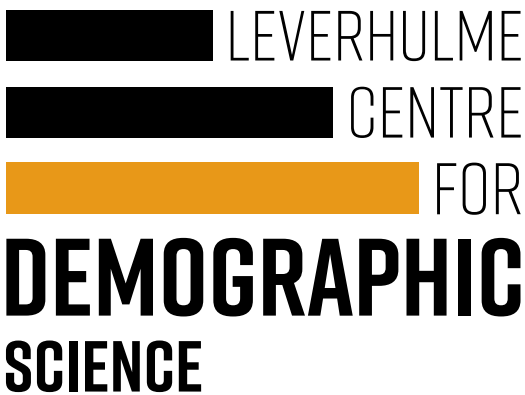
On the Responsible use of Pseudo-Random Number Generators in Scientific Research

Charlie Rahal

LCDS and Nuffield College, University of Oxford

Ox | Ber 2023



Pseudo-Random Number Generation

An invisible source of uncertainty in the scientific record: PRNGs!

Pseudo-Random Number Generation

An invisible source of uncertainty in the scientific record: PRNGs!

- **Q**: Has anyone ran a program twice, with different results?

Pseudo-Random Number Generation

An invisible source of uncertainty in the scientific record: PRNGs!

- **Q**: Has anyone ran a program twice, with different results?
- **Q**: Has anyone here ever set a 'seed'? Which seed?

Pseudo-Random Number Generation

An invisible source of uncertainty in the scientific record: PRNGs!

- **Q**: Has anyone ran a program twice, with different results?
- **Q**: Has anyone here ever set a 'seed'? Which seed?
- Maybe to eliminate variation in algorithms with PRNGs?
  - This seems to be the current 'best practice' advice.

Pseudo-Random Number Generation

An invisible source of uncertainty in the scientific record: PRNGs!

- **Q**: Has anyone ran a program twice, with different results?
- **Q**: Has anyone here ever set a 'seed'? Which seed?
- Maybe to eliminate variation in algorithms with PRNGs?
  - This seems to be the current 'best practice' advice.
- We argue this is the opposite of what we want!
  - We propose you pre-specify **multiple** (complex) seeds.

Pseudo-Random Number Generation (Cont.)

- We want to consider possible variation independent of arbitrary variation in seed choice.

Pseudo-Random Number Generation (Cont.)

- We want to consider possible variation independent of arbitrary variation in seed choice.
- This is an extremely important and scarcely research problem.

Pseudo-Random Number Generation (Cont.)

- We want to consider possible variation independent of arbitrary variation in seed choice.
- This is an extremely important and scarcely research problem.
- Pseudo-Random Number Generators occur everywhere!

Pseudo-Random Number Generation (Cont.)

- We want to consider possible variation independent of arbitrary variation in seed choice.
- This is an extremely important and scarcely research problem.
- Pseudo-Random Number Generators occur everywhere!
- The variation in estimand can be **huge**.

Pseudo-Random Number Generation (Cont.)

- We want to consider possible variation independent of arbitrary variation in seed choice.
- This is an extremely important and scarcely research problem.
- Pseudo-Random Number Generators occur everywhere!
- The variation in estimand can be **huge**.
- We bring attention to this through multiple types of replications.
  - Simulations, machine learning, and inferential research.

- **Problem Statement**: By setting one seed, we have no idea about the potential key variation of our estimand as a function of how random numbers were generated. This is computationally un-intensive, but scientifically dour.

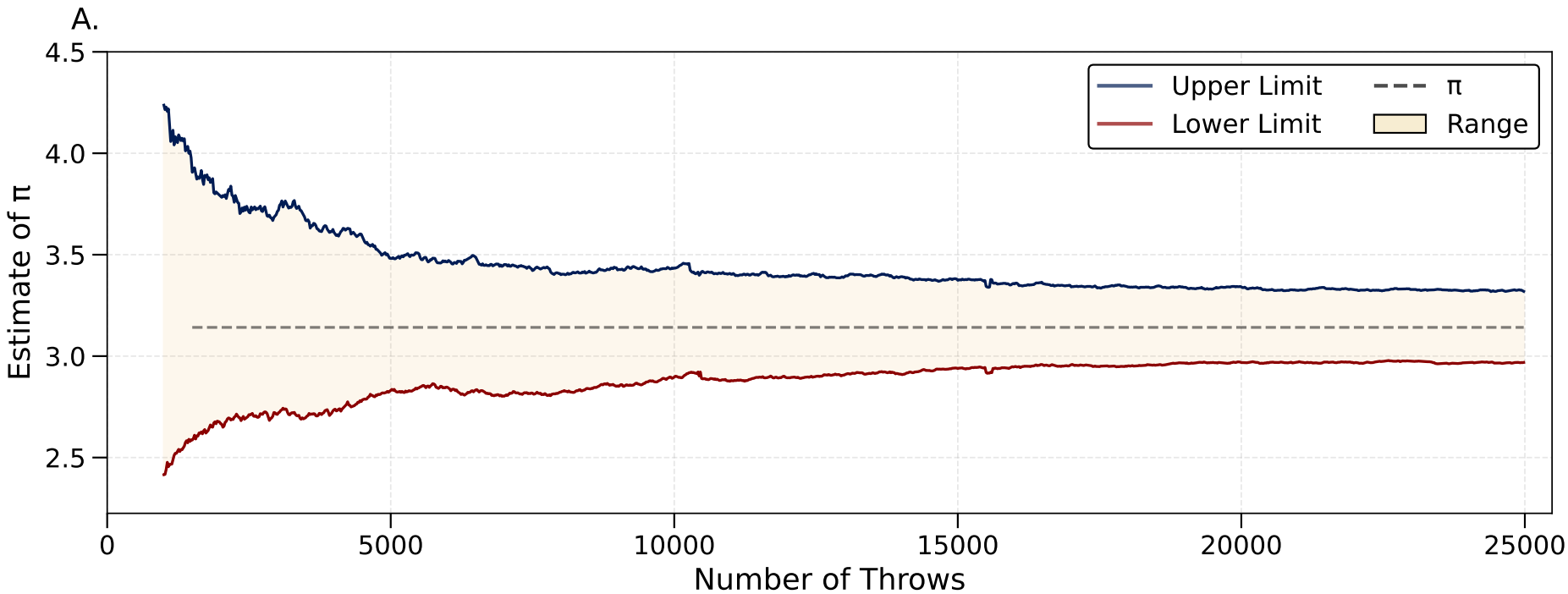
- **Problem Statement**: By setting one seed, we have no idea about the potential key variation of our estimand as a function of how random numbers were generated. This is computationally un-intensive, but scientifically dour.

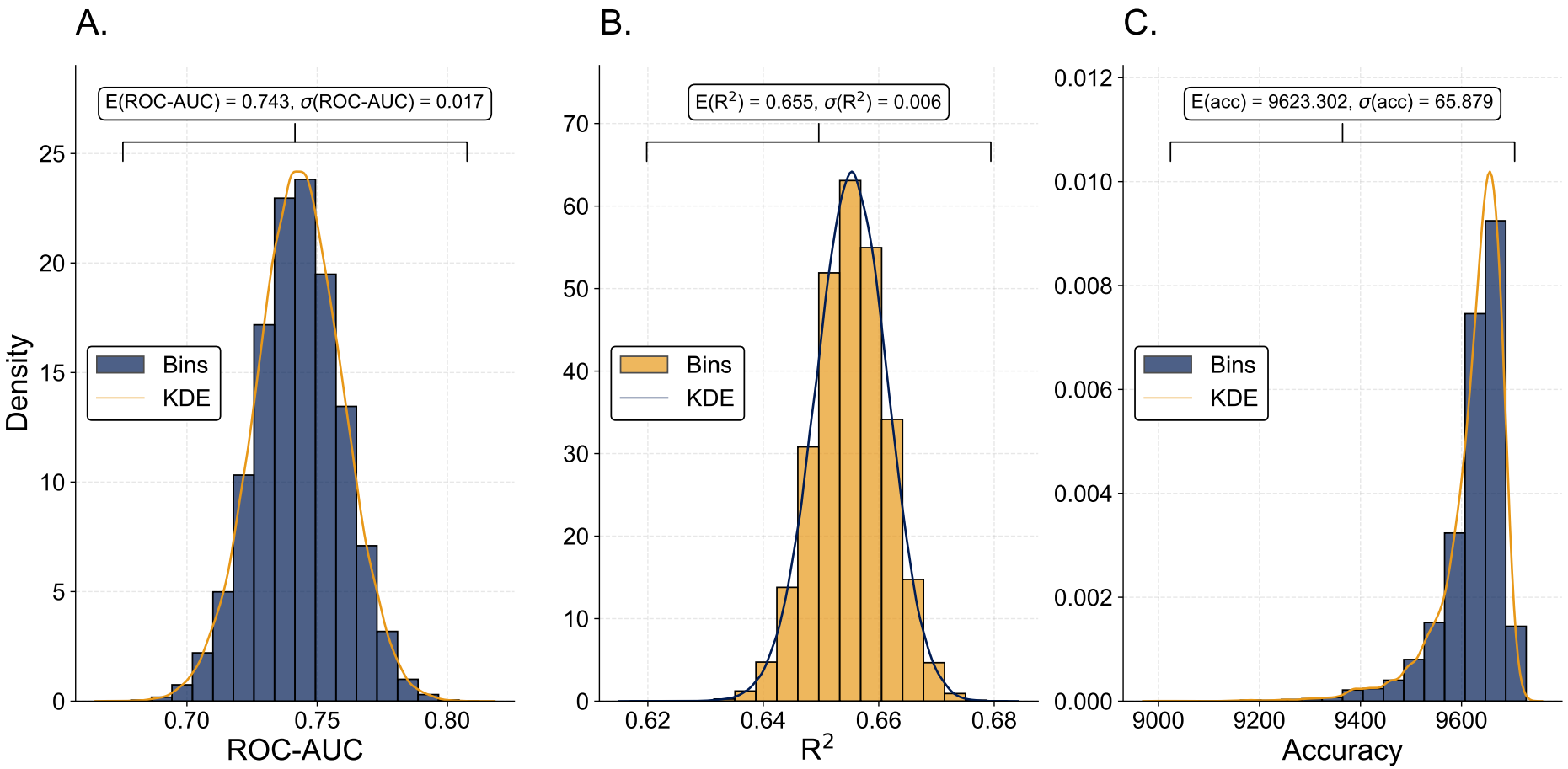
- **Solution**: Visualize the outcome space of a of large number (10k? 100k?) seeds simultaneously. This is computationally intensive, but scientifically faithful.

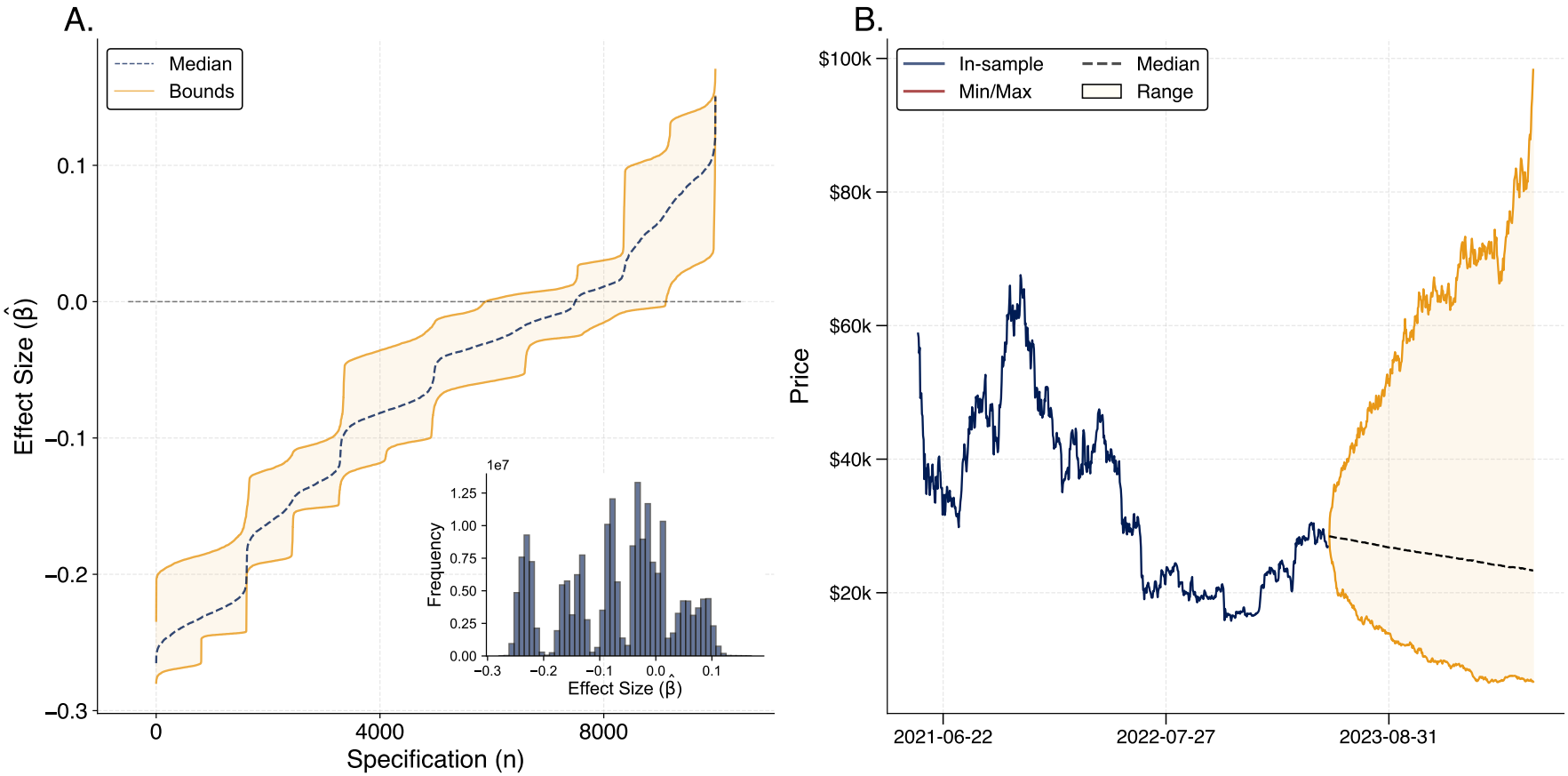
- **Problem Statement**: By setting one seed, we have no idea about the potential key variation of our estimand as a function of how random numbers were generated. This is computationally un-intensive, but scientifically dour.

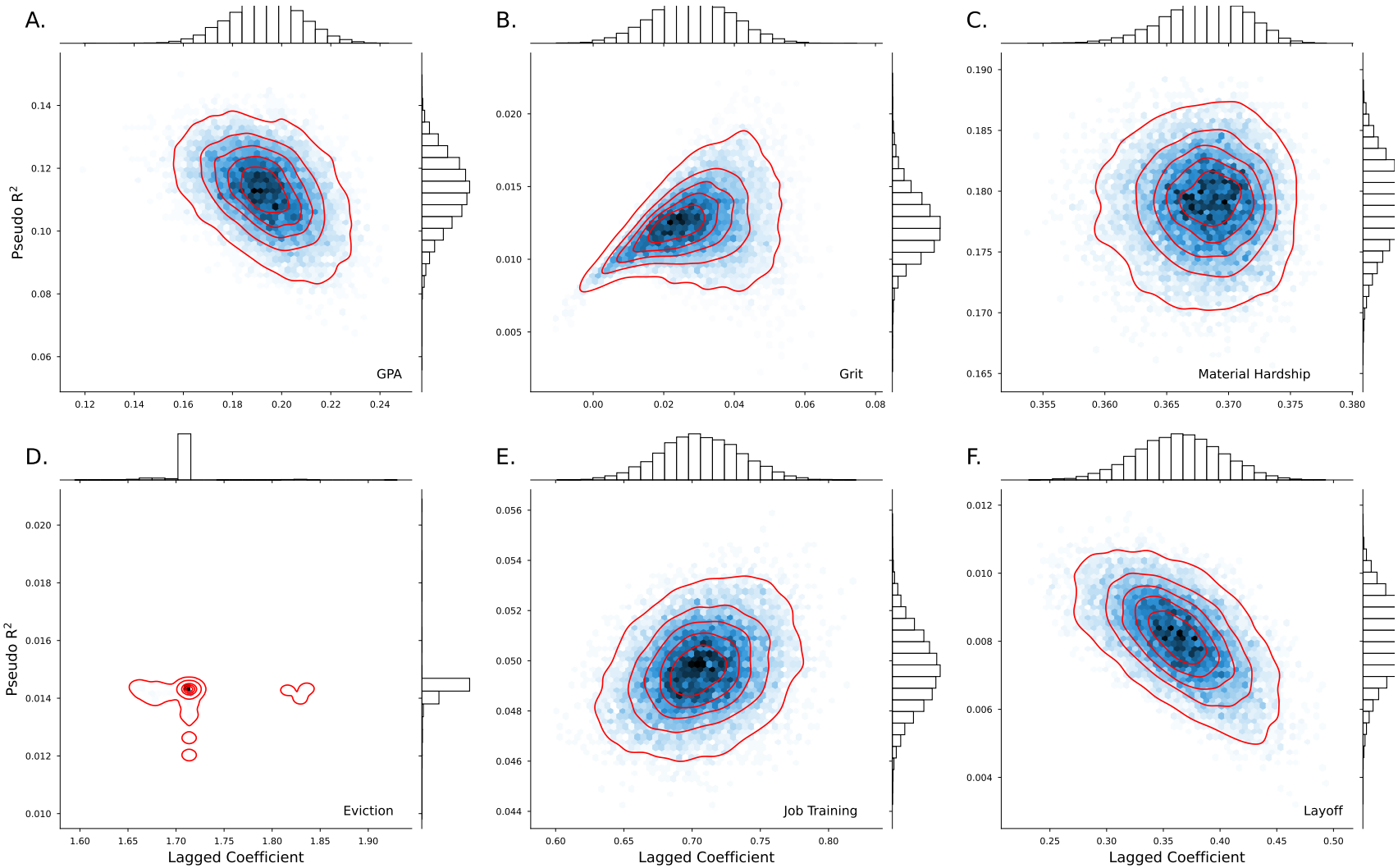
- **Solution**: Visualize the outcome space of a of large number (10k? 100k?) seeds simultaneously. This is computationally intensive, but scientifically faithful.

- **Replication**: We can consider whether an original result is in the tail of our distribution (or IQR?) or not.


A.
Upper Limit
Lower Limit
π
Range
Estimate of π
4.5
4.0
3.5
3.0
2.5
0
5000
10000
15000
20000
25000
Number of Throws


A.
E(ROC-AUC) = 0.743, σ(ROC-AUC) = 0.017
Bins
KDE
Density
ROC-AUC
B.
E(R²) = 0.655, σ(R²) = 0.006
Bins
KDE
R²
C.
E(acc) = 9623.302, σ(acc) = 65.879
Bins
KDE
Accuracy


A.
Median
Bounds
0.1
0.0
−0.1
−0.2
−0.3
Effect Size (β̂)
0
4000
8000
Specification (n)
1e7
1.25
1.00
0.75
0.50
0.25
0.00
Frequency
−0.3
−0.2
−0.1
0.0
0.1
Effect Size (β̂)
B.
In-sample
Min/Max
Median
Range
$100k
$80k
$60k
$40k
$20k
Price
2021-06-22
2022-07-27
2023-08-31


A.
B.
C.
D.
E.
F.
Pseudo R^2
GPA
Grit
Material Hardship
Eviction
Job Training
Layoff
Lagged Coefficient

What's Next? Indexing Seed Variability

- The next thing to do is to formalize the replication project.
  - Does anyone have ideas for other types of seed variability?
- We are seeking to make a project website, which:
  1. Provides (a set of) seeds (already downloadable on GitHub).
  2. Indexes examples of seed variability in existent papers.
- The first improves the future of the scientific record.
- The second makes the historic scientific record more robust.