RobustiPy Hackathon

27th June, 2024

Daniel Valdenegro Ibarra, Charlie Rahal, and Jiani Yan

Said Business School, University of Oxford

UNIVERSITY OF OXFORD
OXFORD POPULATION HEALTH
DSU

LEVERHULME CENTRE FOR DEMOGRAPHIC SCIENCE

Centre for Care

Maybe we can introduce ourselves?

- ◇ Welcome everyone to the RobustiPy hackathon!
- ◇ Maybe we can do a quick round of introductions?
  1. What is your name?
  2. What are you interested in?
  3. How comfortable are you with ideas of model uncertainty?
  4. How much Python programming have you done before?

Useful Links

This slide contains all the useful links you might need:

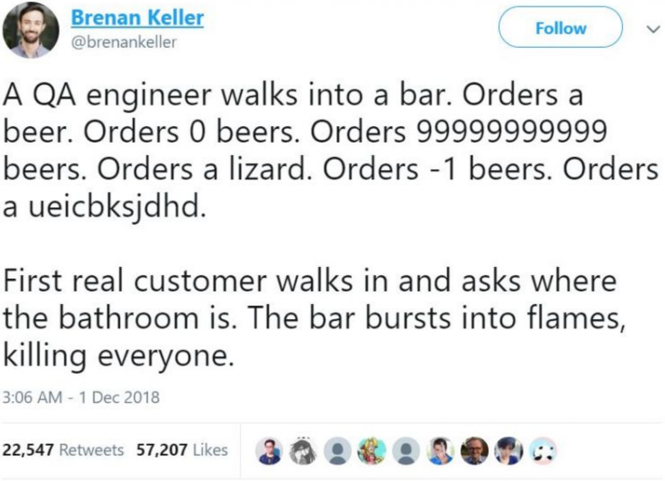
- For the RobustiPy website, see here!
- For an open Google Doc for you to drop issues into, see here!
- For a Health and Safety video for Said, click here.
- For the organisational RobustiPy GitHub account (inc. examples), see here!
- For the RobustiPy Python *repository*, see here!
- For a place to upload *your* slides, please click here!
- For the 'Improving RobustiPy' Jamboard session, see here!

A couple of useful tools!

- Hopefully we've been able to get you all set up and running with RobustiPy!
- There are three things we'd especially recommend for days like today:
  1. A text editor (**PyCharm**, **VSCode**).
  2. A **GitHub** account (very quick to register).
  3. A Google **Colab** account (very quick to register).
- For those of you who feel comfortable:
  - Please feel free to fork, make issues, and raise PRs throughout the day!
- For those less comfortable with Git, please drop issues into our Google Doc tracker!

What is GitHub?

- **Version Control:** Web-based, used for managing/tracking code changes using Git.
- **Collaboration:** Teamwork with features for code review, issue tracking, etc.
- **Repository Hosting:** Stores code and project files in public or private repositories.
- **Open Source Community:** Centralises contributing to/managing projects.
- **Documentation:** Create and host documentation/wikis within repositories.
- **Integrations:** Connects with various development tools/services.



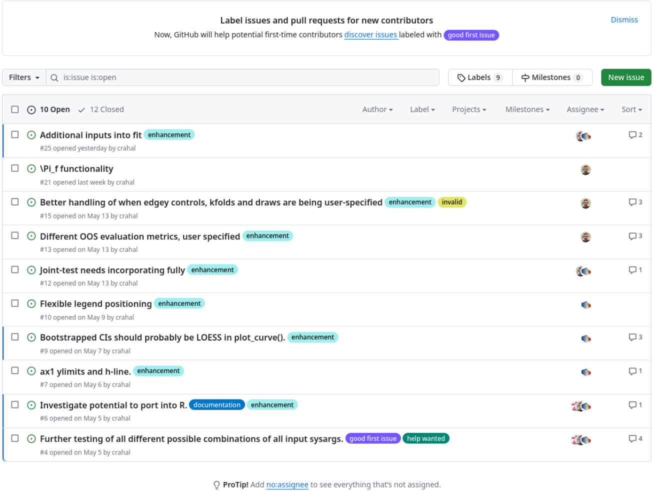
Brenan Keller
@brenankeller

A QA engineer walks into a bar. Orders a beer. Orders 0 beers. Orders 99999999999 beers. Orders a lizard. Orders -1 beers. Orders a ueicbksjdhd.

First real customer walks in and asks where the bathroom is. The bar bursts into flames, killing everyone.

3:06 AM - 1 Dec 2018

22,547 Retweets 57,207 Likes



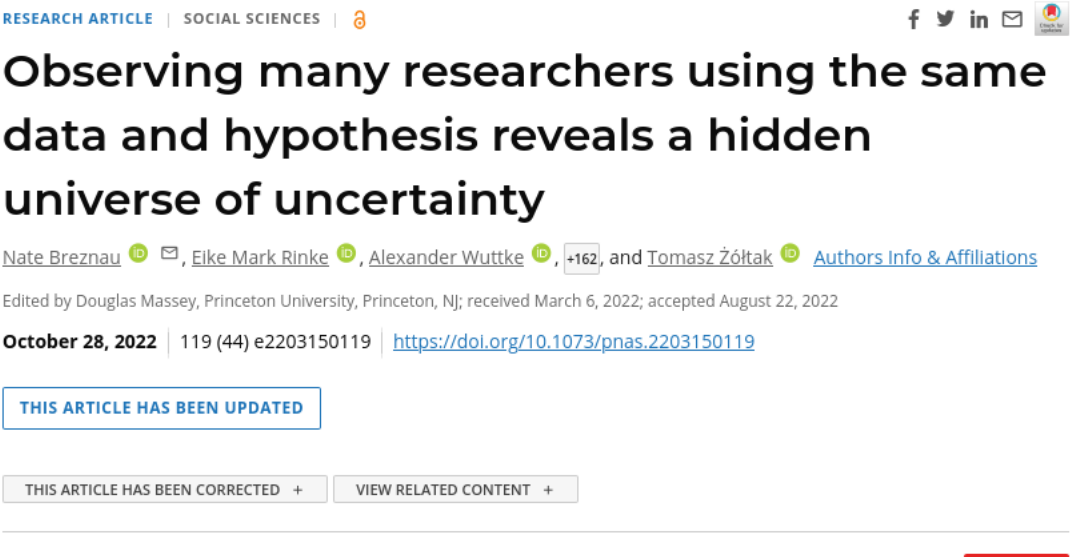
Label issues and pull requests for new contributors

Now, GitHub will help potential first-time contributors discover issues labeled with good first issue

○ 10 Open ✓ 12 Closed

- □ ○ **Additional inputs into fit** enhancement — #25 opened yesterday by crahal — 2
- □ ○ **\Pi_f functionality** — #21 opened last week by crahal
- □ ○ **Better handling of when edgey controls, kfolds and draws are being user-specified** enhancement invalid — #15 opened on May 13 by crahal — 3
- □ ○ **Different OOS evaluation metrics, user specified** enhancement — #13 opened on May 13 by crahal — 3
- □ ○ **Joint-test needs incorporating fully** enhancement — #12 opened on May 13 by crahal — 1
- □ ○ **Flexible legend positioning** enhancement — #10 opened on May 9 by crahal
- □ ○ **Bootstrapped CIs should probably be LOESS in plot_curve().** enhancement — #9 opened on May 7 by crahal — 3
- □ ○ **ax1 ylimits and h-line.** enhancement — #7 opened on May 6 by crahal — 1
- □ ○ **Investigate potential to port into R.** documentation enhancement — #6 opened on May 5 by crahal — 1
- □ ○ **Further testing of all different possible combinations of all input sysargs.** good first issue help wanted — #4 opened on May 5 by crahal — 4

ProTip! Add no:assignee to see everything that's not assigned.



RESEARCH ARTICLE | SOCIAL SCIENCES |

Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty

Nate Breznau, Eike Mark Rinke, Alexander Wuttke, +162, and Tomasz Żółtak Authors Info & Affiliations

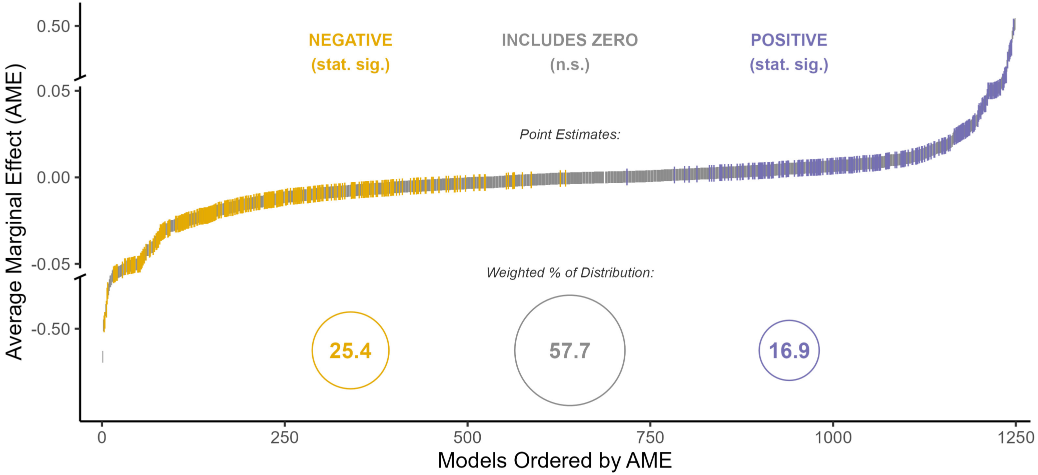
Edited by Douglas Massey, Princeton University, Princeton, NJ; received March 6, 2022; accepted August 22, 2022

October 28, 2022 | 119 (44) e2203150119 | https://doi.org/10.1073/pnas.2203150119

THIS ARTICLE HAS BEEN UPDATED

THIS ARTICLE HAS BEEN CORRECTED +

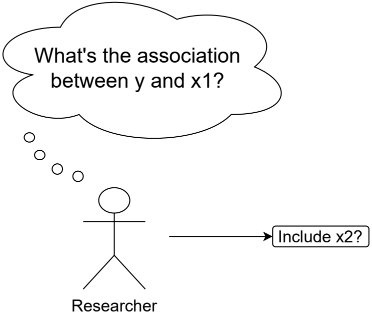
VIEW RELATED CONTENT +

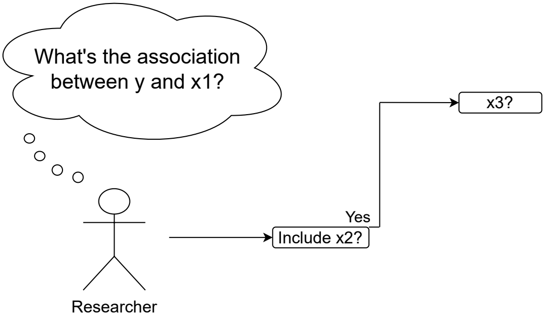

NEGATIVE
(stat. sig.)
INCLUDES ZERO
(n.s.)
POSITIVE
(stat. sig.)
Point Estimates:
Weighted % of Distribution:
25.4
57.7
16.9
Average Marginal Effect (AME)
0.50
0.05
0.00
-0.05
-0.50
0
250
500
750
1000
1250
Models Ordered by AME

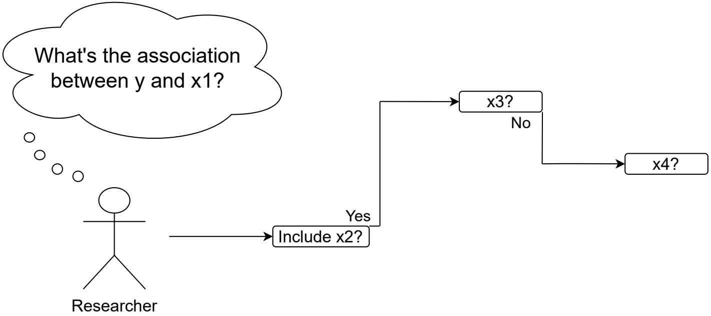
The problem: Types of Model Uncertainty

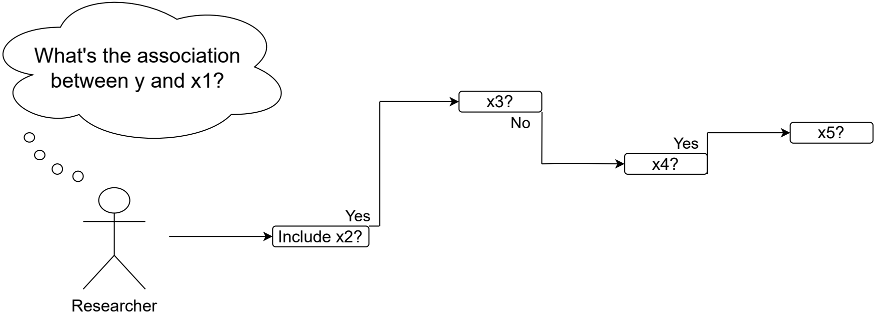
- **Parameter Uncertainty**:
  - Variability due to estimation from limited data.
- **Structural Uncertainty**:
  - Inadequate model to capture system complexities fully.
  - Simplifications that may oversimplify or miss critical aspects.
- **Data Uncertainty**:
  - Errors in data collection or measurement.
  - Outliers affecting model stability or accuracy.
- **Model Formulation Uncertainty**:
  - Choice of mathematical model structure.
  - Assumptions about relationships between variables.

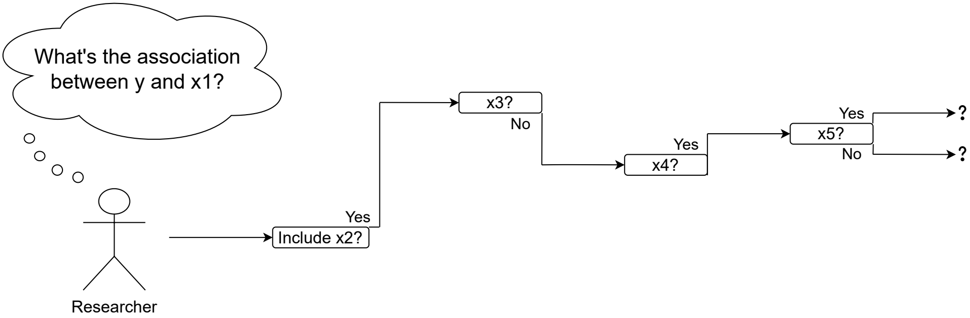
Some types of uncertainty can be handled by applied researchers, some can't.

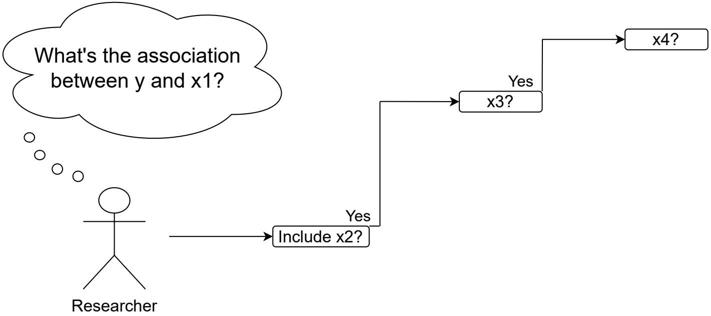

What's the association between y and x1?
Researcher
Include x2?

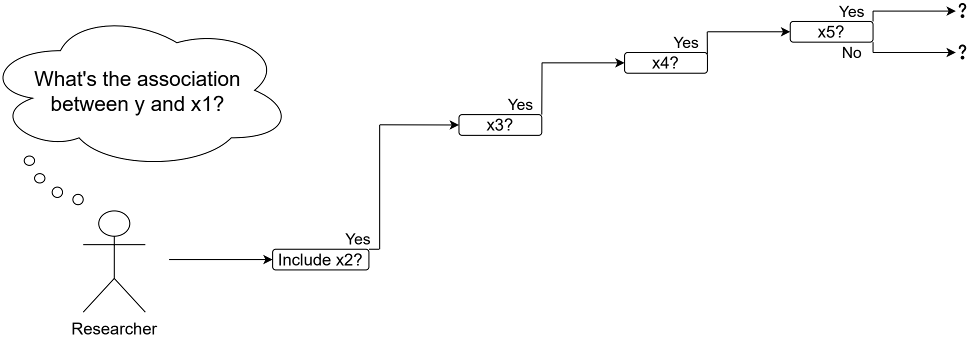

What's the association between y and x1?
Researcher
Include x2?
Yes
x3?

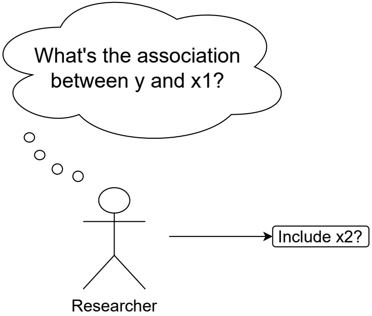

What's the association between y and x1?
Researcher
Include x2?
Yes
x3?
No
x4?

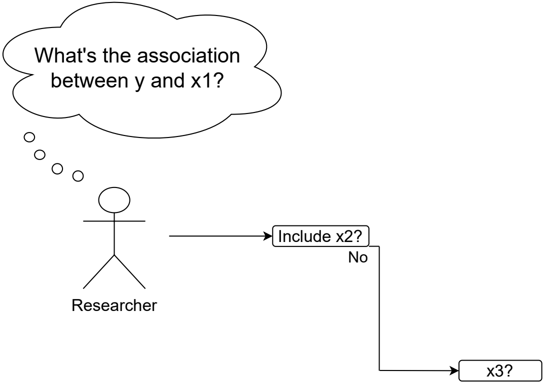

What's the association between y and x1?
Researcher
Include x2?
Yes
x3?
No
x4?
Yes
x5?

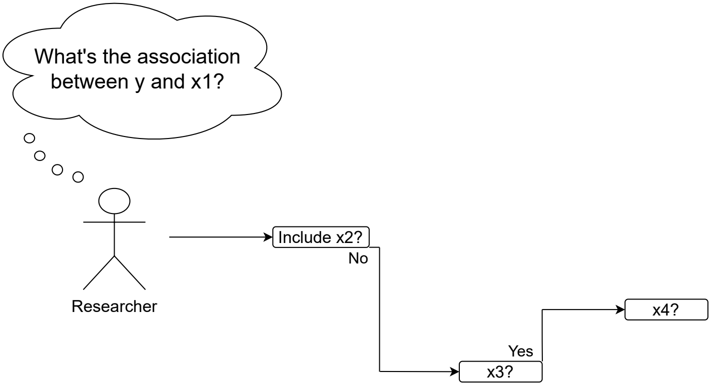

What's the association between y and x1?
Researcher
Include x2?
Yes
x3?
No
x4?
Yes
x5?
Yes
No
?
?

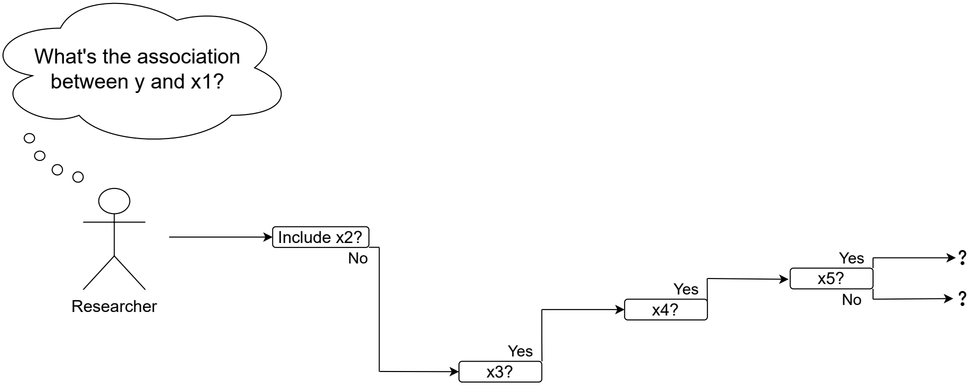

What's the association between y and x1?
Researcher
Include x2?
Yes
x3?
Yes
x4?

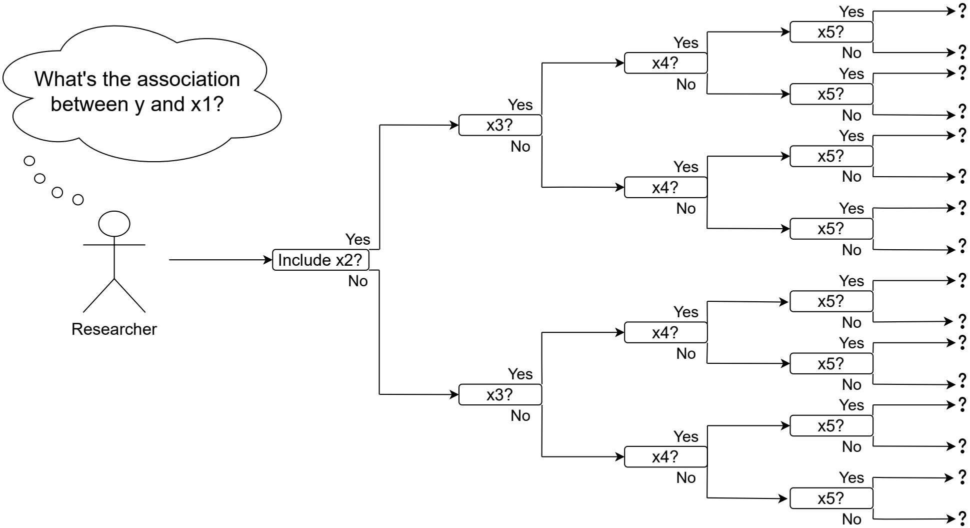

What's the association between y and x1?
Researcher
Include x2?
Yes
x3?
Yes
x4?
Yes
x5?
Yes
No
?
?


What's the association between y and x1?
Researcher
Include x2?


What's the association between y and x1?
Researcher
Include x2?
No
x3?


What's the association between y and x1?
Researcher
Include x2?
No
x3?
Yes
x4?


What's the association between y and x1?
Researcher
Include x2?
No
x3?
Yes
x4?
Yes
x5?
Yes
?
No
?


What's the association between y and x1?
Researcher
Include x2?
Yes
No
x3?
Yes
No
x3?
Yes
No
x4?
Yes
No
x4?
Yes
No
x4?
Yes
No
x4?
Yes
No
x5?
Yes
No
x5?
Yes
No
x5?
Yes
No
x5?
Yes
No
x5?
Yes
No
x5?
Yes
No
x5?
Yes
No
x5?
Yes
No
?
?
?
?
?
?
?
?
?
?
?
?
?
?
?
?

Introducing RobustiPy

- Multiverse Analysis is ideal to tackle P-Hacking and Harking.
  - Both are big issues for the reproducibility of scientific findings.
- However, multiverse analysis is new: no *established* tool exists to conduct it reliably.
- Additionally, not all researchers are, nor need to be expert programmers.
- So, we set ourselves the task:

Can we create a reliable tool to conduct this/adjacent analysis?

RobustiPy: The formal problem

Consider the key association between variables Y and X, where a set of covariates **Z** can influence the relationship between the former as follows:

$$Y = F(X, \mathbf{Z}) + \epsilon. \tag{1}$$

Let's now observe that:

- Usually Y and X are imprecisely defined latent variables.
- Likewise, the set of covariates **Z** can also be composed of imprecisely defined variables from which the number of them can be unknown and/or non-finite.
- Finally, $F()$ is an unknown data generating function.

RobustiPy: The formal problem

Let's define the set of *reasonable* operationalisations of Y, X, $\mathbf{Z}$ and $F()$ as $\overleftrightarrow{Y}$, $\overleftrightarrow{Y}$, $\overleftrightarrow{\mathbf{Z}}$ and $\overleftrightarrow{F()}$. Then, we have:

$$\overleftrightarrow{Y}_{k_Y} = \overleftrightarrow{F}_{k_f}(\overleftrightarrow{X}_{k_X}, \overleftrightarrow{\mathbf{Z}}_{k_{\mathbf{Z}}}) + \epsilon. \tag{2}$$

- Eq. 2 corresponds to a single possible *specification* of the Π set of all possible specifications.
- The total number of specifications can then be calculated as 2^N where N is $n_{\overleftrightarrow{Y}} + n_{\overleftrightarrow{X}} + n_{\overleftrightarrow{\mathbf{Z}}} + n_{\overleftrightarrow{F()}}$.
- For example, an analysis with two functional forms, two target variables, two predictors and five covariates will yield a specification space Π of 2^{10} or 1024 possible specifications.

RobustiPy: The formal problem

We can also split the computation based on the operationalised variables in Eq. 2 by creating a ‘reasonable specification space‘ for each: $\Pi_{\overleftrightarrow{Y}}, \Pi_{\overleftrightarrow{F()}}, \Pi_{\overleftrightarrow{X}}, \Pi_{\overleftrightarrow{Z}}$. The whole specification space can be obtained again as follows:

$$\Pi = \Pi_{\overleftrightarrow{Y}} \times \Pi_{\overleftrightarrow{F()}} \times \Pi_{\overleftrightarrow{X}} \times \Pi_{\overleftrightarrow{\mathbf{Z}}} \tag{3}$$

- Any random and independently selected sample π from Π would lead to a reasonable approximation of $Y = F(X, \mathbf{Z})$.
- The main problem with current research practice is that the sample of Π reported by the researchers is not random.
- **The goal of RobustiPy is to automate this process to make it as easy as possible.**

RobustiPy in Action

Link to live **demo is here**!

- Things to discuss:
  - How do we load RobustiPy? OLSRobust or LRobust?
  - What are the *essential* inputs to these Python Classes?
  - What's this about `y`, `x`, and `c`? What object types should they be?
  - What are the optional inputs? Why is this taking so long?
  - How do I get the results? What can I do with them?

Empirical Example Time!

- Please find here a link to some empirical examples!

1. This is our canonical ‘Union’ example **here**.
   - Compare results to Young and Holsteen (2016) if you like!
2. There is an example of how LRobust runs **here**.
   - Is it faster or slower than OLSRobust?
3. There is an example of how the grouping functionality runs can be found **here**.
   - Could or should this work with LRobust?
4. Although we can’t easily share the data with you, we have an example in the Adult Social Care space which combines grouping and multiple `x` **here**.
   - Why do we *need* `x` here?

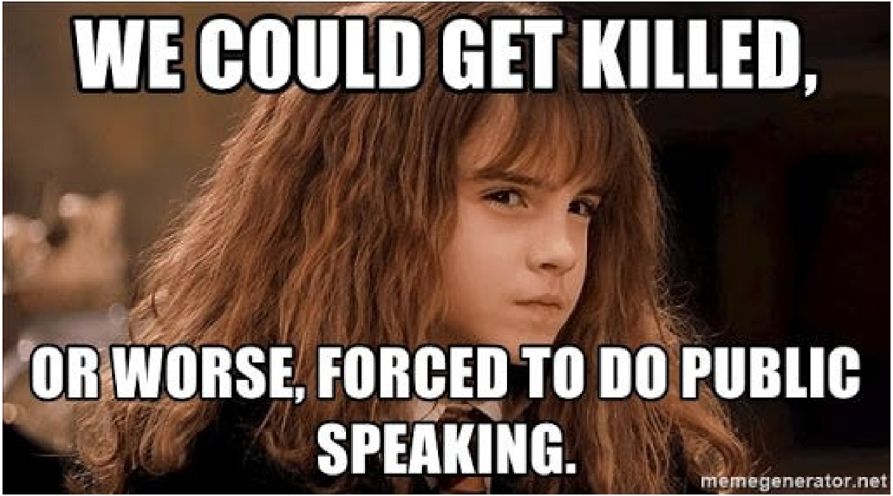
Simulated Example Time!

- A link to the first simulation can be found **here**!
  - This is a simple example which has three control variables and uses OLSRobust.
- A link to the second simulation can be found **here**!
  - This uses grouping data.
- A link to the third simulation can be found **here**!
  - This uses a different LRobust class and a different out-of-sample metric.

Lets take a break and get some pizza!

Bring Your Own Dataset!

- Now it's time to Bring Your Own Dataset!
- Choose to work individually, or in groups, depending on how many datasets there are.
- For anyone *without* a dataset, why not try the **infamous titanic dataset**?
  - Take your dataset, and process it as necessary.
  - Determine your `y`, x and `c` fields.
  - Run OLSRobust/LRobust as necessary depending on your dataset's dependent variable.
  - Set your key variables, such as number of `draws`, or `kfolds`.
  - Visualise your results, with some alternative specifications.
  - Please upload your slides **here**!


WE COULD GET KILLED,
OR WORSE, FORCED TO DO PUBLIC SPEAKING.
memegenerator.net

WINNERS, WINNERS
EVERYWHERE

How can we improve RobustiPy?

Please use the Jamboard: thoughts/ideas on how can we improve the RobustiPy library!

Also: How can we best *present* RobustiPy, including sample datasets/examples?

Discussion Questions 1: Defining Model Uncertainty

- How do we currently define and quantify model uncertainty in academic research?
- What types of uncertainty does RobustiPy capture?
- How does the necessity of capturing such things vary by academic field?
- What are the major challenges researchers face?
  - Are they more about accurate modelling, or communicating uncertainty?

Discussion Questions 2: Impact on Research Validity and Reproducibility

- How does model uncertainty impact the validity and reproducibility of research findings?
- What steps can researchers take to mitigate the effects of uncertainty on their results?
- How should researchers communicate model uncertainty?
- What role does transparency in uncertainty play in the credibility of academic research?

Discussion Questions 3: Educational and Training Needs

- What is the current status of teaching model uncertainty?
- What is needed to improve understanding and management of uncertainty?
- What role do open-source frameworks and collaborative platforms play?
- How commonly utilised are Open Science tools and ideas in *your* discipline?
    - How widely are *they* taught?

Discussion Questions 4: Ethical Considerations

- How trusted are academic ‘experts’ right now in general?
- Are there ethical implications related to how model uncertainty is managed and communicated in academic research?
- Is RobustiPy a tool that can be used for evil, as well as good?
- How should uncertainty be addressed when research findings influence policy or public perception?

Discussion Questions 5: Impact of Emerging Technologies

- How might emerging technologies (e.g., quantum computing, advanced simulations) impact our ability to handle and reduce model uncertainty?
- How can advancements in AI and machine learning contribute to improving uncertainty estimation?
- How can interdisciplinary collaboration drive innovation in uncertainty modelling and management?
- What are the potential risks and opportunities associated with these technologies in uncertainty modelling?

Thank You!

Thank you so much all for coming!

And importantly, thanks to Bradley, Hannah, and Dan for all of the logistical help!