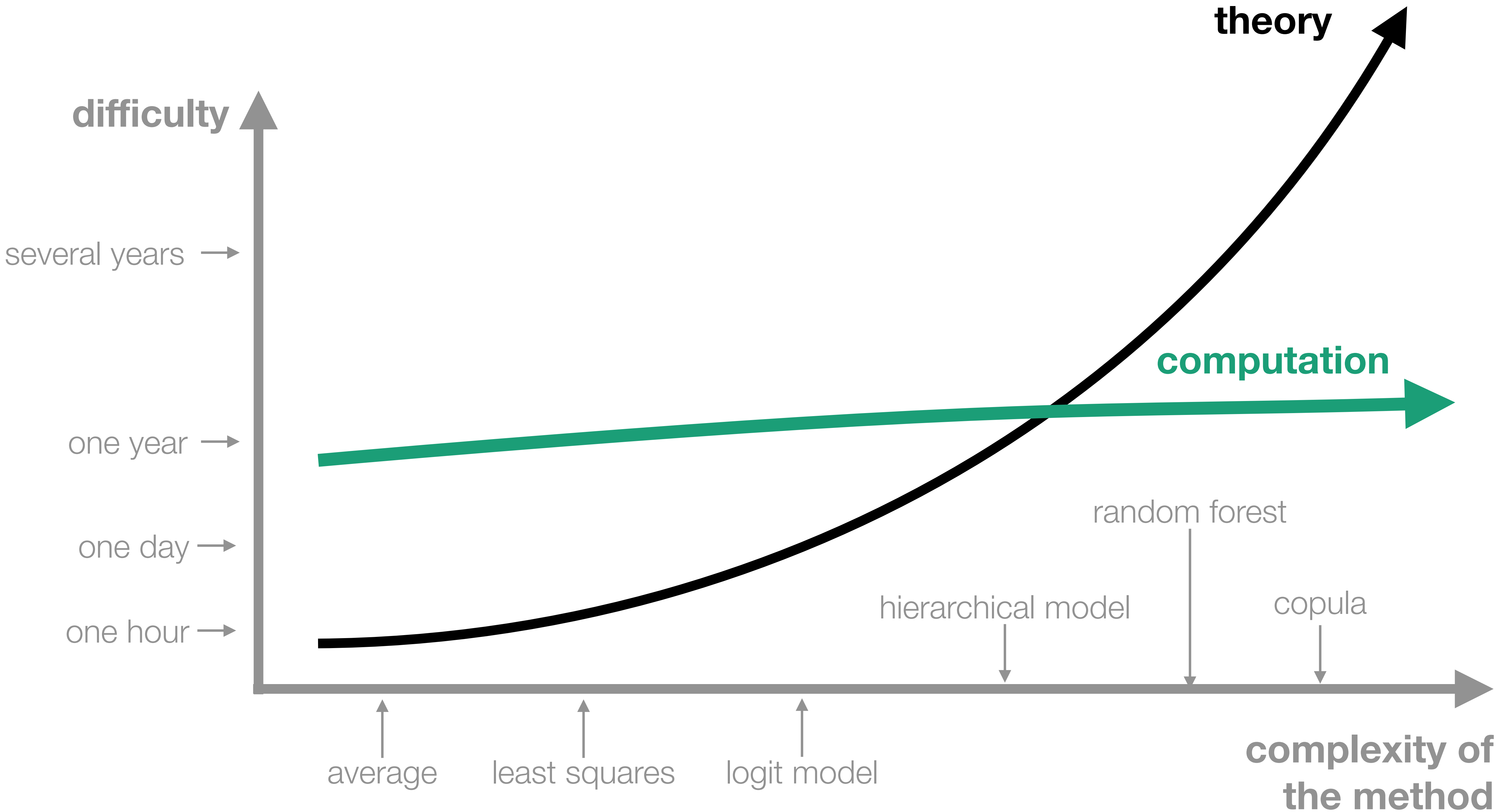

theory
difficulty
several years
computation
one year
random forest
one day
hierarchical model
copula
one hour
average
least squares
logit model
complexity of
the method



.fitted
price
0
5000
10000
15000

```r
# load packages
library(tidyverse)

# loads a default data set in R
data(diamonds)

# quick look at the data set
glimpse(diamonds)

# find average for each cut, color, and clarity
sum_df <- diamonds %>%
  group_by(cut, color, clarity) %>%
  summarize(avg_price = mean(price))
sum_df

# plot averages
ggplot(sum_df, aes(x = cut, y = avg_price, color = clarity)) +
  geom_point() +
  facet_wrap( ~ color) +
  theme_bw()

```

```r
# load packages
library(tidyverse)
library(randomForest)

# loads a default data set in R
data(diamonds)

# quick look at the data set
glimpse(diamonds)

# fit regression model using all predictors
fit <- randomForest(price ~ ., data = diamonds)

# tidy fit
diamonds$.fitted <- predict(fit)

# plot predictions
ggplot(diamonds, aes(x = price, y = .fitted)) +
  geom_point()

```



theory
difficulty
several years
computation
one year
random forest
one day
hierarchical model
copula
one hour
average
least squares
logit model
complexity of
the method

What does a research project look like?

think + write + discuss

What are the three essential components of a great research* project?

*empirical, computational

raw data

the process

manuscript


Electoral Studies

Strategic mobilization: Why proportional representation decreases voter mobilization

Carlisle Rainey

1. Introduction

raw data

the process

100s of decisions

1000s of lines of code

manuscript


Electoral Studies

Strategic mobilization: Why proportional representation decreases voter mobilization

Carlisle Rainey

1. Introduction

think + write + discuss

What characteristics should the process* have?

*from raw data to published paper

raw data

the process

100s of decisions
1000s of lines of code

tidying
modeling
plotting

principled
implemented
documented

manuscript

principled

if you made correct decisions

implemented

if you did what you decided to do

documented

if you can check that you did what you decided to do


I'M EXHAUSTED FROM ALL OF THE BASIC RESEARCH I'M DOING.
Dilbert.com DilbertCartoonist@gmail.com
IT'S TOO BAD THAT THE VALUE OF MY WORK WON'T BE QUANTIFIABLE FOR ANOTHER TEN YEARS.
5-28-10 © 2010 Scott Adams, Inc./Dist. by UFS, Inc.
I'D LIKE TO SEE YOUR LAB REPORT.
SO. . . THE NEW RULE IS THAT WE WRITE DOWN STUFF?

think + write + discuss

principled

implemented

documented

Rank these from least to most challenging.

What makes a research project compelling?

The Process

"The process never ends until we die. And the choices we make are ultimately our own responsibility."

–Eleanor Roosevelt

principled

if you made correct decisions

implemented

if you did what you decided to do

documented

if you can check that you did what you decided to do

principled

sharp intuition

some math

implemented

two sources of errors

errors in a script

software errors

user errors

```
# Make sure that working directory is set properly
# setwd("~/Dropbox/projects/strategic-mobilization/")

# Clear workspace
rm(list = ls())

# Read in the raw data from the CSES Module 2 data set
cses2 <- read.csv("data/cses2_rawdata.txt")

#  Pull out variables of interest
mycses2 <- c("B1004", "B2001", "B2002", "B2003", "B2004", "B2005", "B2020", "B2023", "B2030", "B2031", "B3001_2", "B3002_2",
"B3003", "B3004_1", "B3014", "B3016", "B3028", "B3045", "B3047_1", "B3047_2", "B3047_3", "B4001", "B4002", "B4003", "B4004_A",
"B4004_B", "B4004_C", "B4004_D", "B4004_E", "B4004_F", "B4005", "B5043_1")
cses2  <- cses2[, mycses2]

#   Change the variable names
names(cses2) <- c("Alpha.Polity", "Age", "Male", "Education", "Married", "Union.Member", "Household.Income",
"Religious.Attendance", "Urban", "District", "Campaign.Activities", "Freq.Campaign", "Contacted", "Cast.Ballot", "Vote.Matters",
"Cast.Ballot.Previous", "Close.To.Party", "Ideology", "Know1", "Know2", "Know3", "Number.Seats", "Number.Candidates",
"Number.Lists", "VoteA",  "VoteB", "VoteC", "VoteD", "VoteE", "VoteF", "District.Turnout", "Electoral.Formula")

# Drop countries for which there is not information about the electoral district
cses2 <- cses2[cses2$District!= 99999, ]
cses2 <- cses2[cses2$Number.Seats != 999, ]


#### Recode and Create Variables

# Alpha.Polity

cses2$Alpha.Polity <- as.character(cses2$Alpha.Polity)

cses2$Alpha.Polity[cses2$Alpha.Polity=="CAN_2004"] <- "Canada"
cses2$Alpha.Polity[cses2$Alpha.Polity=="FIN_2003"] <- "Finland"
cses2$Alpha.Polity[cses2$Alpha.Polity=="GBR_2005"] <- "Great Britain"
cses2$Alpha.Polity[cses2$Alpha.Polity=="PRT_2002"] <- "Portugal 2002"
cses2$Alpha.Polity[cses2$Alpha.Polity=="PRT_2005"] <- "Portugal 2005"

cses2 <- cses2[cses2$Alpha.Polity == "Canada" |
                 cses2$Alpha.Polity == "Finland"   |
                 cses2$Alpha.Polity == "Great Britain"  |
                 cses2$Alpha.Polity == "Portugal 2002"|
                 cses2$Alpha.Polity == "Portugal 2005", ]

cses2$Alpha.Polity <- as.factor(cses2$Alpha.Polity)

# Age
```

Pr(correct) < 1

```
cses2$District.Country <- paste(cses2$Alpha.Polity, cses2$District, sep = "")
cses2$District.Country <- as.factor(cses2$District.Country)

District.Names <- sort(unique(cses2$District.Country))
for (i in 1:length(District.Names)) {
  cses2$District[cses2$District.Country == District.Names[i]] <- i
}

######################################
## Save datasets as .csv files       ##
######################################
cses2$District <- as.numeric(as.character(cses2$District))
cses2$Country <- as.numeric(cses2$Alpha.Polity)

# Save a listwise-deleted data set.
ld.vars <- c("Contacted", "Age", "Male", "Education", "Married", "Union.Member", "Household.Income", "Urban", "Close.To.Party",
"District.Competitiveness", "ENEP", "PR", "Alpha.Polity", "District", "Country", "District.Country")
ld.data <- cses2[, ld.vars]
ld.data <- na.omit(ld.data)
write.csv(ld.data, "output/ld-data.csv")

# Save a data set [illegible]issing va[illegible]es for multiple imputation.
mi.vars <- c("Alph[illegible]Pol[illegible]", "[illegible]", "Ma[illegible]", "Ed[illegible]ion", "Ma[illegible]ed", "Union.Member[illegible]", "[illegible]usehold.Income", "R[illegible]ious.Attendance",
"Urban", "District[illegible]pa[illegible].A[illegible]iv[illegible]es", "[illegible]eq.[illegible]am[illegible]gn", "Co[illegible]act[illegible]d", "Ca[illegible].B[illegible]lot[illegible]", "Vote[illegible]ers", "Ca[illegible].Ballot.Previous",
"Close.To.Party", "[illegible]deology", "K[illegible]w1", "Know[illegible]", "K[illegible]w[illegible]", "D[illegible]stri[illegible].Compe[illegible]tivenes[illegible]", "[illegible]R", "N[illegible]er.Seats", "[illegible]EP", "Country",
"District")
mi.data <- cses2[, mi.vars]
write.csv(mi.data, "output/mi-data.csv")



# Create the district-level data
get.first <- function(x) {
  return(x[1])
}

district.data <- cses2[, c("Alpha.Polity", "Country", "District", "District.Competitiveness", "PR")]
district.data <- aggregate(district.data, by = list(cses2$District), FUN = get.first)
district.data$SMDP <- 1 - district.data$PR
write.csv(district.data, "output/district-data.csv")

# Create the country-level data
country.data <- cses2[, c("Alpha.Polity", "Country", "PR")]
country.data <- aggregate(country.data, by = list(cses2$Country), FUN = get.first)
country.data$SMDP <- 1 - country.data$PR
write.csv(country.data, "output/country-data.csv")
```

two sources of errors

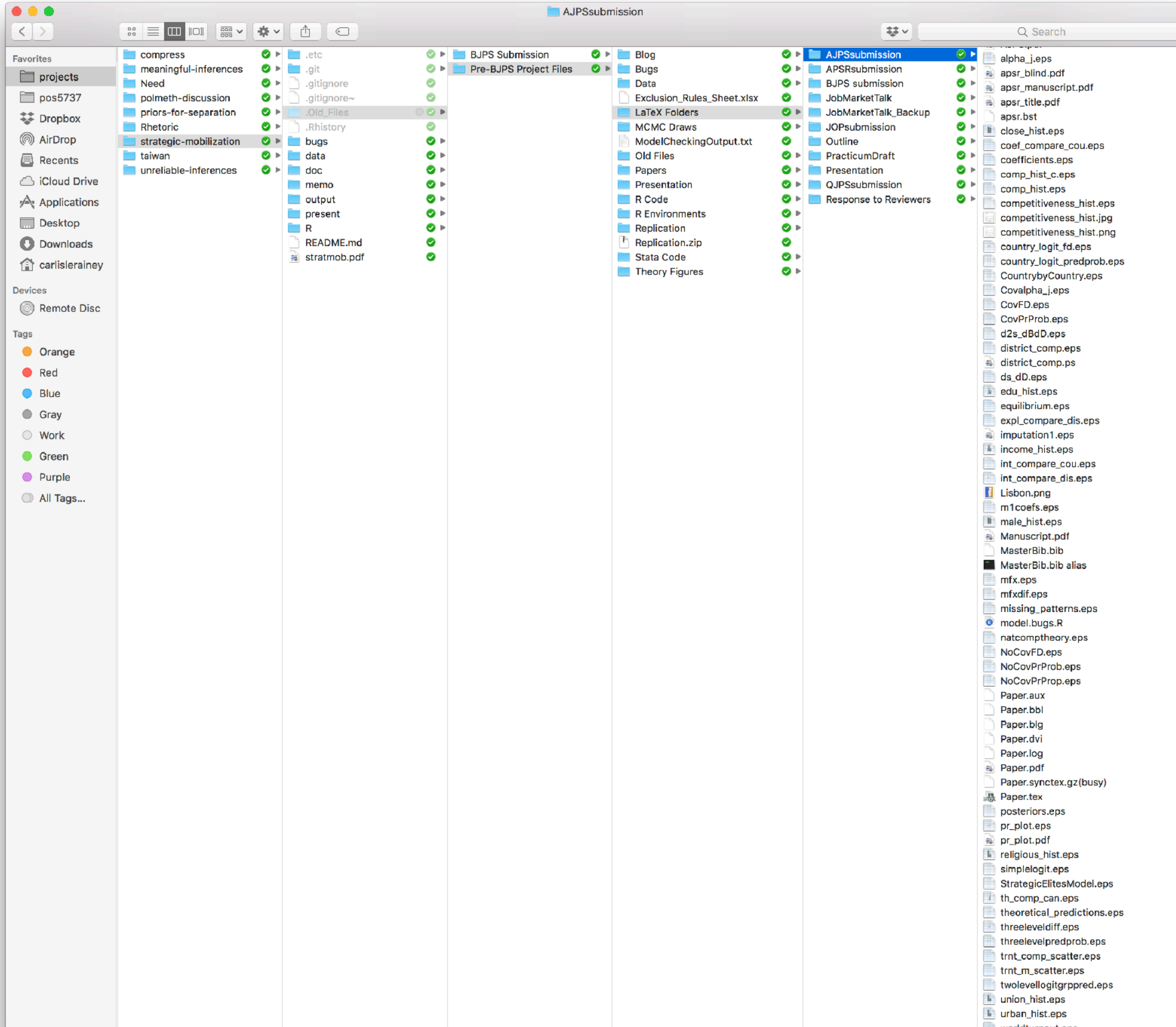
errors in a script

software errors

user errors

mismanage your files

mismanage versions


AJPSsubmission
Search
Favorites
projects
pos5737
Dropbox
AirDrop
Recents
iCloud Drive
Applications
Desktop
Downloads
carlislerainey
Devices
Remote Disc
Tags
Orange
Red
Blue
Gray
Work
Green
Purple
All Tags...
compress
meaningful-inferences
Need
polmeth-discussion
priors-for-separation
Rhetoric
strategic-mobilization
taiwan
unreliable-inferences
.etc
.git
.gitignore
.gitignore~
.Old_Files
.Rhistory
bugs
data
doc
memo
output
present
R
README.md
stratmob.pdf
BJPS Submission
Pre-BJPS Project Files
Blog
Bugs
Data
Exclusion_Rules_Sheet.xlsx
LaTeX Folders
MCMC Draws
ModelCheckingOutput.txt
Old Files
Papers
Presentation
R Code
R Environments
Replication
Replication.zip
Stata Code
Theory Figures
AJPSsubmission
APSRsubmission
BJPS submission
JobMarketTalk
JobMarketTalk_Backup
JOPsubmission
Outline
PracticumDraft
Presentation
QJPSsubmission
Response to Reviewers
alpha_j.eps
apsr_blind.pdf
apsr_manuscript.pdf
apsr_title.pdf
apsr.bst
close_hist.eps
coef_compare_cou.eps
coefficients.eps
comp_hist_c.eps
comp_hist.eps
competitiveness_hist.eps
competitiveness_hist.jpg
competitiveness_hist.png
country_logit_fd.eps
country_logit_predprob.eps
CountrybyCountry.eps
Covalpha_j.eps
CovFD.eps
CovPrProb.eps
d2s_dBdD.eps
district_comp.eps
district_comp.ps
ds_dD.eps
edu_hist.eps
equilibrium.eps
expl_compare_dis.eps
imputation1.eps
income_hist.eps
int_compare_cou.eps
int_compare_dis.eps
Lisbon.png
m1coefs.eps
male_hist.eps
Manuscript.pdf
MasterBib.bib
MasterBib.bib alias
mfx.eps
mfxdif.eps
missing_patterns.eps
model.bugs.R
natcomptheory.eps
NoCovFD.eps
NoCovPrProb.eps
NoCovPrProp.eps
Paper.aux
Paper.bbl
Paper.blg
Paper.dvi
Paper.log
Paper.pdf
Paper.synctex.gz(busy)
Paper.tex
posteriors.eps
pr_plot.eps
pr_plot.pdf
religious_hist.eps
simplelogit.eps
StrategicElitesModel.eps
th_comp_can.eps
theoretical_predictions.eps
threeleveldiff.eps
threelevelpredprob.eps
trnt_comp_scatter.eps
trnt_m_scatter.eps
twolevellogitgrppred.eps
union_hist.eps
urban_hist.eps
worldturnout.eps

fit-model-v4.2_withBayes-add_GDP-APSRrevisions-(for Scott).R

two sources of errors

errors in a script

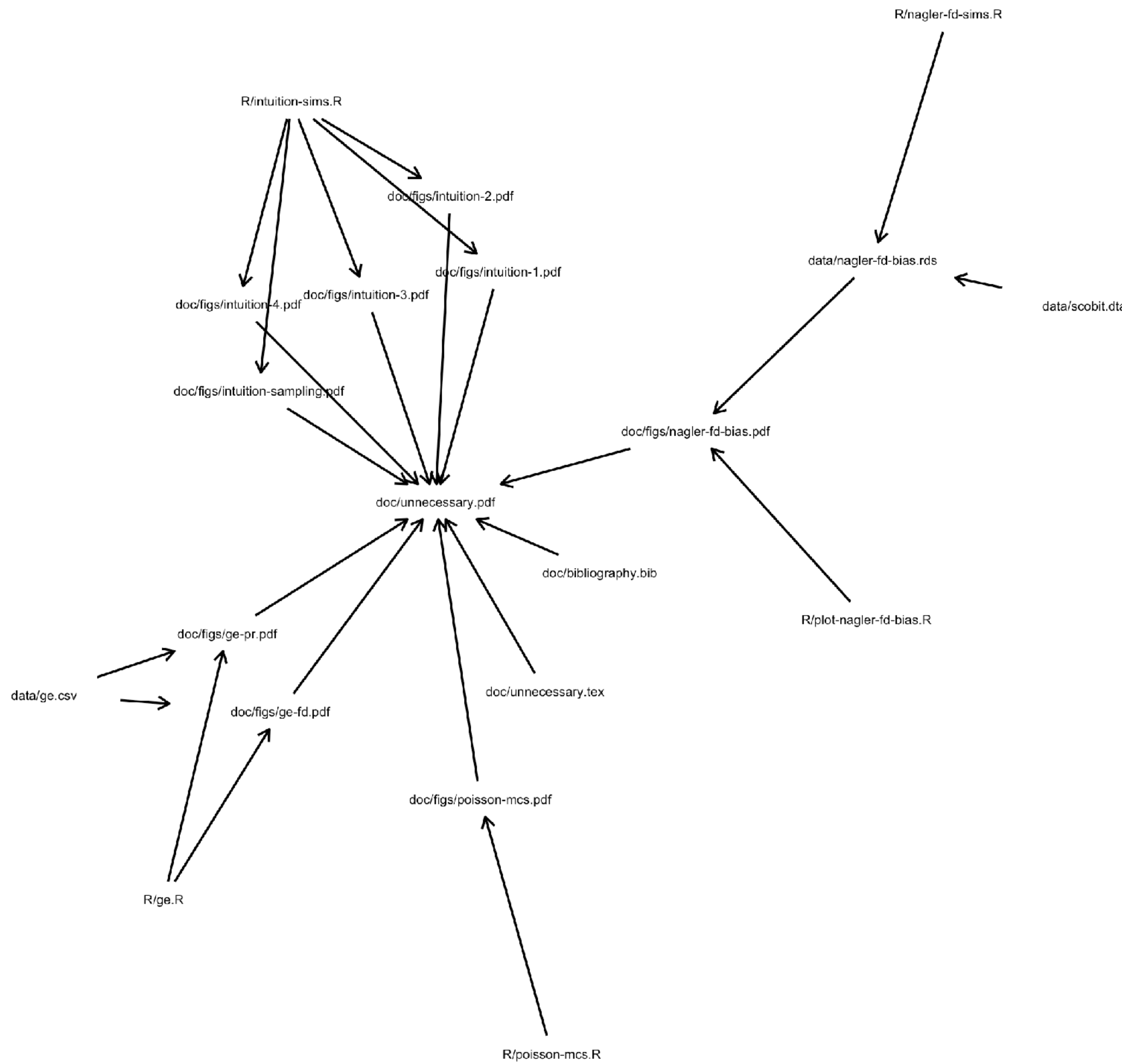
software errors

user errors

mismanage your files

mismanage versions

mismanage dependencies


R/nagler-fd-sims.R
R/intuition-sims.R
doc/figs/intuition-2.pdf
data/nagler-fd-bias.rds
doc/figs/intuition-1.pdf
doc/figs/intuition-3.pdf
doc/figs/intuition-4.pdf
data/scobit.dta
doc/figs/intuition-sampling.pdf
doc/figs/nagler-fd-bias.pdf
doc/unnecessary.pdf
doc/bibliography.bib
R/plot-nagler-fd-bias.R
doc/figs/ge-pr.pdf
data/ge.csv
doc/unnecessary.tex
doc/figs/ge-fd.pdf
doc/figs/poisson-mcs.pdf
R/ge.R
R/poisson-mcs.R

documented

share your work

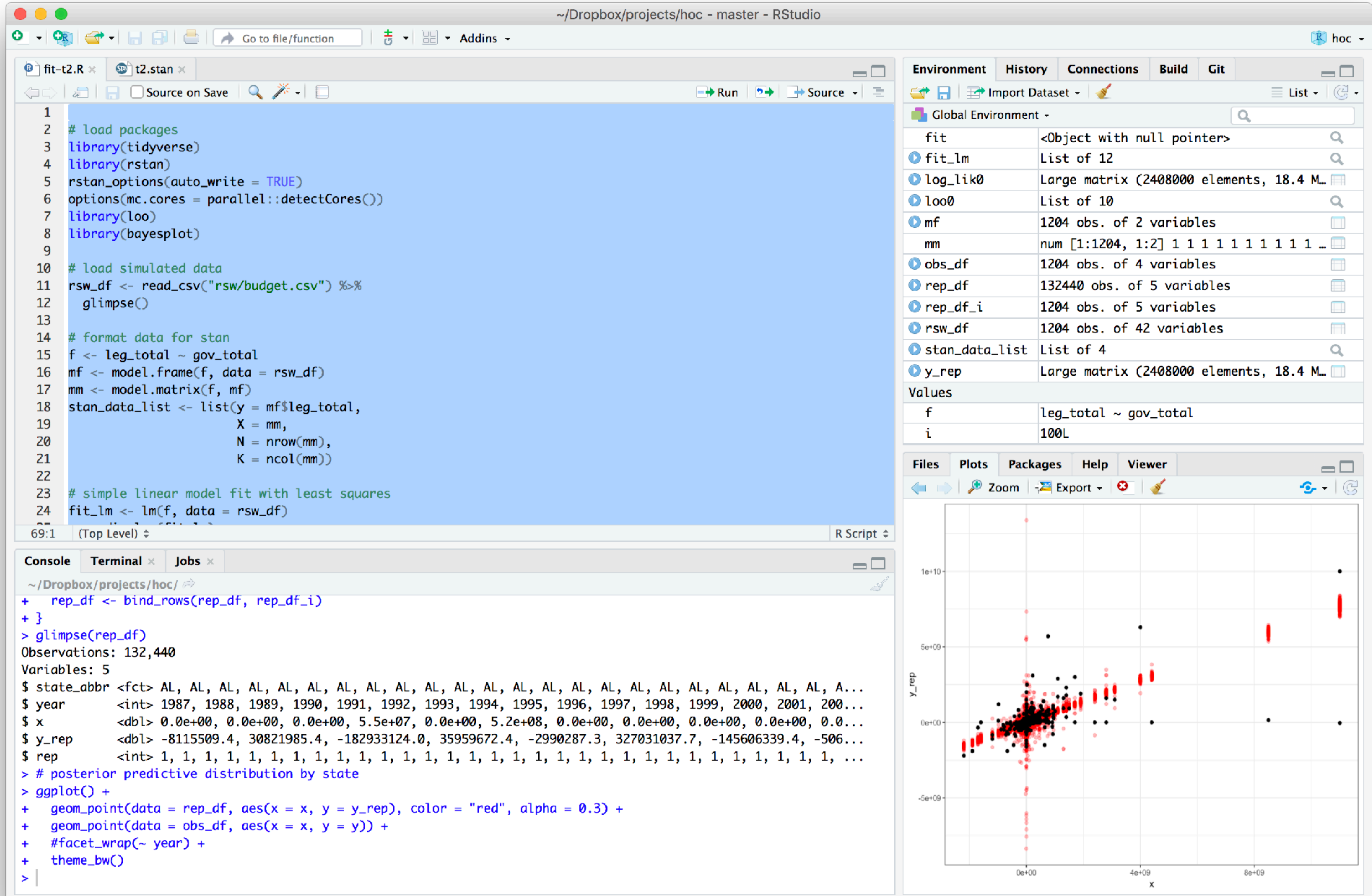
Carlisle's Fundamental Theorem of Implementation (CFTI)
The same strategies that allow others to easily check your work (1) allow you to easily check your work and (2) ensure that you implement your decisions correctly in the first place.

What tools allow me to execute a compelling research project?

Our Tools

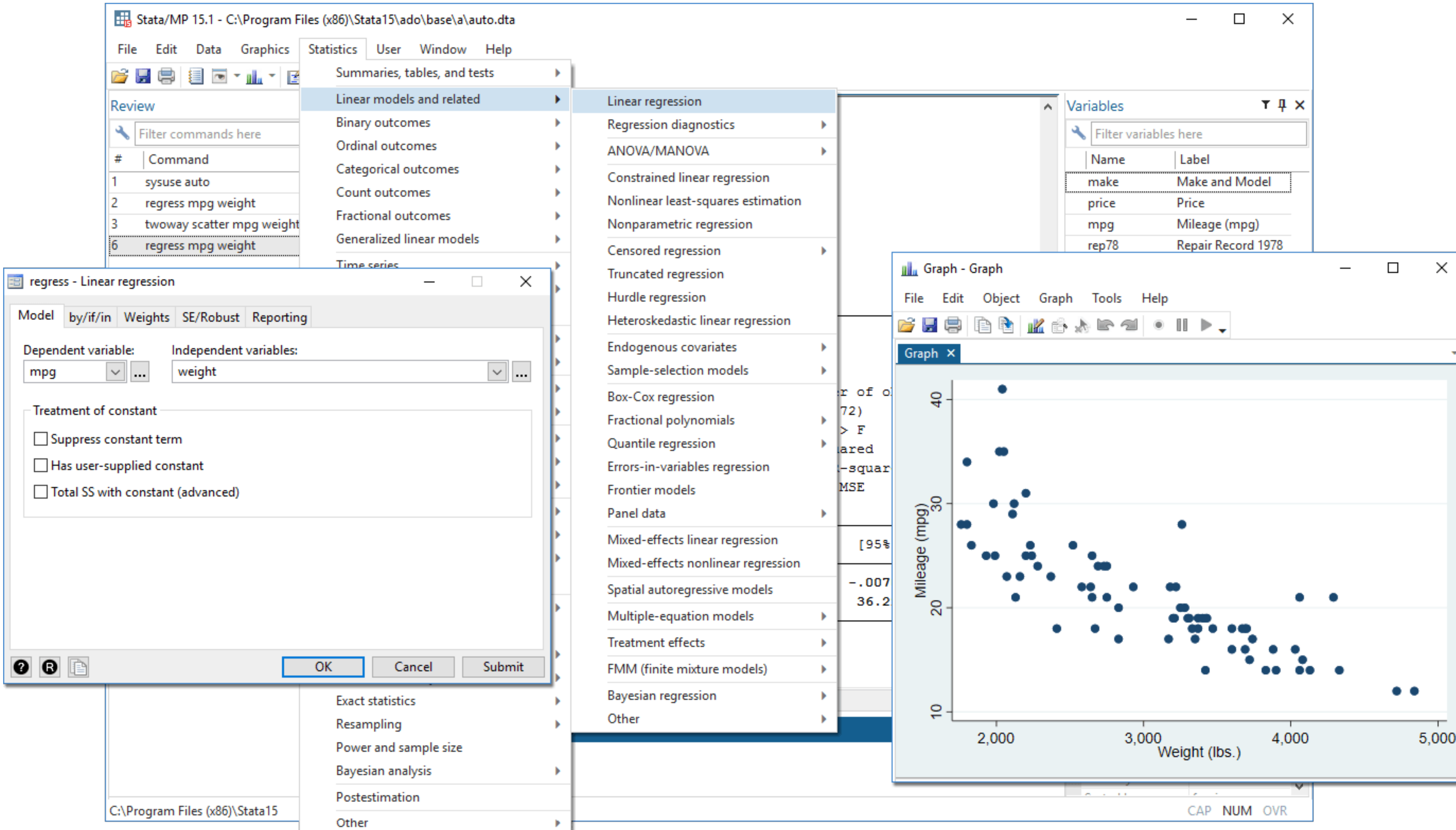
statistical computing

RStudio®


~/Dropbox/projects/hoc - master - RStudio
Go to file/function
Addins
hoc
fit-t2.R
t2.stan
Source on Save
Run
Source
# load packages
library(tidyverse)
library(rstan)
rstan_options(auto_write = TRUE)
options(mc.cores = parallel::detectCores())
library(loo)
library(bayesplot)
# load simulated data
rsw_df <- read_csv("rsw/budget.csv") %>%
glimpse()
# format data for stan
f <- leg_total ~ gov_total
mf <- model.frame(f, data = rsw_df)
mm <- model.matrix(f, mf)
stan_data_list <- list(y = mf$leg_total,
X = mm,
N = nrow(mm),
K = ncol(mm))
# simple linear model fit with least squares
fit_lm <- lm(f, data = rsw_df)
69:1
(Top Level)
R Script
Environment
History
Connections
Build
Git
Import Dataset
List
Global Environment
fit <Object with null pointer>
fit_lm List of 12
log_lik0 Large matrix (2408000 elements, 18.4 M…
loo0 List of 10
mf 1204 obs. of 2 variables
mm num [1:1204, 1:2] 1 1 1 1 1 1 1 1 1 1 …
obs_df 1204 obs. of 4 variables
rep_df 132440 obs. of 5 variables
rep_df_i 1204 obs. of 5 variables
rsw_df 1204 obs. of 42 variables
stan_data_list List of 4
y_rep Large matrix (2408000 elements, 18.4 M…
Values
f leg_total ~ gov_total
i 100L
Files
Plots
Packages
Help
Viewer
Zoom
Export
Console
Terminal
Jobs
~/Dropbox/projects/hoc/
+ rep_df <- bind_rows(rep_df, rep_df_i)
+ }
> glimpse(rep_df)
Observations: 132,440
Variables: 5
$ state_abbr <fct> AL, AL, AL, AL, AL, AL, AL, AL, AL, AL, AL, AL, AL, AL, AL, AL, AL, AL, AL, AL, AL, AL, AL, A...
$ year <int> 1987, 1988, 1989, 1990, 1991, 1992, 1993, 1994, 1995, 1996, 1997, 1998, 1999, 2000, 2001, 200...
$ x <dbl> 0.0e+00, 0.0e+00, 0.0e+00, 5.5e+07, 0.0e+00, 5.2e+08, 0.0e+00, 0.0e+00, 0.0e+00, 0.0e+00, 0.0...
$ y_rep <dbl> -8115509.4, 30821985.4, -182933124.0, 35959672.4, -2990287.3, 327031037.7, -145606339.4, -506...
$ rep <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
> # posterior predictive distribution by state
> ggplot() +
+ geom_point(data = rep_df, aes(x = x, y = y_rep), color = "red", alpha = 0.3) +
+ geom_point(data = obs_df, aes(x = x, y = y)) +
+ #facet_wrap(~ year) +
+ theme_bw()
>
1e+10
5e+09
0e+00
-5e+09
y_rep
0e+00
4e+09
8e+09
x


readr
haven
stringr
readxl
R Studio
%>%
Ceci n'est pas un pipe.
forcats
broom
tidyr
dplyr
lubridate
ggplot2

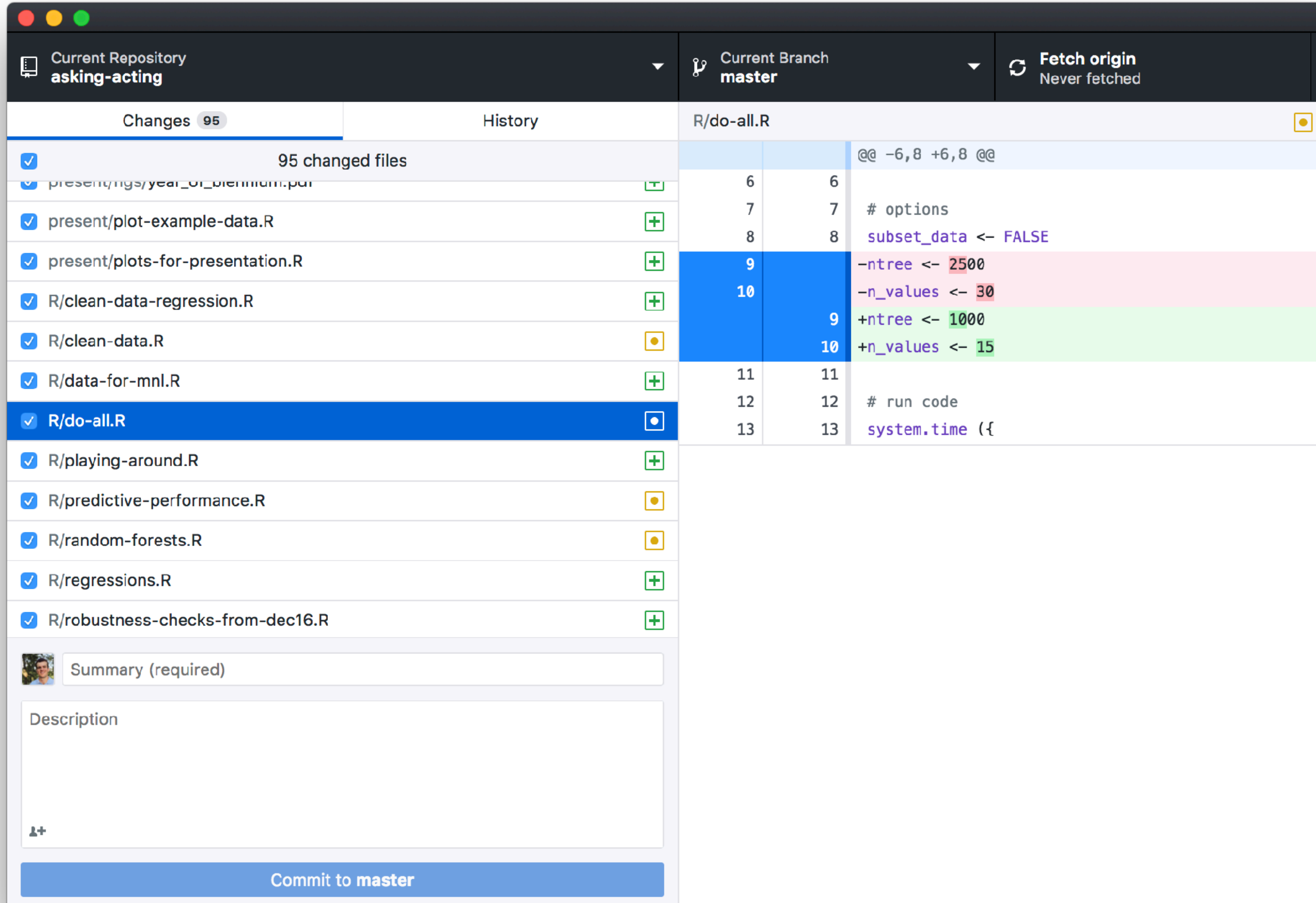
STATA


Stata/MP 15.1 - C:\Program Files (x86)\Stata15\ado\base\a\auto.dta
File
Edit
Data
Graphics
Statistics
User
Window
Help
Review
Filter commands here
#
Command
1 sysuse auto
2 regress mpg weight
3 twoway scatter mpg weight
6 regress mpg weight
Summaries, tables, and tests
Linear models and related
Binary outcomes
Ordinal outcomes
Categorical outcomes
Count outcomes
Fractional outcomes
Generalized linear models
Exact statistics
Resampling
Power and sample size
Bayesian analysis
Postestimation
Other
Linear regression
Regression diagnostics
ANOVA/MANOVA
Constrained linear regression
Nonlinear least-squares estimation
Nonparametric regression
Censored regression
Truncated regression
Hurdle regression
Heteroskedastic linear regression
Endogenous covariates
Sample-selection models
Box-Cox regression
Fractional polynomials
Quantile regression
Errors-in-variables regression
Frontier models
Panel data
Mixed-effects linear regression
Mixed-effects nonlinear regression
Spatial autoregressive models
Multiple-equation models
Treatment effects
FMM (finite mixture models)
Bayesian regression
Other
Variables
Filter variables here
Name
Label
make Make and Model
price Price
mpg Mileage (mpg)
rep78 Repair Record 1978
regress - Linear regression
Model
by/if/in
Weights
SE/Robust
Reporting
Dependent variable:
mpg
Independent variables:
weight
Treatment of constant
Suppress constant term
Has user-supplied constant
Total SS with constant (advanced)
OK
Cancel
Submit
Graph - Graph
File
Edit
Object
Graph
Tools
Help
Graph
Mileage (mpg)
Weight (lbs.)
10
20
30
40
2,000
3,000
4,000
5,000
C:\Program Files (x86)\Stata15
CAP
NUM
OVR

version-controlling

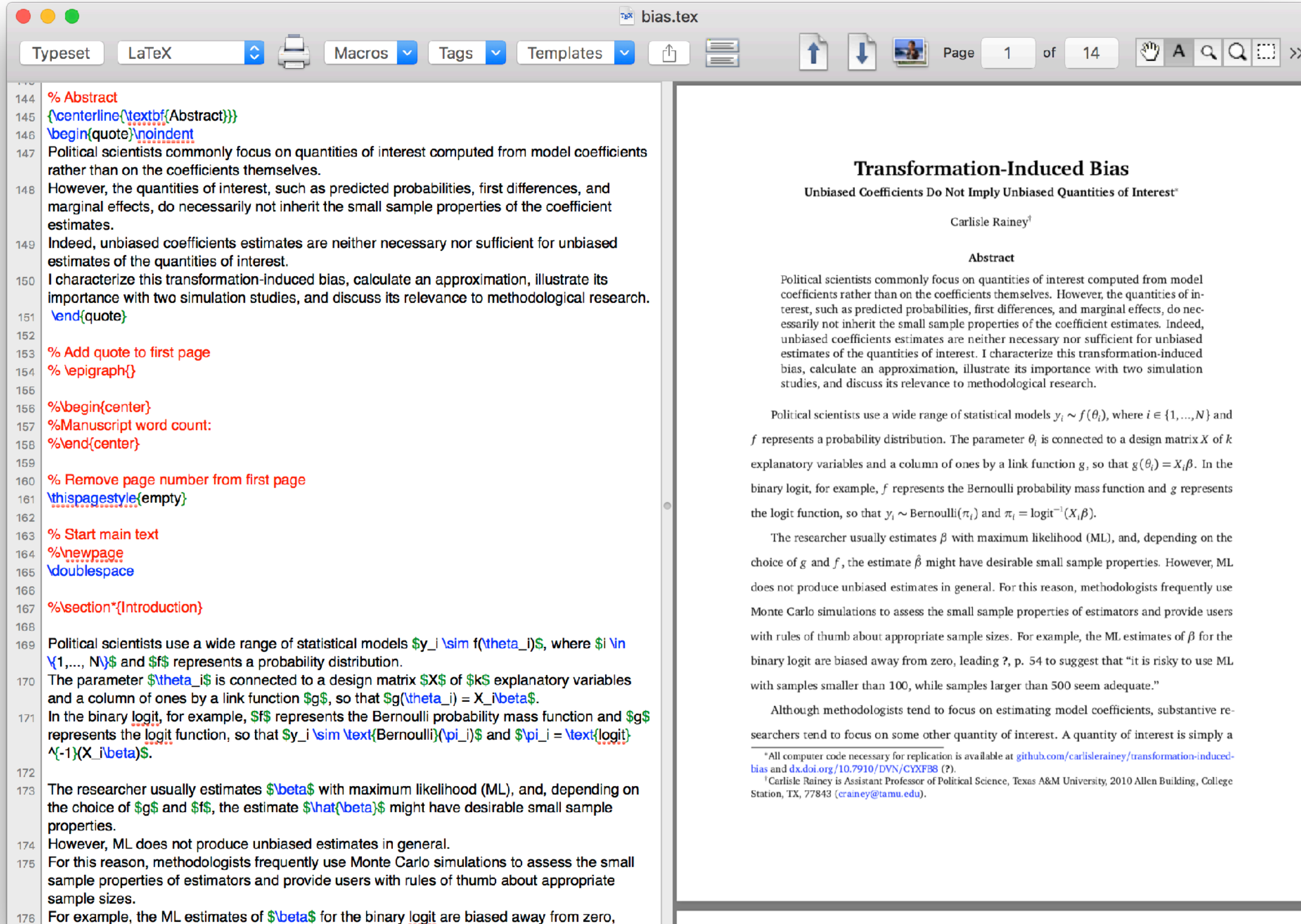
git

GitHub Desktop


Current Repository
asking-acting
Current Branch
master
Fetch origin
Never fetched
Changes 95
History
R/do-all.R
95 changed files
present/plot-example-data.R
present/plots-for-presentation.R
R/clean-data-regression.R
R/clean-data.R
R/data-for-mnl.R
R/do-all.R
R/playing-around.R
R/predictive-performance.R
R/random-forests.R
R/regressions.R
R/robustness-checks-from-dec16.R
Summary (required)
Description
Commit to master
@@ -6,8 +6,8 @@
# options
subset_data <- FALSE
-ntree <- 2500
-n_values <- 30
+ntree <- 1000
+n_values <- 15
# run code
system.time ({

creating documents

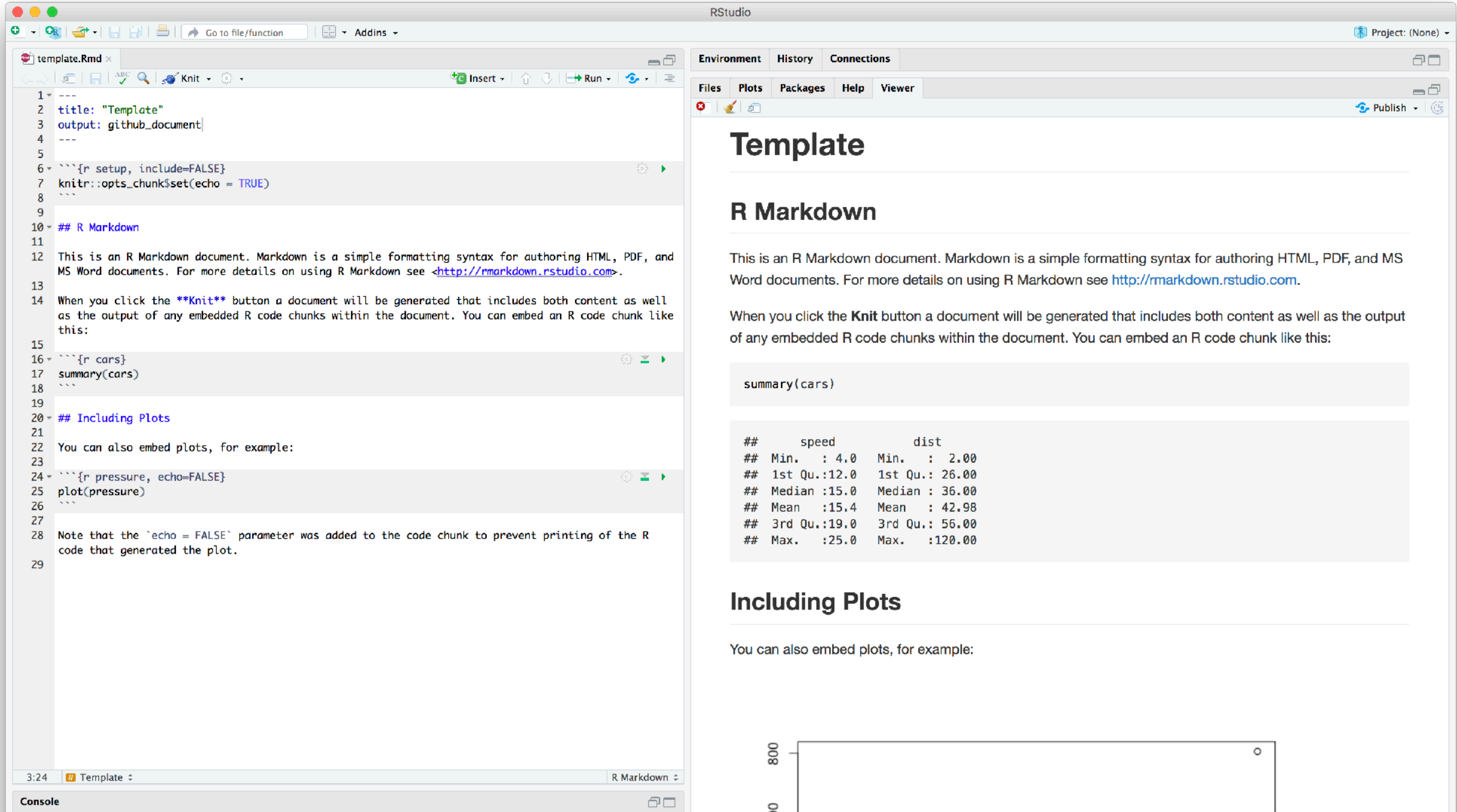
LaTeX



bias.tex
Typeset
LaTeX
Macros
Tags
Templates
Page 1 of 14
% Abstract
{\centerline{\textbf{Abstract}}}
\begin{quote}\noindent
Political scientists commonly focus on quantities of interest computed from model coefficients rather than on the coefficients themselves.
However, the quantities of interest, such as predicted probabilities, first differences, and marginal effects, do necessarily not inherit the small sample properties of the coefficient estimates.
Indeed, unbiased coefficients estimates are neither necessary nor sufficient for unbiased estimates of the quantities of interest.
I characterize this transformation-induced bias, calculate an approximation, illustrate its importance with two simulation studies, and discuss its relevance to methodological research.
\end{quote}
% Add quote to first page
% \epigraph{}
%\begin{center}
%Manuscript word count:
%\end{center}
% Remove page number from first page
\thispagestyle{empty}
% Start main text
%\newpage
\doublespace
%\section*{Introduction}
Political scientists use a wide range of statistical models $y_i \sim f(\theta_i)$, where $i \in \{1,..., N\}$ and f represents a probability distribution.
The parameter θ_i is connected to a design matrix X of k explanatory variables and a column of ones by a link function g, so that $g(\theta_i) = X_i\beta$.
In the binary logit, for example, f represents the Bernoulli probability mass function and g represents the logit function, so that $y_i \sim \text{Bernoulli}(\pi_i)$ and $\pi_i = \text{logit}^{-1}(X_i\beta)$.
The researcher usually estimates β with maximum likelihood (ML), and, depending on the choice of g and f, the estimate $\hat{\beta}$ might have desirable small sample properties.
However, ML does not produce unbiased estimates in general.
For this reason, methodologists frequently use Monte Carlo simulations to assess the small sample properties of estimators and provide users with rules of thumb about appropriate sample sizes.
For example, the ML estimates of β for the binary logit are biased away from zero,
Transformation-Induced Bias
Unbiased Coefficients Do Not Imply Unbiased Quantities of Interest*
Carlisle Rainey†
Abstract
Political scientists commonly focus on quantities of interest computed from model coefficients rather than on the coefficients themselves. However, the quantities of interest, such as predicted probabilities, first differences, and marginal effects, do necessarily not inherit the small sample properties of the coefficient estimates. Indeed, unbiased coefficients estimates are neither necessary nor sufficient for unbiased estimates of the quantities of interest. I characterize this transformation-induced bias, calculate an approximation, illustrate its importance with two simulation studies, and discuss its relevance to methodological research.
Political scientists use a wide range of statistical models $y_i \sim f(\theta_i)$, where $i \in \{1, \ldots, N\}$ and f represents a probability distribution. The parameter θ_i is connected to a design matrix X of k explanatory variables and a column of ones by a link function g, so that $g(\theta_i) = X_i\beta$. In the binary logit, for example, f represents the Bernoulli probability mass function and g represents the logit function, so that $y_i \sim \text{Bernoulli}(\pi_i)$ and $\pi_i = \text{logit}^{-1}(X_i\beta)$.
The researcher usually estimates β with maximum likelihood (ML), and, depending on the choice of g and f, the estimate $\hat{\beta}$ might have desirable small sample properties. However, ML does not produce unbiased estimates in general. For this reason, methodologists frequently use Monte Carlo simulations to assess the small sample properties of estimators and provide users with rules of thumb about appropriate sample sizes. For example, the ML estimates of β for the binary logit are biased away from zero, leading ?, p. 54 to suggest that “it is risky to use ML with samples smaller than 100, while samples larger than 500 seem adequate.”
Although methodologists tend to focus on estimating model coefficients, substantive researchers tend to focus on some other quantity of interest. A quantity of interest is simply a
*All computer code necessary for replication is available at github.com/carlislerainey/transformation-induced-bias and dx.doi.org/10.7910/DVN/CYXFB8 (?).
†Carlisle Rainey is Assistant Professor of Political Science, Texas A&M University, 2010 Allen Building, College Station, TX, 77843 (crainey@tamu.edu).

rmarkdown

www.rstudio.com


RStudio
Go to file/function
Addins
Project: (None)
template.Rmd
Knit
Insert
Run
---
title: "Template"
output: github_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)
```
Including Plots
You can also embed plots, for example:
```{r pressure, echo=FALSE}
plot(pressure)
```
Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.
3:24
Template
R Markdown
Console
Environment
History
Connections
Files
Plots
Packages
Help
Viewer
Publish
Template
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:
800

W

Alright, what’s the first homework?

Homework 1: Intro

- **Conceptual Homework**: Several readings and exercises; data sets, research design, computational research
- **Computational Homework**
  - Part 1: Installing and testing software (long and tedious)
  - Part 2: Practice making a data set
  - Part 3: Practice loading a data set
- **Reflection**: What did you learn?

At some point, come to my office and ask a question.