Improving Reproducibility in Quantitative Social Sciences

Establishing a Simulation-Based Workflow Enhanced with Large Language Models

Rohan Alexander, 6 March 2024, Schwartz Reisman Institute for Technology and Society

Thanks to Ryan Briggs, Carlisle Rainey, and Inessa De Angelis for helpful pointers.

contact@rohanalexander.com rohanalexander.com

Agenda

Part 1: *Background*

In the quantitative social sciences we attempt to use data to help understand some aspect of society. A credibility turn occurred about a generation ago, and since then quantitative social sciences disciplines have experienced increasing homogeneity. There have been attempts to address common issues that previously hampered the believability of results, including making data and code available through replication packages, and insisting on methodological transparency through pre-registration.

Part 2: *Workflows*

However, the centrality of code has not yet been fully internalized in quantitative social sciences. To adjust and enable best-practices from software engineering, the way that quantitative social scientists approach their work needs to change. The use of realistic simulated datasets enables test-driven practices and validation of scientific conclusions in quantitative social sciences.

Part 3: *Adopting AI*

Quantitative social scientists are not unique in being slow to change. The use of LLMs can help drive the adoption of better practice despite cultural issues that might otherwise take generations to resolve.

Key take-aways

1. Quantitative social sciences, such as economics and political science are now as code-dependent as any other software engineering application.

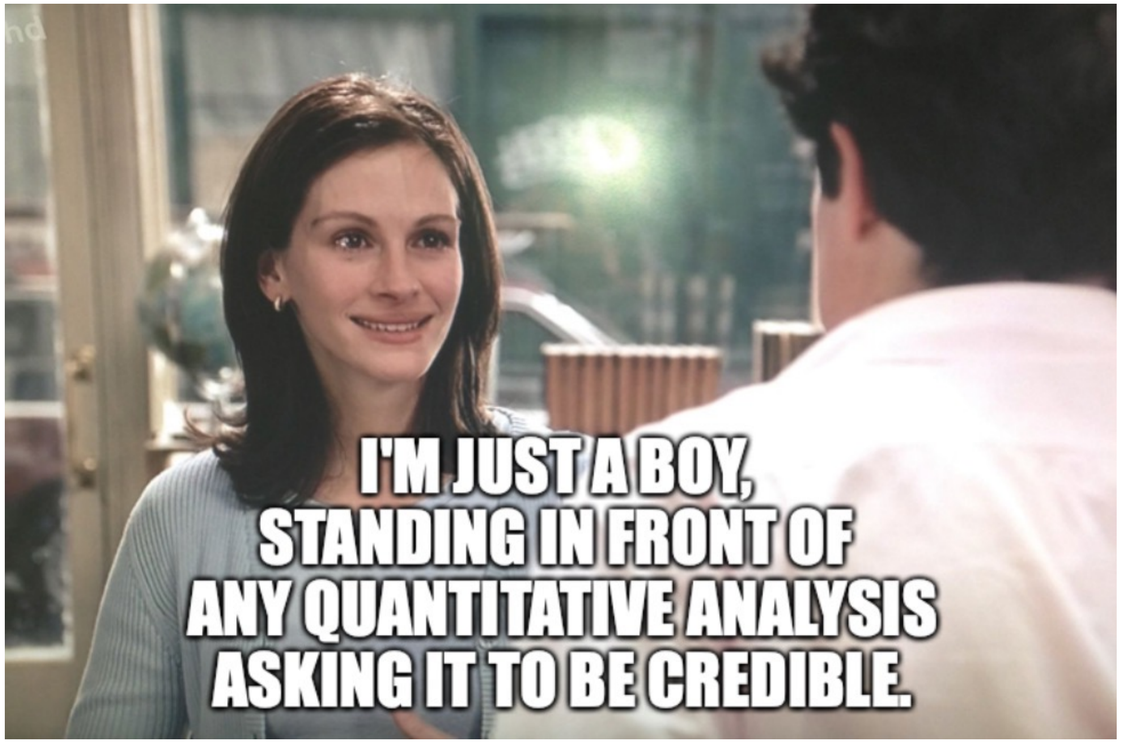
2. Considerable improvements in reproducibility including pre-registration and replication packages, have occurred over the past decade, largely driven by a handful of individuals. There is evidence suggesting these have improved things.

3. But practice and training around code has not evolved, largely due to culture and incentives.

4. We can use a multidisciplinary, AI-informed, approach to help improve things.

Some caveats

- My PhD is in Australian economic history.
- I have an outsider's view of a bunch of disciplines, but I am **not** a biostatistician, computer scientist, demographer, economist, political scientist, statistician, etc.
- These are just my opinions—please tell me where I'm wrong!



Part 1 - Background

“Between the idea
And the reality…
Falls the Shadow”

T.S. Eliot, The Hollow Men

What are we trying to achieve in social sciences?

- Social scientists—anthropologists, economists, geographers, historians, linguists, political scientists, psychologists, sociologists, etc—are interested in understanding society and human behaviour (Imai, 2017, p. 1).

- *Quantitative* social scientists do this using numbers, rather than stories (although the latter is very powerful, makes extensive contributions, and has a robust tradition e.g. Vaughan (1996), Oliphant (2021), etc).

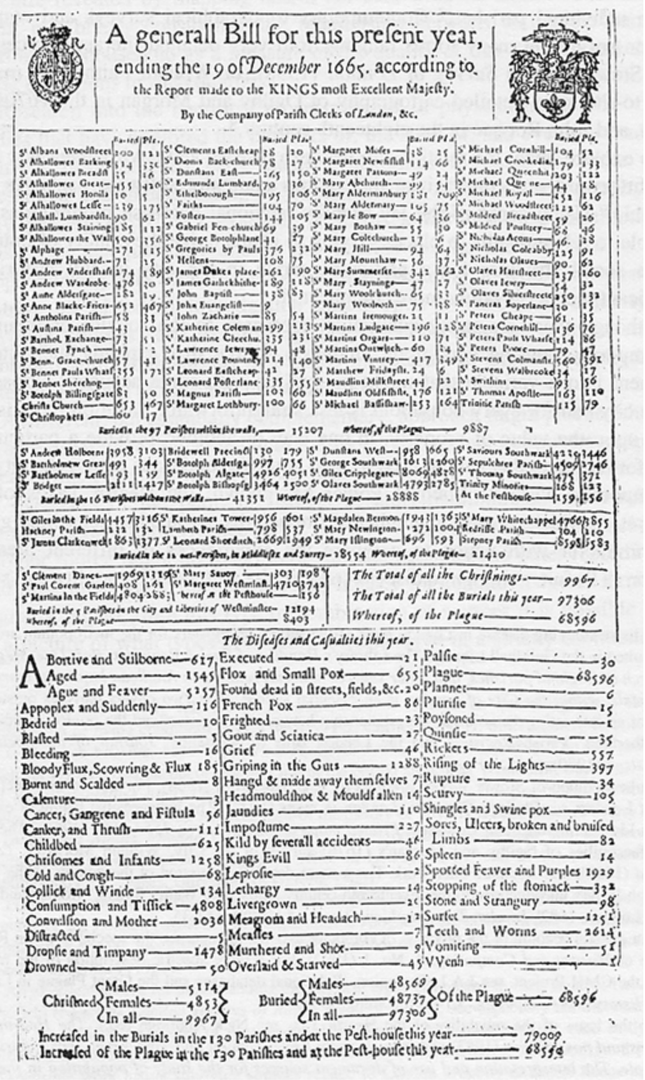
- Considering an individual as an observation in a dataset began with the census. Earliest surviving written records are from the Yellow River Valley (Whitby, 2020).

- The simplification of the world into data necessary involves measurement, which will always reflect the priorities of those in society with power. For instance, Section 127 of the Australian Constitution (repealed in 1967) stated: “In reckoning the numbers of the people of the Commonwealth, or of a State or other part of the Commonwealth, aboriginal natives shall not be counted.”

The rise of quantitative approaches

- The London plague of the 1500s see an emphasis on “the number of dead over the identities of the dead” (Otis, p. 131), and these turn into the 1600s Bills of Mortality, which famously John Graunt (1662), the founder of demography, analyzes.

- Least squares, a way to fit linear models, was associated with foundational problems in astronomy in the 1700s, such as determining the motion of the moon and reconciling the non-periodic motion of Jupiter and Saturn. It took longer for social scientists to become comfortable with it possibly because they were hesitant to group together data they worried were not alike (Stigler, 1986, p. 16).



A generall Bill for this present year, ending the 19 of December 1665. according to the Report made to the KINGS most Excellent Majesty.

By the Company of Parish Clerks of London, &c.

The Diseases and Casualties this year.

Abortive and Stilborne	617	Executed	21	Palsie	30
Aged	1545	Flox and Small Pox	655	Plague	68596
Ague and Feaver	5257	Found dead in ſtreets, fields, &c.	20	Planet	6
Appoplex and Suddenly	116	French Pox	86	Plurifie	15
Bedrid	10	Frighted	23	Poyſoned	1
Blaſted	5	Gout and Sciatica	27	Quinſie	35
Bleeding	16	Grief	46	Rickets	557
Bloody Flux, Scowring & Flux	185	Griping in the Guts	1288	Riſing of the Lighes	397
Burnt and Scalded	8	Hangd & made away themſelves	7	Rupture	34
Calenture	3	Headmouldſhot & Mouldfalken	14	Scurvy	105
Cancer, Gangrene and Fiſtula	56	Jaundies	110	Shingles and Swine pox	2
Canker, and Thruſh	111	Impoſtume	227	Sores, Ulcers, broken and bruiſed Limbs	82
Childbed	625	Kild by ſeverall accidents	46	Spleen	14
Chriſomes and Infants	1258	Kings Evill	86	Spotted Feaver and Purples	1929
Cold and Cough	68	Leproſie	2	Stopping of the ſtomack	332
Collick and Winde	134	Lethargy	14	Stone and Strangury	98
Conſumption and Tiſſick	4808	Livergrown	2	Surfet	1251
Convulſion and Mother	2036	Meagrom and Headach	12	Teeth and Worms	2614
Diſtracted	5	Meaſles	7	Vomiting	55
Dropſie and Timpany	1478	Murthered and Shot	9	Wenn	5
Drowned	50	Overlaid & Starved	45		

Christned: Males 5114, Females 4853, In all 9967. Buried: Males 48569, Females 48737, In all 97306. Of the Plague 68596.

Increaſed in the Burials in the 130 Pariſhes and at the Peſt-houſe this year 79009.

Increaſed of the Plague in the 130 Pariſhes and at the Peſt-houſe this year 68559.

Image source: https://en.wikipedia.org/wiki/Bills_of_mortality

The rise of quantitative approaches (cont.)

- Ratio estimators were used in 1802 by Pierre-Simon Laplace to estimate the total population of France based on scaling certain communes (Lohr [1999] 2022). Similarly, Adolphe Quetelet, proposed an approach based on counts in specific geographies, which could then be scaled up to the whole country.
- It would be Francis Galton and Ronald Fisher, noted eugenists, in the late 1800s and early 1900s who would put in place the way we use correlation, ANOVA, null hypothesis significance testing, and p-values, all of which is the foundation for approaches which came to dominate quantitative social sciences.
- Between 1950 and 1974 the percentage of economics articles using quantitative methods increased from 27% to 84%, and the percentage using “prose only” decreased from 49% to 13% (Kamerschen, 1977).

Cross-discipline homogeneity

- If there was once a separation of social sciences into disciplines — anthropology, economics, geography, history, linguistics, political science, psychology, sociology, etc — there has been a gradual coming together of approaches to the extent that, for researchers on the quantitative frontier, disciplinary groupings no longer make sense.

- This was catalyzed by the “credibility revolution” of the 2010s which focused quantitative social scientists on separating correlation and causation (Ashworth, Berry, Bueno De Mesquita, 2021). The criticisms it addressed in one discipline were applicable to all quantitative social science discipline.

- Best practice is now generally agreed by those on the frontier of constituent quantitative social science disciplines, and the methods they use would fit into any other social science discipline (modulo the literature review!). Joint appointments in a social science home and a statistics department are now common. Some quantitative social scientists even just call themselves applied statisticians!

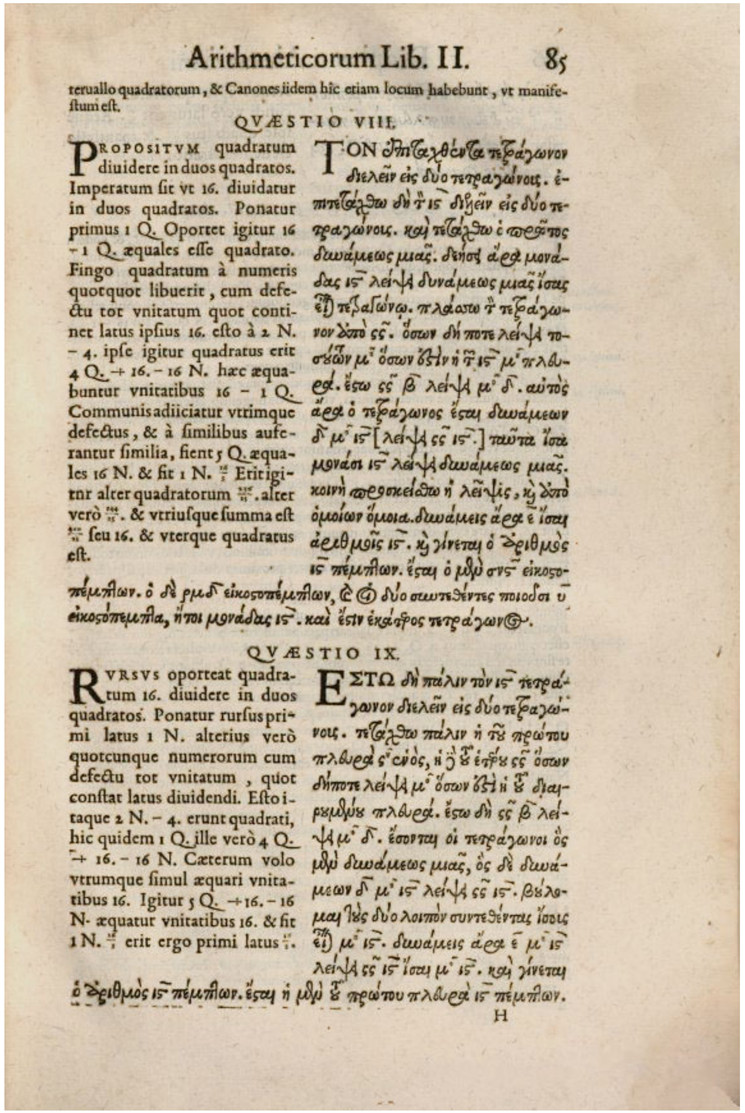
Recent changes and challenges

1) Use-of-methods improvements; and 2) reproducibility

1. **Arel-Bundock, Briggs, et al (2022):**
   - "Collating over 16,000 hypothesis tests from about 2,000 articles. The median analysis has about 10% power, and only about 1 in 10 tests have at least 80% power to detect the consensus effects reported in the literature."
2. **Brodeur, Cook, and Heyes (2020):**
   - "Applying multiple approaches to over 21,000 hypothesis tests published in 25 leading economics journals, we find that the extent of p-hacking and publication bias varies greatly by method. IV (and to a lesser extent DID) are particularly problematic."
3. **Mellon (2023):**
   - "A review of 288 studies reveals 192 variables previously linked to weather: all representing potential exclusion violations. Using sensitivity analysis, I show that the magnitude of many of these violations is sufficient to overturn numerous existing IV results."
4. **Stommes, Aronow, and Sävje (2023):**
   - "We collect all RD-based findings published in top political science journals from 2009--2018. The findings exhibit pathological features; estimates tend to bunch just above the conventional level of statistical significance…. The issues we uncover… cause concern that many published findings using the RD design are exaggerated, if not entirely spurious."
5. **Trisovic, Lau, Pasquier, and Crosas (2022):**
   - "We retrieve and analyze more than 2000 replication datasets with over 9000 unique R files published from 2010 to 2020. 74% of R files failed to complete without error in the initial execution, while 56% failed when code cleaning was applied"

Reproducibility

- About 10 years ago there was a big push on reproducibility across many quantitative social sciences.
- Partly this was due to research practice, which we would now consider shoddy, but was acceptable at the time. For instance, Herndon, Ash, and Pollin (2014), identify errors, caused by the poor use of Excel by Reinhart and Rogoff used to justify austerity following the financial crisis of 2007/08.
- Such changes were not (and are not) universally approved, but there is now broad, general, consensus about the benefits of moving away from null hypothesis significance testing, sharing code and data, and pre-registering analysis (even if it's not always acted on).
- These, and other, changes are making quantitative social science scientific.



Arithmeticorum Lib. II. 85

teruallo quadratorum, & Canones iidem hic etiam locum habebunt, vt manifestum est.

QVÆSTIO VIII.

PROPOSITVM quadratum diuidere in duos quadratos. Imperatum sit vt 16. diuidatur in duos quadratos. Ponatur primus 1 Q. Oportet igitur 16 − 1 Q. æquales esse quadrato. Fingo quadratum à numeris quotquot libuerit, cum defectu tot vnitatum quot continet latus ipsius 16. esto à 2 N. − 4. ipse igitur quadratus erit 4 Q. + 16. − 16 N. hæc æquabuntur vnitatibus 16 − 1 Q. Communis adiiciatur vtrimque defectus, & à similibus auferantur similia, fient 5 Q. æquales 16 N. & fit 1 N. $\frac{4}{5}$. Et itigitur alter quadratorum $\frac{16}{25}$. alter verò $\frac{9}{25}$. & vtriusque summa est $\frac{25}{25}$ seu 16. & vterque quadratus est.

ΤΟΝ δὲ τετραγωνον διελεῖν εἰς δύο τετραγώνους. ἐπιτεταχθω δὴ τὸν ις διῆσθαι εἰς δύο τετραγώνους. καὶ τεταχθω ὁ τῶν β τος δυνάμεως μιᾶς. δήσει ἄρα μονάδας ις λείψει δυνάμεως μιᾶς ἴσους ἑξ πλαισίῳ. πλάσσω τὸ τετράγωνον ἀπὸ ς. ὅσων δὴ ποτε λείψει πόσων μ ὅσων ἔστιν ὁ τὸν ις μ πλευρά. ἔστω ς β λείψει μ δ. αὐτὸς ἄρα ὁ τετράγωνος ἔσται δυνάμεων δ μ ις [λείψει ς ις.] παῦται ἴσαι μονάσι ις λείψει δυνάμεως μιᾶς. κοινὴ προσκείσθω ἡ λεῖψις, καὶ ἀπὸ ὁμοίων ὅμοια. δυνάμεις ἄρα ε ἴσαι ἀριθμοῖς ις. καὶ γίνεται ὁ ἀριθμὸς ις πέμπτων. ἔσται ὁ μὲν στος εἰκοστὰ πέμπτων. ὁ δὲ ρμδ εἰκοσοπέμπτων, ᾧ δύο στωτεθέντες ποιοῦσι τὸν εἰκοσοπεμπτα, ἥτοι μονάδας ις. καὶ ἔστιν ἑκάτερος τετράγωνος.

QVÆSTIO IX.

RVRSVS oporteat quadratum 16. diuidere in duos quadratos. Ponatur rursus primi latus 1 N. alterius verò quotcunque numerorum cum defectu tot vnitatum, quot constat latus diuidendi. Esto itaque 2 N. − 4. erunt quadrati, hic quidem 1 Q. ille verò 4 Q. + 16. − 16 N. Cæterum volo vtrumque simul æquari vnitatibus 16. Igitur 5 Q. + 16. − 16 N. æquatur vnitatibus 16. & fit 1 N. $\frac{4}{5}$. erit ergo primi latus $\frac{4}{5}$.

ΕΣΤΩ δὴ πάλιν τὸν ις τετράγωνον διελεῖν εἰς δύο τετραγώνους. τεταχθω πάλιν ἡ τῶν πρώτου πλευρὰ ς σ ὁ ὑπ ἐπίταγμα ς ὅσων δήποτε λείψει μ ὅσων ἔστιν ἡ τὸν διαιρεμένου πλευρά. ἔστω δὴ ς β λείψει μ δ. ἔσονται οἱ τετράγωνοι ὡς μὲν δυνάμεως μιᾶς, ὡς δὲ δυνάμεων δ μ ις λείψει ς ις. βέλομαι γοῦν δύο λοιπὸν συντεθέντας ἴσους ἑξ μ ις. δυνάμεις ἄρα ε μ ις λείψει ς ις ἴσοι μ ις. καὶ γίνεται ὁ ἀριθμὸς ις πέμπτων. ἔσται ἡ μὲν τοῦ πρώτου πλευρὰ ις πέμπτων.

H

Image source: https://en.wikipedia.org/wiki/Fermat's_Last_Theorem

Case study 1: Code and data availability

- Despite reproducibility being a bedrock of scientific practice, the American Economic Review (AER) only implemented a data (and code) availability policy in 2004. While innovative, replications based on those tended to fail.
- The American Economics Association (AEA) journals then appointed a data editor in 2018, and required pre-publication reproducibility checks (Vilhuber, 2020). The AEA has since verified (does the code produce what is in the manuscript) >1,000 articles at the “conditionally-accepted” stage (Vilhuber, Son, et al, 2022).
- Some challenges include non-public data and extensive computational requirements. (N.B. verification does not check whether a result is right!)

Case study 2: Pre-registration

- Researcher reliance on null hypothesis significance testing cannot be changed immediately, but pre-registration can help deal with HARKing (where a hypothesis is developed after results), p-hacking (where a researcher changes aspects of the analysis to get a significant finding) and file drawers (Rubenson, 2021).

- Typically this requires pre-registration of hypotheses and data analysis plans (Brodeur et al 2024), but can ranges from just answering a few questions through to fully writing out the code that will be used.

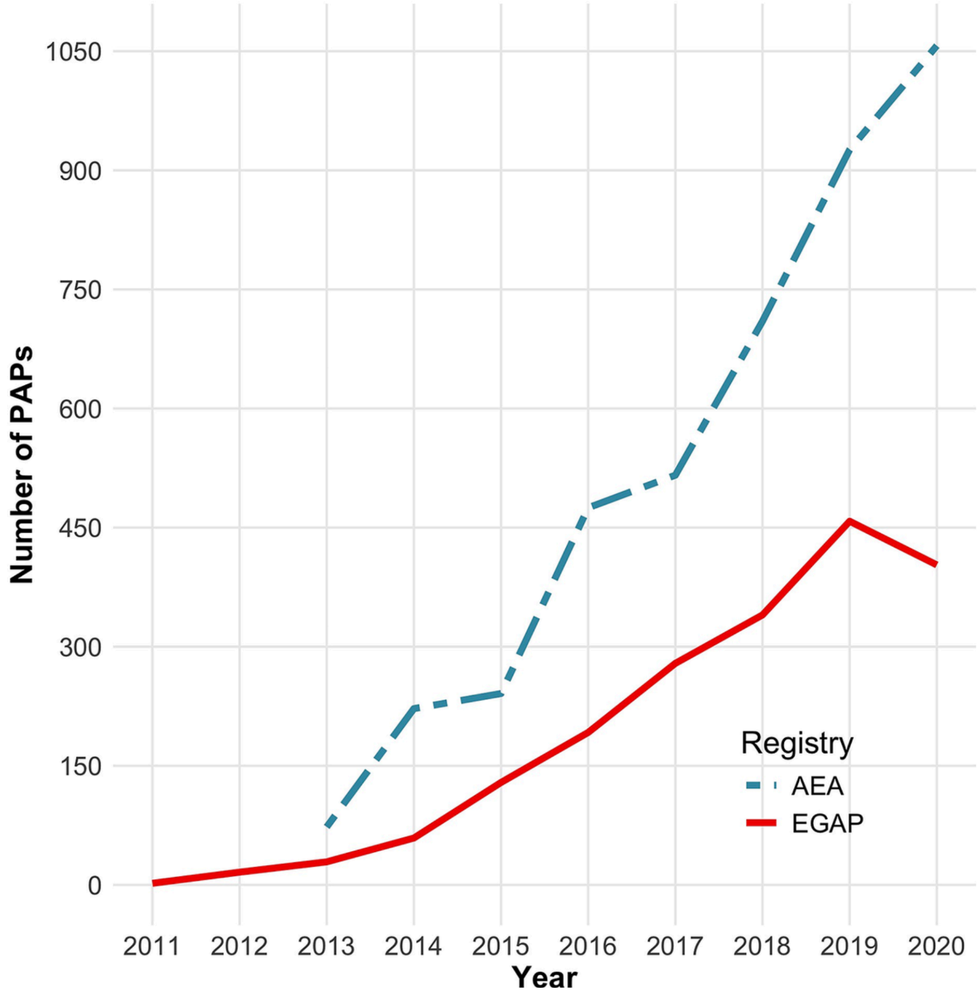
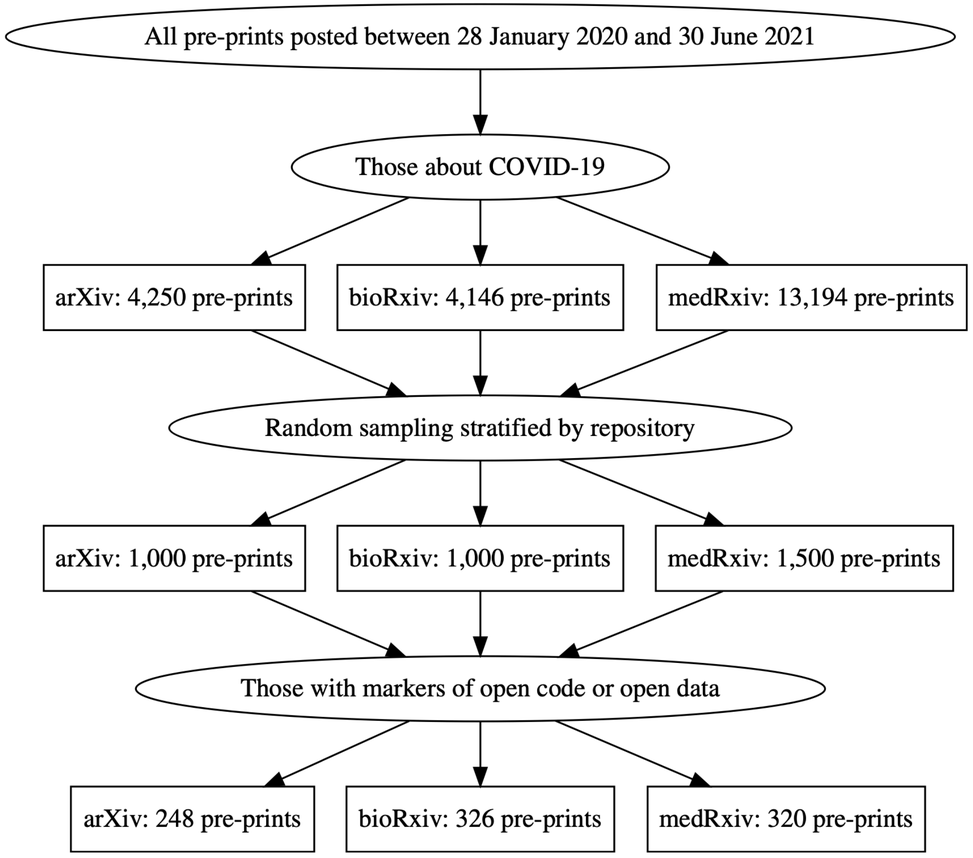


Image source: Rubenson (2021)

Where do we stand now?

- In 2014 only around 12% of political science articles included a data and code repository, and by 2022 this increased to 31% (Rainey, Roe, Wang, and Zhou, 2024).
- Pre-prints are similar, with Collins and Alexander (2022) unable to find markers of either open data or open code for 75% of those on arXiv.
- While this is shifting slowly, it is shifting. There is (soon-to-be publicly released) evidence from Replication Games that papers since 2022 can at least be re-run.



Source: Collins, Annie, and **Rohan Alexander**, 2022, "Reproducibility of COVID-19 pre-prints", *Scientometrics*, 4 July, https://doi.org/10.1007/s11192-022-04418-2.

Part 2 - Workflows

“Sunlight is said to be the best of disinfectants; electric light the most efficient policeman.”

Louis Brandeis

Workflows

A workflow is a “system for managing repetitive processes and tasks which occur in a particular order.” (IBM, 2024).For instance (drawing from statistics rather than quantitative social science) Gelman et al (2020), establish a (Bayesian) workflow:

An expanded (Bayesian) workflow:

The issue is that there is still considerable manual work. We want a solution where we can use computers (and eventually AI) to do some of this for us. Drawing from software engineering, this requires the ability to write automated tests, which implies a need for simulation.

The implied role of simulation

- Simulation is widely used in statistics, but less established in quantitative social sciences. It should be the first step of any project.
- Why simulate?
  - It forces us into the details.
  - It helps with cleaning and preparing the dataset
  - It provides us with clear features that our real dataset should satisfy.
  - It allows us to establish exceptions about the bounds of coefficients.
  - We can share simulated data even when we can’t share the real data.
- In the case of workflow, we need simulation to give something to test.

Adaptive control design

Mindell (2008)

- One of the innovations associated with the development of the Apollo program (specifically the X-15) was a two-stage control process.
- A computer—MH-96—which was on the plane, continuously generated a model of an ideal X-15 flying under ideal conditions.
- When the pilot moved the stick to make a change, this was actually just an input to the model of how the model should change. The computer then made the changes necessary to implement this in the real plane.
- In this way, the real plane to responded like the idealized model.

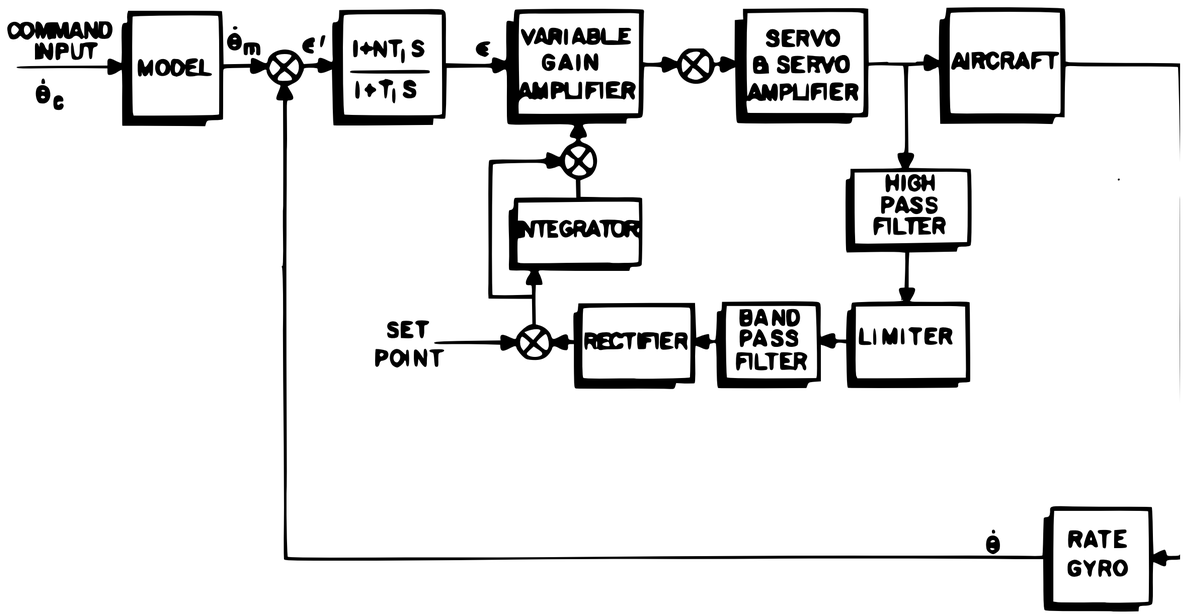


Image source: Draper Laboratory, NESC Request No: TI-13-00847.

Case study 3: The case of the missing sodium

Can data testing save your life? Lessons from Pou-Prom (2022)

- St Mike's has a Data Science & Advanced Analytics group who implemented an early warning system for patients.
- They were moving from historical data to live data and were concerned about external factors, data entry issues, and selection bias. They tested the data and noticed that sodium labs were missing.
- What do you think happened?

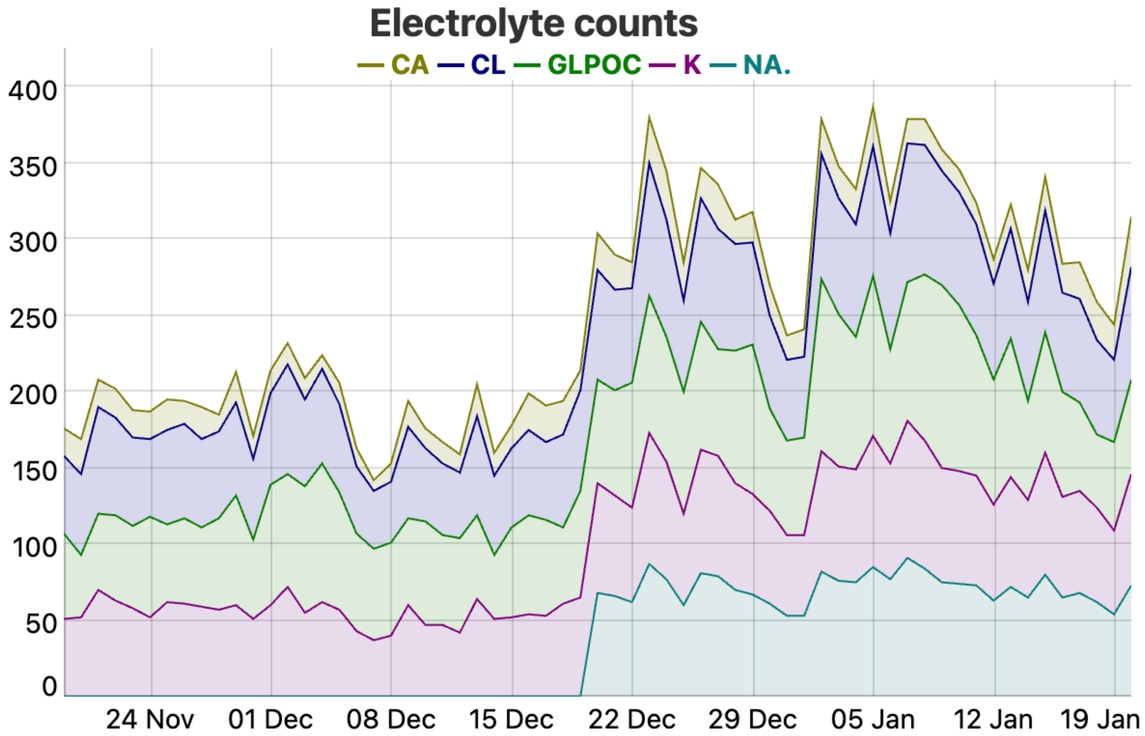


Image source: Pou-Prom (2022).

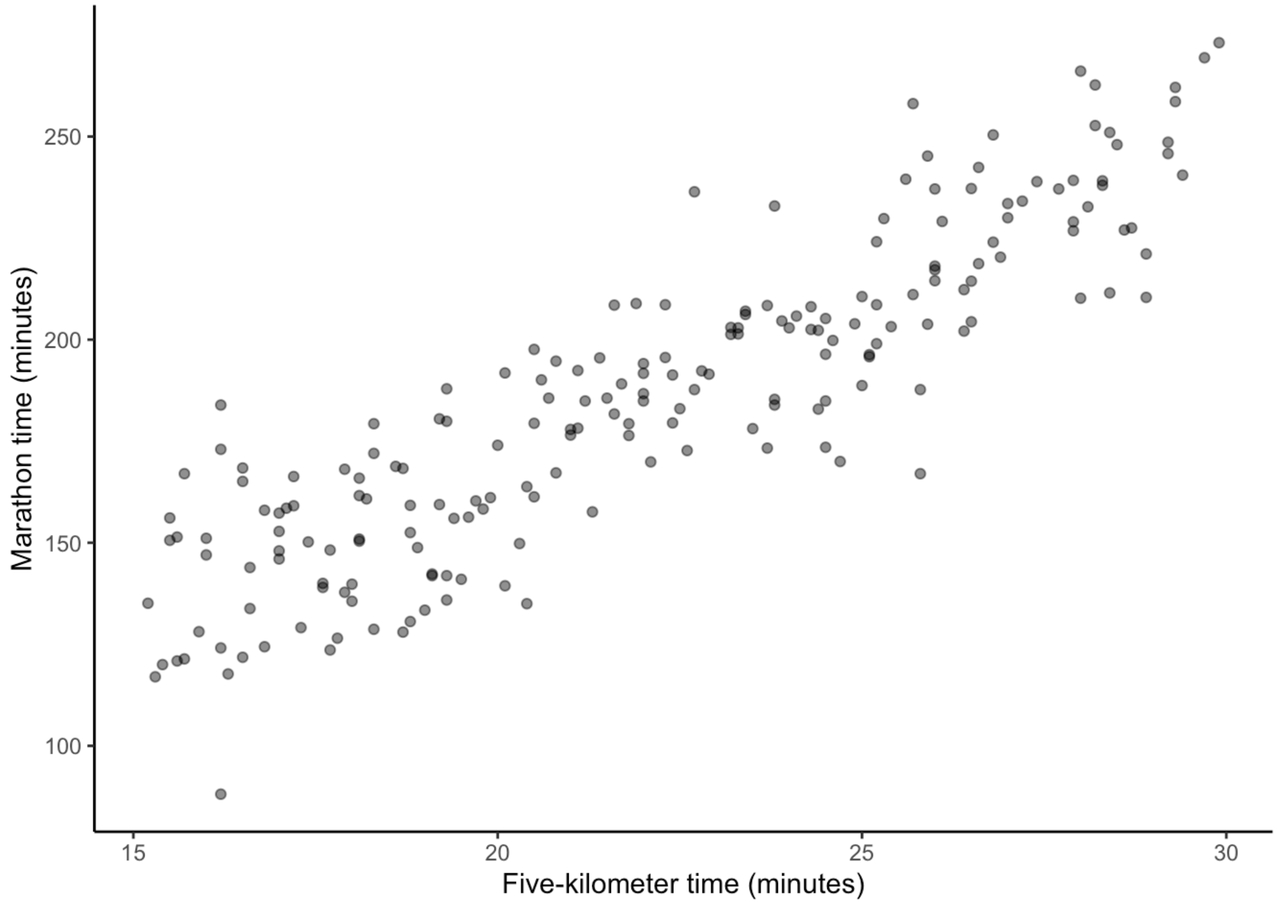
What does this look like in practice?

Marathon time prediction given 5km time

```
set.seed(853)

number_of_observations <- 200
dis_multiplier <- 8.4
very_fast_5km_time <- 15
good_enough_5km_time <- 30

sim_run_data <-
  tibble(
    five_km_time =
      runif(
        n = number_of_observations,
        min = very_fast_5km_time,
        max = good_enough_5km_time
      ),
    noise = rnorm(n = number_of_observations,
                  mean = 0,
                  sd = 20),
    marathon_time =
      five_km_time * dis_multiplier + noise
  )
```



Code is too complicated for humans

- Quantitative social science analysis code these days commonly amounts to thousands or tens of thousands of lines.
    - This means we need computers (and humans) to run checks.
    - Computers can run thousands of small checks.
    - But only if we have the infrastructure.
- We need to convince another person that it is fit for purpose (without requiring that they go through it line by line).
- In quantitative social science our tests differ to software engineering in that we are especially focused on models.

What does this look like in practice?

Marathon time prediction given 5km time

"test_class.R"

```
test_that("Check class", {
  expect_type(sim_run_data$marathon_time, "double")
  expect_type(sim_run_data$five_km_time, "double")
  expect_type(sim_run_data$was_raining, "character")
})
```

"test_observations.R"

```
test_that("Check number of observations is correct", {
  expect_equal(nrow(sim_run_data), 200)
})

test_that("Check complete", {
  expect_true(all(complete.cases(sim_run_data)))
})
```

"test_coefficient_estimates.R"

```
test_that("Check coefficients", {
  expect_gt(sim_run_data_rain_model$coefficients[3], 0)
  expect_lt(sim_run_data_rain_model$coefficients[3], 20)
})
```

What does this look like in practice?

Marathon time prediction given 5km time

"test_class.R"

```
test_that("Check class", {
  expect_type(sim_run_data$marathon_time, "double")
  expect_type(sim_run_data$five_km_time, "double")
  expect_type(sim_run_data$was_raining, "character")
})
```

"test_observations.R"

```
test_that("Check number of observations is correct", {
  expect_equal(nrow(sim_run_data), 200)
})

test_that("Check complete", {
  expect_true(all(complete.cases(sim_run_data)))
})
```

"test_coefficient_estimates.R"

```
test_that("Check coefficients", {
  expect_gt(sim_run_data_rain_model$coefficients[3], 0)
  expect_lt(sim_run_data_rain_model$coefficients[3], 20)
})
```

```
library(testthat)

test_file("tests/test_observations.R")
test_file("tests/test_class.R")
```

```
sim_run_data_rain_model <-
  lm(
    marathon_time ~ five_km_time + was_raining,
    data = sim_run_data
  )
```

```
test_file("tests/test_coefficient_estimates.R")
```

Case study 4: Excess death

Lessons from Knutson et al. (2023)

- Excess death models are needed to estimated the effect of COVID-19. Knutson et al. (2023) build a complicated Bayesian model for every country.
- Model checks typically involve going through each country and looking at the graph but difficult to do.
- After initial estimates were publicized, there was attention on the Germany estimates— 195K (UI 161K, 229K. Knutson et al. (2023) re-examined those estimates and made changes to the model. This revised the estimate to 122K (UI 101K, 143K).
- How do you think this was identified?

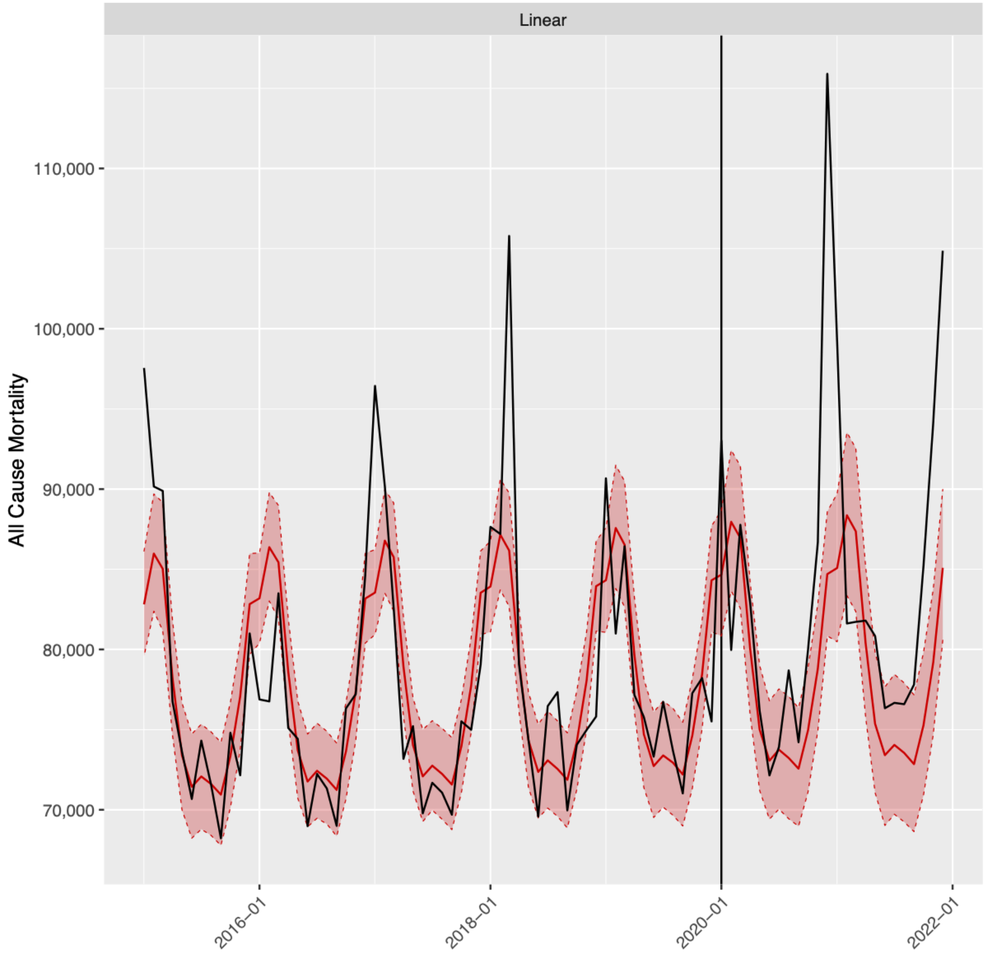


Image source: Msemburi, et al (2023) supplementary material.

What are our tests doing in quantitative social science?

Tests document our assumptions about the data and our expectations about our model

- Based on the simulated dataset, we develop a suite of tests. We then apply these tests to our actual dataset and model.
- These tests can be shared, even if the data/model cannot.
- They prevent known errors.
- They ensure our models satisfy our expectations.
- These are all checks that we do in our heads, but no one can verify that we have done them right and there are too many things to think about.

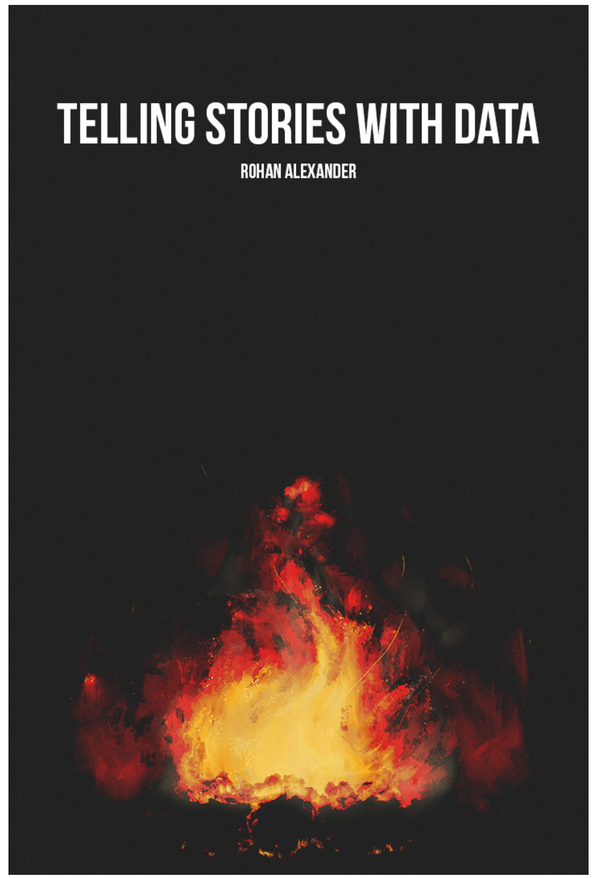


Image source: **Alexander, Rohan** (2023)

The critical role of testing

Open questions

- What does an effective test suite look like in quantitative social sciences?
- How can tests address the issue of (lack of) data sharing?
- What does a test checklist look like in different social sciences?
- How do we develop expectations of coefficients before running the model?
- How can we integrate model tests with notions of priors?
- How can we automate the creation of model tests?
- How can we automate the creation of data tests?

The critical role of testing

Open questions

- What does an effective test suite look like in quantitative social sciences?
- How can tests address the issue of (lack of) data sharing?
- What does a test checklist look like in different social sciences?
- How do we develop expectations of coefficients before running the model?
- How can we integrate model tests with notions of priors?
- How can we automate the creation of model tests? LLMs.
- How can we automate the creation of data tests? LLMs.

Part 3 - Adopting AI

"A new scientific truth does not triumph by convincing its opponents and making them see the light, but rather because its opponents eventually die…."

Max Planck

What role for AI?

Given this workflow, we can adopt AI, specifically LLMs, to help address cultural issues that might otherwise slow down the adoption of better practices, specifically:

1. Monolingualism.
2. Path dependency.
3. Lack of incentive to change.

Monolingualism

- Quantitative social sciences are dominated by a handful of propriety languages—Stata, SAS, SPSS, Excel—e.g. as at 2019 around 60% of AEA economics papers used Stata (Vilhuber, Turrito, and Welch, 2020).
- The main issue is not the use of any particular language, but instead monolingualism.
- This allows errors unknowingly propagate and means the strengths of different languages are not taken advantage of.

Thanks to Ash for being a good sport and allowing his tongue in cheek tweet to be used.

Monolingualism

“Language was always the companion of empire, and as such, together they begin, grow, and flourish. And later, together, they fall.”

Antonio de Nebrija (as quoted by RF Kuang in *Babel*)

- Can we use LLMs to translate code from one language—typically Stata or R—into Python?
- In general, “yes”: Talented undergraduates can use LLMs to quickly translate replication packages in a few hours. They quickly identify:
  - Aspects that cannot be the actual final code.
  - Minors errors.
- Struggles with:
  - Language specific packages such as binscatter.
  - Inconsistencies between different files.
  - Differences between what is in the paper and what is in the code.
- The key is the “talented undergraduates” bit! They quickly iterate and improve the results of the LLMs.

Sarah Xie, Allyson Cui, Inessa De Angelis, **Rohan Alexander**, “Using LLMs to translate replication packages to enhance credibility” (In-progress) https://github.com/InessaDeAngelis/AER_code_translation.

Monolingualism

Stata results:

	Illiteracy							
	Argentina, Brazil, and Paraguay		Brazil		Argentina		Paraguay	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Mission distance	0.0105***	0.0112**	0.0200***	0.0313***	0.0157**	0.0669***	0.00451	0.0138
	(0.004)	(0.005)	(0.007)	(0.010)	(0.007)	(0.022)	(0.012)	(0.027)

Python results:

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
VARIABLES	illiteracy	illiteracy	illiteracy	illiteracy	illiteracy	illiteracy	illiteracy	illiteracy
distmiss	0.0105***	0.0112**	0.0200***	0.0313***	0.0157**	0.0669***	0.00451	0.0138
	(0.00368)	(0.00495)	(0.00653)	(0.00985)	(0.00735)	(0.0220)	(0.0120)	(0.0270)
Constant	-35.33***	-53.74*	1,725	731.8	69.26*	-41.06	8.673***	-80.72*
	(11.80)	(32.50)	(1,159)	(1,299)	(36.79)	(54.63)	(0.694)	(43.89)
Observations	549	548	467	467	42	42	40	39
R-squared	0.042	0.073	0.056	0.095	0.165	0.669	0.004	0.251

Sarah Xie, Allyson Cui, Inessa De Angelis, **Rohan Alexander**, "Using LLMs to translate replication packages to enhance credibility" (In-progress) https://github.com/InessaDeAngelis/AER_code_translation.

Monolingualism

Typos: The dta is “Structural Transformation Brazil.dta”

```
292   *Panel B
293   *Brazil
294   *use Brazil Structural Transformation.dta.dta
295   qui: reg agroper distmiss distfran lat1 long1 area temp alt rain rugg river coast slope meso, r
296   qui: outreg2 using TableIXB, append keep(distmiss)
297   qui: reg induper distmiss distfran lat1 long1 area temp alt rain rugg river coast slope meso, r
298   qui: outreg2 using TableIXB, append keep(distmiss)
299   qui: reg commper distmiss distfran lat1 long1 area temp alt rain rugg river coast slope meso, r
300   qui: outreg2 using TableIXB, append keep(distmiss)
301   *Conley errors calculated using Conley Standard Errors.do
302   *Within R-squared calculated manually
```

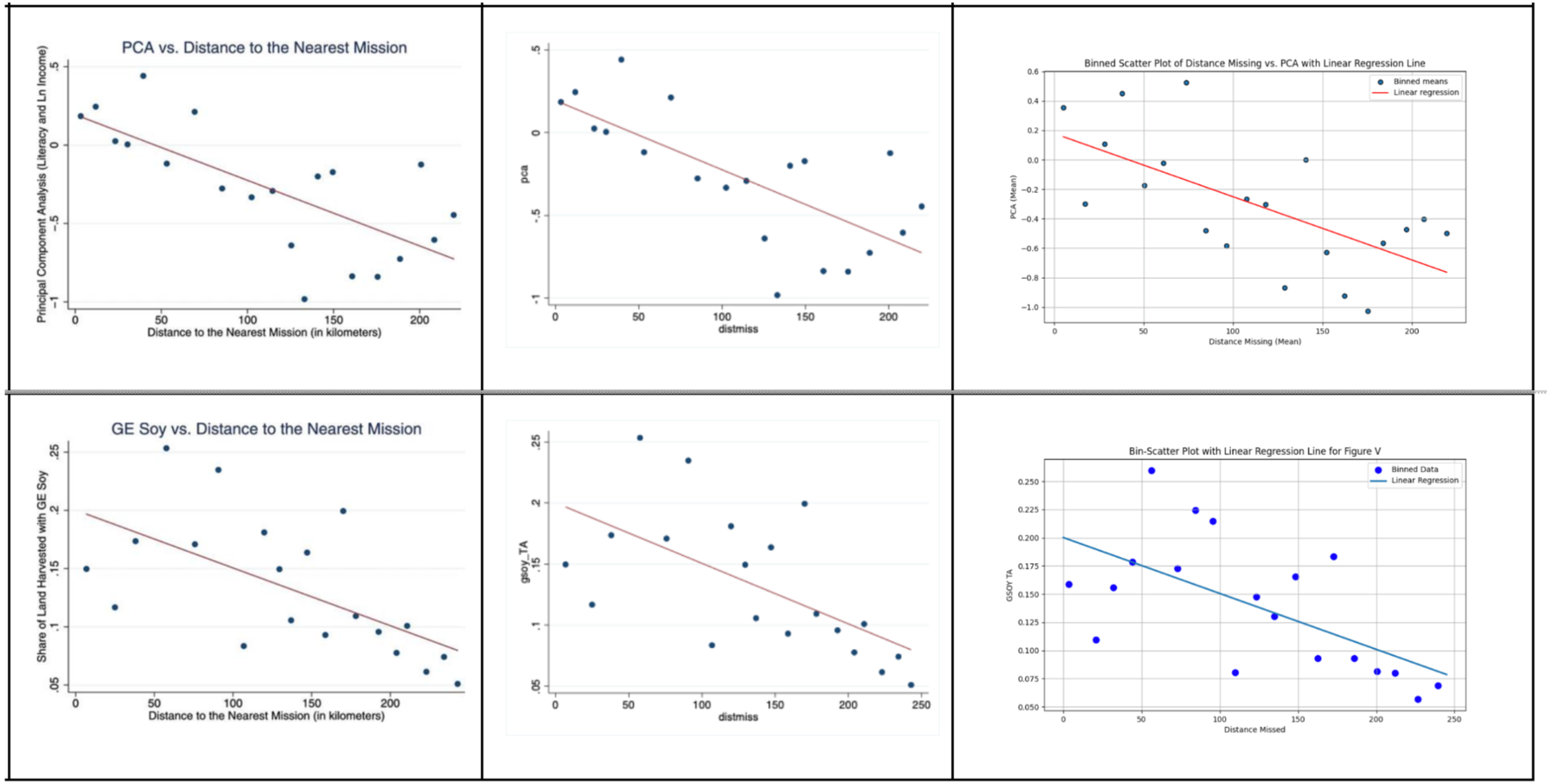
Syntax issues: Line 394 is missing command

```
388   *use Pre-Colonial Population Density.dta
389   *Data is from Maloney and Valencia Caicedo (2016)
390   *Available at: https://onlinelibrary.wiley.com/doi/full/10.1111/ecoj.12276
391   qui: iis countrynum
392   qui: xtreg popd mission temp_avg rainfall alti landlocked altisq tempsq rainsq , r fe
393   qui: outreg2 using TableXIIA, append keep(mission)
394   popd mission temp_avg rainfall alti landlocked altisq tempsq rainsq
395   *reg popd mission temp_avg rainfall alti landlocked altisq tempsq rainsq arg bra, cl (country)
396   *Within R-squared calculated manually
```

Sarah Xie, Allyson Cui, Inessa De Angelis, **Rohan Alexander**, “Using LLMs to translate replication packages to enhance credibility” (In-progress) https://github.com/InessaDeAngelis/AER_code_translation.

Monolingualism

Python did not adapt well with binscatter (packages unique to the language)



Sarah Xie, Allyson Cui, Inessa De Angelis, **Rohan Alexander**, "Using LLMs to translate replication packages to enhance credibility" (In-progress) https://github.com/InessaDeAngelis/AER_code_translation.

Path dependency

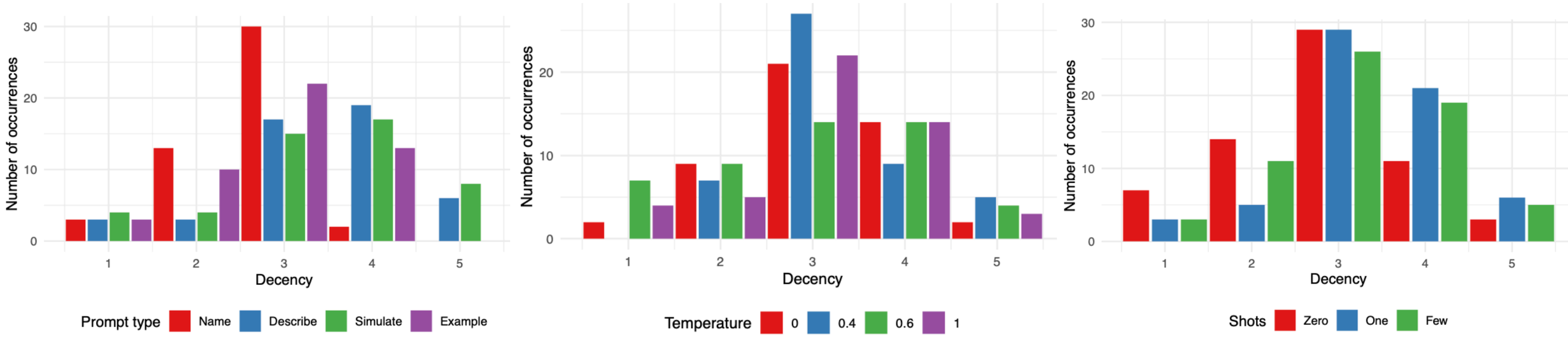
Can LLMs write data tests so that you don't have to? Yes.

- The IJF is a nonprofit news media outlet that is centred around public interest journalism. One of the IJFs eight databases, and the focus of this work, is the political donations database. While the IJF's database is available in a clean, user-friendly format, the original records upon which it was created were not all accessible in this way.
- To what extent can LLMs develop a suite of data validation tests that is similar to those written by an experienced expert data scientist?
- Four prompt scenarios; three learning modes; four temperature settings; two roles; two models: 1) GPT-3.5 and 2) GPT-4.

Rohan Alexander, Lindsay Katz, Callandra Moore, Zane Schwartz, "Evaluating the Decency and Consistency of Data Validation Tests Generated by LLMs", https://arxiv.org/abs/2310.01402.

Path dependency

- These are evaluated by human coders.
- Our work is consistent with previous literature demonstrating that LLM performance on complex, user-defined tasks is sensitive to prompt engineering.



Rohan Alexander, Lindsay Katz, Callandra Moore, Zane Schwartz, "Evaluating the Decency and Consistency of Data Validation Tests Generated by LLMs", https://arxiv.org/abs/2310.01402.

Lack of incentive to change

Can LLMs improve quantitative social science code, so that you don't have to? Yes.

- "Spellcheck for data science"
- Unique combination of:
  - AI expertise: Chris Maddison (SRI Affiliate)
  - Business expertise: Zane Schwartz
  - Data science expertise: Annie Collins and Michael Chong.

HuonResearch

Demo: https://youtu.be/Jhk-EK_7g9U?si=oGhqPpfeNQq_RYG-&t=25

Before:

```
# Documentation: https://www.sbert.net/docs/pretrained_models.html
# Quick start: https://www.sbert.net/docs/quickstart.html
# paper: https://arxiv.org/pdf/1908.10084.pdf
# blog: https://blog.ml6.eu/decoding-sentence-encoders-37e63244ae00
# hugging face: https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

from sentence_transformers import SentenceTransformer, util
import pandas as pd
from sklearn.cluster import KMeans
import numpy as np
from IPython.display import display

def get_sentence_embeddings(fName, model):
    articles = pd.read_csv(fName)
    headline_list = articles['title'].tolist()
    embeddings = np.array(model.encode(headline_list))
    print(type(embeddings[0]))
    print("vector length", len(embeddings[0]))

    return embeddings, headline_list

def assign_headlines(headlines, centers, transformer_model):
    '''assign a list of headlines (list of str) to clusters (list of str)'''
    headline_embeddings = transformer_model.encode(headlines)
    center_embeddings = transformer_model.encode(centers)

    # Compute cosine similarity between all embeddings and centroids
    cos_sim = util.cos_sim(headline_embeddings, center_embeddings)

    # find closest centroid for each headline
    cos_sim = cos_sim.numpy()
    # get max cosine similarity for each row, ie. closest cluster for each headline
    preds = np.argmax(cos_sim, axis=1)

    # write cluster results to file
    clusters_df = pd.DataFrame({"title": headlines,
                                "cluster_label": preds})
    return clusters_df

def main():
    model = SentenceTransformer('all-MiniLM-L6-v2')
    # Note: initial clustering used only articles from 2021 and 2022
    # To replicate results, "all_relevant_2021_2022.csv" should contain only
    # a subset of headlines from "all_relevant.csv" from 2021 and 2022
    embeddings, sentences = get_sentence_embeddings("development/python/example_data/all_relevant_2021_2022.csv", model)
    kmeans = KMeans(n_clusters=100, random_state=0, n_init="auto")
    kmeans.fit(embeddings)
    centroids = kmeans.cluster_centers_
    labels = kmeans.labels_
    # write cluster results to file
    clusters_df = pd.DataFrame({"title": sentences,
                                "cluster_label": labels})
    clusters_df.to_csv('cluster_results_TEST.csv', index=False)

    # Compute cosine similarity between all embeddings and centroids
    # to find 100 closest to each centroid
    cos_sim = util.cos_sim(embeddings, centroids)
    cos_sim = cos_sim.numpy()
    # get max cosine similarity for each column, ie. closest sentence for each cluster
```

After (2 min):

```
import os
from sentence_transformers import SentenceTransformer, util
import pandas as pd
from sklearn.cluster import KMeans
import numpy as np


def get_sentence_embeddings(file_path, model):
    '''
    Get sentence embeddings using a pre-trained model.

    Arguments:
    file_path -- a string representing the file path of the input data
    model -- the Sentence Transformer model

    Returns:
    embeddings -- an array of sentence embeddings
    headline_list -- a list of headlines
    '''
    articles = pd.read_csv(file_path)
    headline_list = articles['title'].tolist()
    embeddings = np.array(model.encode(headline_list))
    print(type(embeddings[0]))
    print("vector length", len(embeddings[0]))

    return embeddings, headline_list


def assign_headlines_to_clusters(headlines, centers, transformer_model):
    '''
    Assign a list of headlines to clusters.

    Arguments:
    headlines -- a list of headlines (list of strings)
    centers -- a list of cluster centers (list of strings)
    transformer_model -- the Sentence Transformer model

    Returns:
    clusters_df -- a DataFrame with assigned cluster labels
    '''
    headline_embeddings = transformer_model.encode(headlines)
    center_embeddings = transformer_model.encode(centers)

    cos_sim = util.cos_sim(headline_embeddings, center_embeddings)
    cos_sim = cos_sim.numpy()
    preds = np.argmax(cos_sim, axis=1)

    clusters_df = pd.DataFrame({"title": headlines, "cluster_label": preds})
    return clusters_df


def main():
    model = SentenceTransformer('all-MiniLM-L6-v2')
    embeddings, sentences = get_sentence_embeddings(os.path.join("development", "python", "example_data", "al
    kmeans = KMeans(n_clusters=100, random_state=0, n_init
    kmeans.fit(embeddings)
    centroids = kmeans.cluster_centers_
    labels = kmeans.labels_
```

Concluding remarks

- A unique multidisciplinary, AI-informed, approach can be used to move quantitative social sciences from credible to scientific.
- The key will be building exchanges of knowledge between AI-experts and quantitative social scientists.
- There is limited scope for training in both—each one is too complicated! But it is also difficult to get AI-experts interested in social science problems given the other exciting areas they can work on.
- But establishing incentives for the adoption of new technologies (journals, academic acknowledgement) will allow us to move forward.

contact@rohanalexander.com

References

- Alexander, Rohan (2023), *Telling Stories with Data*, Chapman and Hall/CRC.
- Arel-Bundock, Vincent; Briggs, Ryan C.; Doucouliagos, Hristos; Mendoza Aviña, Marco; Stanley, Tom D. (2022) : Quantitative Political Science Research is Greatly Underpowered, I4R Discussion Paper Series, No. 6, Institute for Replication (I4R), s.l.
- Ashworth, Berry, Bueno De Mesquita, 2021
- Brodeur, Abel, Nikolai Cook, and Anthony Heyes. 2020. "Methods Matter: p-Hacking and Publication Bias in Causal Analysis in Economics." American Economic Review, 110 (11): 3634-60.
- Brodeur, A., Esterling, K., Ankel-Peters, J., Bueno, N. S., Desposato, S., Dreber, A., Genovese, F., Green, D. P., Hepplewhite, M., Hoces de la Guardia, F., Johannesson, M., Kotsadam, A., Miguel, E., Velez, Y. R., & Young, L. (2024). Promoting Reproducibility and Replicability in Political Science. Research & Politics, 11(1). https://doi.org/10.1177/20531680241233439
- Gelman, Andrew. 2016. “What has happened down here is the winds have changed,” September. https://statmodeling.stat.columbia.edu/2016/09/21/what-has-happened-down-here-is-the-winds-have-changed/.
- Gelman, Andrew, Aki Vehtari, Daniel Simpson, Charles Margossian, Bob Carpenter, Yuling Yao, Lauren Kennedy, Jonah Gabry, Paul-Christian Bürkner, and Martin Modrák. 2020. “Bayesian Workflow.” arXiv. https://doi.org/10.48550/arXiv.2011.01808.
- Herndon, Thomas, Michael Ash, and Robert Pollin. 2014. “Does High Public Debt Consistently Stifle Economic Growth? A Critique of Reinhart and Rogoff.” *Cambridge Journal of Economics* 38 (2): 257–79. https://doi.org/10.1093/cje/bet075.
- Imai, Kosuke (2017) *Quantitative Social Science: An Introduction*, Princeton University Press.
- Kamerschen, David R., 1977, “The Rise of ‘Quantitative Methods’ in Economics Journals, 1950-1974”, *The Journal of Economic Education*, Vol. 9, No. 1, pp. 51-53.
- Knutson, Victoria, Serge Aleshin-Guendel, Ariel Karlinsky, William Msemburi, and Jon Wakefield. 2022. “Estimating Global and Country-Specific Excess Mortality During the COVID-19 Pandemic,” May. https://cdn.who.int/media/docs/default-source/world-health-data-platform/covid-19-excessmortality/covid-methods-paper-revision.pdf.
- Lohr, Sharon (1999) 2022. *Sampling: Design and Analysis*. 3rd ed. Chapman Hall/CRC.
- Mellon, Jonathan. 2023. “Rain, Rain, Go Away: 195 Potential Exclusion-Restriction Violations for Studies Using Weather as an Instrumental Variable.” SocArXiv. https://doi.org/10.31235/osf.io/9qj4f.
- Mindell, David. 2008. Digital Apollo: Human and Machine in Spaceflight. 1st ed. New York: The MIT Press.
- Msemburi, William, Ariel Karlinsky, Victoria Knutson, Serge Aleshin-Guendel, Somnath Chatterji & Jon Wakefield (2023) “The WHO estimates of excess mortality associated with the COVID-19 pandemic”, *Nature*, 613, 130–137.
- Oliphant, Elayne (2021) *The Privilege of Being Banal*, The University of Chicago Press.
- Otis, Jessica Marie (2024) *By The Numbers*, Oxford University Press.
- Pou-Prom, Chloe (2022), “Ooh na na… where are my sodium labs?”, 9 May, https://lks-chart.github.io/blog/posts/2022-05-09-ooh-na-na-where-are-my-sodium-labs/
- Rainey, C., Roe, H., Wang, Q., & Zhou, H. (2024, January 21). Data and Code Availability in Political Science Publications from 1995 to 2022. https://doi.org/10.31235/osf.io/a5yxe
- Rubenson D. Tie my hands loosely: Pre-analysis plans in political science. Politics and the Life Sciences. 2021;40(2):142-151. doi:10.1017/pls.2021.23
- Stigler, Stephen, 1986, *The History of Statistics*. Belknap Harvard.
- Stommes, D., Aronow, P. M., & Sävje, F. (2023). On the reliability of published findings using the regression discontinuity design in political science. Research & Politics, 10(2). https://doi.org/10.1177/20531680231166457
- Trisovic, Ana, Matthew Lau, Thomas Pasquier, and Mercè Crosas. 2022. “A Large-Scale Study on Research Code Quality and Execution.” *Scientific Data* 9 (1). https://doi.org/10.1038/s41597-022-01143-6.
- Vaughan, Diane (1996) *The Challenger Launch Decision*, The University of Chicago Press.
- Vilhuber, Lars, James Turrito, and Keesler Welch. 2020. “Report by the AEA Data Editor” *AEA Papers and Proceedings*, 110: 764-75.
- Vilhuber, Lars, (2020). “Reproducibility and Replicability in Economics”. *Harvard Data Science Review*, 2(4). https://doi.org/10.1162/99608f92.4f6b9e67.
- Vilhuber, Lars, Hyuk Harry Son, Meredith Welch, David N. Wasser & Michael Darisse (2022) “Teaching for Large-Scale Reproducibility Verification”, *Journal of Statistics and Data Science Education*,