1

The Many Worlds of Analysis

> A fragile inference is not worth taking seriously.
>
> —Edward Leamer (1985)

> The best defense against subjectivity in science is to expose it.
>
> —Silberzahn et al. (2018: 354)

Statistical models are, at best, only approximations of reality. Econometric theory is built on imperfect assumptions and provides only "inexact … guidance about how to do empirical research" (Solon, Haider, and Wooldridge 2015: 311). The link between social theories and statistical models is often vague, open to debate, and dependent on many auxiliary assumptions (Western 1996; Strevens 2020). Testing a hypothesis – estimating a coefficient – requires taking many approximate inputs from social and statistical theory and turning them into a single, exact regression model. Out of many fuzzy things, one.

There are many ways of conducting an analysis, but most studies report only a few carefully curated estimates. Behind the curtain, in the backstage realm of research, lie many worlds of alternative analyses that could have been conducted: alternative models and alternative results. Multiverse analysis explores and reports on this often-hidden world between theory and data. From one published analysis, we can imagine many feasible alternatives.

The principle of robustness is central to modern science. In the most general sense, robustness refers to "situations in which something is stable under variations of something else" (Basso 2017: 57). This book is about model robustness: where an estimate – a regression coefficient of

Downloaded from https://www.cambridge.org/core. IP address: 216.73.216.25, on 25 Jul 2025 at 07:10:45, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009003391.002

interest – is (or is not) stable under different variations of the model. And model robustness is fundamental to the credibility of research.

Researchers want to be able to say: "This is not my opinion, this is what the evidence says." Multiverse analysis gives us a tool to *show* how much results are driven by the evidence rather than by subjective researcher choices and assumptions. Statistical models involve so many unique decisions that they become a "garden of forking paths" (Borges 1941; Gelman and Loken 2014). In theory, there is a single "true" model of the data generation process, but that model is almost never known. In practice, a single-path analysis represents a bundle of assumptions: ideas about the correct choice of controls, functional form, estimation command, variable definitions, and more, which are not yet proven to be true. A single-path point estimate reflects just one ad hoc route through the forking paths world. Different researchers studying the same question almost never use the same models (Breznau et al. 2022). Sometimes a different model would give a similar answer, but other times it might diverge dramatically. It can be difficult for readers to know if a result is driven more by the data or by the author's model assumptions. The raw data are often external to a researcher and must be accepted as given, but the model assumptions are not. Typically, researchers decide on their model assumptions with the data in hand, and they can see which assumptions favor their hypothesis. This is a problem of asymmetric information between analyst and reader: Analysts with data in hand know much more about the sensitivity of results than do readers, who have access only to the curated results published in the paper.

The inspiration and language of the multiverse come from quantum physics and cosmology (Gribbin 2009; Carroll 2019a). In a multiverse, there is more universe and there are more worlds than we can currently see. We know the universe extends beyond the cosmic horizon of our best telescopes and instruments, but we can only guess what might be out there. The universe is what we can *see*, while the multiverse is everything that *exists*. When applied to methodology, the multiverse means there are many more ways of estimating a parameter than what any one study shows. Individual papers tend to offer only a narrow horizon into the plausible model space. But scholars are awash with computational power and can easily estimate a vast number of models prior to selecting a careful few for publication. The multiverse software we used in preparing this book can estimate 1,000 unique model specifications in a matter of seconds using a normal computer. In an hour, one can see typically results from 100,000 model variations. The most ambitious multiverse

Downloaded from https://www.cambridge.org/core. IP address: 216.73.216.25, on 25 Jul 2025 at 07:10:45, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009003391.002

analysis on record ran more than nine billion regressions (Muñoz and Young 2018). With such staggering computational power in the hands of analysts, thinking in terms of one estimate is anachronistic at best.

There are many ways of thinking about model robustness, and including some version of robustness analysis has become increasingly common in quantitative research papers. Social scientists often publish tables with only a few specifications but also have "robustness footnotes" mentioning other models that, inevitably, are said to "show the same results" (Young and Holsteen 2017). These footnotes are weakly transparent, but they at least acknowledge the existence of other plausible model assumptions. In meta-analysis studies of the existing literature, we see a kind of multiverse of models that have been used in past research. And meta-analysis routinely reports that individual studies are a poor guide to the true range of results that multiple studies show (Stanley and Doucouliagos 2012). Each individual study reports a point estimate and a confidence interval, and in theory that confidence interval shows the range of results that should occur 95 percent of the time. But it is very common for the next study to show an estimate entirely outside that interval. Compared to the wide range of results seen in most social science literatures, the individual studies making up the literature all seem very *overconfident*. This is because confidence intervals do not take into account *model error* or the possibility that other studies will make different assumptions and use different methods.

One response to the troubling range of results in the published literature is to organize an adversarial collaboration: Researchers with rival views and prior beliefs agree to jointly analyze one dataset (Mellers, Hertwig, and Kahneman 2001; Clark et al. 2022). The resulting publication shows the strongest possible results from each side in the debate. This elegantly shows (1) how much *common support* the data provide to each side of a debate and (2) how much the modeling assumptions shape what each side can claim. Often, adversarial collaborations do not result in agreement between the different sides, but rather they help clarify which auxiliary assumptions drive their disagreement and help build a future research agenda for new data collection and new empirical testing.

Many-analysts studies expand on this approach by drawing in modeling expertise from larger and more diverse groups of researchers, all studying the same question with the same raw dataset. An emerging consensus from these many-analysts trials is that no two researchers ever use the same model specification nor ever get exactly the same results (Silberzahn et al. 2018; Schweinsberg et al. 2021; Breznau et al. 2022). Participants

Downloaded from https://www.cambridge.org/core. IP address: 216.73.216.25, on 25 Jul 2025 at 07:10:45, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009003391.002

are routinely surprised by the variation across other participants' estimates. A number of studies have asked researchers, after they completed their own analysis of the data, to predict the range of results from other research teams. Researchers almost uniformly underestimate the range of models that other intelligent people will think of. "Individual scientists do not appreciate how different their peers' analytical choices are and how much results will be affected" (Camerer 2022: 3). When you consider a model specification but eventually decide "nobody would run that model," you are likely wrong. In crowdsourcing studies, knowledgeable scholars as a collective seem willing to run almost any plausible model – and the diversity of methods and results is not explained by researcher training, experience, publication record, or even peer evaluations of quality. This is not a world where "bad scholars" use "bad models," while "good scholars" use "good models." Scholars should embrace thinking in the gray – the gray zone between one's own first-choice method and alternative methods that could be defended by others. Between those two points are a range of methodological strategies that deserve attention.

WHERE DO "MANY MODELS" COME FROM?

Ideal-Type Approaches

There are two ideal-type illustrations of how to develop a large set of plausible models that define the model space. The first approach is what we call the "super log file" approach, which captures any model a researcher ever estimated or looked at in a project. The second approach uses a task force of experts representing theory competition and adversarial collaboration. Neither of these approaches are practical for day-to-day work, but computational methods aim to approximate their best features.

The Super Log File Approach

An interesting feature of Excel files is that they remember every computation that was ever conducted, with or without the author knowing or wanting it.[1] In contrast, when researchers use Stata or R, they have to choose what parts of their work get recorded and saved for others to see.

[1] This feature of Excel has been used to identify evidence of manually tampering with data to generate supporting evidence in social science publications (https://datacolada.org/109).

Downloaded from https://www.cambridge.org/core. IP address: 216.73.216.25, on 25 Jul 2025 at 07:10:45, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009003391.002

There is an unknown selection process to what researchers chose to publish. However, imagine that statistical software kept a super log file that automatically captured the results of every unique regression a researcher ever ran in the course of studying their data and preparing an article. Once the project is finalized, the log program generates a graph showing every unique regression result an author ever looked at.

The philosophy here is that any model a researcher considered worth running is also worth reporting (even if the model could be criticized – as all can be). This is full disclosure of all results the author has ever seen. If an author chooses to run a model specification, it becomes part of a permanent record available to all readers. We like this thought experiment for two reasons: (1) It allows authors to disclose the many ways they a priori think a model could be credibly specified and (2) it equalizes the information asymmetry between authors and readers – authors can see an estimate only if they are also willing to show it to their readers.

The Task Force Approach

Another ideal-type way to develop the model space is to convene a task force of specialists to study an important social question. The task force would reflect on a range of disciplinary and political perspectives, ensuring a healthy dose of theory competition and adversarial collaboration (Mellers et al. 2001; Doucouliagos and Stanley 2013). Any model specification that a task force member credibly argues for becomes part of the model space. There might be one model and estimate that gets the most votes by the majority of the task force, but a graphical display shows what results can be found by serious scholars using credible alternative methods. Dissenting votes and rejected model specifications are part of the public record. The final report might include any number of different specifications that best reflect the methodological views among the task force. One example is the American Psychological Association task force on the relationships between race, genetics, and intelligence, published as "Intelligence: Knowns and Unknowns" (Neisser et al. 1996). The task force sought to "make clear what has been scientifically established, what is presently in dispute, and what is still unknown" (Neisser et al. 1996: 77). These kinds of prestige task forces are rare, but they provide an ideal of how to elicit a wide range of analytical views from top scholars in a field. In recent years, crowdsourcing studies have sought to emulate the task force approach, recruiting many scholars to analyze a specific

Downloaded from https://www.cambridge.org/core. IP address: 216.73.216.25, on 25 Jul 2025 at 07:10:45, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009003391.002

question using a shared dataset, with each participant sending back their preferred specification, code, and estimate.

The Computational Multiverse

Both of the aforementioned approaches involve running many unique models and reporting a distribution of results in graphical form. Computational model robustness aims to incorporate features of both the super log file and the task force approaches. The objective is to reduce the discretion of authors to pick an exactly preferred model and result (the strength of the super log file approach) while expanding the range of models and results that any one author considers (the merit of the task force approach). The method involves specifying a set of plausible model ingredients (including possible controls, variable definitions, estimation commands, and standard error calculations) and estimating all possible combinations of those model ingredients. The principle is to use only vetted, credible model inputs, as any author would do when selecting a single estimate, but then report back every estimate that can be obtained from those inputs. It perturbates the model using a combinations algorithm while also reporting how much each modeling input (or assumption) matters for the results.

To be clear, a computational approach can only *aspire to* the breadth of insight available in an expert task force assembled for adversarial collaboration. It requires users to specify credible alternatives for each model input. But the checklist is valuable for any author to work through. For each control variable, is the variable strictly necessary or is it possibly a bad or unnecessary control? What arguments could be made against including a control? For each equation, is there another credible functional form – another way to link the left- and right-hand sides of the model? For any variable in the system, could it be defined or coded in a different and possibly better way? After working through these questions, the resulting *modeling distribution* shows what estimates are possible, while *model influence* shows how these decisions affect the results. As we will see in Chapter 9, Figure 9.5, applying this set of questions to a project generates a multiverse of possible models that is at least similar to the range of models in a many-analysts study. And in a multiverse analysis of intergenerational mobility Engzell and Mood (2023) showed how this process is constructive, developmental, and informative and how working through the many decisions a researcher inevitably has to make can yield unexpected insight.

Downloaded from https://www.cambridge.org/core. IP address: 216.73.216.25, on 25 Jul 2025 at 07:10:45, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009003391.002

The Central Theorem of Multiverse Analysis

The core principles of multiverse analysis are as follows: Confidence intervals never show the true range of credible results. Every analysis depends on untested assumptions that are never exactly correct. Every analysis is a rough approximation of the true model. There is always model error; it is just rarely acknowledged.

The aim of multiverse methods is to reduce the discretion of authors to pick an exactly preferred model and result while expanding the range of models and results under consideration. The method involves specifying a set of plausible model ingredients (including possible controls, variable definitions, estimation commands, and standard error calculations) and estimating all possible combinations of those model ingredients. Acknowledging those other paths allows multiple plausible models and yields a modeling *distribution* of estimates. A single-path point estimate is a "best guess" starting place to enter the multiverse: a reference point from which to define alternative assumptions and to see how different the alternative estimates are from the author's *first choice* of model specification. From here, we leave the point estimate behind and think primarily in terms of distributions: What is the range of plausible estimates from alternative models? How many model assumptions can be relaxed without overturning an empirical conclusion? Which model assumptions affect the results the most?

The multiverse approach goes far beyond simply generating possible models; it demands careful thinking about model specifications and their underlying assumptions. It calls for prudent interpretation and highlights what methods, techniques, and assumptions need rigorous evaluation before they can be considered a credible part of the analysis. When a finding lacks robustness to model specification, this introduces a *methodological scope condition* that not only shows under what conditions a result holds but also serves as a guide to further deliberation and research.

Notes on the Multiverse Metaphor

The term "multiverse analysis" was first used by Steegen et al. (2016); we assume, but do not know for sure, that this imagery comes from Andrew Gelman, a coauthor of the study and a statistician who has a great gift for vivid writing. This language, in our work, now replaces less inspired but perhaps more descriptive terms like "multi-model analysis" or a framework for "model uncertainty and robustness" (e.g., Young and Holsteen

Downloaded from https://www.cambridge.org/core. IP address: 216.73.216.25, on 25 Jul 2025 at 07:10:45, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009003391.002

2017). The multiverse concept encapsulates the problem of uncertainty and the solution of robustness, all while being packaged in a metaphor that sparks the imagination.

At the same time, we must acknowledge that some scholars have reservations about the term due to its ascendance in popular culture, science fiction circles, and superhero movies. This can lead to misconceptions or oversimplifications of the idea when applied to scholarly discourse. We recognize the concern but push back in part because we welcome a newer style of methodological terminology. Classical statistics and econometrics have a dismal and stodgy record of naming new methods. In the formative years of statistics, new concepts were given intimidating, polysyllabic names derived from Greek and Latin: heteroskedasticity, autocorrelation, multicollinearity, nonparametric, kurtosis, and endogeneity. What is distinctive about these terms is that they have no common meaning in English and serve as purely technical constructs that intimidate outsiders. In this style, the problem of "small sample size" could be given more scientific gravitas by calling it "micro-numerosity."

With the rise of data science, developments in statistical methods are given more informal and vivid, often playful, names. Early examples are the bootstrap and the jackknife: resampling methods that invoked folk terminology to suggest their underlying logic (e.g., the jackknife uses a one-at-a-time resampling method, in analogy to how a Swiss army knife has many blades that can be taken out one at a time). More recent data science tools come with names like neural networks, decision trees, and random forests, which are all variants of machine learning. The spirit of data science naming conventions has been to make the language more vivid, approachable, and even fun to talk about.

The language and empirical imagery of the multiverse in social science is catching on fast. To illustrate, we list disciplines that have, following Steegen et al. (2016), published studies that embrace the language of multiverse analysis. The language and methods of multiverse analysis are experiencing rapid take-up in the social sciences.

RECENT ARTICLES USING MULTIVERSE LANGUAGE AND METHODS

Sociology: Engzell and Mood (2023); Young and Stewart (2021); Auspurg and Brüderl (2021)

Computer Science: Liu et al. (2021); Hall et al. (2022); Sarma et al. (2023)

Downloaded from https://www.cambridge.org/core. IP address: 216.73.216.25, on 25 Jul 2025 at 07:10:45, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009003391.002

Political Science: Saraceno, Hansen, and Treul (2021)
Psychology: Harder (2020); Modecki et al. (2020); Olsson-Collentine et al. (2023)
Public Policy: Breznau et al. (2022)
Education: Robitzsch (2022); Herrala (2023); Neuendorf and Jansen (2023)
Organizational Behavior: Schweinsberg et al. (2021)
Religion: Hanel and Zarzeczna (2023)
Health and Epidemiology: Cantone and Tomaselli (2023); Levitt, Zonta, and Ioannidis (2023); Rengasamy et al. (2023)

OUTLINE OF THE BOOK

This book walks readers through every aspect of a rigorous multiverse analysis, drawing on real-life datasets and providing code for others to use in their own work (and to replicate our work). In this process we believe that almost everyone's beliefs about modeling assumptions will be deeply challenged. The goal is to better understand how data and assumptions work together to produce empirical estimates. In Chapter 2, we round out the introduction of this book by discussing the multiverse as a philosophy of science.

As we move on to Part II, The Computational Multiverse, we start with a vivid empirical case: research claiming that female hurricanes are deadlier than male hurricanes (Chapter 3). We demonstrate multiverse analysis using analytical inputs from many scholars in a high-profile empirical debate. The original claims appear remarkably weak: 99.7 percent of alternative models show weaker results, and 88 percent of models report null findings. From here, we cover the core methodology of multiverse methods across five chapters (Chapters 4–8). First we aim at understanding that the modeling distribution is distinct from the sampling distribution and at applying the method to multiple datasets. Next we discuss the second pillar of multiverse methods: influence analysis, which documents how different features of model specification (such as individual controls) affect the results. Part II mostly builds the foundations of multiverse methods using assumptions about control variables and discusses in depth the complexity and difficulty of assuming that a control variable belongs or does not belong in a model (Chapter 7, "Good and Bad Controls").

Part III, Expanding the Multiverse, explores the next two dimensions of modeling assumptions: functional forms that link the left- and

Downloaded from https://www.cambridge.org/core. IP address: 216.73.216.25, on 25 Jul 2025 at 07:10:45, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009003391.002

right-hand sides of a regression model and data processing choices such as cleaning, coding, and categorizing variables. In Chapter 9 (coauthored with Sheridan Stewart) we develop multiverse analyses that compare estimation commands such as OLS, logit, probit, inverse probability weighting, and two different matching algorithms. Does the use of these different link functions or algorithms lead to different empirical findings? Do some yield more stable and reliable estimates than others? In Chapters 10–12, we show that data processing is a large world of model uncertainty, where there is little clear guidance for practice and where researcher degrees of freedom are often nearly invisible to readers. Chapter 10 digs into theories of how data processing influences results. Chapter 11 illustrates a complex data processing multiverse with a reanalysis of a highly contentious work by Regnerus (2012a). Chapter 12 reviews a series of social science cases where articles were retracted due to errors in data processing that undermined their analyses – powerful lessons about the centrality of data processing in analytical work. Chapter 13 explores a frontier question of whether, or how well, one could weight models by their probability of being the true model. This is a challenging task and involves a fundamental tradeoff between transparency and model selection.

In Chapter 14, we revisit the key conclusions and insights from multiverse analysis we found along the way. We emphasize that computational power has transformed social science in both positive and negative ways: It has greatly expanded the capacity for empirical research but also created a large information asymmetry between analyst and reader that lies at the core of the crisis in science. Computational power makes multiverse analysis feasible and, we believe, inevitable. However, computational power does not replace the need for human knowledge and judgment. The best multiverse analyses will come from scholars with advanced statistical training, rich field-area knowledge about the research question, and a great capacity to understand and appreciate rival scientific views.

Downloaded from https://www.cambridge.org/core. IP address: 216.73.216.25, on 25 Jul 2025 at 07:10:45, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009003391.002