Forking paths in empirical studies

Guillaume Coqueret*

February 3, 2022

Abstract

In this article, we propose a theoretical framework that characterizes the range of outcomes in empirical studies, depending on the nature and number of design choices that researchers make. We provide a simple method that evaluates if particular choices have a significant impact on the distribution of these outcomes. We also discuss ways to exploit multitudes of such outputs (e.g., t-statistics and p-values) in order to improve the robustness and validity of empirical conclusions. Our ideas are illustrated via an exercise of equity premium prediction in which we conclude that net equity expansion and the book-to-market ratio are solid predictors of the aggregate excess return.

1 Introduction

"*Because the empirical economist must deal with nature in all her complexity, it is optimistic in the extreme to hope or believe that standard parametric economic models or probability models are sufficiently adequate to capture this complexity.*" - *White (1996)*

Empirical studies are built on many choices. Recently, two separate studies (Huntington-Klein et al. (2021) and Menkveld et al. (2021)) have revealed that, given the same replication task, independent researchers can reach vastly different conclusions, including on trivial items such as sample sizes. Such divergences are the consequence of the large number of design options that are left to the appreciation of the empiricist.[1] Examples of these choices are for instance listed in Mitton (2021b) for the field of corporate finance.

In the best case scenario, minor shifts will lead to small adjustments, and solid conclusions will not be altered. Nevertheless, sometimes, articles may inspire policy choices based on erroneous findings (Reinhart and Rogoff (2010)), or be retracted because the results can be contradicted or are simply not robust enough (Rampini et al. (2021)).[2] The replication of results has become a strong imperative in modern research, so that data and code sharing policies have for instance been enforced in most major journals in finance and economics.[3]

*EMLYON Business School, 23 avenue Guy de Collongue, 69130 Ecully, FRANCE. coqueret@em-lyon.com

[1] An enlightening exercise is provided by FivethirtyEight, in its Science is not broken article, in the political science discipline.

[2] Retractions are closely followed by a handful of researchers (https://retractionwatch.com) who have compiled and update a database of scientific articles that have been retracted from their journals: http://retractiondatabase.org

[3] In **finance**, we can list: Journal of Finance, Review of Financial Studies, Journal of Financial Economics, and Review of Finance. In **economics**, there are at least the following: Econometrica, journals of the American Economic Association (including the American Economic Review), Quarterly Journal of Economics, Review of Economic Studies.

The sensitivity of results to design choices is well-known in the research community, which is why many papers incorporate robustness checks in which analyses are replicated with small alterations in the original empirical protocol. For instance, researchers can reproduce their results on sub-samples, or with alternative estimators, or when testing different values in parametric methods and techniques.

Given the publication bias towards positive results,[4] authors can be, knowingly or not, incentivized to produce low p-values in order to demonstrate that their results are *significant*, and hence, worthy of publication.[5] Equipped with a large palette of adjustments (i.e., leeway) in the research designs, researchers may be tempted to mostly report those that will increase odds of acceptance by journal editors.

The troubles of false discoveries are also potent and problematic outside academia. In the money management industry, quantitative researchers compete to discover profitable trading strategies. Unfortunately, many of the latter which perform well in-sample end up disappointing out-of-sample (Bailey et al. (2014), De Prado (2018), Chen and Velikov (2021)). This has spurred a debate on whether there is a crisis of reproducibility in the field (Bailey and Lopez de Prado (2021), Chen (2021), Harvey and Liu (2021) Harvey (2021)) - and this debate is also ongoing in the medical sciences (see Ioannidis (2005) and Leek and Jager (2017), as well as the discussion in Fanelli (2018)). These issues have prompted the research community to propose identification and detection methods for false discoveries and p-hacking (Simonsohn et al. (2014a,b), Elliott et al. (2021)), and even solutions via correction measures (Andrews and Kasy (2019)), noise (Echenique and He (2021)), or Bayesian publication decisions (Frankel and Kasy (2022)).

In the present paper, similarly to Fabozzi and de Prado (2018), we argue that one solution, albeit a costly one, is to report the outcomes of a large number of forking paths in empirical studies. Large scale experiments are for example also advised in Milkman et al. (2021) in the field of behavioral science. This is likely to dramatically increase the transparency of the research process and to strengthen the robustness of findings. Unfortunately, this is only feasible under two conditions. First, the number of research hypotheses must be small, because the corresponding results cannot be presented succinctly, e.g., with one or two numbers, as is often the norm with a coefficient estimate along its t-statistic. The second condition pertains to computation times. Some empirical protocols are resource-intensive and cannot be duplicated a large number of times, each time with a minor alteration.

Our recommendation is built on the modelling of the research process as an iteration of mappings, akin to a feed-forward neural network structure. As long as mappings are somewhat continuous, their compositions can also be characterized as continuous. Hence, small shifts in

[4] This phenomenon, also known as the *file drawer problem*, has been widely documented in many fields, especially **psychology** (Rosenthal (1979)), **medicine** (Dickersin et al. (1987), Begg and Berlin (1988), Olson et al. (2002) to cite only three references), **economics** (Leamer and Leonard (1983), De Long and Lang (1992), Stanley (2005), Doucouliagos and Stanley (2009), Doucouliagos and Stanley (2013), Brodeur et al. (2016), Ioannidis et al. (2017), Brodeur et al. (2020) and Kasy (2021)), **finance** (Lo and MacKinlay (1990), Harvey (2017), Morey and Yadav (2018) and Harvey and Liu (2021)), and **accounting** (Chang et al. (2021)). A related issue is bias in research (Fanelli et al. (2017)).

[5] The topic of p-hacking is now widely documented in many fields since the seminal work of Sterling (1959), and we point to a few articles on the matter, among many others: Head et al. (2015), and Christensen and Miguel (2018).

inputs or choices, at any stage in the process, are expected to yield minor changes in the final results. Furthermore, if design decisions are viewed as (or parametrized by) random variables, the reporting of many outcomes (t-statistics, p-values, or other metrics - see Mitton (2021a) for further examples) provides a much richer characterization of the effects under investigation.

The generation of multiple outputs, based on small variations of similar datasets, shares some similarities with re-sampling techniques, as well as data augmentation and bagging, all of which are sometimes used in machine learning. The premises is that a model which relies on a diversified set of sub-models will benefit from a *wisdom of the crowds* effect, as long as each individual model is relevant (loosely speaking) and that correlations between models are not too high. The best situation is when diversification operates and outcomes of forking paths reveal complementary facets of the initial problem. These ideas have blossomed in the frequentist (Hansen (2007), Zhang (2015), Zhang and Liu (2019) and Zhu et al. (2021)) and Bayesian (Draper (1995), Raftery et al. (1997)) circles. For instance, Bayesian averaging has recently been used in Avramov et al. (2022) to cope with model uncertainty. We refer to Steel (2020) for a survey on model averaging in economics.

In addition to improving the robustness of reported results, framing empirical work as successive mappings helps organize code more neatly into a well-structured research pipeline. Each mapping has its own module, which opportunely prevents potential errors in lengthy scripts written in one block. It also forces to reflect upon the computational cost of each step of the research project and how to optimize it.[6] Consequently, an exhaustive approach to the reporting of results forces the analyst to focus on the first order choices of the research process and to filter out the unnecessary artifices.

We illustrate our framework and recommendations with a study on the prediction of the equity premium, a well-documented research question in financial economics. We consider a large number of ways to run the empirical protocol and report the distributions of test statistics. The latter allow us to determine which design choices alter the average of the statistics and are hence strong drivers thereof. Under technical assumptions on the dependence between p-values, it is possible to approximate the average rejection rate of the null. While this remains an imperfect metric, it is arguably one that is more compelling than a single p-value.

The remainder of the paper is structured as follows. Section 2 lays out a static representation of research studies as compositions of operators. Many examples of such mappings are presented therein. In Section 3, we introduce randomness into the operators, thereby implying that researchers can randomly choose between options, i.e., what we refer to as forking paths. We propose a way to characterize which of those are significant for a given research output. Section 4 presents our numerical application on the prediction of the equity premium. Finally, Section 5 concludes.

[6] Even if modern computers allow for the parallelization of tasks, the complexity of most pipelines outweigh the CPU (and GPU/TPU) capabilities of standard machines. This is likely to limit the exploration of potentially promising but untested questions and configurations.

2 Deterministic framework: compositions of Lipschitz operators

We start with a few comments on notation. This section relies heavily on norms. For vectors, we work with L^p norms: $\|\boldsymbol{v}\|_p^p = \sum_{n=1}^N |v_n|^p$ and for $(N \times K)$ matrices we will consider the following:

- $\|\boldsymbol{M}\|_2^2 = \sum_{n=1}^N \sum_{k=1}^K M_{n,k}^2$: Frobenius Norm;
- $\|\boldsymbol{M}\|_1 = \max_k \sum_{n=1}^N |M_{n,k}|$: maximum absolute column sum;
- $\|\boldsymbol{M}\|_\infty = \max_n \sum_{k=1}^K |M_{n,k}|$: maximum absolute column row.

Unless otherwise stated, the integer N will always be the length of the vectors and the number of rows of matrices. Henceforth, lowercase bold letters $\mathbb{d} = \{\mathbb{d}_1, \ldots, \mathbb{d}_N\}$ will denote vectors and uppercase bold letters matrices or tables. For the latter, we adopt the **tidy data** convention of Wickham (2014): rows are observations and columns are variables. Finally, we will sometimes (when there is little ambiguity) use the simplified notation $f(\mathbb{d})$ for the vector $[f(\mathbb{d}_1), \ldots f(\mathbb{d}_N)]$. Moreover, of a matrix $\boldsymbol{M}$ or a vector $\boldsymbol{v}$, $\boldsymbol{M}'$ and $\boldsymbol{v}'$ will denote their transpose.

In addition, we will often compare two alternative inputs. Readers are accustomed to $\boldsymbol{X}$ and $\boldsymbol{y}$ for modelling purposes. To avoid any confusion, we work with the letters $\mathbb{D}$ and $\boldsymbol{D}$, which will stand for two versions of some data that is collected by the researcher. This choice of notation is disconcerting at first, but imperative because we will resort to the $\boldsymbol{X}$ and $\boldsymbol{y}$ letters for linear models later on.

2.1 Theory

We assume that the empirical part of research process starts with some input which we call $\mathbb{D}$ and can be thought of as the initial version of the data that is collected. The empirical study is modelled as a sequence of operations f_j that occur successively so that the reference research output (e.g., one t-statistic) is such that

$$o_J(\mathbb{D}) = \left[\bigcirc_{j=J}^1 f_j\right](\mathbb{D}) = f_J \circ f_{J-1} \circ \cdots \circ f_1(\mathbb{D}), \tag{1}$$

where $f_j : S_j \mapsto S_{j+1}$, with S_1 and S_{J+1} encompassing the sets of feasible input $\mathbb{D}$ and output values, respectively. For simplicity, we can assume that o_J is simply a real number, but it may be a more complex object (e.g., a vector (confidence interval), or a matrix). The index J indicates that the output is the result of J successive operations, which Gelman and Loken (2014) refer to as *forking paths*.

One interesting question pertains to the sensitivity of the output o_J to a change in initial input $\mathbb{D}$. In order to derive theoretical results, we must impose some conditions on the mappings f_j and we choose to work with Lipschitz smoothness. More precisely, we assume that for $\mathbb{D}, \boldsymbol{D} \in S_j$, there exists some constant $c_j > 0$ such that

$$\|f_j(\mathbb{D}) - f_j(\boldsymbol{D})\| \leq c_j \|\mathbb{D} - \boldsymbol{D}\|, \tag{2}$$

for some norms which are implicitly defined on S_{j+1} and S_j. Given the number of norms we will use subsequently, specific notations would be cumbersome. We recall that when inputs belong to $\mathbb{R}^n$, we will work with the L^p norms: $\|\mathbb{d}\|_p = (|\mathbb{d}_1|^p + \dots |\mathbb{d}_n|^p)^{1/p}$.

In all generality, the object $\mathbb{D}$ can comprise several data types, categorical features notably (ordinal or nominal). Handling distances with such features is complex, though not impossible (see Boriah et al. (2008)), but for the sake of simplicity, the exposé will often assume that $\mathbb{D}$ is a matrix of real numbers.

Composing two operators yields

$$\begin{aligned}\|f_{j+1} \circ f_j(\mathbb{D}) - f_{j+1} \circ f_j(\boldsymbol{D})\| &\leq c_{j+1}\|f_j(\mathbb{D}) - f_j(\boldsymbol{D}\| \\ &\leq c_j c_{j+1}\|\mathbb{D} - \boldsymbol{D}\|.\end{aligned}$$

Iterating this inequality leads to

$$\left\| \left[\bigcirc_{j=J}^{1} f_j\right](\mathbb{D}) - \left[\bigcirc_{j=J}^{1} f_j\right](\boldsymbol{D}) \right\| \leq \|\mathbb{D} - \boldsymbol{D}\| \prod_{j=1}^{J} c_j. \tag{3}$$

More generally, the accumulation of shifts may not start at the initial data sample $\mathbb{D}$, but at a later stage, say at o_K, after K steps, for $K < J$. It is easy to prove the following lemma.

Lemma 1. *If o_J is given by Equation (1) and the mappings f_j satisfy (2), then for $1 \leq K < J$,*

$$\left\| \left[\bigcirc_{j=J}^{K+1} f_j\right](o_K(\mathbb{D})) - \left[\bigcirc_{j=J}^{K+1} f_j\right](o_K(\boldsymbol{D})) \right\| \leq \|\mathbb{D} - \boldsymbol{D}\| \prod_{j=K+1}^{J} c_j. \tag{4}$$

It is obvious that the constants c_j are the main drivers of the error bounds. In practice, a lower bound for the c_j is often 1, meaning that their compounded effect can be sizeable, theoretically, especially if many c_j are such that $c_j \gg 1$. Thus, adding steps in the design increases the amplitude of the difference between outcomes. In the remainder of the section, we provide many illustrations Lipschitz constants. Few are exactly sharp and in some cases, they even depend on $\mathbb{D}$ or $\boldsymbol{D}$.

2.2 Examples: descriptive statistics

Sample moments and other mainstream metrics play an important in empirical studies. We begin our journey of illustrations with the simplest of them all: the sample mean. In the sequel, we will use the notation f as generic mapping. For example, if f is the **sample mean** operation, we have, via Hölder's inequality in the last inequality,

$$\|f(\mathbb{d}) - f(\boldsymbol{d})\|_p = \left| \frac{1}{N}\sum_{n=1}^{N} \mathbb{d}_n - \frac{1}{N}\sum_{n=1}^{N} d_n \right| \leq \frac{1}{N}\sum_{n=1}^{N} |\mathbb{d}_n - d_n| \leq N^{-1/p}\|\mathbb{d} - \boldsymbol{d}\|_p, \tag{5}$$

so that in this case the Lipschitz constant is $N^{-1/p}$. The case of the (biased) **sample variance** is more tricky:

$$\begin{aligned}
\|f(\mathbb{d}) - f(\boldsymbol{d})\|_p &= \left| \frac{1}{N}\sum_{n=1}^{N}\left(\mathbb{d}_n - \frac{1}{N}\sum_{n=1}^{N}\mathbb{d}_n\right)^2 - \frac{1}{N}\sum_{n=1}^{N}\left(d_n - \frac{1}{N}\sum_{n=1}^{N} d_n\right)^2 \right| \\
&= \frac{1}{N}\left| N(\bar{d}^2 - \bar{\mathbb{d}}^2) + \sum_{n=1}^{N}\mathbb{d}_n^2 - d_n^2 \right| \\
&= \frac{1}{N}\left| N(\bar{d} - \bar{\mathbb{d}})(\bar{d} + \bar{\mathbb{d}}) + \sum_{n=1}^{N}(\mathbb{d}_n + d_n)(\mathbb{d}_n - d_n) \right| \\
&\leq |\bar{\mathbb{d}} + \bar{d}| \times |\bar{\mathbb{d}} - \bar{d}| + \frac{d^*}{N}\left|\sum_{n=1}^{N}(\mathbb{d}_n - d_n)\right| \\
&\leq c_p \|\mathbb{d} - \boldsymbol{d}\|_p, \qquad (6)
\end{aligned}$$

where $c_p = N^{-1/p}(|\bar{\mathbb{d}} + \bar{d}| + d^*)$, with $d^* = \max_n |\mathbb{d}_n + d_n|$. $\bar{\mathbb{d}}$ and $\bar{d}$ are the sample means. The last inequality comes from (21). In this case, and as will be recurrent, the constant depends on the magnitude of the series. To remove this dependence, it is imperative to specify some properties of the vectors (e.g., if they belong to the unit sphere, or if their range is restricted to particular intervals). This comment holds for the remainder of the paper, as many constants will be input-dependent below.

Typically, for the sample **covariance**, we have that

$$\begin{aligned}
|f(\mathbb{d}_1, \boldsymbol{d}_1) - f(\mathbb{d}_2, \boldsymbol{d}_2)| &= |(\mathbb{d}_1 - \bar{\mathbb{d}}_1)'(\boldsymbol{d}_1 - \bar{\boldsymbol{d}}_1) - (\mathbb{d}_2 - \bar{\mathbb{d}}_2)'(\boldsymbol{d}_2 - \bar{\boldsymbol{d}}_2)| \\
&= |(\mathbb{d}_1 - \bar{\mathbb{d}}_1 - (\mathbb{d}_2 - \bar{\mathbb{d}}_2))'(\boldsymbol{d}_1 - \bar{\boldsymbol{d}}_1) - (\mathbb{d}_2 - \bar{\mathbb{d}}_2)'(\boldsymbol{d}_2 - \bar{\boldsymbol{d}}_2 - (\boldsymbol{d}_1 - \bar{\boldsymbol{d}}_1))| \\
&\leq \frac{\|\boldsymbol{d}_1 - \bar{\boldsymbol{d}}_1\|_1}{N}(|\mathbb{d}_1 - \mathbb{d}_2| + |\bar{\mathbb{d}}_1 - \bar{\mathbb{d}}_2|) + \frac{\|\mathbb{d}_2 - \bar{\mathbb{d}}_2\|_1}{N}(|\boldsymbol{d}_1 - \boldsymbol{d}_2| + |\bar{\boldsymbol{d}}_1 - \bar{\boldsymbol{d}}_1|)
\end{aligned}$$

which is again a similar form.

Let us now mention the **maximum** of vectors. We have that

$$\max_n \mathbb{d}_n = \max_n [\mathbb{d}_n - d_n + d_n] \leq \max_n |\mathbb{d}_n - d_n| + \max_n d_n,$$

so that

$$\|\max_n \mathbb{d}_n - \max_n d_n\|_\infty = |\max_n \mathbb{d}_n - \max_n d_n| \leq \max_n |\mathbb{d}_n - d_n| = \|\mathbb{d} - \boldsymbol{d}\|_\infty,$$

i.e, the Lipschitz constant in this case is one. Straightforwardly, the same applies to the minimum operator.

2.3 Other examples of Lipschitz constants

In empirical studies, the **data collection** stage is the hardest to model, because of its heterogeneity. It can be quite constrained if data comes from a provider (e.g., WRDS, Bloomberg,

etc.), in which case the researcher has a few degrees of freedom: which variables to import, for which universe, at which frequency, over which time frame, etc. Providers often update their data so that downloading a sample at two different periods may generate discrepancies if series are not kept point-in-time. This has been recently shown for the Fama-French factors in Akey et al. (2021). There is also relatively little room for initiative in economics or physics when working with official series, such as GDP output, inflation, unemployment, CO_2 concentration, temperatures, etc.

However, when the study is based on surveys, the researcher has more latitude, and we for instance refer to the guide of Bergman et al. (2020) for an overview of the range of options in that case. In qualitative studies, there is also an important coding phase (see, e.g., the review by Basit (2003)), which is difficult to model neatly and efficiently.

For all these reasons, we commence this section with the step that comes right *after* data collection, namely **data cleaning**. We underline that Mitton (2021b) reports that "*the methodological decisions that affect statistical significance the most are dependent variable selection, variable transformation, and outlier treatment*". In this subsection, we tackle all of these elements. Lastly, the purpose of the section is to show that most classical operations on data can be represented as Lipschitz mappings. It is not to provide sharp constants for these mappings.

2.3.1 Data cleaning

The first issue that most, if not all, researchers encounter, is **missing data**. When working with time-series, a common practice is to impute with the most recent well-defined point prior to the missing value, if it exists. Interpolation is usually avoided because it introduces a forward-looking bias. In this subsection, to ease the exposition, we make strong assumptions on the vectors on which the imputation mapping will operate.

Formally, we consider two vectors $\mathtt{d}$ and $\boldsymbol{d}$ such that S is the *common* set of indices for which a value is missing. The fact that S is common to the two vectors comes from the fact that we need $\|\mathtt{d}-\boldsymbol{d}\|$ to be well defined. In the above norm, two points that are not defined are assumed to be equal, but if one value is defined and the other is not, there is no unambiguous way to proceed. For simplicity, we assume that 1 does not belong to the S set, so that the imputation values will always be defined. In addition, we impose that the indices in S are never consecutive numbers, though this assumption can be relaxed easily. We then have

$$\|f(\mathtt{d}) - f(\boldsymbol{d})\|_p^p = \sum_{n\in S} |\mathtt{d}_{n-1} - d_{n-1}|^p + \sum_{n\notin S} |\mathtt{d}_n - d_n|^p \leq 2\|\mathtt{d} - \boldsymbol{d}\|_p^p$$

where the last inequality comes from the fact that the values that precede missing points get counted twice. The Lipschitz constant is not very sharp in this case.

Other examples of methods include cross-sectional imputation, whereby a missing value is replaced by the cross-sectional mean (or median) across other observations. In this case, via inequality (21) it is also possible to derive a Lipshitz constant for mean-driven imputation. Parametric imputation based on some distributional assumption follow the same logic, though their treatment is substantially more involved.

One extreme solution when facing missing data is simply the **removal of observations**. To illustrate this issue, we consider two matrices of numerical data $\mathbb{D}$, $\boldsymbol{D}$ with equal sizes, N rows and M columns. In line with the above assumptions, we write S for the (common) indices of their rows which contain missing points. Naturally, we again assume that the cardinal of S is much smaller than the total number of rows N. In this case, it is straightforward that for the usual matrix norms (Frobenius, $\|\cdot\|_1$ and $\|\cdot\|_\infty$), the Lipschitz constant is one at most, i.e., that

$$\|f(\mathbb{D}) - f(\boldsymbol{D})\| \leq \|\mathbb{D} - \boldsymbol{D}\|.$$

Similarly, the researcher may want to remove columns (instead of rows) of the data because of **co-linearity** issues. The Lipschitz constant of such an operation is also one at most.

Another important stage in data processing is **outlier management**. One of the most frequently used tools to this purpose is **winsorization**, whereby extreme values are replaced by given quantiles, often at the 1% and 99% levels.

Without loss of generality, let us assume that the vector $\mathbb{d}$ is ordered, i.e., that $\mathbb{d}_1 < \cdots < \mathbb{d}_N$. For a given integer $k \ll N/2$, the winsorization operator is defined as:

$$f(\mathbb{d}_n) = \begin{cases} \mathbb{d}_n & \text{if } n \in [k+1, N-k] \\ \mathbb{d}_{k+1} & \text{if } n \leq k \\ \mathbb{d}_{N-k} & \text{if } n > N-k \end{cases}, \tag{7}$$

so that exactly $2k$ values are replaced: the most extreme k values in both tails. If we abusively write $f(\mathbb{d})$ for the vector of $f(\mathbb{d}_n)$ values, it holds that

$$\begin{aligned} \|f(\mathbb{d}) - f(\boldsymbol{d})\|_1 &= \sum_{n=1}^{k} |\mathbb{d}_{k+1} - d_{k+1}| + \sum_{n=k+1}^{N-K} |\mathbb{d}_n - d_n| + \sum_{n=N-K+1}^{N} |\mathbb{d}_{N-k} - d_{N-k}| \\ &= K(|\mathbb{d}_{k+1} - d_{k+1}| + |\mathbb{d}_{N-k} - d_{N-k}|) + \sum_{n=k+1}^{N-K} |\mathbb{d}_n - d_n| \\ &\leq (1+K)\|\mathbb{d} - \boldsymbol{d}\|, \end{aligned}$$

where the constant is clearly sub-optimal. It could be improved but would then rely on the properties of the underlying vectors.

2.3.2 Variable engineering

Once the data has been cleaned, the researcher will often perform additional adjustments. We list a few below.

Normalization is a common step in data preparation: it ensures that all variables have roughly the same scales. This is convenient when one wants to compare effect sizes for example. There are several ways to proceed, such as standardization, or min-max rescaling. Let us analyze

the former:

$$\begin{aligned}
\|f(\mathbb{d}) - f(d)\|_2 &= \left\| \frac{\mathbb{d} - \boldsymbol{m}_{\mathbb{d}}}{\sigma_{\mathbb{d}}} - \frac{d - \boldsymbol{m}_d}{\sigma_d} \right\|_2 = \sigma_{\mathbb{d}}^{-1}\sigma_d^{-1}\|\sigma_d(\mathbb{d} - \boldsymbol{m}_{\mathbb{d}}) - \sigma_x(d - \boldsymbol{m}_d)\|_2 \\
&= \sigma_{\mathbb{d}}^{-1}\sigma_d^{-1}\|(\sigma_d - \sigma_{\mathbb{d}} + \sigma_{\mathbb{d}})(\mathbb{d} - \boldsymbol{m}_{\mathbb{d}}) - \sigma_{\mathbb{d}}(d - \mathbb{d} + \mathbb{d} - \boldsymbol{m}_)\|_2 \\
&\leq \sigma_{\mathbb{d}}^{-1}\sigma_d^{-1}|\sigma_d - \sigma_{\mathbb{d}}| \times \|\mathbb{d} - \boldsymbol{m}_{\mathbb{d}}\|_2 + \sigma_d^{-1}\|\mathbb{d} - d\|_2 + \sigma_d^{-1}\|\boldsymbol{m}_d - \boldsymbol{m}_{\mathbb{d}}\|_2 \qquad (8) \\
&\leq \sigma_d^{-1}\sqrt{N}\frac{|\sigma_x^2 - \sigma_d^2|}{\sigma_{\mathbb{d}} + \sigma_d} + \sigma_d^{-1}\|\mathbb{d} - d\|_2 + \sigma_d^{-1}N^{-1/2}\|\mathbb{d} - d\|_2 \qquad (9) \\
&\leq c\|\mathbb{d} - d\|_2
\end{aligned}$$

where $\boldsymbol{m}_{\mathbb{d}}$ is the constant mean vector of $\mathbb{d}$, $\sigma_{\mathbb{d}}$ its standard deviation and

$$c = \sigma_d^{-1}(1 + N^{-1/2} + \sqrt{N}c_1(\sigma_{\mathbb{d}} + \sigma_d)^{-1}),$$

the constant c_1 being the one from Inequation (6) for $p = 1$. We have used that $\|\mathbb{d} - \boldsymbol{m}_{\mathbb{d}}\|_2 = \sqrt{N}\sigma_{\mathbb{d}}$ and applied (21) and (6) in lines (8) and (9), respectively.

Sometimes, when working with time-series, the model requires stationary variables, but the collected data is integrated and has unit roots. In other contexts, the level of the independent variable may matter less than its variations from a predictive standpoints. Thus it is relevant to consider variable differences in these settings too. In any case, the solution is **differentiation**:

$$f(\mathbb{d}_n) = \begin{cases} \text{NA} & \text{if } n = 1 \\ \mathbb{d}_n - \mathbb{d}_{n-1} & \text{otherwise} \end{cases} . \qquad (10)$$

In practice, the first missing point is often removed so that the resulting vector has length $N-1$. For two numerical vectors with no missing points $\mathbb{d}$ and d,

$$\begin{aligned}
\|f(\mathbb{d}) - f(d)\|_1 &= \sum_{n=2}^{N} |\mathbb{d}_n - \mathbb{d}_{n-1} - d_n + d_{n-1}| \\
&\leq \sum_{n=2}^{N} |\mathbb{d}_n - d_n| + |\mathbb{d}_{n-1} - d_{n-1}| \leq 2\|\mathbb{d} - d\|_1
\end{aligned}$$

Sometimes, researchers seek to explain long term effects. For instance, in the predictability literature, there is a debate between short-term and long-term predictability. At a first order approximation, long term returns can be viewed as **cumulative sums** of shorter horizon returns, which is why we briefly mention the topic below. For a given well-defined numerical vector, we have in this case, for $n > 0$,

$$f(\mathbb{d}_n) = \sum_{k=1}^{n} \mathbb{d}_k,$$

and

$$\|f(\mathbb{d}) - f(\boldsymbol{d})\|_1 = \sum_{n=1}^{N} \left| \sum_{k=1}^{n} \mathbb{d}_k - \sum_{k=1}^{n} d_k \right| \leq \sum_{n=1}^{N} \sum_{k=1}^{n} |\mathbb{d}_k - d_k| \leq N\|\mathbb{d} - \boldsymbol{d}\|_1,$$

where the bound may seem loose, but can be sharp if $\mathbb{d}_n = d_n = 0$ for $n > 1$ for instance.

To conclude this subsection, we acknowledge that many more operations exist in the data preparation phase and some would require a lengthy treatment. For instance, joining procedures that merge two tables according to a common key are a widespread practice. They are however more complex to handle and we leave their analysis to future work.

2.3.3 Testing

An ubiquitous tool in the researcher's arsenal is the **linear regression**. Given a matrix of independent variables $\boldsymbol{X}$ and the vector of dependent variable $\boldsymbol{y}$, the ordinary least square (OLS) estimator $\boldsymbol{b}$ that minimizes the quadratic error

$$e^2(\boldsymbol{X}, \boldsymbol{y}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}) \tag{11}$$

is

$$\boldsymbol{b}(\boldsymbol{X}, \boldsymbol{y}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \tag{12}$$

where $\boldsymbol{v}'$ denotes the transpose of $\boldsymbol{v}$ (vector or matrix). In the sequel, we will always assume that the inverse matrix is well-defined. The issue is then that there are two inputs. Given one of them, it is possible to intuitively deduce data-specific Lipschitz constants. For instance, if $\boldsymbol{X}$ is fixed, then factorizing the $\boldsymbol{X}$-dependent matrices yields

$$\|\boldsymbol{b}(\boldsymbol{X}, \boldsymbol{y}) - \boldsymbol{b}(\boldsymbol{X}, \boldsymbol{z})\| \leq c_X \|\boldsymbol{y} - \boldsymbol{z}\|. \tag{13}$$

The case when $\boldsymbol{y}$ is fixed is less straightforward but can be handled with suitable norms. The general case when both $\boldsymbol{X}$ and $\boldsymbol{y}$ are subject to perturbation is more intricate. It is reviewed in Section 5 of Grcar (2003). One foundational result (Golub and Wilkinson (1966)) is that if $\|\boldsymbol{X}\|_2 = \|\boldsymbol{y}\|_2 = 1$, then

$$\|\boldsymbol{b}(\boldsymbol{X}, \boldsymbol{y}) - \boldsymbol{b}(\boldsymbol{Z}, \boldsymbol{v})\|_2 \leq c(\|\boldsymbol{X} - \boldsymbol{Z}\|_2 + \|\boldsymbol{y} - \boldsymbol{v}\|_2) + R, \tag{14}$$

where c depends on the smallest singular value of $\boldsymbol{X}$, on $\|\boldsymbol{b}\|_2$, and on the quadratic error e^2 defined in Equation (12). The matrix norms are of Frobenius type: $\|\boldsymbol{X}\|_2^2 = \mathrm{tr}(\boldsymbol{X}\boldsymbol{X}')$, where $\mathrm{tr}(\cdot)$ is the trace operator. The above result holds in the case when the norms are arbitrarily small and the residual term R is a second order term which is quadratic in the maximum of the two norms.

While the coefficients in linear regressions (or more general models) are undoubtedly analyzed by researchers, it is their **statistical significance** which often matters more because it will determine if the effect revealed by the study is strong enough.

In the case of a linear model, the expressions for the t-statistics are $t_k = b_k/\sqrt{s^2 S_k}$, where S_k is the k^{th} diagonal element of $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ and $s^2 = e^2/(N - K)$, with K being the number of columns of $\boldsymbol{X}$ (Equation 4-47 in Greene (2018)). Lipschitz numbers can be obtained for S_k (see e.g., Demmel (1992)) and for $s^2 = (N - K)^{-1}(\boldsymbol{y}'\boldsymbol{y} - \boldsymbol{y}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y})$ as well. From these numbers, it is possible to derive a Lipschitz constant similar to (14) for the statistic, based on norms that depend on the inputs.

Often, models are tested with several control variables and model permutations, in order to check the robustness of the initial specification. It is even possible to aggregate coefficients of multiple specifications via weighting, as in Hansen (2007) and Zhang and Liu (2019). In linear models, adding columns to the matrix $\boldsymbol{X}$ is often handled via the Frish-Waugh-Lovell theorem. The Lipschitz continuity of this operation with respect to estimates and t-statistics is currently out of the scope of the paper (it requires to replace the concatenation $[\boldsymbol{X}_1\ \boldsymbol{V}_1]$ with $[\boldsymbol{X}_2\ \boldsymbol{V}_2]$ as the independent matrix in Equation (12)). In the field of randomized experiments, Muralidharan et al. (2022) analyze the change in significance for coefficients of short versus long models.

Naturally, modern studies rely on much more complex apparatus, including structural equations, difference-in-differences, dynamic panels, improved estimators (White (1980), Newey and West (1987) to cite the most commonly used), etc. Given the exhaustiveness and complexity of the related methods, we cannot treat them exhaustively, but it is possible that many of them can be described as Lipschitz operators.

2.4 Model design vs. protocol shifts

Not all mappings are equal. For instance, the choice of independent variable may very well be a strong modelling assumption. Consider the two alternative questions:

- Does variable X predict variable Y?
- Can variable Y be predicted?

In the first case, it the focus is clearly on the predictive ability of variable X, hence picking it as independent variable is not a design choice, it is an obligation. In the second option, choosing variable X or Z, or W is left to the appreciation of the researcher, and, in fact, it is conceivable to mine as much data as possible to find the few predictors that may indeed predict Y. Typically, in the debate on the predictability of the equity premium, many variables have been proposed. If many are studied, statistical significance must be corrected for **multiple testing**. Because we propose to generate series of outcomes, the ideas presented in the present paper are undoubtedly linked to this notion, a theme that goes back at least to Bonferroni (1936), and which is useful in many settings, including medicine (Farcomeni (2008)), economics (Viviano et al. (2021)), finance (Harvey and Liu (2020), Harvey et al. (2020), Giglio et al. (2021)), generic model discrimination (Hansen et al. (2011)) and statistics more generally (Fan and Han (2017), Wang et al. (2017) to cite but a few). In many cases, as in Romano and Wolf (2005, 2010) or Wilson (2019), the methods take as input series of test statistics (or p-values). Recently, several studies in finance have approached research questions by resorting to large scale tests in which many predictors are considered (Yan and Zheng (2017), Chordia et al. (2020), Giglio et al. (2021) and Jensen et al. (2021)).

The approach we advocate here is slightly different. We assume that the researcher has a precise research question in mind, but that there are many different ways to approach and answer it, thanks to small shifts, or tweaks, in the empirical protocol, exactly as in Huntington-Klein et al. (2021) and Menkveld et al. (2021). This is also the spirit of model averaging, or

multi-model inference (Burnham and Anderson (2004)). One common way of extracting multiple coefficients is to consider several combinations of control variables in regressions, as in Zhang and Liu (2019) and many references therein.

One way to put it is to consider that exhaustive robustness checks must become the baseline result. The difference with multiple testing is illustrated in Figure 1. In the left graph, the space of models is wide, and all hypotheses are tested with the same unique protocol (e.g., simple portfolio sorts, or linear models). In the right plot, the scope is narrower, but the methods used to reach conclusions are heterogeneous. In financial economics, the first type can be found in Jensen et al. (2021), in which the authors approach the topic of asset pricing anomalies via a large-scale study encompassing thousands of firms worldwide. The portfolios are all constructed using the same methodology. Two examples of the second type of studies are Asness and Frazzini (2013) and Amenc et al. (2020), wherein the authors focus solely on the **value** anomaly, but propose alternative ways to construct value factors. Similar analyses have been carried out for the **momentum** anomaly (see Novy-Marx (2012) and Gong et al. (2015)).

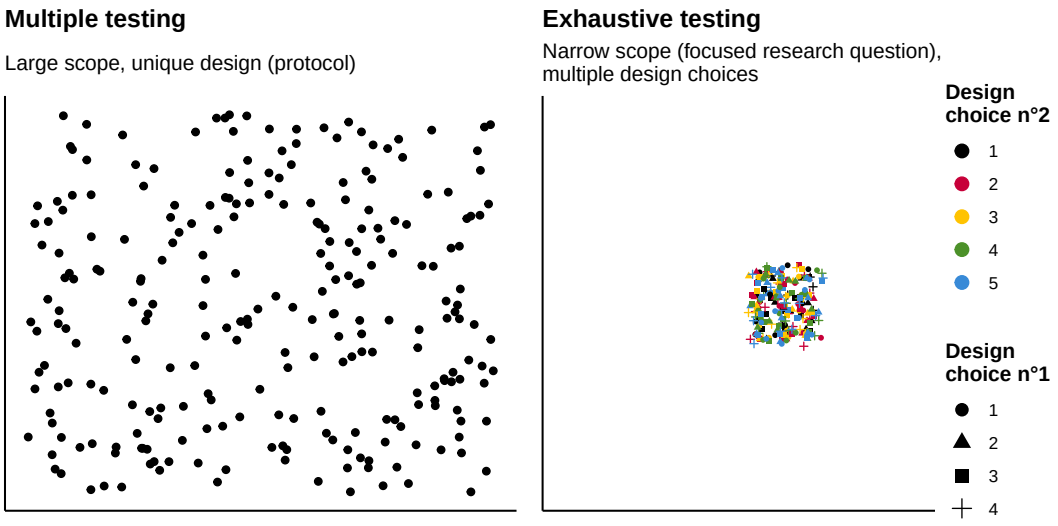

Figure 1: Illustration of **model rich and protocol poor** versus **model poor and protocol rich** studies.

The variety of strata in the design of empirical studies is such that most intermediate steps cannot be listed in the present paper - one reason being that many of them can be discipline-specific.

3 Parametric random operators

3.1 Presentation

The deterministic framework presented in Section 2 suggests, via Lemma 1, that many changes in the empirical protocol can be compounded and lead to very different outcomes.[7] This comes from the upper bound in the lemma, which depends on products of constants that can often be larger than one. In practice however, two choices at different stages of the study may have opposite effects, thereby cancelling each other.

In this section, we propose another, somewhat similar, approach, designed for operationalization. First, as we do not model the data extraction phase, we can consider the initial object $\mathbb{D}$ to be random or deterministic. Second, we impose that each step of the research process which we model as f_j is *random*, and can be parametrized by a vector $\boldsymbol{p}_j$. The rationale is that, for any stage of the process, the researcher faces options and instead of setting choices in stone, we allow for each option to have a probability of materialization. The vector $\boldsymbol{p}_j$ is used to code these probabilities. We illustrate this idea in Figure 2 below.

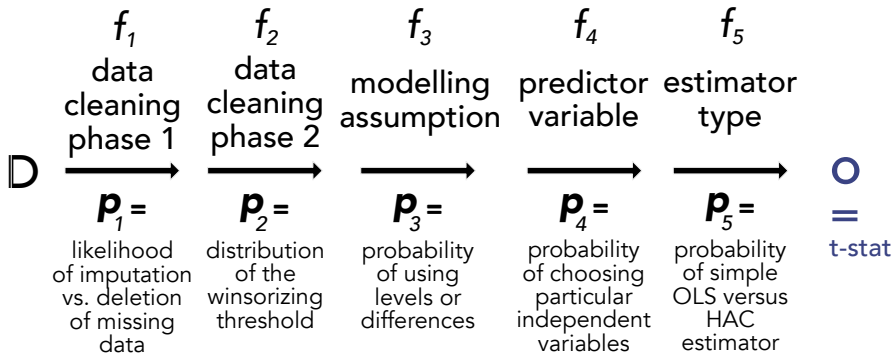

Figure 2: **Illustrative diagram**. We provide a schematic view of a simple five-step empirical protocol with associated parameters $\boldsymbol{p}_j$.

Given an original dataset $\mathbb{D}$ (which could also be random), the researcher:

1. first decides on how to deal with missing data. For simplicity, we consider two options: imputation (say from previous non-missing value) and deletion;
2. then removes outliers via winsorization. The thresholds can be drawn from uniform distributions in the intervals [0,0.02] and [0.98,1];
3. chooses whether to use original series as independent variables, or differences (this may depend on the econometric properties of the variables);
4. picks the dependent variables (e.g., short-term versus long-term);
5. selects an estimator for coefficients and variances (simple OLS (iid) versus HAC).

[7] The dispersion of outcomes can be measured as the standard variation of these outputs. This is what Menkveld et al. (2021) refer to as *non standard errors*.

We thus update the core sequence of the paper to:

$$o_J(\mathbb{D}, \boldsymbol{P}) = \left[\bigcirc_{j=J}^{1} f_{j,\boldsymbol{p}_j}\right](\mathbb{D}) = f_{J,\boldsymbol{p}_J} \circ \cdots \circ f_{1,\boldsymbol{p}_1}(\mathbb{D}), \tag{15}$$

where $\boldsymbol{P}$ encompasses all parameter sets $\boldsymbol{p}_j$. The output o_J is a random variable that depends on the realizations of the operators f_j - and possibly on that of $\mathbb{D}$ if stochastic initial samples are allowed. In many studies, only one realization is reported, and robustness checks correspond to other realizations. In all generality, the realization of f_j may depend on those of prior operators. For example, some predictors may or may not be relevant if they are differentiated.

The formulation (15), because it relies on parameters at each layer, resembles feed-forward **neural networks** (NNs), though there are several notable differences between the two concepts. First, NNs are not traditionally viewed as random objects. In stochastic gradient descent, the randomness comes from the samples that are drawn, akin to $\mathbb{D}$, in our framework. Second, there is no supervision in Equation (15). The output is not known in advance (labelled), and we do not seek to optimize on the set $\boldsymbol{P}$ to obtain a desired value (or distribution) for o_J.[8] Lastly, in all generality, the operators are not differentiable, meaning that back-propagation is not possible.

3.2 Are mappings drivers of outcomes?

In traditional inferential statistics, the randomness of metrics (e.g., t-statistics) comes from the data generating process (DGP) ex-ante, and from estimation errors ex-post. In the equation $\boldsymbol{y} = g(\boldsymbol{X}) + \boldsymbol{\epsilon}$, the stochastic term is often the error $\boldsymbol{\epsilon}$, and assumptions are made on its distribution (notably with respect to $\boldsymbol{X}$) in order to derive properties of potential estimators of g. Thus, the theoretical law of a t-statistic (e.g., Student) stems from the hypothesized behaviour of the terms on the right-hand side of the modelling equation.

In Equation (15), the randomness comes from the original sample $\boldsymbol{X}$ (which is linked to the DGP), but also, and more importantly, from all modelling steps f_j. One interesting extension pertains to the random variables $o_J|f_j$, which are the output values, conditional on the knowledge of one operator, say, f_j. This is notably useful to determine if the operator f_j has an important impact on the distribution of the outcome.

A priori, we cannot make any distributional assumption on the o_J. Thus, in order to test if one operator has an impact on the outcome, it is feasible to devise a test à la Kolmogorov-Smirnov on $o_J^{\{\mathbb{f}\}}$ and $o_J^{\{f\}}$, which we define as the random variables $o_J|f_j = \mathbb{f}$ and $o_J|f_j = f$, respectively - for two different mappings $\mathbb{f} \neq f$.

This approach is somewhat cumbersome, so we take a simpler route and test for average outcome values only. This is particularly relevant when the parameter $\boldsymbol{p}_j$ is the probability between choosing two alternatives (e.g., imputation versus deletion), but can be defined whenever the mapping f_j has any finite number of possible realizations. In order to evaluate the impact of one particular mapping, we will resort to a paired test operator, which we define below.

Definition 2. *Given two vectors $\boldsymbol{v}$ and $\boldsymbol{w}$ of equal size, we call $p(\boldsymbol{v}, \boldsymbol{w})$ the p-value of the paired Wilcoxon signed-rank test comparing the averages of $\boldsymbol{v}$ and $\boldsymbol{w}$.*

[8] Going forth with this analogy, **data snooping** could be viewed as some form of supervised selection: each new forward pass in the network would only be retained if the outcome produces statistically significant results.

Now, we assume that the mapping f_j has r_j possible realizations $\mathbb{f}_{j,r}$ for $r = 1, \dots r_j$. For example, this can be alternative ways of handling missing data (deletion versus imputation), or the set of possible combinations of independent variables, in which case r_j is the cardinal of this set (all permutations that are relevant for the study). We write $\boldsymbol{v}_{j,m}$ for the vector of all realizations of $o_J^{\{\mathbb{f}_{j,m}\}}$. It corresponds to all observed outcomes (e.g., t-statistics) over all permutations of all mappings f_i for $i \neq j$ when the mapping f_j is fixed to its alternative $\mathbb{f}_{j,m}$.

Heuristically, we posit that a mapping f_j is a driver of the outcome o_J if variations in f_j yield significant changes in the average values observed for o_J. We formalize this notion below.

Definition 3. *The mapping f_j is a driver of the scalar outcome o_J at the level $\alpha \in (0, 1)$ if*

$$\min_{m \neq n} p(\boldsymbol{v}_{j,m}, \boldsymbol{v}_{j,n}) < \alpha \quad \text{(strong driver, case } r_j > 2) \tag{16}$$

$$\max_{m \neq n} p(\boldsymbol{v}_{j,m}, \boldsymbol{v}_{j,n}) < \alpha \quad \text{(weak driver, case } r_j > 2) \tag{17}$$

$$p(\boldsymbol{v}_{j,1}, \boldsymbol{v}_{j,2}) < \alpha \quad \text{(driver, case } r_j = 2) \tag{18}$$

This notion of DOSO (driver of scalar outcome) can be further refined, conditionally on some other mapping being fixed, so as to focus on sub-parts of the empirical study. We leave this refinement aside for the time being to avoid unnecessary heaviness in the notations.

3.3 What about statistical significance?

Many scientists argue in favor of redefining the notion of *statistical significance* (Harvey (2017), Benjamin et al. (2018)), or even propose abandoning it purely and simply (Carver (1978), Amrhein et al. (2019), McShane et al. (2019)) - mainly for two reasons. First, because people sometimes confuse p-values it with the probability of the null being true, given the data. And second, because p-values become the objective of the research project so that researchers can be, consciously or not, incentivized to tilt their empirical protocols in order to produce the sought output (i.e., p-values below some chosen threshold, usually 1% or 5%). Without p-values, no more p-hacking.

However, the framing of the problem here is different: we are not interested in multiple hypothesis testing, but rather in testing the same hypothesis multiple times, via many variations on the protocol. Simply put, in traditional inference, the randomness of the t-statistic comes from the estimated errors - and we observe only one value. In a garden of forking paths, we observe many, and their randomness comes from the multitude of choice made by the modeller. In a simple linear model, $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{e}$, inference usually comes from $\boldsymbol{e}$. Subject to random paths, the premise is the opposite because the starting point is the generation random inputs ($\boldsymbol{y}$ and $\boldsymbol{X}$) which are all particular realizations of the world.

Under precise modelling assumptions (see the discussion in Subsection 2.4), forking paths can be viewed as robustness checks that provide information about the distribution of the metric or the effect under consideration. Similarly to the improvement brought by random forest to decision trees, gardens of forking paths enhance the inferential power of single paths when determining the significance of a given effect.

Unfortunately, the true distribution of the effect is usually unknown and there are at least two non-exclusive routes the researcher can take to try to characterize it. The first route is to be transparent about the protocol (*all* paths) and the generated outcomes, and to disclose their full distribution, via a histogram, an empirical cdf, a boxplot, or simply via descriptive statistics.

The second route involves additional hypotheses. Below, we list two possible approaches.

1. If it is critical to determine the significance of a true effect b, the most natural option is to compute a weighted average of estimators. For instance, if M paths have generated a series of outputs (effects) b_m such that $\mathbb{E}[b_m] = b$, two families of techniques are the Bayesian model averages (BMA, see Chapter 3 of Steel (2020) and Section 2 of Moral-Benito (2015)) and the frequentist model averages (FMA, see Chapter 4 of Steel (2020) and Section 3 of Moral-Benito (2015)) - though some models combine both (Magnus and De Luca (2016)).

   One crucial issue is that given the proximity of some paths, the assumption of non-correlation between the random variables b_m may be unrealistic. This problem is discussed in Buckland et al. (1997), wherein the authors acknowledge that the most conservative stance is to assume perfect correlation. Benjamini and Yekutieli (2001) discuss the evaluation of the false discovery rate (FDR) in a multiple testing framework with correlated test statistics. Another route is to invoke ergodic results, such as Theorem 5.16 in White (2001). In our framework, connecting the dots would require to model the sequences $b_m - b$ as adapted mixingales.

   One alternative approach would rely on the results of Chen and Doerge (2020). The main theorem therein proves a strong law of large numbers for the average number of rejections (from the p-values) and average number of false discoveries. The results hold under fairly general dependence structure, as long the L^1 norm of the correlation matrix does not increase too rapidly with the number of outcomes. This assumption is hard to verify in practice, but assuming it holds, the proportion of rejections at the α level is approximated by $M^{-1} \sum_{m=1}^{M} 1_{\{p_m < \alpha\}}$, where the p_m are the p-values associated with the b_m coefficients.

2. If the model under consideration can be formulated in a simple uni-dimensional fashion (observations have only one index, i), a panel approach can be considered in order to build a meta-model. In this case, the two dimensions are defined as follows: $y_{n,i}$ are $\boldsymbol{x}_{n,i}$ are the i^{th} observations of path number n. The model becomes longitudinal along the paths. Again, correlation issues arise, and the literature provides some solutions to handle them. The most direct technique is to assume a factor structure in the errors, as in Pesaran (2006). Extensions to non-linear models can be found in Su and Jin (2012) and Cai et al. (2020).

3.4 Exhaustiveness and selectivity

One major challenge with forking paths is computation times. If, at step j, there are η_j options which take τ_j time to compute (assuming homogeneity across options), then the total time to run all paths is $\prod_{j=1}^{J} \eta_j \tau_j$. Here we consider that all paths have the same probability for simplicity. Let us lay out a simple example. We assume $J = 10$ and half of the steps have two options,

while the other half have three. This makes $2^5 \times 3^5 = 7,776$ paths in total, an amount which is rather *exhaustive*. If each stage takes 10 seconds, it will take 21.6 hours to generate all paths.

This implies that the forking paths must be chosen carefully if some steps require long treatments: some robustness checks (mappings) are more impactful than others. In addition, the reporting of outcomes of forking paths forces to focus on a limited number of metrics to disclose. In some fields, it is customary to gather dozens, if not hundreds, of values inside tables. Showing distributional properties of outcomes requires more space than one estimate and one p-value, which forces the researcher to be more *selective* with the information that is shown in academic articles.

4 Application: equity premium prediction

4.1 Data

For the sake of reproducibility, the illustration of the concepts of the paper rely on a public dataset as well as on a problem which is widely documented in the literature.[9] In financial economics, an old, still unresolved, question pertains to whether aggregate stock returns can be predicted by macro-economics indicators. The debate is impossible to settle, but recent results suggest a contingency on return horizon (Bandi et al. (2019)), even if long-term predictability is biased by construction (Boudoukh et al. (2008), Boudoukh et al. (2021)).

A critical view on the matter is the seminal article by Welch and Goyal (2008), in which the authors document the poor forecasting ability of traditional predictors. A favorable feature of the study is that the data is public, and has even been updated in the follow-up paper Goyal et al. (2021). It is this material that we use for our application.

4.2 Forking paths

In order to generate enough metrics, we work with nine stages which are depicted in Figure 3 below. We briefly comment on each.

1. **data frequency** determines the horizon of returns, hence the left-hand side of the equation. In addition, this has a major effect on sample sizes, as annual samples are 12 times smaller, compared to monthly ones.
2. **handling missing points** boils to two options. The first is to remove rows of missing points, which means all regressors will start at the same point in time (1937). The second option (imputation) allows predictor-dependent sample sizes and some of them are available in 1871. Thus, this stage impacts sample depth and samples with fewer than 30 observations are discarded from the analysis.
3. **winsorization** defines the cutoff threshold for the taming of outliers, from none (0%) to 3%. The data is standard, thus all values are trustworthy, so this step could theoretically be omitted. But it participates to increase the number of outputs, hence we keep it for the sake of exhaustiveness.

[9] The code used to generate all results is available at https://www.gcoqueret.com/files/misc/forking_paths.html

4. **variable engineering** decides whether or not to use levels or differences in the regressions.
5. **independent variable** sets the predictor. Six options are possible and all are available across the three frequencies (monthly to annually).[10]
6. **horizon** fixes the number of periods that are used to compute the future return (dependent variable). We underline that three periods have different meanings depending on the original data frequency (chosen in step 1).
7. **starting point** determines if the sample commences at its first point, or at its middle point. This option leaves room for sub-sampling (on the two halves of each original sample).
8. **end point** is either the end of the sample, or its middle point. The latter option is not possible if it also corresponds to the starting point.
9. **estimator type** chooses between iid errors, or the improved HAC variance estimator of Newey and West (1987). The regression model is simply

$$y_{t+1} = a + bx_t + e_{t+1}, \tag{19}$$

where y_{t+1} is the equity premium, x_t the lagged predictor and e_{t+1} the residual.

Code-wise, each path is generated by a call to chained modules, akin to the first version of Keras API (keras.io), as shown in Table 1. The paths are determined by the combination of parameters which are the arguments of the modules. Resorting to functional programming and parallelization eases the syntax and accelerates the production of outputs.

```
module_sheet(prob_sheet = prob_sheet) %>%                # Step 1: select data frequency
  module_cleaning(prob_remove = prob_remove) %>%         # Step 2: clean the data
  module_winsorization(threshold = threshold) %>%        # Step 3: manage outliers
  module_levels(prob_levels = prob_levels) %>%           # Step 4: decide between levels versus differences
  module_independent(variable = variable) %>%            # Step 5: choose predictor
  module_horizon(horizon = horizon) %>%                  # Step 6: pick return horizon (dependent variable)
  module_start(start_quantile = start_quantile) %>%      # Step 7: starting point (for subsampling)
  module_stop(stop_quantile = stop_quantile) %>%         # Step 8: ending point (for subsampling)
  module_estimator(prob_HAC = prob_HAC)                  # Step 9: select estimator type
```

Table 1: **Illustration via R code**

There are $\prod_{j=1}^{9} r_j = 6,912$ possible paths from the data to the output. For simplicity, each is equi-probable so that we only need to consider each combination once.

[10] **payout** is the difference between the log of dividends and the log of earnings, **b/m** is the the ratio of book value to market value for the Dow Jones Industrial Average, **svar** is the sum of squared daily returns on S&P 500, **dfr** is the difference between the return on long-term corporate bonds and returns on the long-term government bonds, **dfy** is is the difference between BAA- and AAA- rated corporate bond yields, and **ntis** is the ratio of twelve-month moving sums of net issues by NYSE listed stocks divided by the total market capitalization of NYSE stocks.

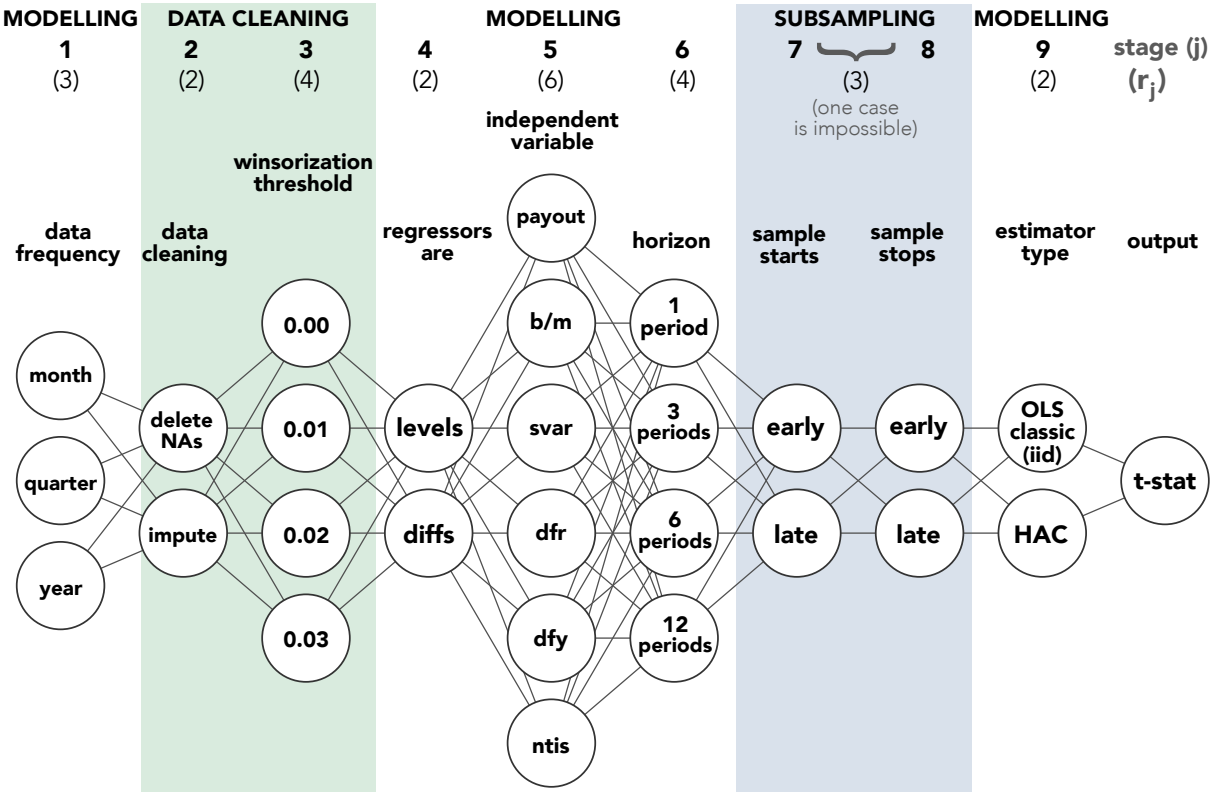

Figure 3: **Diagram of empirical protocol** (forking paths). The graph, akin to a neural network structure, depicts the nine-step algorithm used to produce the t-statistic in the study. Each path has the same probability of realization (uniform distribution). The number of mapping options, r_j is reported between brackets below the stage number.

4.3 Baseline results

In Figure 4, we plot the distribution of the p-values obtained across all paths. The restriction of the distribution, to the extreme left of its support, known as the *p-curve*, (Simonsohn et al. (2014a)) is downward sloping (or right skewed). The opposite shape (left-skewed) is usually associated with attempts of p-hacking. In addition, we show the cumulative distribution function (cdf) of p-values associated to each independent variable. The latter choice is an important modelling decision which is why we consider predictors separately. The cdf can be a tool for p-hacking detection, especially if it is not concave (Elliott et al. (2021)). The corresponding curve for the *ntis* predictor seems to be concave, which signals a potentially strong effect of this variable over the equity premium.

The impact of mappings is displayed in Figures 5 (histograms for dual choices) and 6 (box-plots for mappings with more options). In the first case, while distributions are not necessarily very close, it is hard to argue that the mappings profoundly alter the average of the result.

In Figure 6 however, two phenomena are at work. In the right panel, the horizon of the

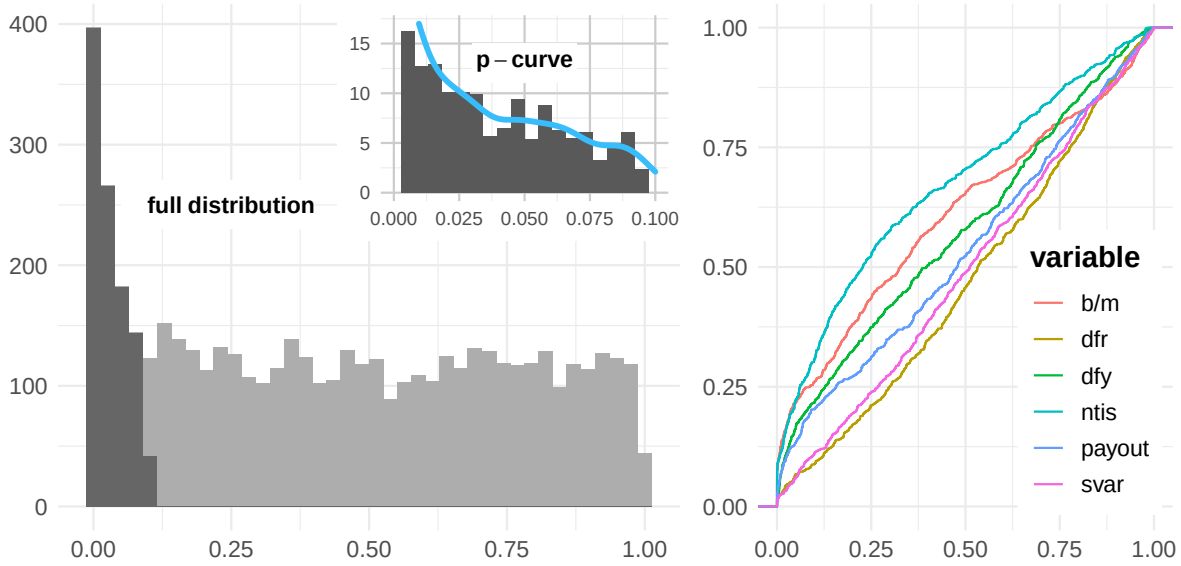

Figure 4: **Distribution of p-values**. In the left panel, we plot the histogram of all p-values, as well as the p-curve (Simonsohn et al. (2014a)), which is the restriction of the distribution to the interval of significant values (which we take to be [0,0.1]). In the right panel, we show the cdf of the p-values, sorted by independent variable. Results for regressions with fewer than 30 observations are discarded.

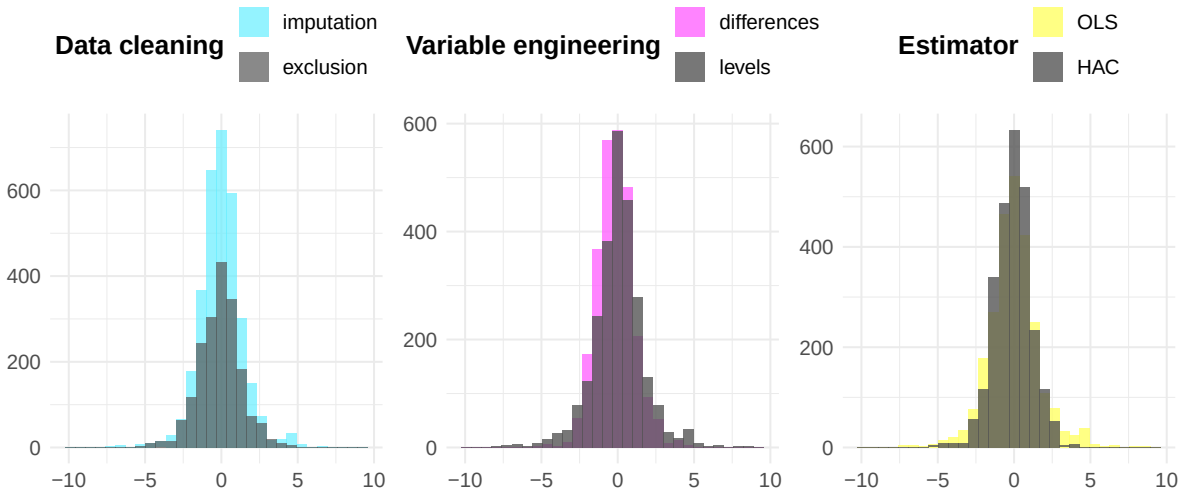

Figure 5: **Impact of mappings: robustness checks**. We report the distribution of t-statistics for three binary choices in mappings. Results for regressions with fewer than 30 observations are discarded.

dependent variable generates heterogeneity in the dispersion of t-statistics. The latter have

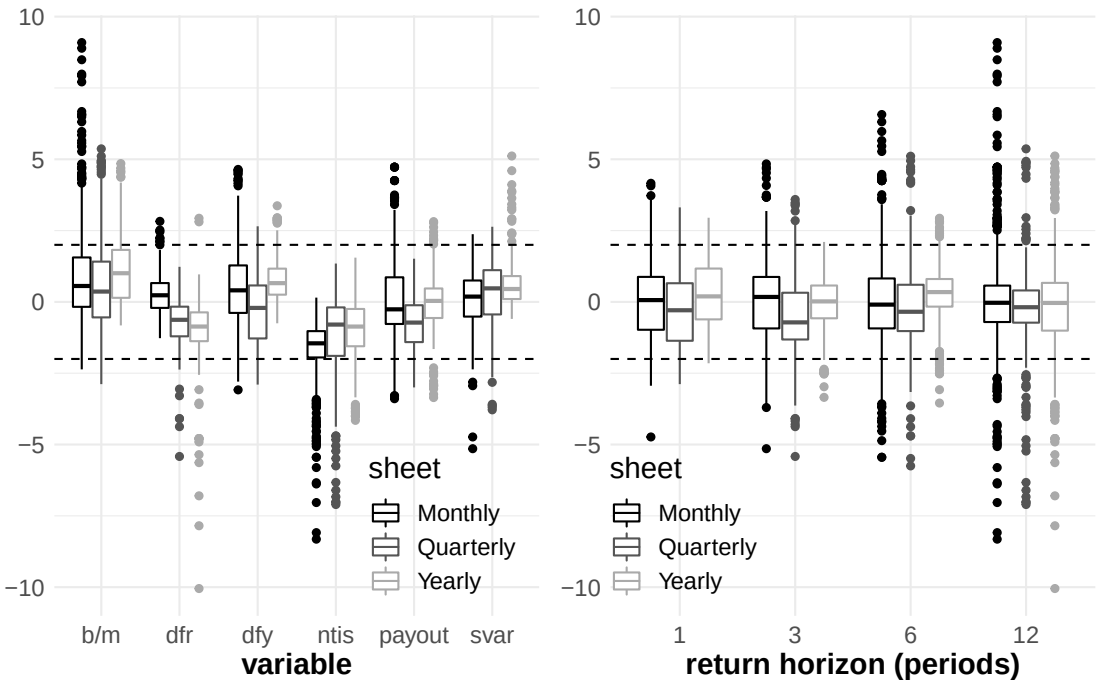

Figure 6: **Drivers of scalar output: modelling assumptions**. We report the distribution of t-statistics for two important modelling choices: the independent variable (left panel), and the return horizon of the dependent variable. Results for regressions with fewer than 30 observations are discarded.

more outliers beyond the 5% threshold for the long-term returns. This is somewhat in line with the findings of Bandi et al. (2019), if we only focus on the magnitude of the statistics (and not their sign). We note, however, that the inter-quartile ranges remain inside the *insignificance band* between the dotted lines.

The left panel of Figure 6 clearly illustrates the contingency of results on the choice of the predictor. This is true both for the signs of t-statistics (rather positive for *b/m*, rather negative for *ntis*), but also for their dispersion (*dfy* seems to have few outliers).

4.4 Drivers of outcomes

To illustrate the concept of drivers of outcomes proposed in Definition 3, we run batches of paired Wilcoxon tests between two types of scalars obtained in the predictive regressions: the t-statistics and the corresponding p-values. The main difference between the two comes from the importance from the sign. Two series with t-statistics of -4 and +4 will have similar (negligible) p-values, but the test will reject equality between the statistics.

In Table 2, we provide the p-values of the Wilcoxon tests for both outcome types (t-statistics in the upper triangle and p-values in the lower triangle). The 0.281 value in the middle of the first row means that the t-statistics pertaining to the *payout* and *dfr* variable have means that

are not significantly different. Thus, switching from one variable to the other is not impactful for the conclusion of the study. This is confirmed in the left panel of Figure 6, from which we infer that both means are slightly negative (-0.21 and -0.37), and indeed not very far apart.

		pair-wise tests						simple test
		payout	b/m	svar	dfr	dfy	ntis	
pair-wise tests	payout		0.000	0.000	**0.335**	0.000	0.000	0.000
	b/m	0.000		0.000	0.000	0.000	0.000	0.000
	svar	0.003	0.000		0.000	**0.281**	0.000	0.000
	dfr	0.000	0.000	**0.057**		0.000	0.000	0.000
	dfy	0.008	**0.054**	0.000	0.000		0.000	0.000
	ntis	0.000	0.000	0.000	0.000	0.000		0.000
	simple test	0.011	0.000	0.000	0.000	0.043	0.000	

Table 2: **Test for the variable choice**. We evaluate if the choice of independent variable is a mapping that is a driver of scalar output, in the sense of Definition 3. The reported numbers in the bulk of the table are the p-values of the Wilcoxon test for all pairs of variables. The upper triangle of the matrix pertains to the values obtains when comparing the series of t-statistics, while the lower triangle relates to comparisons of p-values (both t-statistics and p-values are scalar outcomes). In the last column and row, we report the p-values of the simple mean tests of each predictor outcome versus all other predictor values (performed on t-statistics and p-values, respectively). Results for regressions with fewer than 30 observations were beforehand discarded.

In both triangles of the inner matrix of Table 2, two p-values lie above the 5% threshold (though marginally above in the lower triangle). According to Definition 3, the choice of the variable is therefore a weak driver of both outcomes at the 5% level. At the 6% level however, the mapping that chooses the variable becomes a **strong** driver of p-values: the series of p-values have significantly different means (across variables).

The last row and column of Table 2 include the p-values of the mean tests when comparing p-values (row) or t-statistics (column) of each predictor versus all others. The tests forcefully reject the null for the t-statistics. For the p-values, some predictors (*payout* and *dfy*) have results above 1% , implying that their average p-values are not exceedingly far from the average of all other predictors.

4.5 A closer look at *b/m* and *ntis*

The evaluation of the impact of one mapping at the over-arching level, as shown in Figure 5 can hide important local differences. For instance, the choice of levels versus differences in the central panel seems to have little impact on the full distribution of t-statistics. In Figure 7, we use a jitter plot to locate the t-statistics of the seemingly two best predictors of the equity premium. They are shown in panels of data frequency (x-axis inside plots) and of forecasting horizon (from left to right).

For the *b/m* variable, the differentiation mapping has undoubtedly a strong impact. The two colors are hardly intertwined in many panels, meaning that the average t-statistic for levels and that for differences are significantly different, on average. The associated p-values are zero,

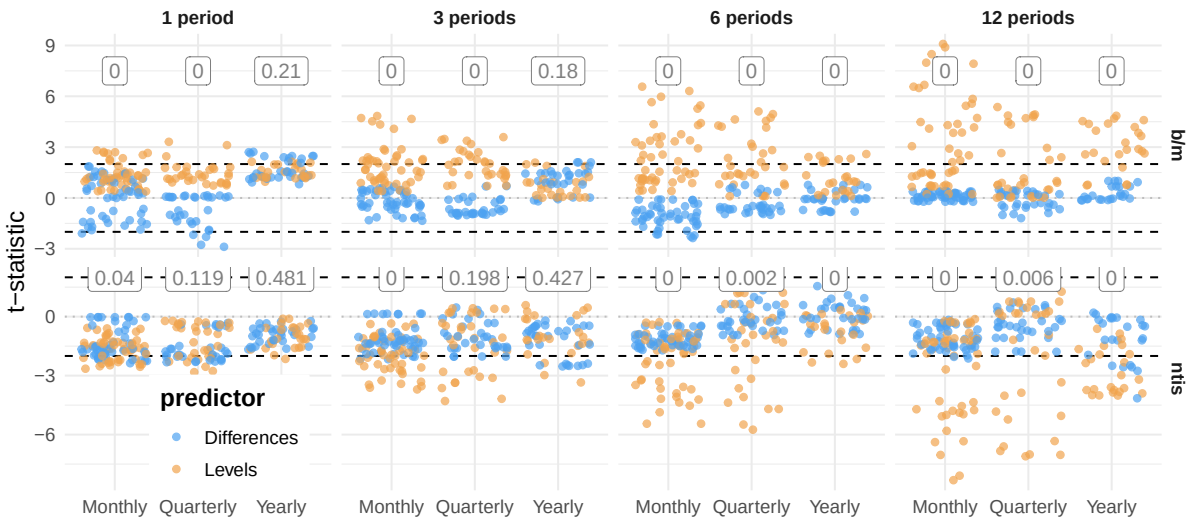

Figure 7: **Focus on *b/m* and *ntis***. We plot the values of t-statistics for two predictors: b/m (upper panel) and *ntis* (lower panel). These values are shown as functions of data frequency (x-axis inside plots) and of forecasting horizon (from left to right). Colors mark the difference between levels and differences for predictors and the numbers in the rounded rectangles are the p-values of the t-test between the two series (levels versus differences). Results for regressions with fewer than 30 observations are discarded.

except for two clusters (1 and 3 periods for annual data). For *ntis*, the discrepancy between the two subsets is more ambiguous, though it is clearer for long forecasting horizons (all p-values are below 1%).

4.6 Model averaging

After the generation of multiple estimated effects, the natural next step is to aggregate them into a meta-estimator and the most straightforward option is a linear combination:

$$\hat{b}_* = \sum_{j=1}^{J} w_j \hat{b}_j, \tag{20}$$

where the main issue is the determination of the weights w_j. There are several ways to proceed in a **frequentist** fashion (see Moral-Benito (2015), Zhang and Liu (2019) and Steel (2020)), and for simplicity, we will stick with the with the definition that relies on likelihood through information criteria:

$$w_j = \frac{e^{-\Delta_j/2}}{\sum_{k=1}^{J} e^{-\Delta_k/2}}, \quad \Delta_j = AIC_j - \min_j AIC_j,$$

where AIC_j is the Akaike Information Criterion of model j. For the estimation of the variance of the aggregate estimator, we follow Equation (1) in Burnham and Anderson (2004):

$$\hat{\sigma}_*^2 = \left(\sum_{j=1}^{J} w_j \sqrt{\hat{\sigma}_j + (\hat{b}_* - \hat{b}_j)^2} \right)^2 .$$

In Figure 8, we show the averaged coefficients within their 95% confidence interval. We split the analysis along three axes: variable, level versus difference, and sampling frequency. The latter is important because it is determinant in the sample sizes which are used to compute the width of the intervals. For a given frequency and variable, they are homogeneous, though not exactly equal, and we use their weighted average $T_* = \sum_{j=1}^{J} w_j T_j$. The impact of sampling frequency on the width of intervals is obvious. Intervals pertaining to monthly data coincide with the average estimators, whereas intervals linked to annual samples are fairly large.

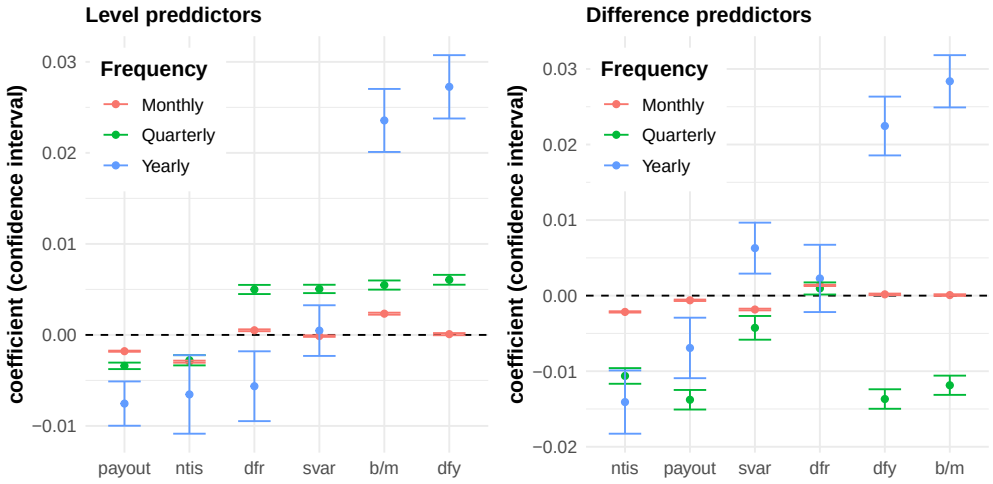

Figure 8: **Frequentist model averaging**. We display average coefficients within their 95% confidence interval. Coefficients stem from Equation (20). Confidence intervals are defined by $[\hat{b}_* - 1.96\sigma_*^2/\sqrt{T_*}, \hat{b}_* + 1.96\sigma_*^2/\sqrt{T_*}]$, where $T_* = \sum_{j=1}^{J} w_j T_j$, with T_j being the sample size of model j. The left panel displays results when predictors are levels, while the right one focuses on differences of variables. To allow comparisons, all predictors are scaled to have unit variance before estimation.

As a confirmation of our previous results, we obtain that *ntis* yields only negative coefficients, and the intervals do not overlap with zero. The *b/m* variable has mostly positive estimates, with one exception of the quarterly data in the right panel. Surprisingly, the *dfy* variable also stands out with coefficients which are large in magnitude for quarterly and annual samples. For quarterly variables, the effect cancels out between levels (positive coefficients) and differences (negative ones). This partly explains why the variable was not previously identified as a potent driver of the equity premium.

For the sake of completeness, we also propose an analysis from the perspective of **Bayesian** model averaging. We follow the standard nomenclature, as is for instance laid out in Hoeting et al. (1999). The quantity of interest is b, with posterior probability given the data D equal to

$$\mathbb{P}[b|D] = \sum_{m=1}^{M} \mathbb{P}[b|M_m, D]\mathbb{P}[M_m|D],$$

where $\mathbb{M} = \{M_m, m = 1, \ldots, M\}$ is the set of models under consideration. In this paper, one model corresponds to one complete path. Notably, the above equation translates to the following conditional average and variance:

$$\mathbb{E}[b|D] = \sum_{m=1}^{M} \hat{b}_m \mathbb{P}[M_m|D] \tag{21}$$

$$\mathbb{V}[b|D] = \sum_{m=1}^{M} \left(\mathbb{V}(b|M_m, D) + \hat{b}_m^2\right) \mathbb{P}[M_m|D] - (\mathbb{E}[b|D])^2 \tag{22}$$

where $\hat{b}_m$ is the estimated effect in model m. The posterior model probabilities are given by

$$\mathbb{P}[M_m|D] = \left(\sum_{j=1}^{M} \frac{\mathbb{P}[M_j]}{\mathbb{P}[M_m]} \frac{l_D(M_j)}{l_D(M_m)}\right)^{-1}, \tag{23}$$

with $l_D(M_j)$ being the marginal likelihood of model j. Because we are agnostic with respect to which model is more realistic, we will set the prior odds $\frac{\mathbb{P}[M_j]}{\mathbb{P}[M_m]}$ equal to one. Note that in this case, the posterior probabilities are then simply proportional to the likelihoods. The remaining Bayes factor is by far the most complex and we follow the recommendations of Fernandez et al. (2001) (Equation (2.16), adapted for inhomogeneous sample sizes):

$$\frac{l_D(M_j)}{l_D(M_m)} = \left(\frac{n_j}{n_j+1}\right)^{\frac{k_j}{2}} \left(\frac{n_m+1}{n_m}\right)^{\frac{k_m}{2}} \frac{\left(\frac{s_m+n_m v_m}{n_m+1}\right)^{(n_m-1)/2}}{\left(\frac{s_j+n_j v_j}{n_j+1}\right)^{(n_j-1)/2}}, \tag{24}$$

where n_j is the inverse of the number of observations used in model j and k_j is the number of predictors in this model, omitting the constant ($k_j = 1$ in our case). Moreover, $n_j v_j$ is the sample variance of the dependent variable in model j. Finally, s_j is the sum of squared residuals under model j.

In Figure 9, we plot the average coefficients computed according to Equation (21), along with ad-hoc confidence intervals. We only report results for annually sampled variables in order to reduce sample sizes. Indeed, their exponentiation in some term of Equation (24) are problematic when two models have significantly contrasting sample sizes. This issue is circumvented with annual samples.

The intervals in Figure 9 tend to confirm those obtained for the frequentist averages. Both *ntis* and *b/m* are associated with intervals that do not overlap over zero. In fact, for difference predictors, the Bayesian averages are indistinguishable from their frequentist counterparts.

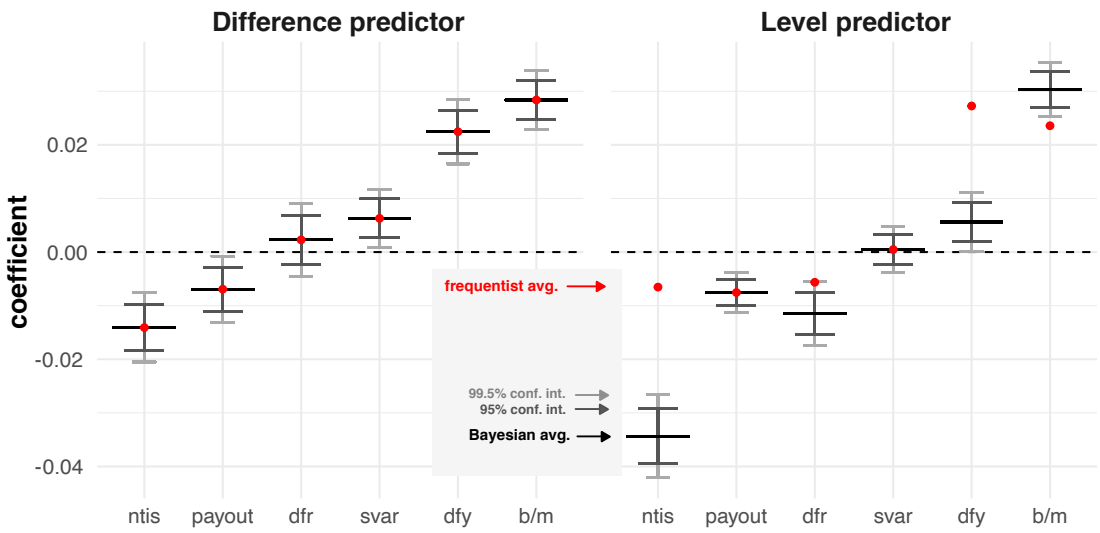

Figure 9: **Bayesian model averaging** (annual data). We display average coefficients within their 95% and 99.5% confidence intervals. Averages stem from Equation (21). The bounds of the confidence intervals are defined by $\mathbb{E}[b|D] \pm \alpha\sqrt{\mathbb{V}[b|D]/T_*}$, where $T_* = \sum_{j=1}^{J} w_j T_j$, with T_j being the sample size of model j and w_j the posterior model probabilities. α relates to the confidence level. The left panel displays results when predictors are differences, while the right one focuses on levels. To allow comparisons, all predictors are scaled to have unit variance before estimation.

4.7 Rejection rates

Our final analysis pertains to the average rejection rate proposed in Chen and Doerge (2020). In the left panel of Figure 10, we plot these rates for each individual predictor, as well as for all variables taken together (ALL, in black). For some variables (*ntis* and *b/m* notably), it is plain that the rejection rate is well above the theoretical decision threshold (under the null), which seems to suggest some kind of predictability. In the right panel, we consider the subcases when the predictors are only taken as levels, and not differences. The rejection rates at the 1% level then surpass 20% for these two variables.

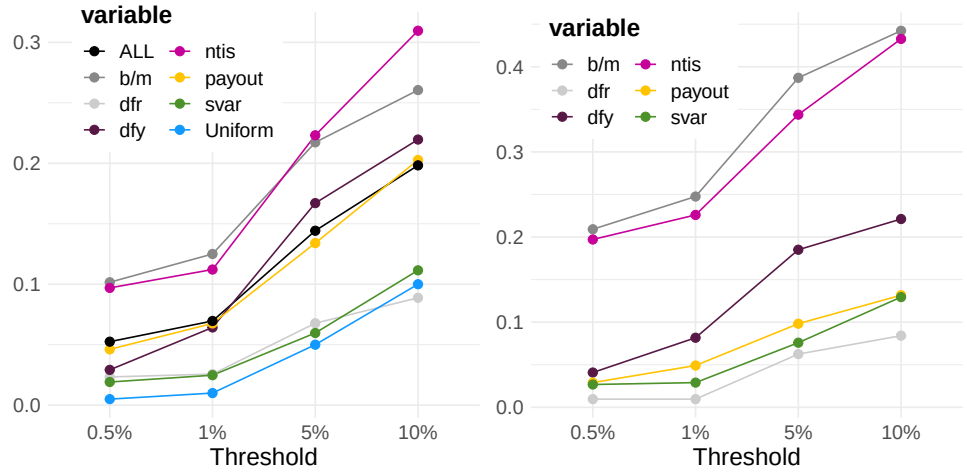

Figure 10: **Rejection rates**. Following Chen and Doerge (2020), we compute the average rejection rate across all forking paths (left), and all forking paths with level predictors (right). We report the numbers for 4 significance levels (x-axis), and split the analysis on a variable-by-variable basis. The theoretical rate under the null is shown in light blue.

5 Conclusion

Amid debates in some scientific communities about the validity of empirical results, we make the case for an exhaustive approach. Namely, we suggest to report results for a large number of design choices, as if extensive robustness checks were in fact constituent of the baseline research protocol. Small variations in designs allow the generation of many estimates and test statistics. The distribution of these statistics can help figure out if one configuration yielded a favorable outlier, or if the sought effect is indeed statistically strong. Moreover, having many coefficient or statistics at one's disposal allows to resort to aggregation so as to obtain more robust values and confidence intervals.

The application of these ideas to equity premium prediction shows that two variables, net equity expansion and aggregate book-to-market ratio, have strong explanatory power over future excess returns.

There are of course several limitations to our suggestion. First, it is possible to push the limits of data-snooping to the extreme by reporting only the combinations of design choices that fit a particular narrative. Second, given the amount of time required to generate comprehensive results, the research question must be inherently simple. Each path should not take more than a handful of minutes, so that hundreds, or thousands, of them can be generated in less than one day. The aim of the paper is clearly not to increase the carbon footprint of researchers. Long computation times may contribute to this footprint and we refer to Mariette et al. (2021) for a discussion on this matter. This is why a precise framing of the research question, as well as its relevant ramifications, is imperative to avoid superfluous digressions.

References

Akey, P., A. Robertson, and M. Simutin (2021). Noisy factors. *SSRN Working Paper 3930228*.

Amenc, N., F. Goltz, and B. Luyten (2020). Intangible capital and the value factor: Has your value definition just expired? *Journal of Portfolio Management 46*(7), 83–99.

Amrhein, V., S. Greenland, and B. McShane (2019). Scientists rise up against statistical significance. *Nature 567*, 305–307.

Andrews, I. and M. Kasy (2019). Identification of and correction for publication bias. *American Economic Review 109*(8), 2766–94.

Asness, C. and A. Frazzini (2013). The devil in HML's details. *Journal of Portfolio Management 39*(4), 49–68.

Avramov, D., S. Cheng, L. Metzker, and S. Voigt (2022). Integrating factor models. *Journal of Finance Forthcoming*.

Bailey, D. H., J. Borwein, M. Lopez de Prado, and Q. J. Zhu (2014). Pseudo-mathematics and financial charlatanism: The effects of backtest overfitting on out-of-sample performance. *Notices of the American Mathematical Society 61*(5), 458–471.

Bailey, D. H. and M. Lopez de Prado (2021). Finance is not excused: Why finance should not flout basic principles of statistics. *Significance (Royal Statistical Society) Forthcoming*.

Bandi, F. M., B. Perron, A. Tamoni, and C. Tebaldi (2019). The scale of predictability. *Journal of Econometrics 208*(1), 120–140.

Basit, T. (2003). Manual or electronic? The role of coding in qualitative data analysis. *Educational research 45*(2), 143–154.

Begg, C. B. and J. A. Berlin (1988). Publication bias: a problem in interpreting medical data. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 151*(3), 419–445.

Benjamin, D. J., J. O. Berger, M. Johannesson, B. A. Nosek, E.-J. Wagenmakers, R. Berk, K. A. Bollen, B. Brembs, L. Brown, C. Camerer, et al. (2018). Redefine statistical significance. *Nature Human Behaviour 2*(1), 6–10.

Benjamini, Y. and D. Yekutieli (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics 29*(4), 1165–1188.

Bergman, A., A. Chinco, S. M. Hartzmark, and A. B. Sussman (2020). Survey curious? Start-up guide and best practices for running surveys and experiments online. *SSRN Working Paper 3701330*.

Bonferroni, C. (1936). Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze 8*, 3–62.

Boriah, S., V. Chandola, and V. Kumar (2008). Similarity measures for categorical data: A comparative evaluation. In *Proceedings of the 2008 SIAM International Conference on Data Mining*, pp. 243–254. SIAM.

Boudoukh, J., R. Israel, and M. Richardson (2021). Biases in long-horizon predictive regressions. *Journal of Financial Economics Forthcoming*.

Boudoukh, J., M. Richardson, and R. F. Whitelaw (2008). The myth of long-horizon predictability. *Review of Financial Studies 21*(4), 1577–1605.

Brodeur, A., N. Cook, and A. Heyes (2020). Methods matter: P-hacking and publication bias in causal analysis in economics. *American Economic Review 110*(11), 3634–60.

Brodeur, A., M. Lé, M. Sangnier, and Y. Zylberberg (2016). Star wars: The empirics strike back. *American Economic Journal: Applied Economics 8*(1), 1–32.

Buckland, S. T., K. P. Burnham, and N. H. Augustin (1997). Model selection: an integral part of inference. *Biometrics 53*(2), 603–618.

Burnham, K. P. and D. R. Anderson (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research 33*(2), 261–304.

Cai, Z., Y. Fang, and Q. Xu (2020). Testing capital asset pricing models using functional-coefficient panel data models with cross-sectional dependence. *Journal of Econometrics Forthcoming*.

Carver, R. (1978). The case against statistical significance testing. *Harvard Educational Review 48*(3), 378–399.

Chang, X. S., H. Gao, and W. Li (2021). P-hacking in experimental accounting studies. *SSRN Working Paper 3762342*.

Chen, A. Y. (2021). Most claimed statistical findings in cross-sectional return predictability are likely true. *SSRN Working Paper 3912915*.

Chen, A. Y. and M. Velikov (2021). Zeroing in on the expected returns of anomalies. *SSRN Working Paper 3073681*.

Chen, X. and R. W. Doerge (2020). A strong law of large numbers related to multiple testing normal means. *Statistics & Probability Letters 159*, 108693.

Chordia, T., A. Goyal, and A. Saretto (2020). Anomalies and false rejections. *Review of Financial Studies 33*(5), 2134–2179.

Christensen, G. and E. Miguel (2018). Transparency, reproducibility, and the credibility of economics research. *Journal of Economic Literature 56*(3), 920–80.

De Long, J. B. and K. Lang (1992). Are all economic hypotheses false? *Journal of Political Economy 100*(6), 1257–1272.

De Prado, M. L. (2018). The 10 reasons most machine learning funds fail. *Journal of Portfolio Management 44*(6), 120–133.

Demmel, J. (1992). The componentwise distance to the nearest singular matrix. *SIAM Journal on Matrix Analysis and Applications 13*(1), 10–19.

Dickersin, K., S. Chan, T. Chalmersx, H. Sacks, and H. Smith Jr (1987). Publication bias and clinical trials. *Controlled Clinical Trials 8*(4), 343–353.

Doucouliagos, C. and T. D. Stanley (2013). Are all economic facts greatly exaggerated? Theory competition and selectivity. *Journal of Economic Surveys 27*(2), 316–339.

Doucouliagos, H. and T. D. Stanley (2009). Publication selection bias in minimum-wage research? A meta-regression analysis. *British Journal of Industrial Relations 47*(2), 406–428.

Draper, D. (1995). Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society: Series B (Methodological) 57*(1), 45–70.

Echenique, F. and K. He (2021). Screening p-hackers: Dissemination noise as bait. *arXiv Preprint* (2103.09164).

Elliott, G., N. Kudrin, and K. Wuthrich (2021). Detecting p-hacking. *Econometrica Forthcoming*.

Fabozzi, F. J. and M. L. de Prado (2018). Being honest in backtest reporting: A template for disclosing multiple tests. *Journal of Portfolio Management 45*(1), 141–147.

Fan, J. and X. Han (2017). Estimation of the false discovery proportion with unknown dependence. *Journal of the Royal Statistical Society. Series B, Statistical methodology 79*(4), 1143.

Fanelli, D. (2018). Opinion: Is science really facing a reproducibility crisis, and do we need it to? *Proceedings of the National Academy of Sciences 115*(11), 2628–2631.

Fanelli, D., R. Costas, and J. P. Ioannidis (2017). Meta-assessment of bias in science. *Proceedings of the National Academy of Sciences 114*(14), 3714–3719.

Farcomeni, A. (2008). A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. *Statistical methods in medical research 17*(4), 347–388.

Fernandez, C., E. Ley, and M. F. Steel (2001). Benchmark priors for bayesian model averaging. *Journal of Econometrics 100*(2), 381–427.

Frankel, A. and M. Kasy (2022). Which findings should be published? *American Economic Journal: Microeconomics 14*(1), 1–38.

Gelman, A. and E. Loken (2014). The statistical crisis in science. *American Scientist 102*, 460–465.

Giglio, S., Y. Liao, and D. Xiu (2021). Thousands of alpha tests. *Review of Financial Studies 34*(7), 3456–3496.

Golub, G. H. and J. H. Wilkinson (1966). Note on the iterative refinement of least squares solution. *Numerische Mathematik 9*(2), 139–148.

Gong, Q., M. Liu, and Q. Liu (2015). Momentum is really short-term momentum. *Journal of Banking & Finance 50*, 169–182.

Goyal, A., I. Welch, and A. Zafirov (2021). A comprehensive look at the empirical performance ofequity premium prediction ii. *SSRN Working Paper 3929119*.

Grcar, J. F. (2003). Optimal sensitivity analysis of linear least squares. *Lawrence Berkeley National Laboratory, Report LBNL-52434 99*.

Greene, W. H. (2018). *Econometric analysis - Eighth Edition*. Pearson Education India.

Hansen, B. E. (2007). Least squares model averaging. *Econometrica 75*(4), 1175–1189.

Hansen, P. R., A. Lunde, and J. M. Nason (2011). The model confidence set. *Econometrica 79*(2), 453–497.

Harvey, C. R. (2017). Presidential address: The scientific outlook in financial economics. *Journal of Finance 72*(4), 1399–1440.

Harvey, C. R. (2021). Be skeptical of asset management research. *Available at SSRN Working Paper 3906277*.

Harvey, C. R. and Y. Liu (2020). False (and missed) discoveries in financial economics. *Journal of Finance 75*(5), 2503–2553.

Harvey, C. R. and Y. Liu (2021). Uncovering the iceberg from its tip: A model of publication bias and p-hacking. *SSRN Working Paper 3865813*.

Harvey, C. R., Y. Liu, and A. Saretto (2020). An evaluation of alternative multiple testing methods for finance applications. *Review of Asset Pricing Studies 10*(2), 199–248.

Head, M. L., L. Holman, R. Lanfear, A. T. Kahn, and M. D. Jennions (2015). The extent and consequences of p-hacking in science. *PLoS biology 13*(3), e1002106.

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky (1999). Bayesian model averaging: A tutorial. *Statistical Science 14*(4), 382–417.

Huntington-Klein, N., A. Arenas, E. Beam, M. Bertoni, J. R. Bloem, P. Burli, N. Chen, P. Grieco, G. Ekpe, and T. Pugatch (2021). The influence of hidden researcher decisions in applied microeconomics. *Economic Inquiry 59*, 944–960.

Ioannidis, J., T. Stanley, and H. Doucouliagos (2017). The power of bias in economics research. *Economic Journal 127*(605), F236–F265.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS medicine 2*(8), e124.

Jensen, T. I., B. T. Kelly, and L. H. Pedersen (2021). Is there a replication crisis in finance? *Journal of Finance Forthcoming*.

Kasy, M. (2021). Of forking paths and tied hands: Selective publication of findings, and what economists should do about it. *Journal of Economic Perspectives 35*(3), 175–92.

Leamer, E. and H. Leonard (1983). Reporting the fragility of regression estimates. *Review of Economics and Statistics 65*(2), 306–317.

Leek, J. T. and L. R. Jager (2017). Is most published research really false? *Annual Review of Statistics and Its Application 4*, 109–122.

Lo, A. W. and A. C. MacKinlay (1990). Data-snooping biases in tests of financial asset pricing models. *Review of Financial Studies 3*(3), 431–467.

Magnus, J. R. and G. De Luca (2016). Weighted-average least squares (WALS): a survey. *Journal of Economic Surveys 30*(1), 117–148.

Mariette, J., O. Blanchard, O. Berné, and T. B. Ari (2021). An open-source tool to assess the carbon footprint of research. *arXiv Preprint* (2101.10124).

McShane, B. B., D. Gal, A. Gelman, C. Robert, and J. L. Tackett (2019). Abandon statistical significance. *American Statistician 73*, 235–245.

Menkveld, A., A. Dreber, F. Holzmeister, M. Johannesson, J. Huber, M. Kirchler, S. Neususs, M. Razen, and U. Weitzel (2021). Non-standard errors. *SSRN Working Paper 3961574*.

Milkman, K. et al. (2021). Megastudies improve the impact of applied behavioral science. *Nature*.

Mitton, T. (2021a). Economic significance in corporate finance. *SSRN Working Paper 3667830*.

Mitton, T. (2021b). Methodological variation in empirical corporate finance. *Review of Financial Studies Forthcoming*.

Moral-Benito, E. (2015). Model averaging in economics: An overview. *Journal of Economic Surveys 29*(1), 46–75.

Morey, M. R. and S. Yadav (2018). Documentation of the file drawer problem in academic finance journals. *Journal of Investing 27*(1), 143–147.

Muralidharan, K., M. Romero, and K. Wüthrich (2022). Factorial designs, model selection, and (incorrect) inference in randomized experiments. *SSRN Working Paper 3551804*.

Newey, W. and K. West (1987). A simple, positive semi-definite, heteroskedasticity and auto-correlation consistent covariance matrix. *Econometrica 55*(3), 703–708.

Novy-Marx, R. (2012). Is momentum really momentum? *Journal of Financial Economics 103*(3), 429–453.

Olson, C. M., D. Rennie, D. Cook, K. Dickersin, A. Flanagin, J. W. Hogan, Q. Zhu, J. Reiling, and B. Pace (2002). Publication bias in editorial decision making. *Journal of the American Medical Association 287*(21), 2825–2828.

Pesaran, M. H. (2006). Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica 74*(4), 967–1012.

Raftery, A. E., D. Madigan, and J. A. Hoeting (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association 92*(437), 179–191.

Rampini, A. A., S. Viswanathan, and G. Vuillemey (2021). Risk management in financial institutions. *Journal of Finance 75*(2), Retracted.

Reinhart, C. M. and K. S. Rogoff (2010). Growth in a time of debt. *American Economic Review 100*(2), 573–78.

Romano, J. P. and M. Wolf (2005). Stepwise multiple testing as formalized data snooping. *Econometrica 73*(4), 1237–1282.

Romano, J. P. and M. Wolf (2010). Balanced control of generalized error rates. *Annals of Statistics 38*(1), 598–633.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological bulletin 86*(3), 638–641.

Simonsohn, U., L. D. Nelson, and J. P. Simmons (2014a). P-curve: a key to the file-drawer. *Journal of Experimental Psychology: General 143*(2), 534.

Simonsohn, U., L. D. Nelson, and J. P. Simmons (2014b). p-curve and effect size: Correcting for publication bias using only significant results. *Perspectives on Psychological Science 9*(6), 666–681.

Stanley, T. D. (2005). Beyond publication bias. *Journal of Economic Surveys 19*(3), 309–345.

Steel, M. F. (2020). Model averaging and its use in economics. *Journal of Economic Literature 58*(3), 644–719.

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *Journal of the American Statistical Association 54*(285), 30–34.

Su, L. and S. Jin (2012). Sieve estimation of panel data models with cross section dependence. *Journal of Econometrics 169*(1), 34–47.

Viviano, D., K. Wuthrich, and P. Niehaus (2021). (When) should you adjust inferences for multiple hypothesis testing? *arXiv Preprint* (2104.13367).

Wang, J., Q. Zhao, T. Hastie, and A. B. Owen (2017). Confounder adjustment in multiple hypothesis testing. *Annals of Statistics 45*(5), 1863.

Welch, I. and A. Goyal (2008). A comprehensive look at the empirical performance of equity premium prediction. *Review of Financial Studies 21*(4), 1455–1508.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 817–838.

White, H. (1996). *Estimation, inference and specification analysis*. Number 22. Cambridge University Press.

White, H. (2001). *Asymptotic theory for econometricians*. Academic Press.

Wickham, H. (2014). Tidy data. *Journal of Statistical Software 59*(1), 1–23.

Wilson, D. J. (2019). The harmonic mean p-value for combining dependent tests. *Proceedings of the National Academy of Sciences 116*(4), 1195–1200.

Yan, X. S. and L. Zheng (2017). Fundamental analysis and the cross-section of stock returns: A data-mining approach. *Review of Financial Studies 30*(4), 1382–1423.

Zhang, X. (2015). Consistency of model averaging estimators. *Economics Letters 130*, 120–123.

Zhang, X. and C.-A. Liu (2019). Inference after model averaging in linear regression models. *Econometric Theory 35*(4), 816–841.

Zhu, R., X. Zhang, A. T. Wan, and G. Zou (2021). Kernel averaging estimators. *Journal of Business & Economic Statistics Forthcoming*, 1–28.