UNIVERSITY OF CALIFORNIA SAN DIEGO

Theoretical Contributions to Meta-Analysis for Research Transparency

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Economics

by

Nikolay Kudrin

Committee in charge:

Professor Graham Elliott, Chair
Professor Ery Arias-Castro
Professor Yixiao Sun
Professor Alexis Akira Toda
Professor Kaspar Wüthrich

2023

The Dissertation of Nikolay Kudrin is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2023

DEDICATION

To my mother, Valentina, and in loving memory of my father, Vladimir.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

This work and my entire academic path up to this point would have been impossible without my parents, their unconditional love and support. I am also grateful to my sister, Olga, who has always been there for me and helped proofreading my application to UCSD.

I would like to thank my advisor, Professor Graham Elliott, who was extremely supportive over these years. Every single meeting we had gave me a tremendous amount of positive motivation. I feel grateful and lucky for having Graham's guidance. I am indebted to Professor Kaspar Wüthrich for all the support, countless meetings, conversations and excellent feedback. Thank you to Professors Yixiao Sun, Alexis Akira Toda, Ery Arias-Castro and Xinwei Ma for great insights and suggestions.

At UCSD, I was lucky to meet new friends who made my time in San Diego even more enjoyable. In particular, I would like to thank Daniel, Evgenii, Linyan, Tjeerd, Wendy and Yu-Chang for sharing large parts of this journey with me.

Finally, I would like to thank Andrey Maksimov, my undergraduate advisor at HSE, who gave rise to my passion for econometrics. I am grateful to Stanislav Anatolyev for further inspiration at NES.

Chapter 1, in full, is a reprint of the material as it appears in Econometrica 2022. Elliott, Graham; Kudrin, Nikolay; Wüthrich, Kaspar. The dissertation author was a primary author of this material.

Chapter 2, in full, is currently being prepared for submission for publication of the material. It is joint work with Graham Elliott and Kaspar Wüthrich. The dissertation author is a primary author of this material.

Chapter 3, in part, is currently being prepared for submission for publication of the material. The dissertation author is the sole author of this material.

VITA

2013 Bachelor of Arts, Higher School of Economics, Nizhny Novgorod, Russia

2015 Master of Arts, New Economic School, Moscow, Russia

2023 Doctor of Philosophy, University of California San Diego

ABSTRACT OF THE DISSERTATION

Theoretical Contributions to Meta-Analysis for Research Transparency

by

Nikolay Kudrin

Doctor of Philosophy in Economics

University of California San Diego, 2023

Professor Graham Elliott, Chair

p-Hacking can undermine the validity of empirical research. The central focus of this dissertation is on analyzing existing and developing new statistical methods for detecting *p*-hacking based on the empirical distribution of reported results across studies.

In Chapter 1 we theoretically analyze the problem of testing for *p*-hacking based on distributions of *p*-values across multiple studies. We provide general results for when such distributions have testable restrictions (are non-increasing) under the null of no *p*-hacking. We find novel additional testable shape restrictions for *p*-values based on *t*-tests. These testable restrictions result in more powerful tests for the null hypothesis of no *p*-hacking. When there is also publication bias, our tests are joint tests for *p*-hacking and publication bias. A reanalysis of

two prominent datasets shows the usefulness of our new tests.

Chapter 2 provides a careful understanding of the power of methods used to detect different types of p-hacking discussed in Chapter 1. We theoretically study the implications of likely forms of p-hacking on the distribution of reported p-values and the power of existing methods for detecting it. Power can be quite low, depending crucially on the particular p-hacking strategy and the distribution of actual effects tested by the studies. We relate the power of the tests to the costs of p-hacking and show that power tends to be larger when p-hacking is very costly.

Chapter 3 studies Caliper tests that are widely used to test for the presence of p-hacking and publication bias based on the distribution of the z-statistics across studies. We show that without additional restrictions on the distribution of true effects, Caliper tests may suffer from substantial size distortions. We propose a modification of the existing Caliper test, referred to as the Robust Caliper test, which is shown to control size irrespective of the true effect distribution. We also propose a way of correcting the regression-based version of the Caliper test that allows for the inclusion of additional covariates.

Chapter 1

Detecting *p*-hacking

Abstract

We theoretically analyze the problem of testing for *p*-hacking based on distributions of *p*-values across multiple studies. We provide general results for when such distributions have testable restrictions (are non-increasing) under the null of no *p*-hacking. We find novel additional testable restrictions for *p*-values based on *t*-tests. Specifically, the shape of the power functions results in both complete monotonicity as well as bounds on the distribution of *p*-values. These testable restrictions result in more powerful tests for the null hypothesis of no *p*-hacking. When there is also publication bias, our tests are joint tests for *p*-hacking and publication bias. A reanalysis of two prominent datasets shows the usefulness of our new tests.

1.1 Introduction

A researcher's ability to explore various ways of analyzing and manipulating data and then selectively report the ones that yield better-looking results, commonly referred to as *p-hacking*, compromises the reliability of research and undermines the scientific credibility of reported results. Absent systematic replication studies or meta analyses, a popular approach for assessing the extent of *p*-hacking is to examine distributions of *p*-values across studies, referred to as *p-curves* (Simonsohn et al., 2014); see Section 2 in Christensen and Miguel (2018) for a

review.[1]

We consider the problem of testing the *null hypothesis of no p-hacking* against the *alternative hypothesis of p-hacking* and provide theoretical foundations for developing tests for p-hacking. We characterize analytically under general assumptions the null set of distributions of p-values implied in the absence of p-hacking and provide general sufficient conditions under which, for any distribution of the true effects, the p-curve is non-increasing and continuous in the absence of p-hacking. These conditions are shown to hold for many, but not all popular approaches to testing for effects.

For the leading case where p-curves are based on t-tests, we derive additional previously unknown testable restrictions. Specifically, the p-curves based on t-tests are completely monotone in the absence of p-hacking, and their magnitude and the magnitude of their derivatives are restricted by upper bounds. These restrictions are particularly useful when p-hacking fails to induce an increasing p-curve—for example when researchers engage in specification search across independent tests. In such cases tests based on non-increasingness have no power.

Our theoretical results allow us to develop more powerful statistical tests for p-hacking, which we apply to two large datasets of p-values. We find evidence for p-hacking in settings where the existing tests do not reject the null of no p-hacking.

When there is publication bias, our results characterize the p-curve under the null hypothesis of *no p-hacking and no publication bias*. Our tests become joint tests for p-hacking and publication bias, complementing available methods for identifying publication bias (see, e.g., Andrews and Kasy, 2019, and the references therein).

[1]Examples include: Masicampo and Lalande (2012), Leggett et al. (2013), Simonsohn et al. (2014, 2015), Head et al. (2015), de Winter and Dodou (2015), and Snyder and Zhuo (2018). Another strand of the literature uses the distribution of t-statistics to test for p-hacking (e.g., Gerber and Malhotra, 2008a; Brodeur et al., 2016b, 2020a; Bruns et al., 2019; Vivalt, 2019).

1.2 The p-curve based on general tests

Here we provide general sufficient conditions under which the p-curve is non-increasing under the null hypothesis of no p-hacking. These results are useful because tests for p-hacking often assume non-increasingness of the p-curve (e.g., Simonsohn et al., 2014, 2015; Head et al., 2015). This assumption has been justified through analytical and numerical examples, which rely on specific choices of tests and distributions of true effects being tested (e.g., Hung et al., 1997; Simonsohn et al., 2014; Ulrich and Miller, 2018). However, such analyses are not sufficient for guaranteeing size control of statistical tests for p-hacking since the true effect distribution is never known. Instead, what is required for size control in a wide range of applications is a characterization of the shape of the p-curve for general tests and effect distributions.

1.2.1 Setup

Consider a test statistic T that is distributed according to a distribution with cumulative distribution function (CDF) F_h, where h indexes parameters of either the exact or asymptotic distribution of the test. We assume that the parameters h only contain the parameters of interest. This is suitable for settings with large enough samples and asymptotically pivotal test statistics, which are prevalent in applied research.

Suppose researchers are testing the hypothesis

$$H_0 : h \in \mathscr{H}_0 \qquad \text{against} \qquad H_1 : h \in \mathscr{H}_1, \tag{1.1}$$

where $\mathscr{H}_0 \cap \mathscr{H}_1 = \emptyset$. Let $\mathscr{H} = \mathscr{H}_0 \cup \mathscr{H}_1$. Denote as F the CDF of the chosen null distribution from which critical values are determined. We assume that the test rejects for large values of the test statistic and denote the critical value for a level p test as $cv(p)$. We will focus on settings with a continuous and strictly increasing F (see Assumption 1.1 below) and set $cv(p) = F^{-1}(1-p)$. For any h, we denote by $\beta(p,h) = \Pr(T > cv(p) \mid h) = 1 - F_h(cv(p))$ the rejection rate of a

level p test with parameters h. For $h \in \mathscr{H}_1$, this is the power of the test, and we refer to $\beta(p,h)$ as the *power function*.

For the remainder of the paper, we focus on settings where the tests generating the p-values satisfy Assumption 1.1. This allows us to work with a well-defined density function and provide general results.

Assumption 1.1 (Regularity)**.** *F and F_h are twice continuously differentiable with uniformly bounded first and second derivatives f, f', f_h and f_h'. $f(x) > 0$ for all $x \in \{cv(p) : p \in (0,1)\}$. For $h \in \mathscr{H}$, $supp(f) = supp(f_h)$.*[2]

Assumption 1.1 holds for many tests with parametric F and F_h, including t-tests and Wald-tests. A necessary condition for Assumption 1.1 is the absolute continuity of F and F_h. This is not too restrictive since, in many cases, F and F_h are the asymptotic distributions of test statistics, which typically satisfy this condition. Further, in cases where the test statistics have a discrete distribution, size does not typically equal level, which could lead to p-curves that violate non-increasingness.

Consider the distribution of the p-values across studies, where we compute p-values from a distribution of T given values of h, which themselves are drawn from a probability distribution Π. We refer to Π as the *distribution of true effects*. The CDF of the p-values is

$$G(p) = \int_{\mathscr{H}} \Pr(T > cv(p) \mid h)\, d\Pi(h) = \int_{\mathscr{H}} \beta(p,h)\, d\Pi(h). \tag{1.2}$$

Under Assumption 1.1, define the p-curve as follows.

Definition 1.1 (P-curve)**.** *The density of the p-values, the p-curve, is defined as*

$$g(p) := \int_{\mathscr{H}} \frac{\partial \beta(p,h)}{\partial p} d\Pi(h).$$

[2]For a function φ, we define $\text{supp}(\varphi)$ to be the closure of $\{x : \varphi(x) \neq 0\}$. Boundedness on $\{cv(p) : p \in (0, 1-\varepsilon]\}$ for any $\varepsilon \in (0,1)$ is sufficient for our results.

In Section 1.2.2, we analyze the shape of g for general tests and distributions Π.

1.2.2 Properties of p-curves based on general tests

Here we derive conditions under which the p-curve is non-increasing in the absence of p-hacking for any distribution of true effects. We show that this property holds for most but not all popular statistical tests.

Under Assumption 1.1, the curvature of the p-curve follows from

$$g'(p) := \frac{dg(p)}{dp} = \int_{\mathscr{H}} \frac{\partial^2 \beta(p,h)}{\partial p^2} d\Pi(h).$$

The sign of $g'(p)$ is determined by the second derivative of the rejection probability, $\partial^2\beta(p,h)/\partial p^2$. As we will show in the proof of Theorem 1.1 below, the following condition implies that $\partial^2\beta(p,h)/\partial p^2$ is non-positive for all $h \in \mathscr{H}$.

Assumption 1.2 (Sufficient condition)**.** *For all* $(x,h) \in \{cv(p) : p \in (0,1)\} \times \mathscr{H}$,

$$f_h'(x) f(x) \geq f'(x) f_h(x).$$

Assumption 1.2 is a restriction on how the power function changes when the critical value changes, which is governed by the shape of the density. When $\mathscr{H}_0 = \{0\}$ and $F = F_0$ (as, for example, for one-sided t-tests), Assumption 1.2 is of the form of a monotone likelihood ratio property, which relates the shape of the density of T under the null to the shape of the density of T under alternative h. The next lemma shows that this condition holds for many popular tests. Let Φ denote the CDF of the standard normal distribution.

Lemma 1.1. *Assumption 1.2 holds when*

(i) $F(x) = \Phi(x)$, $F_h(x) = \Phi(x-h)$, $\mathscr{H}_0 = \{0\}$, $\mathscr{H}_1 \subseteq (0,\infty)$ *(e.g., similar one-sided t-test)*

(ii) F *is the CDF of a half-normal distribution with scale parameter* 1, F_h *is the CDF of a*

folded normal distribution with location parameter h and scale parameter 1, $\mathscr{H}_0 = \{0\}$, $\mathscr{H}_1 \subseteq \mathbb{R}\backslash\{0\}$ *(e.g., two-sided t-test)*

(iii) F *is the CDF of a* χ^2 *distribution with degrees of freedom* $d > 0$*,* F_h *is the CDF of a noncentral* χ^2 *distribution with degrees of freedom* $d > 0$ *and noncentrality parameter h,* $\mathscr{H}_0 = \{0\}$*,* $\mathscr{H}_1 \subseteq (0,\infty)$ *(e.g., Wald test*[3]*)*

The following theorem shows that the p-curve is non-increasing and continuously differentiable under the maintained assumptions for any distribution of true effects.

Theorem 1.1 (Testable restrictions for general tests)**.** *Under Assumptions 1.1–1.2, g is continuously differentiable and* $g'(p) \leq 0$ *for* $p \in (0,1)$.

The result in Theorem 1.1 holds for many commonly-used statistical tests such that, in many empirically relevant settings, the p-curve will be non-increasing in the absence of p-hacking. To our knowledge, Theorem 1.1 provides the first general formal justification for the existing tests for p-hacking that exploit non-increasingness of the p-curve. Theorem 1.1 further motivates the use of density discontinuity tests as an alternative to tests based on non-increasingness of the p-curve.

The results can be extended to settings with nuisance parameters. In such settings, h contains both the parameters of interest, h_1, as well as additional nuisance parameters, h_2, such that $h = (h_1, h_2)$. Let $\mathscr{H}^1$ and $\mathscr{H}^2$ denote the supports of h_1 and h_2. Allow the null distribution to depend on h_2 with CDF F_{h_2}. The CDF of p-values becomes

$$G(p) = \int_{\mathscr{H}^1 \times \mathscr{H}^2} \beta\left(p, h_1, h_2\right) d\Pi(h_1, h_2),$$

where $\beta(p, h_1, h_2) = 1 - F_h\left(cv_{h_2}(p)\right)$ and $cv_{h_2}(p) = F_{h_2}^{-1}(1-p)$. The results of Theorem 1.1 extend to the p-curve generated from this distribution after changing the notation to include

[3]For instance, let $\sqrt{N}(\hat{\theta} - \theta) \overset{a}{\sim} \mathscr{N}(0, V)$, where $\hat{\theta}$ is an estimator of θ based on N observations and $V \in \mathbb{R}^{\dim(\theta)\times\dim(\theta)}$ is known (or can be consistently estimated). Consider the problem of testing $H_0 : R\theta = r$ against $H_1 : R\theta \neq r$, where $R \in \mathbb{R}^{q\times\dim(\theta)}$, $r \in \mathbb{R}^q$, and $rank(R) = q$. Set $T = N(R\hat{\theta} - r)'(RVR')^{-1}(R\hat{\theta} - r)$. This fits our framework with $d = q$ and $h := \lambda'(RVR')^{-1}\lambda$, where $\lambda := \sqrt{N}(R\theta - r)$.

the dependence on h_2. For $h_2 \in \mathscr{H}^2$, F_{h_2}, f_{h_2}, f'_{h_2} have the same properties as F, f, f' in Assumption 1.1, and the assumptions on F_h, f_h, f'_h hold for $h = (h_1, h_2)$. Assumption 1.2 becomes $f'_h(cv_{h_2}(p))f_{h_2}(cv_{h_2}(p)) \geq f'_{h_2}(cv_{h_2}(p))f_h(cv_{h_2}(p))$ for $(h_1, h_2) \in \mathscr{H}^1 \times \mathscr{H}^2$. The proof then follows directly from that of Theorem 1.1.

In applications, often only a part of the p-curve is examined. The p-curve over subintervals $\mathscr{I} \subset (0,1)$ is given by $g_{\mathscr{I}}(p) = g(p)/\int_{\mathscr{I}} g(p)dp$ for $p \in \mathscr{I}$. Therefore, the results extend directly to this situation. Moreover, the p-curve constructed from a finite aggregation of different tests satisfying the assumptions of Theorem 1.1 is continuously differentiable and non-increasing.

The assumptions of Theorem 1.1 directly suggest p-curves for which the results of Theorem 1.1 fail. For example, when the tests are non-similar, the p-curve can be non-monotonic in the absence of p-hacking, which arises through a violation of Assumption 1.2. To illustrate, consider testing $H_0 : h \leq 0$ against $H_1 : h > 0$ using a (non-similar) one-sided t-test, where f is the density of the $\mathscr{N}(0,1)$ distribution and f_h is the density of the $\mathscr{N}(h,1)$ distribution. It follows that $f'(x)/f(x) = -x$ and $f'_h(x)/f_h(x) = -(x-h)$, such that Assumption 1.2 holds when $h \geq 0$ but is violated when $h < 0$. Thus, when the weight in Π on $h < 0$ is large enough, the p-curve can be non-monotonic or increasing. For example, suppose that Π is a normal distribution with mean μ and variance 1, which places some mass on $h < 0$, mixing increasing and decreasing p-curves. Figure 1.1 shows that the resulting p-curve is non-increasing when $\mu = 0$ and non-monotonic when $\mu = -2.5$.

1.3 The p-curve based on t-tests

We now show that for the leading case where p-curves are generated from t-tests with exact or asymptotic normal distributions, there are additional previously unknown testable restrictions. These restrictions allow us to develop more powerful statistical tests for p-hacking (see Section 1.4.3). In particular, these tests have power in situations where p-hacking does not

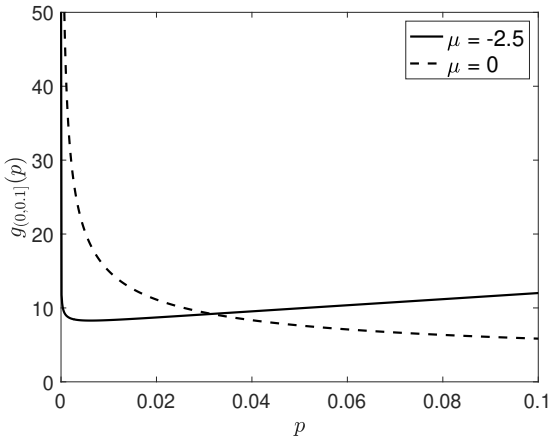

Figure 1.1. *P*-curves based on non-similar one-sided t-tests on $(0, 0.1]$. The distribution of true effects Π is a normal distribution with mean μ and variance 1.

lead to a violation of non-increasingness.

Consider first the problem of testing a one-sided hypothesis

$$H_0 : h = 0 \qquad \text{against} \qquad H_1 : h > 0, \tag{1.3}$$

where h is a scalar, $\mathscr{H}_0 = \{0\}$, and $\mathscr{H}_1 = (0, \infty)$. We assume that $T \sim \mathscr{N}(h, 1)$. This holds when using one-sided t-tests to test a hypothesis concerning a scalar parameter θ: $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$. Let $\sqrt{N}\left(\hat{\theta} - \theta\right) \sim \mathscr{N}\left(0, \sigma^2\right)$, where $\hat{\theta}$ is an estimator of θ based on N observations and σ^2 is assumed to be known. Denote the usual t-statistic as $\hat{t}$ and set $T = \hat{t}$. Defining $h := \sqrt{N}((\theta - \theta_0)/\sigma)$ this fits (1.3). More generally, testing problems with limiting normal experiments employed to test hypotheses of the form (1.3) are common in empirical work (e.g., a one-sided test of a regression parameter using normal critical values).

The chosen null distribution is the standard normal distribution, $F = \Phi$. A level p test rejects the null hypothesis when T is larger than $cv_1(p) := \Phi^{-1}(1 - p)$. Note that $cv_1(p) \geq 0$ for $p \in (0, 1/2]$. Then $\beta(p, h) = 1 - \Phi(cv_1(p) - h)$ and the CDF of p-values is

$$G_1(p) = 1 - \int_{[0,\infty)} \Phi(cv_1(p) - h)\, d\Pi(h). \tag{1.4}$$

We also consider the two-sided version of this test. Here the hypothesis is

$$H_0 : h = 0 \qquad \text{against} \qquad H_1 : h \neq 0 \tag{1.5}$$

with $\mathscr{H}_0 = \{0\}$ and $\mathscr{H}_1 = \mathbb{R}\backslash\{0\}$. The two-sided test statistic T is assumed to have a folded normal distribution. This holds when using a two-sided t-test with $T = |\hat{t}|$ for testing a two-sided hypothesis about $\theta : H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$. More generally, testing problems with limiting normal experiments employed to test hypotheses of the form (1.5) are also common in empirical work.

The chosen null distribution is the half normal distribution with scale parameter 1. A level p test rejects the null hypothesis when T is larger than $cv_2(p) := \Phi^{-1}\left(1 - \frac{p}{2}\right)$. The CDF of the p-values is

$$G_2(p) = 2 - \int_{\mathbb{R}} \left[\Phi\left(cv_2(p) - h\right) + \Phi\left(cv_2(p) + h\right)\right] d\Pi(h). \tag{1.6}$$

In addition to the results of Section 1.2.2, previously unknown testable restrictions for p-curves based on t-tests follow from the shape of the power functions for these tests. These additional restrictions enable us to better pin down the space of potential p-curves when there is no p-hacking, allowing us to construct more powerful statistical tests for p-hacking. They also enable distinguishing non-increasing p-curves, which can arise from certain types of p-hacking, from curves where there is no p-hacking.

The p-curve based on one-sided t-tests testing hypothesis (1.4) is

$$g_1(p) = \int_{[0,\infty)} \exp\left(hcv_1(p) - \frac{h^2}{2}\right) d\Pi(h). \tag{1.7}$$

For two-sided t-tests testing hypothesis (1.6), the p-curve is

$$g_2(p) = \int_{\mathbb{R}} \frac{1}{2}\left[\exp\left(hcv_2(p) - \frac{h^2}{2}\right) + \exp\left(-hcv_2(p) - \frac{h^2}{2}\right)\right] d\Pi(h). \tag{1.8}$$

Our next theorem shows that the p-curves (1.7) and (1.8) are completely monotone. A function ξ is completely monotone on an interval $\mathscr{I}$ if $0 \leq (-1)^k \xi^{(k)}(x)$ for every $x \in \mathscr{I}$ and all $k = 0, 1, 2, \ldots$, where $\xi^{(k)}$ is the kth derivative of ξ.

Theorem 1.2 (Complete monotonicity)**.** *(i) The p-curve g_1 is completely monotone on $(0, 1/2]$. (ii) The p-curve g_2 is completely monotone on $(0, 1)$.*

Complete monotonicity yields additional restrictions that can be exploited to improve the power of statistical tests for p-hacking. Whilst available for one- and two-sided t-tests, not all tests yield completely monotonic p-curves. For example, a direct calculation shows that complete monotonicity may fail for tests based on χ^2 distributions with more than two degrees of freedom (e.g., Wald tests).

The next theorem presents additional testable restrictions in the form of upper bounds on the p-curves and their derivatives.

Theorem 1.3 (Upper bounds)**.**

(i) The p-curves g_1 and g_2 are bounded from above:

$$\begin{aligned} g_1(p) &\leq 1_{\{p \leq 1/2\}} \exp\left(\frac{cv_1(p)^2}{2}\right) + 1_{\{p > 1/2\}} =: \mathscr{B}_1^{(0)}(p), & (1.9) \\ g_2(p) &\leq 1_{\{p < 2(1-\Phi(1))\}} \tilde{\mathscr{B}}_2^{(0)} + 1_{\{p \geq 2(1-\Phi(1))\}} =: \mathscr{B}_2^{(0)}(p), & (1.10) \end{aligned}$$

where

$$\begin{aligned} \tilde{\mathscr{B}}_2^{(0)}(p) &:= \frac{1}{2}\left[\exp\left(h^*(p)cv_2(p) - \frac{h^*(p)^2}{2}\right) + \exp\left(-h^*(p)cv_2(p) - \frac{h^*(p)^2}{2}\right)\right] \\ &\leq \exp\left(\frac{cv_2(p)^2}{2}\right), \end{aligned}$$

and $h^*(p)$ is the non-zero solution to

$$\varphi(cv_2(p),h) := (cv_2(p)-h)\exp(cv_2(p)h) - (cv_2(p)+h)\exp(-cv_2(p)h) = 0.$$

(ii) The derivatives of g_1 and g_2 are bounded from above. For $s = 1,2$ and $k = 1,2,3,\ldots$, then $(-1)^k g_s^{(k)}(p) \leq \mathscr{B}_s^{(k)}(p)$, where $\mathscr{B}_s^{(k)}$ is defined in Appendix A.2.3.

As with the results in Theorem 1.2, the results in Theorem 1.3 yield additional restrictions, allowing more powerful tests for p-hacking.[4] The bounds in Theorem 1.3 do not only rule out large humps around significance cutoffs such as 0.01, 0.05, and 0.1 but also restrict the magnitude of the p-curves near zero. For the two-sided test, tests for p-hacking can be either constructed using the sharper (but not explicit) bound $\tilde{\mathscr{B}}_2^{(0)}(p)$ or the simpler explicit bound $\exp\left(\frac{cv_2(p)^2}{2}\right)$.

The bounds of Theorem 1.3 are particularly useful when p-hacking fails to induce an increasing p-curve, a situation where tests based on non-increasingness of the p-curve have no power. Intuitively we might suspect this happens when all researchers p-hack but this simply shifts mass of the p-curve to the left, rather than inducing humps. A concrete example is when researchers run a finite number of $M > 1$ independent analyses and report the smallest p-value, for example, when engaging in specification search across independent subsamples or data sets. The resulting p-curve under p-hacking is $g^p(p;M) = M(1-G^{np}(p))^{M-1}g^{np}(p)$, where G^{np} and g^{np} are the CDF and density of p-values in the absence of p-hacking.[5] Note that g^p is non-increasing (completely monotone) whenever g^{np} is non-increasing (completely monotone).[6] Thus, g^p will not violate the testable implications of Theorems 1.1–1.2, so tests based on these restrictions do not have power. However, g^p can violate the bounds in Theorem 1.3 whenever $M(1-G^{np}(p))^{M-1} > 1$. For example, consider the one-sided case and let Π be a half-normal

[4] One can use similar arguments as in Theorem 1.3 to derive bounds for p-curves based on other specific tests such as Wald tests.

[5] This generalizes the example in Ulrich and Miller (2015), who studied the special case where all null hypotheses are true such that $G(p) = p$.

[6] Since the products of completely monotone functions are completely monotone, complete monotonicity of $g^p(p;M)$ follows from complete monotonicity of $1-G^{np}(p)$ and $g^{np}(p)$.

distribution with scale parameter 1. Figure 1.2 shows that g^p violates the upper bound in Theorem 1.3 to an extent that depends on M.

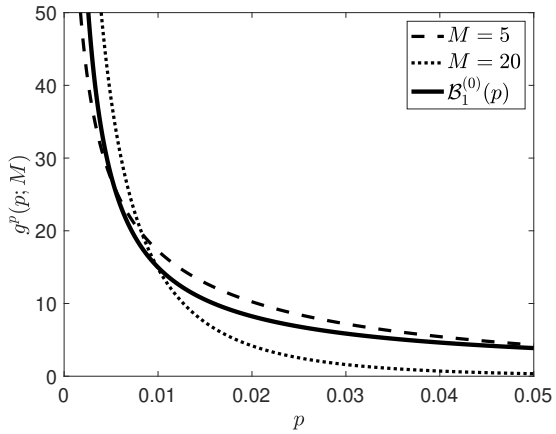

Figure 1.2. Comparison of the p-curve from specification search based on one-sided t-tests and the upper bound in Equation (1.9).

Upper bounds also help with testing for p-hacking with non-similar tests. In Section 1.2.2, we show that non-increasingness may fail for non-similar one-sided t-tests, in which case tests of p-hacking based on non-increasingness may well reject because of non-similarity rather than p-hacking. Since upper bounds can also be derived for non-similar tests, we can still use bounds on the p-curve and its derivatives to test for p-hacking.[7]

Finally, the characterizations in Theorems 1.2–1.3 imply related characterizations of p-curves over subintervals $\mathscr{I} \subset (0,1)$, $g_{s,\mathscr{I}}(p) = g_s(p)/\int_{\mathscr{I}} g_s(p)dp$. In particular, complete monotonicity of g_s implies the complete monotonicity of $g_{s,\mathscr{I}}$, because the sign of $g^{(k)}_{s,\mathscr{I}}$ equals the sign of $g^{(k)}_s$ for $k = 0,1,2\ldots$. Moreover, (conservative) upper bounds on $g_{s,\mathscr{I}}(p)$ for $\mathscr{I} = (0,\alpha]$ are given by the upper bounds in Theorem 1.3, re-scaled by α since $G_s(\alpha) \geq \alpha$ for $s = 1,2$.

1.4 Statistical tests for p-hacking

Here we consider tests for p-hacking based on a sample of n p-values. We consider three types of tests that differ with respect to the specification of the null hypothesis (the null space of

[7]For instance, for $p \leq 1/2$, the upper bound on the p-curve for non-similar one-sided t-tests coincides with that in Part (i) of Theorem 1.3.

p-curves). As a result, the different tests will differ with respect to the violations of the null of no p-hacking that they are able to detect.

In the absence of publication bias, our tests are tests for p-hacking; when there is also publication bias, they are joint tests for p-hacking and publication bias in general.

1.4.1 Tests for non-increasingness of the p-curve

Theorem 1.1 shows that, under general conditions, the p-curve is non-increasing. Consider the following testing problem

$$H_0 : g \text{ is non-increasing} \qquad \text{against} \qquad H_1 : g \text{ is not non-increasing.} \tag{1.11}$$

Popular tests based on hypothesis testing problem (1.11) include the Binomial test (e.g., Simonsohn et al., 2014; Head et al., 2015) and Fisher's test (Simonsohn et al., 2014). Here we describe two alternative and more powerful tests.

Histogram-based tests. Let $0 = x_0 < x_1 < \cdots < x_J = 1$ be an equidistant partition of the unit interval. Define the population proportions as $\pi_j := \int_{x_{j-1}}^{x_j} g(p)dp,\ j = 1,\ldots,J$. When g is non-increasing, $\Delta_j := \pi_{j+1} - \pi_j$ is non-positive for all $j = 1,\ldots,J-1$. Thus, the null hypothesis in testing problem (1.11) can be reformulated as $H_0 : \Delta_j \leq 0$ for all $j = 1,\ldots,J-1$. To test this hypothesis, we apply the conditional chi-squared test of Cox and Shi (2022). We describe the implementation of this test in Section 1.4.3 and Appendix A.1, where we propose more general tests that nest the histogram-based test for non-increasingness.

LCM test based on concavity of the CDF of p-values. Under the null hypothesis (1.11), the CDF of p-values is concave. This observation allows us to apply tests based on the least concave majorant (LCM) (e.g., Carolan and Tebbs, 2005; Beare and Moon, 2015; Fang, 2019). LCM-based tests assess concavity of the CDF based on the distance between the empirical CDF of p-values, $\hat{G}$, and its LCM, $\mathscr{M}\hat{G}$, where $\mathscr{M}$ is the LCM operator.[8] We consider the test statistic

[8]For a function f, the LCM operator is defined as $\mathscr{M}f = \inf\{g : g \text{ is concave and } f \leq g\}$ (e.g., Beare and Moon,

$T = \sqrt{n}\|\mathscr{M}\hat{G} - \hat{G}\|_\infty$. The uniform distribution is least favorable for LCM tests (e.g., Kulikov and Lopuhaä, 2008; Beare, 2021), in which case T converges weakly to $\|\mathscr{M}B - B\|_\infty$, where B is a standard Brownian Bridge on $[0,1]$.

1.4.2 Tests for continuity

Theorem 1.1 shows that the p-curve is continuous in the absence of p-hacking. Tests for continuity of the p-curve at significance thresholds α such as $\alpha = 0.05$, thus, provide an alternative to the tests based on non-increasingness of the p-curve. Consider the following testing problem:

$$H_0 : \lim_{p\uparrow\alpha} g(p) = \lim_{p\downarrow\alpha} g(p) \qquad \text{against} \qquad H_1 : \lim_{p\uparrow\alpha} g(p) \neq \lim_{p\downarrow\alpha} g(p) \tag{1.12}$$

Testing (1.12) requires estimating two densities at the boundary point α. Traditional kernel density estimators are not suitable for this task because they suffer from boundary bias (e.g., Karunamuni and Alberts, 2005). A popular approach to overcome this problem is to use local linear density estimators that rely on prebinning the data (e.g., McCrary, 2008). We apply the density discontinuity test of Cattaneo et al. (2020) with data-driven bandwidth selection (**?**), which is based on boundary adaptive local polynomial density estimators and avoids prebinning.

1.4.3 Tests for K-monotonicity and upper bounds

Theorem 1.2 shows that p-curves based on t-tests are completely monotone, and Theorem 1.3 establishes upper bounds on the p-curves and their derivatives. Here we develop tests based on these testable restrictions.

We say a function ξ is K-monotone on some interval $\mathscr{I}$ if $0 \leq (-1)^k \xi^{(k)}(x)$ for every $x \in \mathscr{I}$ and all $k = 0, 1, \ldots, K$, where $\xi^{(k)}$ is the kth derivative of ξ. By definition, a completely monotone function is K-monotone. Consider the null hypothesis

$$H_0 : g_s \text{ is } K\text{-monotone and } (-1)^k g_s^{(k)} \leq \mathscr{B}_s^{(k)}, \text{ for } k = 0, 1, \ldots, K, \tag{1.13}$$

2015, Definition 2.1).

where $s = 1$ for one-sided t-tests, $s = 2$ for two-sided t-tests, and $\mathscr{B}_s^{(k)}$ is defined in Theorem 1.3. Hypothesis (1.13) implies restrictions on the population proportions $\pi := (\pi_1, \ldots, \pi_J)'$, which can be expressed as $H_0 : A\pi_{-J} \leq b$, where $\pi_{-J} := (\pi_1, \ldots, \pi_{J-1})'$.[9] The matrix A and vector b are defined in Appendix A.1.2.[10]

We estimate π_{-J} using the sample proportions $\hat{\pi}_{-J}$.[11] This estimator is $\sqrt{n}$-consistent and asymptotically normal with mean π_{-J} and non-singular (if all proportions are positive) covariance matrix $\Omega = \text{diag}\{\pi_1, \ldots, \pi_{J-1}\} - \pi_{-J}\pi_{-J}'$. Following Cox and Shi (2022), we test the null by comparing $T = \inf_{q: Aq \leq b} n(\hat{\pi}_{-J} - q)'\hat{\Omega}^{-1}(\hat{\pi}_{-J} - q)$ to the critical value from a χ^2 distribution with $rank(\hat{A})$ degrees of freedom, where $\hat{A}$ is the matrix formed by the rows of A corresponding to active inequalities.

1.5 Empirical applications

The analyses were done using R (R Core Team, 2020) and Stata (StataCorp., 2019).

1.5.1 P-hacking in economics journals

Here we reanalyze the data collected by Brodeur et al. (2016b), which contain information about 50,078 t-tests from 641 papers published in the AER, QJE, and JPE 2005–2011 (Brodeur et al., 2016a). We convert t-statistics into p-values associated with two-sided t-tests based on the standard normal distribution.[12] After excluding observations with missing information, there are 49,838 tests from 640 papers.

Because the p-values may be correlated within papers, we use cluster-robust estimators of the variance of the sample proportions for the Cox and Shi (2022) tests. In addition, we apply

[9] The upper bounds on π implied by hypothesis (1.13) are not sharp in general. Sharp bounds can be obtained by directly extremizing the proportions and their differences; see Appendix A.1.1.

[10] We use π_{-J} because the variance matrix of the estimator of π is singular by construction and we want to express the left-hand side of our moment inequalities as a combination of "core" moments.

[11] Given a sample of n p-values, $\{P_i\}_{i=1}^n$, the sample proportions are defined as $\hat{\pi}_i = \frac{1}{n}\sum_{i=1}^n 1\{x_{i-1} < P_i \leq x_i\}$, $i = 1, \ldots, J$.

[12] The original data contain p-values for less than 10% of observations. Where available, we work with the reported p-values.

all tests to random subsamples with one p-value per paper, allowing us to use exact tests in the presence of within-paper correlation. To test for p-hacking, we focus on p-values smaller than 0.15. We consider a Binomial test on $[0.04, 0.05]$, Fisher's test, a histogram-based test for non-increasingness (CS1), a histogram-based test for 2-monotonicity and bounds on the p-curve and the first two derivatives (CS2B), the LCM test, and a density discontinuity test at 0.05.[13]

Figure 1.3 shows the results before and after de-rounding and based on the full sample and random subsamples. There is a large number of very small p-values, which is sometimes interpreted as indicative of evidential value (e.g., Simonsohn et al. (2014); in our notation, this is a large mass of Π away from zero). The data exhibit a noticeable mass point at $\hat{t} = 2$ (there are 427 such observations), which translates into a mass point in the p-curve at $p = 0.046$.[14] To analyze the impact of rounding, we also apply the tests to the de-rounded data provided by Brodeur et al. (2016b).[15]

In what follows, we say that a test rejects the null of no p-hacking if its p-value is smaller than 0.1. Based on the original raw (rounded) data on all p-values, all tests reject the null except Fisher's test and the density discontinuity test. There are no rejections based on the random subsample, suggesting that the tests may be underpowered in small samples.

We find different results based on the de-rounded data.[16] There are no rejections based on the full sample of p-values. This finding suggests that the rejections based on the raw data are mainly due to the mass point just below 0.05 and shows that de-rounding may substantially affect empirical conclusions.

Based on the random subsample of de-rounded p-values, only the CS2B test rejects the null of no p-hacking. The CS1 test comes close to rejecting ($p = 0.11$). These two tests yield

[13] For the Binomial test, we split $[0.04, 0.05]$ into two subintervals $[0.04, 0.045]$ and $(0.045, 0.05]$. Under the null of no p-hacking, the fraction of p-values in $(0.045, 0.05]$ should be smaller than or equal to 0.5, which we assess using an exact Binomial test. For CS1 and CS2B, we use 30 bins when testing based on all p-values and 15 bins when testing based on random subsamples of p-values.

[14] This mass point could be due to low precision reporting (Brodeur et al., 2016b), but also due to p-hacking, publication bias, or a combination thereof.

[15] The de-rounded data were constructed by randomly redrawing estimates and standard errors; see Section II in Brodeur et al. (2016b) for a detailed description.

[16] Note that the (sub)sample sizes for the rounded and de-rounded data differ due to de-rounding.

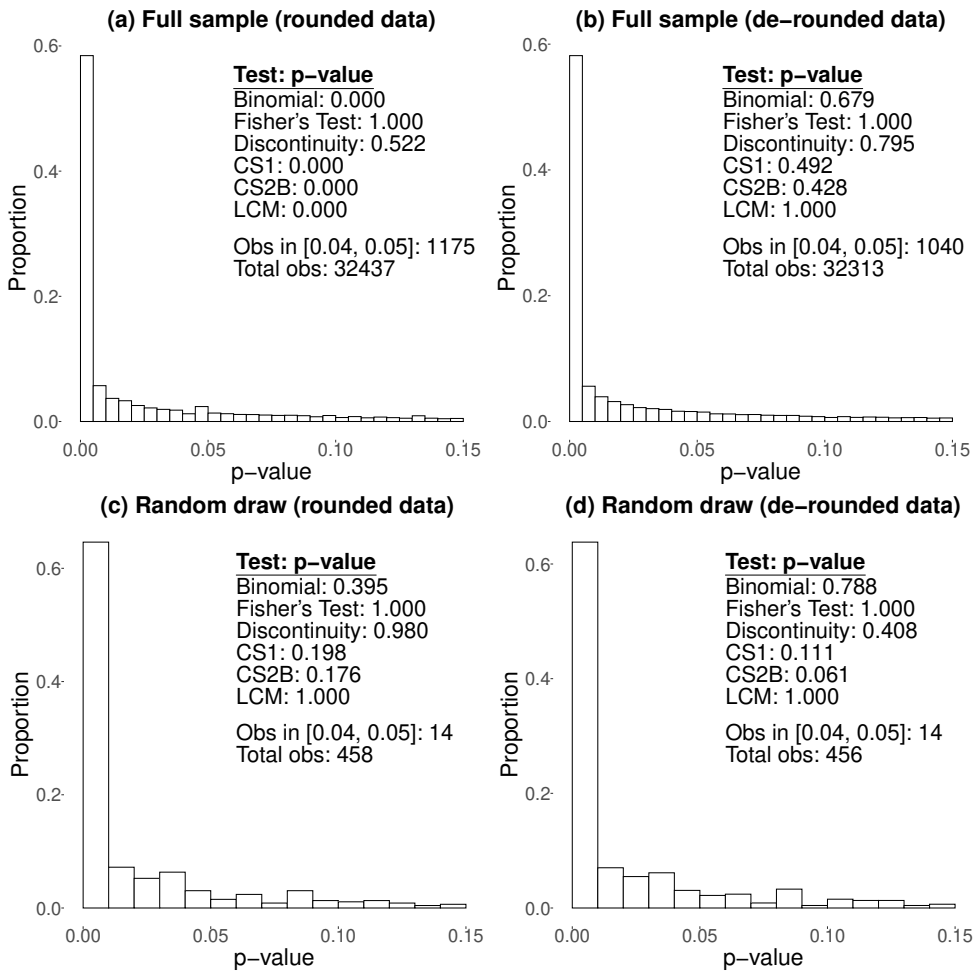

Figure 1.3. *P*-curves and *p*-values from testing for *p*-hacking. The tests for *p*-hacking are described in Section 2.4. Data: Brodeur et al. (2016a).

the smallest *p*-values across all four samples.

1.5.2 P-hacking across different disciplines

Here we reanalyze the data collected by Head et al. (2015), which contain *p*-values obtained from text-mining open access papers in the PubMed database (Head et al., 2016). There are *p*-values from 21 different disciplines. We focus on biology, chemistry, education, engineering, medical and health sciences, and psychology and cognitive science. The data contain *p*-values from the abstracts and the results sections in the main text. We use *p*-values from the results sections, allowing us to work with larger samples and present results for *p*-values

smaller than 0.15.

Since the data do not only contain t-tests, we consider tests based on non-increasingness and continuity of the p-curve (Theorem 1.1): a Binomial test on $[0.04, 0.05]$, Fisher's test, a histogram-based test for non-increasingness (CS1), the LCM test, and a density discontinuity test at 0.05.[17] To account for within-paper dependence of p-values, we use a cluster-robust variance estimator for the CS1 test, and also present results based on random subsamples with one p-value per paper.

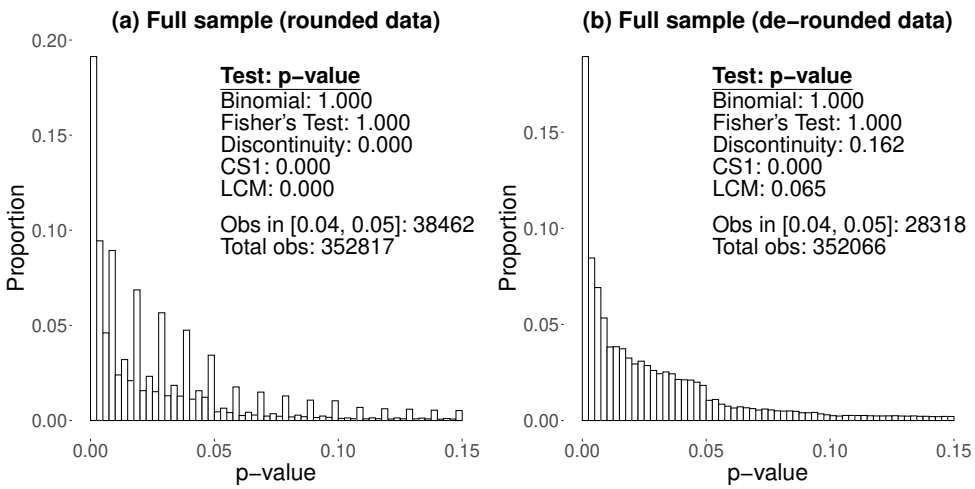

Figure 1.4. *P*-curves and p-values from testing for p-hacking for medical and health sciences. The tests for p-hacking are described in Section 2.4. Data: Head et al. (2016).

The left panel of Figure 1.4 shows a histogram of the raw data on all p-values for the medical and health sciences (the largest subsample). A substantial fraction of p-values is rounded to two decimal places, which results in sizable mass points at $0.01, 0.02, \ldots, 0.15$. Rounding makes the p-curve non-monotonic and discontinuous even in the absence of p-hacking and, thus, invalidates the testable restrictions in Theorem 1.1. Therefore, we also show results based on de-rounded data.[18] In an earlier version of this paper (Elliott et al., 2020), we show that de-rounding restores the non-increasingness but not the continuity of the p-curve. The right panel

[17] For CS1, we use 60 bins (all data) and 30 bins (random subsamples) for biological and medical and health sciences given the large sample sizes, and 30 and 15 bins for the other disciplines.

[18] We de-round the data as follows. To each observed p-value rounded up to the kth decimal point we add a random number generated from the uniform distribution supported on the interval $[\underline{u}, 0.5] \cdot 10^{-k}$, where $\underline{u} = 0$ for zero p-values and $\underline{u} = -0.5$ for non-zero p-values.

of Figure 1.4 shows the impact of de-rounding on the shape of the p-curve. We note that density discontinuity tests are poorly suited here because rounding induces substantial discontinuities, which remain even after de-rounding. This means that rejections of the null can be either due to rounding or due to p-hacking.

In what follows, define a rejection of the null of no p-hacking for p-values smaller than 0.1. Table 1.1 presents the results for the full sample of p-values. For the original (rounded) data, the CS1 and the LCM test reject the null for all disciplines. De-rounding leads to fewer rejections. The CS1 test only rejects for biological sciences, engineering, and medical and health sciences; the LCM test rejects for medical and health sciences. This shows that rounding and de-rounding can substantially affect empirical results. The Binomial and Fisher's test do not reject the null for any discipline, which demonstrates the importance of using our more powerful tests.

Table 1.1. Testing results based on full sample of p-values

Test	Discipline					
	Biological sciences	Chemical sciences	Education	Engineering	Medical and health sciences	Psychology and cognitive sciences
	Rounded					
Binomial	1.000	0.342	0.975	0.999	1.000	1.000
Fisher's Test	1.000	1.000	1.000	1.000	1.000	1.000
Discontinuity	0.000	0.000	0.159	0.000	0.000	0.172
CS1	0.000	0.000	0.000	0.000	0.000	0.000
LCM	0.000	0.000	0.000	0.000	0.000	0.000
Obs in [0.04, 0.05]	7692	296	220	396	38462	1621
Total obs	74746	2631	1993	3262	352817	15189
	De-rounded					
Binomial	0.993	0.133	0.467	0.975	1.000	0.811
Fisher's Test	1.000	1.000	1.000	1.000	1.000	1.000
Discontinuity	0.005	0.117	0.245	0.849	0.162	0.406
CS1	0.028	0.530	0.884	0.084	0.000	0.836
LCM	0.936	1.000	1.000	1.000	0.065	0.653
Obs in [0.04, 0.05]	5720	234	144	250	28318	1161
Total obs	74550	2628	1988	3258	352066	15130

Notes: Table reports p-values from applying different tests for p-hacking based on the full sample of p-values for rounded and de-rounded data. The tests for p-hacking are described in Section 2.4. Data: Head et al. (2016).

Table 1.2 shows the results based on random samples with one p-value per paper. We find that the CS1 test (biological sciences, engineering, medical and health sciences) and the

LCM test (all disciplines except chemical sciences) reject the null based on the rounded data. None of the tests based on non-increasingness rejects the null based on the de-rounded data. A comparison to the results based on all p-values shows that the sample sizes required for detecting p-hacking may be quite large.

Table 1.2. Testing results based on random subsamples of one p-value per paper

Test	Discipline					
	Biological sciences	Chemical sciences	Education	Engineering	Medical and health sciences	Psychology and cognitive sciences
	Rounded					
Binomial	0.510	0.157	0.439	0.904	1.000	0.670
Fisher's Test	1.000	1.000	1.000	1.000	1.000	1.000
Discontinuity	0.113	0.083	0.103	0.000	0.000	0.157
CS1	0.000	0.637	0.232	0.078	0.000	0.734
LCM	0.000	0.265	0.035	0.002	0.000	0.000
Obs in [0.04, 0.05]	1482	63	42	85	6270	185
Total obs	13829	482	366	619	56892	1730
	De-rounded					
Binomial	0.178	0.116	0.286	0.712	0.976	0.465
Fisher's Test	1.000	1.000	1.000	1.000	1.000	1.000
Discontinuity	0.571	0.085	0.997	0.287	0.557	0.637
CS1	0.992	0.688	0.481	0.731	0.872	0.747
LCM	1.000	1.000	1.000	0.999	0.846	1.000
Obs in [0.04, 0.05]	1053	45	28	51	4536	128
Total obs	13788	482	365	619	56753	1716

Notes: Table reports p-values from applying different tests for p-hacking based on random subsamples of p-values for rounded and de-rounded data. The tests for p-hacking are described in Section 2.4. Data: Head et al. (2016).

Finally, the density discontinuity test rejects for at least three disciplines based on the full sample and the random subsamples. After de-rounding, it only rejects for biological sciences (full sample) and chemical sciences (random subsample). These rejections are expected because of the prevalence of rounding-induced discontinuities.

1.6 Conclusion

We provide theoretical foundations for testing for p-hacking based on the distribution of p-values across scientific studies. We establish general results on the p-curve, providing conditions under which a null set of p-curves can be shown to be non-increasing. For p-values

based on t-tests, we derive previously unknown additional restrictions on the p-curve when there is no p-hacking. These restrictions lead to the suggestion of more powerful tests that can be used to test the absence of p-hacking. A reanalysis of two datasets from the literature shows that the new tests based on additional restrictions are useful in testing for p-hacking.

Chapter 1, in full, is a reprint of the material as it appears in Econometrica 2022. Elliott, Graham; Kudrin, Nikolay; Wüthrich, Kaspar. The dissertation author was a primary author of this material.

Chapter 2

(When) Can We Detect p-hacking?

Abstract

p-Hacking can undermine the validity of empirical studies. A flourishing empirical literature investigates the prevalence of p-hacking based on the empirical distribution of reported p-values across studies. Interpreting results in this literature requires a careful understanding of the power of methods used to detect different types of p-hacking. We theoretically study the implications of likely forms of p-hacking on the distribution of reported p-values and the power of existing methods for detecting it. Power can be quite low, depending crucially on the particular p-hacking strategy and the distribution of actual effects tested by the studies. Publication bias can enhance the power for testing the joint null hypothesis of no p-hacking and no publication bias. We relate the power of the tests to the costs of p-hacking and show that power tends to be larger when p-hacking is very costly. Monte Carlo simulations support our theoretical results.

2.1 Introduction

Researchers have a strong incentive to find, report, and publish novel results (e.g., Imbens, 2021, p.158). Translated mathematically, this often results in a strong incentive to find useful results that have small p-values when conducting hypothesis tests examining if the data fits with the current conventional beliefs. Simonsohn et al. (2014) used the term "p-hacking" to encompass decisions made by researchers in conducting their work that are made to improve

the novelty of their results as seen through the lens of the reported p-values. Their work has generated an empirical literature that examines empirically the distribution of p-values across studies (the "p-curve") in an attempt to determine if p-hacking is prevalent or not.[1]

In previous work (Elliott et al., 2022b), we characterized the set of p-curves under the null hypothesis of no p-hacking. Such characterizations are useful for developing tests for detecting p-hacking that control size. However, they are inherently uninformative about the power of the resulting tests. To understand power, we need to understand both how p-hacking impacts the distribution of p-values and how powerful tests are in detecting these impacts. This paper examines theoretically and through Monte Carlo analysis the power of tests available to test for p-hacking using data on p-values across studies.

A careful study of power is relevant to this literature because the implications of p-hacking on the distribution of reported p-values are not clear. When researchers p-hack, the p-curve differs from the null set of p-curves, but there are many ways in which the distribution of p-values can be affected. "Directions" of power depend on precisely how the curve is affected, which in turn will depend on the empirical problem and how the p-hacking is undertaken. This paper places a strong emphasis on considerations of how the distribution might be affected. Many tests sprang from the early intuition that p-hacking would result in "humps" in the distribution of p-values just below common thresholds for size like 5%. But intuition also might suggest that if all researchers p-hack, then this might simply push the distributions to the left. It is also the case that there are limits to how much can be gained by p-hacking; approaches such as searching across regressions with different control variables can help improve p-values but do not allow the researcher to attain any p-value they desire.

Of further interest in examining power is that power is useful if it is directed towards alternatives where the costs of p-hacking are higher. As the ASA notes "Valid scientific con-

[1] See, e.g., Masicampo and Lalande (2012); Simonsohn et al. (2014); Lakens (2015); Simonsohn et al. (2015); Head et al. (2015); Ulrich and Miller (2015) for early applications and further discussions, Havranek et al. (2021); Brodeur et al. (2022a); Malovaná et al. (2022); Yang et al. (2022); Decker and Ottaviani (2023) for recent applications, and Christensen and Miguel (2018) for a review.

clusions based on p-values and related statistics cannot be drawn without at least knowing how many and which analyses were conducted, and how those analyses (including p-values) were selected for reporting." (Wasserstein and Lazar, 2016, p.132) This translated mathematically is that p-hacking has two costs in terms of understanding the statistical results — empirical sizes of tests will be larger than the stated size and in many cases coefficient estimates will be biased in the direction of being more impressive. In our power analyses, we therefore explicitly relate power to the costs of p-hacking.

In order to consider relevant directions of power, we examine two approaches to p-hacking in four situations in which we might think opportunities for p-hacking in economics and other fields commonly arise. The two approaches are what we refer to as a "threshold" approach where a researcher targeting a specific threshold stops if the obvious model rejects at this size and conducts a search over alternative specifications if not and a second approach of simply choosing the best p-value from a set of specifications (denoted the "minimum" approach below). We examine four situations where opportunities for p-hacking arise: (a) searching across linear regression models with different control variables, (b) searching across different choices of instruments in estimating causal effects, (c) searching across datasets, and (d) searching across bandwidth choices in constructing standard errors in time series regressions.[2] We construct theoretical results for the implied distribution of p-values under each approach to p-hacking in a simple model. The point of this exercise is twofold — by seeing how exactly p-hacking affects the distribution we can determine the testing method appropriate for detecting the p-hacking, and also we will be able to determine the features that lead to large or small deviations from the distribution of p-values when there is no p-hacking.[3] We then examine in Monte Carlo analyses

[2]While (a)–(d) are arguably prevalent in empirical research, there are of course many other approaches to p-hacking, such as strategically dropping/including data or excluding outliers, selecting among different econometric methods and empirical strategies, choosing which outcomes to analyze, etc. See, e.g., Simonsohn et al. (2014) and Simonsohn (2020). From an econometric perspective, this implies that the alternative space of the testing problem is very large.

[3]While we focus on the impact of these different types of p-hacking on the shape of the p-curve and the power of tests for detecting p-hacking, such explicit models of p-hacking are also useful in other contexts. For example, McCloskey and Michaillat (2022) use a model of p-hacking to construct incentive compatible critical values.

extensions of these cases.

Our theoretical results and Monte Carlo simulations shed light on how distributions of p-values are impacted by p-hacking and provide a careful understanding of the power of existing tests for detecting p-hacking. The main implications are as follows:

1. From a scientific perspective, the ability of tests to detect p-hacking can be quite low. The threshold approach to p-hacking is more easily detected than when researchers simply take the minimum of the p-values.

2. For the threshold approach, target values for the p-value result in discontinuities in the distribution of p-values as well as violations on upper bounds for this distribution, resulting in tests for these violations having power. It is only in special cases that the intuitive "humps" in this distribution appear, violating the condition that this curve is monotonically non-increasing.

3. When researchers choose the minimum p-value from a set of models that nests the true model, the distribution of p-values is shifted to the left, and only tests based on upper bounds for this distribution have power. For this reason this approach to p-hacking is much harder to detect.

4. The power of different tests for p-hacking depends strongly on where the mass of true values being tested actually lies. If most of the tests are of null hypotheses that are true, tests looking for humps and discontinuity tests can still have power. However, if the majority of the p-values are constructed from tests where the null is false, tests based on upper bounds on the p-curve are more appropriate.

5. The costs of p-hacking in terms of biases through model specification can be quite small. In many cases the estimates and t-statistics are strongly positively correlated across specifications. We show that the power of the tests is positively related to the cost in terms of bias of p-hacking.

6. Publication bias under reasonable models enhances the power of tests for p-hacking, although in this situation it is best to consider the tests as test of the joint null hypothesis of no p-hacking and no publication bias.

In this paper, we focus on the problem of detecting p-hacking based on the distribution of p-values and do not consider the popular Caliper tests.[4] Caliper tests aim to detect p-hacking based on excess mass in the distribution of z-scores right above significance cutoffs. However, since humps in the distribution of z-scores can also be induced by the distribution of true effects, these tests do not control size in general; see Kudrin (2022) for a discussion.

2.2 Setup

2.2.1 The Distribution of p-Values

Elliott et al. (2022b) provided a theoretical characterization of the distribution of p-values across studies in the absence of p-hacking for general distributions of true effects.[5] The notation here follows that work. Individual researchers provide test results of a hypothesis that is reported as a test statistic T, which is distributed according to a distribution with cumulative distribution function (CDF) F_h, where $h \in \mathscr{H}$ indexes parameters of either the exact or asymptotic distribution of the test. Researchers are testing the null hypothesis that $h \in \mathscr{H}_0$ against $h \in \mathscr{H}_1$ with $\mathscr{H}_0 \cap \mathscr{H}_1$ empty. Suppose the test rejects for large values of T and denote by $cv(p)$ the critical value for level p tests. For any individual study the researcher tests a hypothesis at a particular h. We denote the power function of the test for that study by $\beta(p,h) = \Pr(T > cv(p) \mid h)$.

Across researchers, there is a distribution Π of effects h, which is to say that different researchers testing different hypotheses examine different problems that have different "true"

[4] See, e.g, Gerber and Malhotra (2008a,b); Bruns et al. (2019); Vivalt (2019); Brodeur et al. (2020a).

[5] See, e.g., Hung et al. (1997), Simonsohn et al. (2014), and Ulrich and Miller (2018) for numerical and analytical examples of p-curves for specific tests and/or effect distributions.

effects. The resulting CDF of p-values across all these studies is then

$$G(p) = \int_{\mathscr{H}} \Pr\left(T > cv(p) \mid h\right) d\Pi(h) = \int_{\mathscr{H}} \beta\left(p,h\right) d\Pi(h).$$

Under mild regularity assumptions (differentiability of the null and alternative distributions, boundedness and support assumptions; see Elliott et al. (2022b) for details), we can write the p-curve (density of p-values) in the absence of p-hacking as

$$g(p) = \int_{\mathscr{H}} \frac{\partial \beta\left(p,h\right)}{\partial p} d\Pi(h).$$

Next we discuss the properties of $g(p)$, which underlie the statistical tests for p-hacking.

2.2.2 Testable Restrictions

Our goal is to evaluate the power of statistical tests based on the different testable implications derived in the literature. Elliott et al. (2022b) provide general sufficient conditions for when the p-curve is non-increasing, $g' \leq 0$, and continuous when there is no p-hacking, allowing for tests of these properties of the distribution to be interpreted as tests of the null hypothesis of no p-hacking.[6] These conditions hold for many possible distributions F_h that arise in research, for example, normal, folded normal (relevant for two-sided tests), and χ^2 distributions.

When T is normally distributed (for example, tests on means or regression parameters when central limit theorems apply), Elliott et al. (2022b) show that in addition to the non-increasing property the p-curves are completely monotonic (i.e., have derivatives of alternating signs so that $g'' \geq 0$, $g''' \leq 0$, etc.) and there are testable upper bounds on the p-curve and its derivatives.

[6]These results imply that classical approaches for detecting p-hacking based on non-increasingness, such as the Binomial test and Fisher's test (e.g., Simonsohn et al., 2014; Head et al., 2015), are valid in a wide range of empirically relevant settings.

2.2.3 Directions of Power

If researchers do p-hack, the distribution of the reported statistic T differs from that under the null and this affects the distribution of reported p-values. It is then possible that the resulting p-curve violates the properties listed above in one way or another, providing the opportunity to test for p-hacking.

For any form of p-hacking, we consider that there is a set of p-values $\{P_1, P_2, P_3, \dots\}$ that a researcher could report and a method of choosing which p-value P_r to report, i.e.,

$$P_r = d(P_1, P_2, P_3, \dots).$$

The power of tests for p-hacking will be dependent on the functional form of $d(\cdot)$ and the joint distribution of $\{P_1, P_2, P_3, \dots\}$. The relevant functional form is a result of the approach to p-hacking. A discontinuous function arises if researchers search across specifications in order of 'reasonableness' until finding one that is significant at some level such as 0.05. In such situations we might expect humps in the distribution for p-values below such thresholds, as well as discontinuities at that point because of the discontinuity in $d(\cdot)$. For researchers choosing the smallest p-value of those available we are less likely to see humps and unlikely to see discontinuities in the p-curve because of the continuity of the $d(\cdot)$ function, instead the bounds derived in Elliott et al. (2022b) may be violated.

The distribution over possible p-values will depend on the testing situation and the data. It is not the case that researchers can select any p-value they desire, available p-values will be a draw from a distribution. This relates to Simonsohn (2020)'s distinction of "slow p-hacking" and "fast p-hacking": slow p-hacking corresponds to the case where p-values change little across analyses; fast p-hacking refers to settings where p-values change substantially across analyses. Our analysis in each of the cases characterizes the possible distributions analytically thus providing results as to how these distributions affect power.

Ultimately carefully considering the functions $d(\cdot)$ and distributions of p-values allow us to examine power in empirically relevant directions.

2.2.4 Impact of Publication Bias

Our main focus is on the power of testing for various types of p-hacking, however a practical concern is that most studies of reported p-values are restricted to the sample selected set of p-values that appear in papers published in journals, referred to in the literature under the general term of publication bias (see, e.g., Andrews and Kasy, 2019, and the references therein). Publication bias also impacts the distribution of the p-values in ways that tests for p-hacking are designed to detect.

To consider the impact of publication bias, let S denote the publication indicator, where $S = 1$ if the paper is selected for publication and $S = 0$ otherwise. By Bayes' Law, the p-curve conditional on publication, $g_{S=1}(p) := g(p \mid S = 1)$, is given by

$$g_{S=1}(p) = \frac{\Pr(S = 1 \mid p) g^d(p)}{\Pr(S = 1)}. \tag{2.1}$$

Here $\Pr(S = 1 \mid p)$ is the publication probability given p-value $P = p$, and $g^d(p)$ refers to the potentially p-hacked distribution of p-values when there is no publication bias.

If there is no publication bias, the publication probability does not depend on p, $\Pr(S = 1 \mid p) = \Pr(S = 1)$, so that $g_{S=1} = g^d$. If, on the other hand, there is publication bias, the publication probability will depend on the reported p-value so that $\Pr(S = 1 \mid p)$ will not be equal to $\Pr(S = 1)$ for some $p \in (0, 1)$ and $g_{S=1} \neq g^d$. In this case, one can detect p-hacking and/or publication bias if $g_{S=1}$ violates the testable restrictions underlying the statistical tests, so we can regard the tests here as joint tests of the absence of both p-hacking and publication bias.

It is plausible to assume that papers with smaller p-values are more likely to get published so that $\Pr(S = 1 \mid p)$ is decreasing in p. In this case, $g_{S=1}$ is non-increasing in the absence of p-hacking. Selection through publication bias that favors smaller p-values can result in steeper p-

curves violating the bounds derived under the null hypothesis of no p-hacking. Hence rejections of bounds tests may well be exacerbated by publication bias. Discontinuities in $\Pr(S = 1 \mid p)$ can generate discontinuities in the absence of p-hacking, generating power for discontinuity tests. We examine these effects via Monte Carlo analysis in Section 2.5.2.4.

2.3 Implications of p-Hacking

Power of tests will depend on the underlying tests being examined along with the true distribution of the effects (here the distribution of h), the methods used to p-hack, and also the extent to which there are choices in modelling that allow choices of estimates and tests over which p-hacking can be undertaken. Because of this, there is considerable complexity in evaluating the power of tests for p-hacking, and any power evaluation is dependent on the context of the problem and the choices available.

We deal with each of these through the following choices:

1. For the distribution of h, we provide general analytical results for any distribution, for graphical and Monte Carlo purposes we choose either point masses on a particular value of h or well known distributions.
2. For the methods employed in p-hacking, corresponding to the choice of the function $d(\cdot)$, we consider two basic approaches to p-hacking.
    (a) The *threshold approach*, where the researcher constructs a test from their preferred model, accepting this test if it corresponds to a p-value below a target value (for example, 0.05). If the p-value does not achieve this goal value, additional models are considered, and the smallest p-value over the model choices is reported. This is representative of the "intuitive" approach to p-hacking that is discussed in much of the literature on testing for p-hacking, where humps in the p-curve around common critical levels are examined.

(b) The *minimum approach*, where researchers take the smallest p-value from a set of models. Intuitively, we would expect that for this approach the distribution of p-values would shift to the left, be monotonically non-increasing, and there would be no expected hump in the distribution of p-values near commonly reported significance levels. This is true, for example, if the researchers report the minimum p-value across independent tests; see Section 2.3.3.

3. We study the shape of the distribution of p-values under four arguably prevalent empirical problems with the potential for p-hacking. The ability to p-hack depends on the distribution of the p-values conditional on h, the four examples allow us to construct power in these relevant testing situations.

Appendix B.1 presents the analytical derivations underlying the results in this section.

2.3.1 Selecting Control Variables in Linear Regression

Linear regression has been suggested to be particularly prone to p-hacking (e.g., Hendry, 1980; Leamer, 1983; Bruns and Ioannidis, 2016; Bruns, 2017). Researchers usually have available a number of control variables that could be included in a regression along with the variable of interest. Selection of various configurations for the linear model allows multiple chances to obtain a small p-value, perhaps below a threshold such as 0.05. The theoretical results in this section yield a careful understanding of the shape of the p-curve when researchers engage in this type of p-hacking. They clarify which statistical tests can be expected to have power for detecting p-hacking.

2.3.1.1 Shape of the p-Curve

We construct a stylized model and consider the two approaches to p-hacking discussed above in order to provide analytical results that capture the impact of p-hacking. Suppose that the researchers estimate the impact of a scalar regressor X_i on an outcome Y_i. The data are

generated as $Y_i = X_i\beta + U_i$, $i = 1,\ldots,N$, where $U_i \overset{iid}{\sim} \mathcal{N}(0,1)$. For simplicity, we assume that X_i is non-stochastic. The researchers test the hypothesis $H_0 : \beta = 0$ against $H_1 : \beta > 0$.

In addition to X_i, the researchers have access to two additional non-stochastic control variables, Z_{1i} and Z_{2i}.[7] To simplify the exposition, we assume that (X_i, Z_{1i}, Z_{2i}) are scale normalized so that $N^{-1}\sum_{i=1}^N X_i^2 = N^{-1}\sum_{i=1}^N Z_{1i}^2 = N^{-1}\sum_{i=1}^N Z_{2i}^2 = 1$, that $N^{-1}\sum_{i=1}^N Z_{1i}Z_{2i} = \gamma^2$, and that $N^{-1}\sum_{i=1}^N X_i Z_{1i} = N^{-1}\sum_{i=1}^N X_i Z_{2i} = \gamma$, where $|\gamma| \in (0,1)$.[8] Let $h := \sqrt{N}\beta\sqrt{1-\gamma^2}$, where h is drawn from a distribution with support $\mathscr{H} \subseteq [0,\infty)$.

First, consider the threshold form of p-hacking.

1. Researchers regress Y_i on X_i and Z_{1i} and report the resulting p-value, P_1, if $P_1 \leq \alpha$.
2. If $P_1 > \alpha$, researchers regress Y_i on X_i and Z_{2i} instead of Z_{1i} and obtain p-value, P_2. They report $P_r = \min\{P_1, P_2\}$.

Under this threshold form of p-hacking, the reported p-value, P_r, is given by

$$P_r = \begin{cases} P_1, & \text{if } P_1 \leq \alpha, \\ \min\{P_1, P_2\}, & \text{if } P_1 > \alpha. \end{cases}$$

Define $\hat{\beta}_r^t$ to be the OLS estimate from the regression that accords with the chosen p-value. Under the minimum approach, the reported p-value is $P_r = \min\{P_1, P_2\}$. Denote the regression estimate for the chosen model as $\hat{\beta}_r^m$. Each approach results in different distributions of p-values, and, consequently, tests for p-hacking will have different power properties.

In Appendix B.1.1, we show that for the threshold approach the resulting p-curve is

$$g_1^t(p) = \int_{\mathscr{H}} \exp\left(hz_0(p) - \frac{h^2}{2}\right) \Upsilon_1^t(p;\alpha,h,\rho)d\Pi(h),$$

[7] For simplicity, we consider a setting where Z_{1i} and Z_{2i} do not enter the true model so that their omission does not lead to omitted variable biases (unlike, e.g., in Bruns and Ioannidis (2016)). It is straightforward to generalize our results to settings where Z_{1i} and Z_{2i} enter the model: $Y_i = X_i\beta_1 + Z_{1i}\beta_2 + Z_{2i}\beta_3 + U_i$.

[8] We omit $\gamma = 0$, i.e., adding uncorrelated control variables, because in this case the t-statistics and thus p-values for each regression are equivalent and hence there is no opportunity for p-hacking of this form.

where $\rho = 1-\gamma^2$, $z_h(p) = \Phi^{-1}(1-p) - h$, Φ is the standard normal CDF, and

$$\Upsilon_1^t(p;\alpha,h,\rho) = \begin{cases} 1 + \Phi\left(\frac{z_h(\alpha)-\rho z_h(p)}{\sqrt{1-\rho^2}}\right), & \text{if } p \leq \alpha, \\ 2\Phi\left(z_h(p)\sqrt{\frac{1-\rho}{1+\rho}}\right), & \text{if } p > \alpha. \end{cases}$$

In interpreting this result, note that when there is no p-hacking $\Upsilon_1(p;\alpha,h,\rho) = 1$. It follows directly from the properties of Φ that the threshold p-curve lies above the curve when there is no p-hacking for $p \leq \alpha$. We can also see that since $\Phi\left(\frac{z_h(\alpha)-\rho z_h(p)}{\sqrt{1-\rho^2}}\right)$ is decreasing in h that for larger h the difference between the threshold p-curve and the curve without p-hacking becomes smaller. This follows intuitively since for a larger h, the need to p-hack diminishes as most of the studies find an effect without resorting to manipulation.

If researchers simply compute both p-values and report $P_r = \min\{P_1, P_2\}$, the distribution of p-values follows directly from calculations deriving the above result and is equal to

$$g_1^m(p) = 2\int_{\mathscr{H}} \exp\left(hz_0(p) - \frac{h^2}{2}\right)\Phi\left(z_h(p)\sqrt{\frac{1-\rho}{1+\rho}}\right) d\Pi(h).$$

For p-hacking of this form, the entire distribution of p-values is shifted to the left. For p less than one half, the curve lies above the curve when there is no p-hacking. This distribution is monotonically decreasing for all Π, so does not have a hump and remains continuous. Because of this, only the tests based on upper bounds and higher-order monotonicity have any potential for testing the null hypothesis of no p-hacking. If Π is a point mass distribution, there is a range over which $g_1^m(p)$ exceeds the upper bound $\exp(z_0(p)^2/2)$ derived in Elliott et al. (2022b), the upper end (largest p) of which is at $p = 1-\Phi(h)$.

Figure 2.1 shows the theoretical p-curves for various h and γ. In terms of violating the condition that the p-curve is monotonically decreasing, violations for the threshold case can

occur but only for h small enough. For $p < \alpha$, the derivative is

$$g_1^{t\prime}(p) = \int_{\mathscr{H}} \frac{\phi(z_h(p)) \left[\frac{\rho}{\sqrt{1-\rho^2}} \phi\left(\frac{z_h(\alpha) - \rho z_h(p)}{\sqrt{1-\rho^2}} \right) - h\left(1 + \Phi\left(\frac{z_h(\alpha) - \rho z_h(p)}{\sqrt{1-\rho^2}} \right) \right) \right]}{\phi^2(z_0(p))} d\Pi(h),$$

where ϕ is the standard normal probability density function (PDF). Note that ρ is always positive and, when all nulls are true (i.e., when Π assigns probability one to $h = 0$), $g^{t\prime}(p)$ is positive for all $p \in (0, \alpha)$.[9] This can be seen for the dashed line in Figure 2.1 (left panel). However at $h = 1$ this effect no longer holds, and the p-curve is downward sloping. From Figure 2.1 (right panel) we see that violations for monotonicity are larger for smaller γ. When $\gamma = 0.1$, the p-curve even becomes bimodal.

Figure 2.1 indicates that the threshold approach to p-hacking implies a discontinuity at p-values equal to size. The size of the discontinuity is larger for larger h and remains for each γ, although how that translates to power of tests for discontinuity also depends on the shape of the rest of the curve. We examine this in Monte Carlo experiments in Section 3.5.

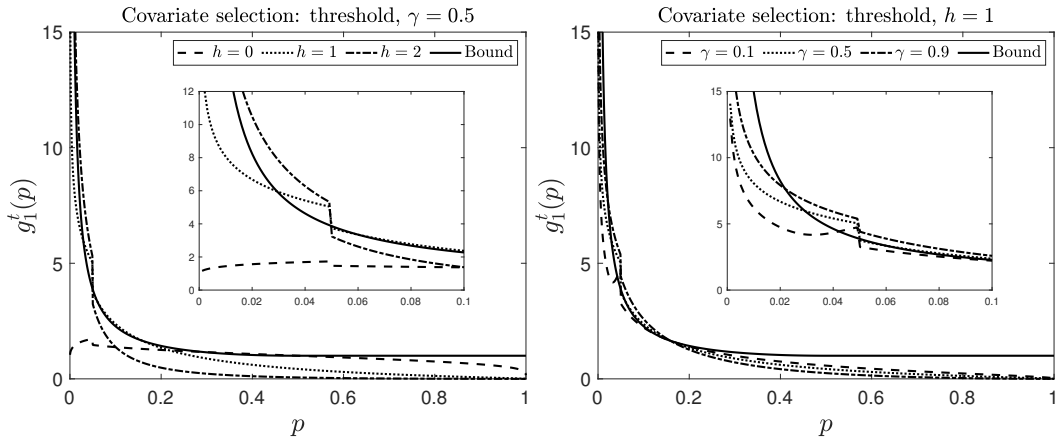

Figure 2.1. p-Curves from covariate selection with thresholding. Left panel: $\gamma = 0.5$. Right panel: $h = 1$.

[9] For $p > \alpha$, the derivative of $g_1^t(p)$ is negative and equal to

$$g_1^{t\prime}(p) = -\int_{\mathscr{H}} \frac{\phi(z_h(p))}{\phi^2(z_0(p))} \left(h + \sqrt{\frac{1-\rho}{1+\rho}} \phi\left(z_h(p) \sqrt{\frac{1-\rho}{1+\rho}} \right) \right).$$

Figure 2.2 examines both the threshold approach to p-hacking as well as the p-hacking approach of directly taking the minimum p-value. Results are presented across the two choices for $h = 1$ and $\gamma = 0.5$, with respect to the bounds under no p-hacking. We also report the no-p-hacking distribution for the same value of h. Simply taking the minimum p-value as a method of p-hacking results in a curve that remains downward sloping and has no discontinuity — tests for these two features will have no power against such p-hacking. But as Figure 2.2 shows, the upper bounds on the p-curve are violated for both methods of p-hacking. The violation in the thresholding case is pronounced; for taking the minimum p-value the violation is much smaller and harder to detect.

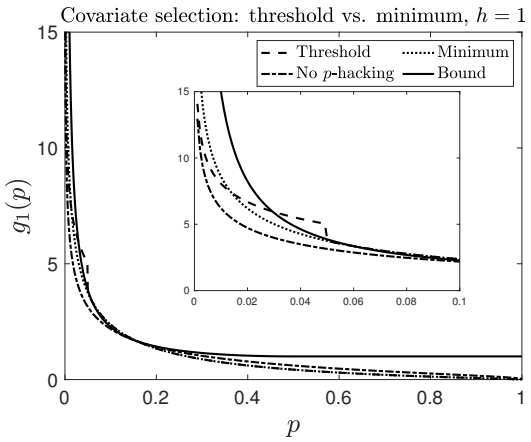

Figure 2.2. p-Curves from covariate selection for threshold and minimum approaches ($h = 1$, $\gamma = 0.5$).

2.3.1.2 Costs of p-Hacking

There are two costs to p-hacking. The first is that when we account for the searching over models, the size of tests when $h = 0$ is understated, larger than the empirical size claimed. The second cost is that the reported estimates will be larger in magnitude and hence biased.

The magnitude of size distortions follows from the derived CDF for the p-hacked curve evaluated at $h = 0$. The size distortion is the same for both the thresholding case and the situation where the researcher simply reports the minimum p-value, since in either case, if there is a rejection at the desired size, each method of p-hacking will use it. Empirical size for any nominal

size is given by

$$G_0(\alpha) = 1 - \Phi_2(z_0(\alpha), z_0(\alpha); \rho),$$

where $\Phi_2(\cdot,\cdot;\rho)$ is the CDF of the bivariate normal distribution with standard marginals and correlation ρ. Figure 2.3 shows the difference between empirical and nominal size. The left panel shows, for nominal size $\alpha = 0.05$, how empirical size varies with γ. For small γ the tests are highly correlated (ρ is close to one), leaving little room for effective p-hacking, and hence there is only a small effect on size. As γ becomes larger, so does the size distortion as it moves towards having an empirical size double that of the nominal size. The right-hand size panel shows, for three choices of γ ($\gamma \in \{0.1, 0.5, 0.9\}$), how the empirical size exceeds nominal size. The lower line is nominal size; the empirical size is larger for each value of γ. Essentially, the result is somewhat uniform over this empirical size range, with size coming close to double empirical size for the largest value of γ.

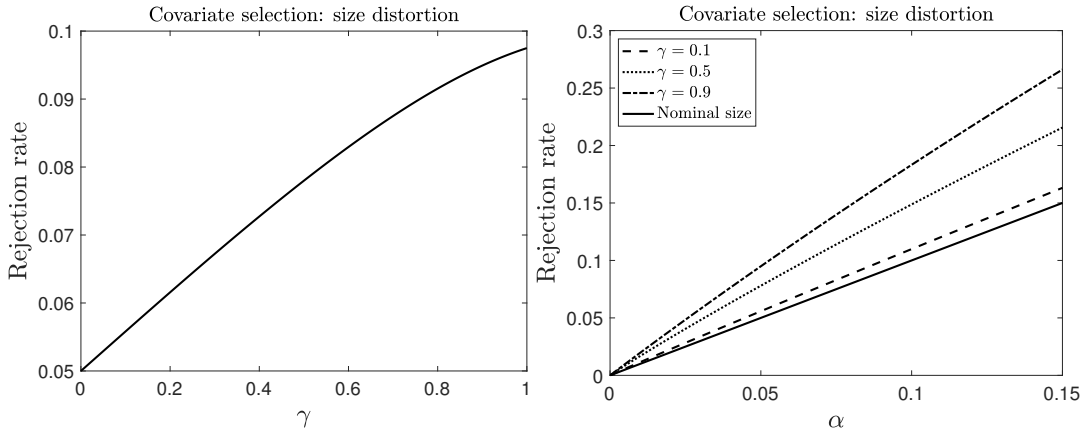

Figure 2.3. Rejection rate under p-hacking. Left panel: rejection rate as a function of γ for $\alpha = 0.05$. Right panel: rejection rate as a function of α for $\gamma \in \{0.1, 0.5, 0.9\}$.

Selectively choosing larger t-statistics results in selectively choosing larger estimated effects. The bias for the threshold case is given by

$$E\hat{\beta}_r^t - \beta = \frac{\left(\sqrt{2(1-\rho)}\phi(0)\Phi\left(\sqrt{\frac{2}{1+\rho}}z_h(\alpha)\right) + (1-\rho)\phi(z_h(\alpha))\left(1-\Phi\left(\sqrt{\frac{1-\rho}{1+\rho}}z_h(\alpha)\right)\right)\right)}{\sqrt{N\rho}},$$

and for the minimum approach it is given by

$$E\hat{\beta}_r^m - \beta = \frac{\sqrt{2(1-\rho)}\phi(0)}{\sqrt{N\rho}}.$$

The bias as a function of h can be seen in Figure 2.4. For the threshold case, most p-hacking occurs when h is small. As a consequence, the bias is larger for small h. A larger γ means a smaller ρ, and hence draws of the estimate and the p-value are less correlated, allowing for larger impacts. For the minimum approach, the bias does not depend on h, and is larger than that for the threshold approach. The reason is that the minimum approach always chooses the largest effect since in our simple setting the standard errors are the same in both regressions.

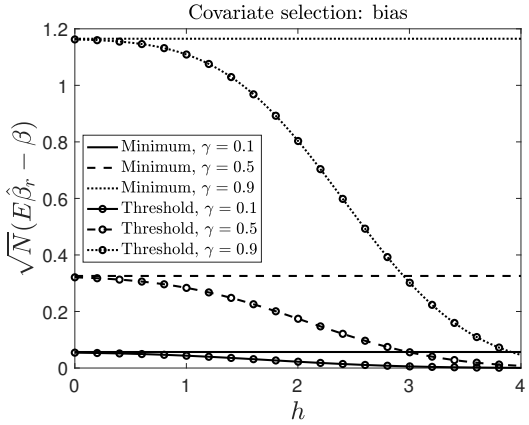

Figure 2.4. Bias from covariate selection for $\gamma \in \{0.1, 0.5, 0.9\}$.

2.3.2 Selecting amongst Instruments in IV Regression

2.3.2.1 Shape of the p-Curve

Suppose that the researchers use an instrumental variables (IV) regression to estimate the causal effect of a scalar regressor X_i on an outcome Y_i. The data are generated as

$$Y_i = X_i\beta + U_i,$$

$$X_i = Z_{1i}\gamma_1 + Z_{2i}\gamma_2 + V_i,$$

where $(U_i, V_i)' \overset{iid}{\sim} \mathcal{N}(0, \Omega)$ with $\Omega_{12} \neq 0$. The instruments are generated as $Z_i \overset{iid}{\sim} \mathcal{N}(0, I_2)$ and independent of (U_i, V_i). The researchers test the hypothesis $H_0 : \beta = 0$ against $H_1 : \beta > 0$. To simplify the exposition, suppose that $\Omega_{11} = \Omega_{22} = 1$ and $\gamma_1 = \gamma_2 = \gamma$, where $|\gamma| \in (0,1)$ is known. We let $h := \sqrt{N}\beta|\gamma|$, where h is drawn from a distribution supported on $\mathscr{H} \subseteq [0, \infty)$.

We again consider the two forms of p-hacking. For the threshold approach, first the researchers run an IV regression of Y_i on X_i using Z_{1i} and Z_{2i} as instruments and report the corresponding p-value, P_{12}, if $P_{12} \leq \alpha$. If $P_{12} > \alpha$, the researchers then run IV regressions of Y_i on X_i using Z_{1i} and Z_{2i} as single instruments and obtain p-values, P_1 and P_2. They report $\min\{P_1, P_2, P_{12}\}$ so that reported p-value, P_r, is

$$P_r = \begin{cases} P_{12}, & \text{if } P_{12} \leq \alpha, \\ \min\{P_1, P_2, P_{12}\}, & \text{if } P_{12} > \alpha. \end{cases}$$

The second approach is to report the $\min\{P_1, P_2, P_{12}\}$, that is to just check for the smallest p-value and report that. Researchers report the estimated effect that accords with the reported p-value in both approaches, defined as $\hat{\beta}_r^t$ and $\hat{\beta}_r^m$, accordingly.[10]

For the threshold approach, the p-curve (see Appendix B.1.2 for derivations) is

$$g_2^t(p) = \int_{\mathscr{H}} \exp\left(hz_0(p) - \frac{h^2}{2}\right) \Upsilon_2^t(p; \alpha, h) d\Pi(h),$$

where

$$\Upsilon_2^t(p; \alpha, h) = \begin{cases} \frac{\phi(z_{\sqrt{2}h}(p))}{\phi(z_h(p))} + 2\Phi(D_h(\alpha) - z_h(p)), & \text{if } 0 < p \leq \alpha, \\ \frac{\phi(z_{\sqrt{2}h}(p))}{\phi(z_h(p))}\zeta(p) + 2\Phi(D_h(p) - z_h(p)), & \text{if } \alpha < p \leq 1/2, \\ 2\Phi(z_h(p)), & \text{if } 1/2 < p < 1, \end{cases}$$

[10] In our stylized model, researchers select the instruments that yield the "best" result. In practice, it is likely that they also select instruments based on the first stage F-statistic exceeding a certain cutoff such as 10 (e.g., Andrews and Kasy, 2019; Brodeur et al., 2020a). We leave the derivation of p-curves under both types of instrument selection for future research.

and $\zeta(p) = 1 - 2\Phi((1-\sqrt{2})z_0(p))$ and $D_h(p) = \sqrt{2}z_0(p) - 2h$.

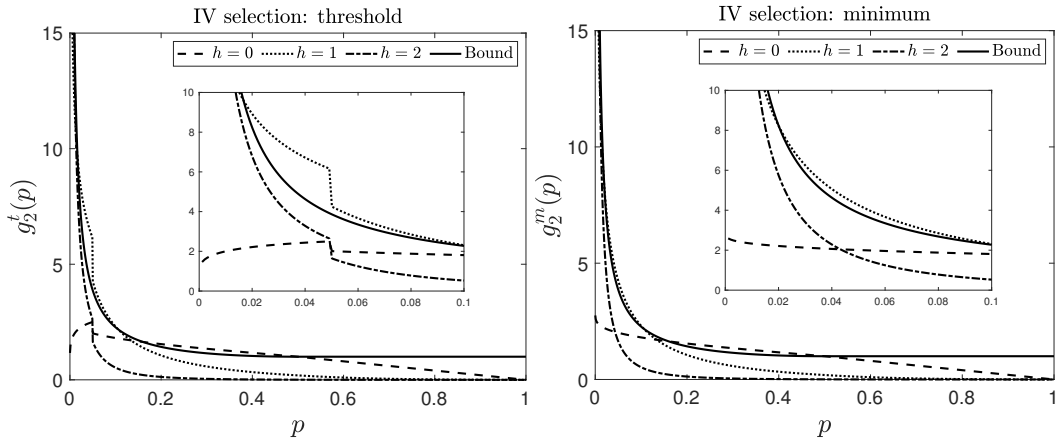

Figure 2.5. p-Curves from IV selection. Left panel: thresholding. Right panel: minimum.

In Figure 2.5 the p-curves for $h \in \{0,1,2\}$ are shown for the threshold approach in the left panel. As in the covariate selection example, it is only for small values of h that we see upward sloping curves and a hump below size. For $h = 1$ and $h = 2$, no such violation of non-increasingness occurs, and tests aimed at detecting such a violation will have no power. The reason is similar to that of the covariate selection problem — when h becomes larger many tests reject anyway, so whilst there is still a possibility to p-hack the actual rejections overwhelm the "hump" building of the p-hacking. For all h there is still a discontinuity in the p-curve arising from p-hacking, so tests for a discontinuity at size will still have power.

If researchers simply report $P_r = \min\{P_1, P_2, P_{12}\}$, the distribution of p-values follows directly from calculations deriving the above result and is equal to

$$g_2^m(p) = \int_{\mathscr{H}} \exp\left(hz_0(p) - \frac{h^2}{2}\right) \Upsilon_2^m(p;\alpha,h) d\Pi(h),$$

where

$$\Upsilon_2^m(p;\alpha,h) = \begin{cases} \frac{\phi(z_{\sqrt{2}h}(p))}{\phi(z_h(p))}\zeta(p) + 2\Phi(D_h(p) - z_h(p)), & \text{if } 0 < p \leq 1/2, \\ 2\Phi(z_h(p)), & \text{if } 1/2 < p < 1. \end{cases}$$

The right hand side panel of Figure 2.5 displays the p-curves for $h \in \{0,1,2\}$ for the

minimum approach. There is no hump as expected, and all the curves are non-increasing. Only tests based on upper bounds for and higher-order monotonicity of the p-curve have the possibility of rejecting the null hypothesis of no p-hacking in this situation. As displayed, the upper bound is violated for large h for low p-values, and is violated for smaller h at larger p-values.

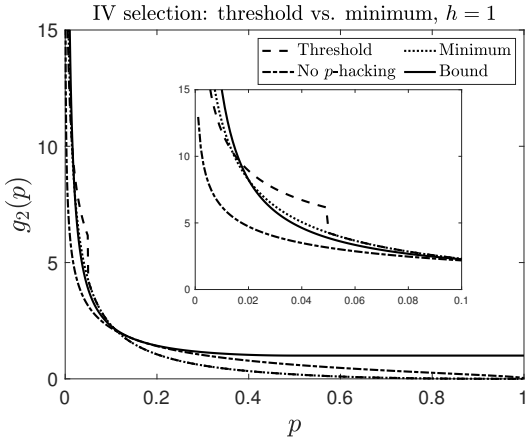

Figure 2.6. p-Curves from IV selection for threshold and minimum approaches ($h = 1$).

Figure 2.6 shows the comparable figure for the IV problem as Figure 2.2 shows for the covariates example. The results are qualitatively similar across these examples, although quantitatively the p-hacked curves in the IV problem are closer to the bounds than in the covariates problem. For $h = 1$, we do find that the p-curves under p-hacking violate the bounds on $(0, 0.1]$, hence suggesting that the tests based on upper bounds can be employed to test the null hypothesis of no p-hacking.

Overall, as with the case of covariate selection, both the relevant tests for p-hacking and their power will depend strongly on the range of h relevant to the studies underlying the data employed for the tests.

2.3.2.2 Costs of p-Hacking

Again, one of the costs of p-hacking is an inflated size of tests when $h = 0$, i.e., when the null hypothesis is true, but the paper hopes to claim it is not. The second is inflated coefficient estimates resulting in a bias of reported results, which occurs to some extent at all values of h.

Size distortions follow from the derivation of the results above. The corresponding CDF for the p-curve evaluated at size α is given by the expression

$$G_0(\alpha) = 1 - \Phi(z_0(\alpha))\Phi((\sqrt{2}-1)z_0(\alpha)) - \int_{(\sqrt{2}-1)z_0(\alpha)}^{z_0(\alpha)} \phi(x)\Phi(\sqrt{2}z_0(\alpha) - x)dx.$$

The expression is the same for both the threshold approach and taking the minimum, for the same reason as in the case of covariate selection. The magnitude of the size distortion is given in Figure 2.7. Size is essentially double the stated size, with the p-hacked size at 11% when nominal size is 5%.

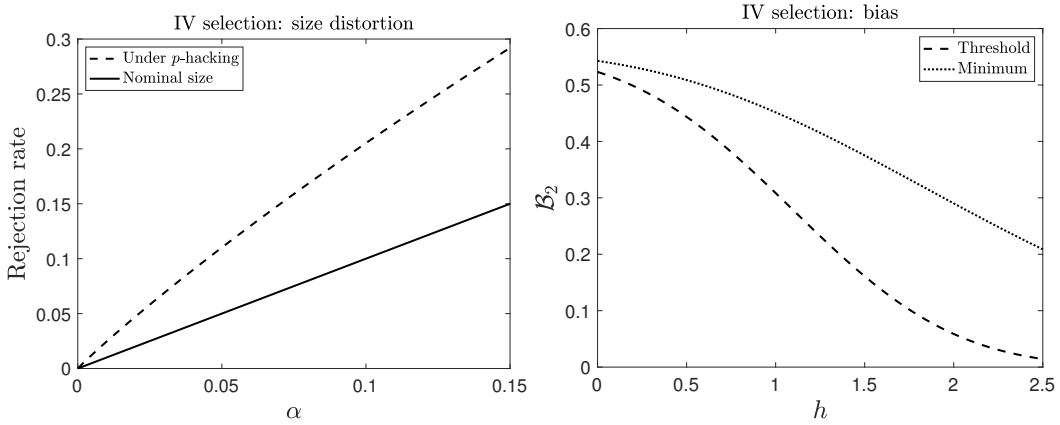

Figure 2.7. Rejection rate under p-hacking. Left panel: rejection rate as a function of α. Right panel: bias from p-hacking for different values of h and $\gamma = 1$.

In terms of the bias induced by p-hacking, distributions over h will induce distributions over the biases since the bias for any study depends on the true model. We report here the bias for different h rather than choose a (non-degenerate) distribution. For the special case example of this section, for the threshold and minimum approaches with $\alpha \leq 1/2$, we can write for any h

the scaled mean (first-order) biases[11], $\mathscr{B}_2^t$ and $\mathscr{B}_2^m$ respectively, as follows

$$\begin{aligned}
\mathscr{B}_2^t &= \mathscr{B}_2^m - \frac{|\gamma|^{-1}}{\sqrt{2-\sqrt{2}}}\phi\left(\sqrt{\frac{\sqrt{2}-1}{\sqrt{2}}}h\right)\Phi\left(\frac{h}{\sqrt{2-\sqrt{2}}} - \sqrt{4-2\sqrt{2}}z_0(\alpha)\right), \\
\mathscr{B}_2^m &= \frac{|\gamma|^{-1}}{\sqrt{2-\sqrt{2}}}\phi\left(\sqrt{\frac{\sqrt{2}-1}{\sqrt{2}}}h\right)\Phi\left(\frac{h}{\sqrt{2-\sqrt{2}}}\right) + |\gamma|^{-1}\sqrt{2}\phi(0)(1-\Phi(\sqrt{2}h).
\end{aligned}$$

The right-hand panel in Figure 2.7 shows the bias as a function of h for both approaches to p-hacking. For the threshold approach, calculations are for tests of size 5%. Estimates are more biased for smaller h. A larger h means that tests are more likely to reject anyway, so there is less likely reason to p-hack. Thus the bias is maximized when the null hypothesis is likely to be true. This indicates that it would be preferable for tests of p-hacking to have higher power when the majority of the underlying studies are examining hypotheses that are more likely to be correct. For the minimum approach, unlike the results of the previous subsection, the bias for the minimum approach is a function of h — this effect is due to the higher power of the test using two instruments resulting in that test being selected more as h is larger. Bias decreases in h as this test statistic becomes more dominant (since it is itself unbiased), but remains higher than that for the threshold approach since taking the minimum results in the researcher being better able to find a smaller p-value.

2.3.3 Selecting across Datasets

Consider a setting where a researcher conducts a finite number of $K > 1$ independent tests over which they can choose the best results. In each case the researcher uses a t-test to test their hypothesis, with test statistic $T_i \sim \mathscr{N}(h,1)$. We assume that the true local effect h is the same across datasets. This gives the researcher K possible p-values to consider, enabling the possibility of p-hacking. For example, a researcher conducting experiments with students, as is

[11] Since the first moments of the IV estimators do not exist in just-identified cases (Kinal, 1980), we define $\mathscr{B}_2^j$ to be the mean of the asymptotic distribution of $\sqrt{N}(\hat{\beta}_r^j - \beta)$, where $\hat{\beta}_r^j$ is the p-hacked estimate and $j \in \{m,t\}$.

common in experimental economics, could have several independent sets of students on which to test a hypothesis. As with the other examples, researchers could simply search over all datasets and report the smallest p-value or engage in a strategy of searching for a low p-value.

Let $K = 2$, and consider a search where first the researchers construct a dataset for their study and compute a p-value for their hypothesis on this dataset, then report this p-value if it is below size. Otherwise, they construct a new dataset and report the smallest of the two possible p-values (threshold approach). For illustration, we assume they use one-sided t-tests to test their hypothesis.

For the threshold approach, the p-curve is given by

$$g_3^t(p) = \int_{\mathscr{H}} \exp\left(hz_0(p) - \frac{h^2}{2}\right) \Upsilon_3^t(p;\alpha,h) d\Pi(h),$$

where

$$\Upsilon_3^t(p;\alpha,h) = \begin{cases} 1 + \Phi\left(z_h(\alpha)\right), & \text{if } p \leq \alpha, \\ 2\Phi\left(z_h(p)\right), & \text{if } p > \alpha. \end{cases}$$

This is a special case of the results in Section 2.3.1 where $\rho = 0$ because of the independence assumption across datasets. If the t-statistics were correlated through dependence between the datasets, then setting ρ equal to that correlation and using the results in Section 2.3.1 would yield the correct distribution of p-values.

Figure 2.8 shows p-curves for $h \in \{0,1,2\}$. For all values of h, no upward sloping p-curves are induced over any range of p. So for this type of p-hacking, even with thresholds such as in this example, tests that look for p-hacking through a lack of monotonically downward sloping p-curves will not have power. This method does suggest that tests for discontinuities in the distribution will have power, but likely only if studies have a large h.

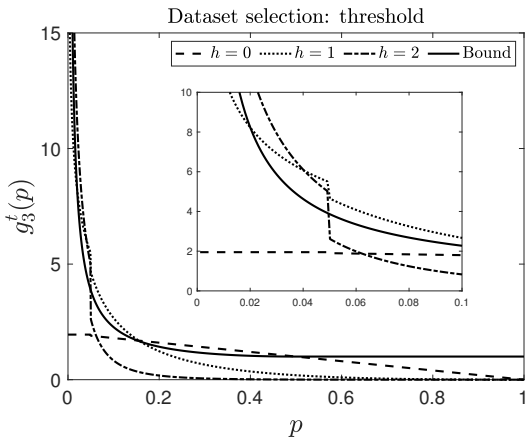

Figure 2.8. p-Curves from dataset selection based on the threshold approach with $\gamma = 0.5$.

An alternative strategy is to simply report the smallest of the p-values across all datasets or subsamples (e.g., Ulrich and Miller, 2015; Elliott et al., 2022b). For general K, the p-curve is given by

$$g_3^m(p;K) = K \int_{\mathscr{H}} \exp\left(hz_0(p) - \frac{h^2}{2}\right) \Phi(z_h(p))^{K-1} d\Pi(h). \tag{2.2}$$

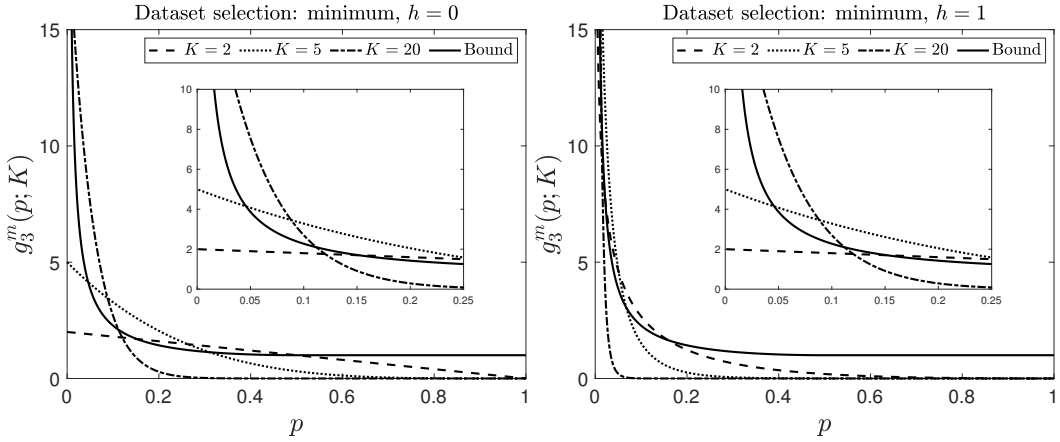

Figure 2.9. p-Curves from dataset selection based on the minimum approach. Left panel: $h = 0$. Right panel: $h = 1$.

The p-curve under p-hacking, g_3^m, is non-increasing (completely monotone) whenever the distribution with no p-hacking is non-increasing (completely monotone) (Elliott et al., 2022b). This can be seen in Figure 2.9 where for various K and h each of the curves are decreasing. Tests for violations of monotonicity will have no power. Similarly, tests for discontinuities will also

not have power. Figure 2.9 also shows (solid line) the bounds under the null hypothesis of no p-hacking. Clearly, each of the curves violates the bounds for some range of p; see also Figure 2 in Elliott et al. (2022b).

Alternatively, the researcher could consider the threshold strategy of first using both datasets, choosing to report this p-value if it is below a threshold and, otherwise, choosing the best of the available p-values. For $K = 2$, this gives three potential p-values to choose between. For many such testing problems (for example, testing a regression coefficient in a linear regression), $T_k \sim \mathcal{N}(h,1)$, $k = 1,2$, approximately so that the t-statistic from the combined samples is $T_{12} \simeq (T_1 + T_2)/\sqrt{2}$. This is precisely the same setup asymptotically as in the IV case presented above, so those results apply directly to this problem. As such, we refer to the discussion there rather than re-present the results.

2.3.4 Variance Bandwidth Selection for Means

In time series regression, sums of random variables such as means or regression coefficients are standardized by an estimate of the spectral density of the relevant series at frequency zero. A number of estimators exist; the most popular in practice is a nonparametric estimator that takes a weighted average of covariances of the data. With this method, researchers are confronted with a choice of the bandwidth for estimation. Different bandwidth choices allow for multiple chances at constructing p-values, hence allowing for the potential for p-hacking.

To examine this analytically, consider the model $Y_t = \beta + U_t$, $t = 1,\ldots,N$, where we assume that $U_t \overset{iid}{\sim} \mathcal{N}(0,1)$. We can consider two statistics for testing the null hypothesis that the mean is zero versus a positive mean. First the usual t-statistic testing the null of zero, $T_0 = \sqrt{N}\bar{Y}_N$, and, secondly, $T_1 = (\sqrt{N}\bar{Y}_N)/\hat{\omega}$, where $\bar{Y}_N = N^{-1}\sum_{t=1}^{N} Y_t$, $\hat{\omega}^2 := \omega^2(\hat{\rho}) := 1 + 2\kappa\hat{\rho}$ and $\hat{\rho} = (N-1)^{-1}\sum_{t=2}^{N} \hat{U}_t\hat{U}_{t-1}$. Here κ is the weight in the spectral density estimator. For example, in the Newey and West (1987) estimator with one lag, $\kappa = 1/2$.

In line with the previous subsections, we consider both a threshold approach to p-hacking as well as simply choosing the best p-value from a set. In the threshold approach, the researcher

constructs T_0 and calculates the corresponding p-value. If it is below $\alpha \leq 1/2$, this p-value is reported. Otherwise, the researcher calculates T_1 and reports the smaller of the p-values from the two t-statistics.[12] In the second approach, the smallest p-value of the two computed is reported.

In Appendix B.1.4, we show that the distribution of p-values has the form

$$g_4^t(p) = \int_{\mathscr{H}} \exp\left(hz_0(p) - \frac{h^2}{2}\right) \Upsilon_4^t(p;\alpha,h,\kappa) d\Pi(h),$$

with $\Upsilon_4(p;\alpha,h,\kappa)$ taking different forms over different parts of the support of the distribution. Define $l(p) = (2\kappa)^{-1}\left(\left(\frac{z_0(\alpha)}{z_0(p)}\right)^2 - 1\right)$ and let H_N and η_N be the CDF and PDF of $\hat{\rho}$, respectively. Then we have

$$\Upsilon_4^t = \begin{cases} 1 + \frac{1}{\phi(z_h(p))} \int_{-(2\kappa)^{-1}}^{l(p)} \omega(r)\phi(z_0(p)\omega(r) - h)\eta_N(r)dr, & \text{if } 0 < p \leq \alpha, \\ 1 - H_N(0) + H_N(-(2\kappa)^{-1}) + \frac{1}{\phi(z_h(p))} \int_{-(2\kappa)^{-1}}^{0} \omega(r)\phi(z_0(p)\omega(r) - h)\eta_N(r)dr, & \text{if } \alpha < p \leq 1/2, \\ H_N(0) + \frac{1}{\phi(z_h(p))} \int_{0}^{\infty} \omega(r)\phi(z_0(p)\omega(r) - h)\eta_N(r)dr, & \text{if } 1/2 < p < 1. \end{cases}$$

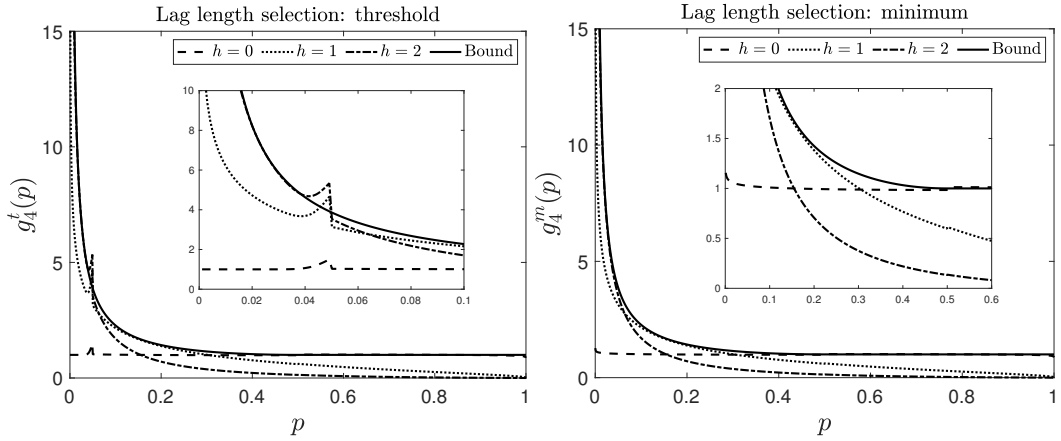

Figure 2.10. p-Curves from lag length selection with $N = 200$ and $\kappa = 1/2$. Left panel: thresholding. Right panel: minimum.

The left-hand side panel in Figure 2.10 presents the p-curves for the thresholding case.

[12] If $\hat{\rho}$ is such that $\hat{\omega}^2$ is negative, the researcher always reports the initial result.

Notice that, unlike the earlier examples, thresholding creates the intuitive hump in the p-curve at the chosen size (here 0.05) for all of the values for h. Thus tests that attempt to find such humps may have power. Discontinuities at the chosen size also occur. For h large enough, the p-curves also violate the bounds for smaller p-values.

When the minimum over the two p-values is chosen, the p-curve is given by

$$g_4^m(p) = \int_{\mathscr{H}} \exp\left(hz_0(p) - \frac{h^2}{2}\right) \Upsilon_4^m(p;\alpha,h,\kappa)d\Pi(h),$$

where

$$\Upsilon_4^m = \begin{cases} 1 - H_N(0) + H_N(-(2\kappa)^{-1}) + \frac{1}{\phi(z_h(p))}\int_{-(2\kappa)^{-1}}^{0} \omega(r)\phi(z_0(p)\omega(r) - h)\eta_N(r)dr, & \text{if } 0 < p \leq 1/2, \\ H_N(0) + \frac{1}{\phi(z_h(p))}\int_{0}^{\infty} \omega(r)\phi(z_0(p)\omega(r) - h)\eta_N(r)dr, & \text{if } 1/2 < p < 1. \end{cases}$$

The right-hand side panel in Figure 2.10 presents the p-curves for the minimum case. When p-hacking works through taking the minimum p-value, as in earlier cases for p-values near commonly used sizes, the impact is to move the distributions towards the left, making the p-curves fall more steeply. The effect here is modest and likely difficult to detect. Of more interest is what happens at $p = 0.5$, where taking the minimum (this effect is also apparent in the thresholding case) results in a discontinuity. The reason for this is that choices over the denominator of the t-statistic used to test the hypothesis cannot change the sign of the t-test. Within each side, the effect is to push the distribution to the left, so this results in a discontinuity at $p = 0.5$. This effect will extend to all methods where p-hacking is based on searching over different choices of variance covariance matrices — for example, different choices in estimators, different choices in the number of clusters, etc. Figure 2.10 shows that for $h = 1,2$ the bound is not reached, and any discontinuity at $p = 0.5$ is very small. For $h = 0$, the bound is slightly below the p-curve after the discontinuity.

For size at $h = 0$ and $p = \alpha$ we have

$$G_0(\alpha) \quad = \quad \alpha + (1-\alpha)(H_N(0) - H_N(-(2\kappa)^{-1})) - \int_{-(2\kappa)^{-1}}^{0} \Phi(z_0(\alpha)\omega(r) - h)\eta_N(r)dr$$

Figure 2.11 shows that the size distortions through this example of p-hacking are quite modest. The reason is that for a reasonable sample size, the estimated first-order correlation is very close to zero. Thus estimated standard errors when an additional lag is included are very close to one, meaning that the two t-statistics are quite similar and very highly correlated. This means that there is not much room to have size distortions due to this p-hacking.

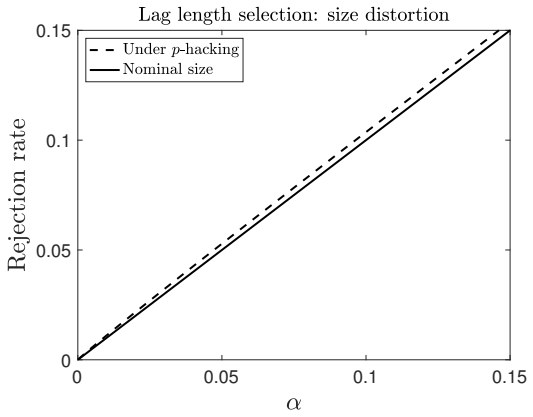

Figure 2.11. Rejection rate under p-hacking from lag length selection with $N = 200$ and $\kappa = 1/2$.

2.4 Statistical Tests for p-Hacking

In this section, we discuss several statistical tests for the null hypothesis of no p-hacking based on a sample of n p-values, $\{P_i\}_{i=1}^n$. We do not consider Caliper tests based on the distribution of t-statistics (Gerber and Malhotra, 2008a,b) because these tests do not control size (Kudrin, 2022).

2.4.1 Histogram-based Tests for Combinations of Restrictions

Histogram-based tests (Elliott et al., 2022b) provide a flexible framework for constructing tests for different combinations of testable restrictions. Let $0 = x_0 < x_1 < \cdots < x_J = 1$ be

an equidistant partition of $[0,1]$ and define the population proportions $\pi_j = \int_{x_{j-1}}^{x_j} g(p)dp$, $j = 1,\ldots,J$. The main idea of histogram-based tests is to express the testable implications of p-hacking in terms of restrictions on the population proportions $(\pi_1,\ldots,\pi_J)$. For instance, non-increasingness of the p-curve implies that $\pi_j - \pi_{j-1} \leq 0$ for $j = 2,\ldots,J$. More generally, Elliott et al. (2022b) show that K-monotonicity[13] restrictions and upper bounds on the p-curve and its derivatives can be expressed as $H_0 : A\pi_{-J} \leq b$, for a matrix A and vector b, where $\pi_{-J} := (\pi_1,\ldots,\pi_{J-1})'$.[14]

To test this hypothesis, we estimate π_{-J} by the vector of sample proportions $\hat{\pi}_{-J}$. The estimator $\hat{\pi}_{-J}$ is asymptotically normal with mean π_{-J} so that the testing problem can be recast as the problem of testing affine inequalities about the mean of a multivariate normal distribution (e.g., Kudo, 1963; Wolak, 1987; Cox and Shi, 2022). Following Elliott et al. (2022b), we use the conditional chi-squared test of Cox and Shi (2022), which is easy to implement and remains computationally tractable when J is moderate or large.

2.4.2 Tests for Non-Increasingness of the p-Curve

A popular test for non-increasingness of the p-curve is the Binomial test (e.g., Simonsohn et al., 2014; Head et al., 2015), where researchers compare the number of p-values in two adjacent bins right below significance cutoffs. Under the null of no p-hacking, the fraction of p-values in the bin closer to the cutoff should be weakly smaller than the fraction in the bin farther away. Implementation is typically based on an exact Binomial test. Binomial tests are "local" tests that ignore information about the shape of the p-curve farther away from the cutoff, which often leads to low power in our simulations. A "global" alternative is Fisher's test (e.g., Simonsohn et al., 2014).[15]

In addition to the classical Binomial test and Fisher's test, we consider tests based on the

[13] A function g is K-monotone if $0 \leq (-1)^k g^{(k)}$ for and all $k = 0,1,\ldots,K$, where $g^{(k)}$ is the kth derivative of g. Complete monotonicity implies K-monotonicity for all K.

[14] Here we incorporate the adding up constraint $\sum_{j=1}^{J} \pi_j = 1$ into the definition of A and b and express the testable implications in terms of the "core moments" $(\pi_1,\ldots,\pi_{J-1})$ instead of $(\pi_1,\ldots,\pi_J)$.

[15] An alternative to Fisher's test is Stouffer's method (Simonsohn et al., 2015).

least concave majorant (LCM) (Elliott et al., 2022b).[16] LCM tests are based on the observation that non-increasingness of g implies that the CDF G is concave. Concavity can be assessed by comparing the empirical CDF of p-values, $\hat{G}$, to its LCM $\mathscr{M}\hat{G}$, where $\mathscr{M}$ is the LCM operator. We choose the test statistic $\sqrt{n}\|\hat{G} - \mathscr{M}\hat{G}\|_\infty$. The uniform distribution is least favorable for the LCM test (Kulikov and Lopuhaä, 2008; Beare, 2021), and critical values can be obtained via simulations.

2.4.3 Tests for Continuity of the p-Curve

Continuity of the p-curve at pre-specified cutoffs $p = \alpha$ can be assessed using standard density discontinuity tests (e.g., McCrary, 2008; Cattaneo et al., 2020). Following Elliott et al. (2022b), we use the approach by Cattaneo et al. (2020) with automatic bandwidth selection implemented in the `R`-package `rddensity` (Cattaneo et al., 2021).

2.5 Monte Carlo Simulations

In this section, we investigate the finite sample properties of the tests in Section 2.4 using a Monte Carlo simulation study. The Monte Carlo study is based on generalizations of the analytical examples of p-hacking in Section 2.3. We do not consider selection across datasets, as this example can be viewed as a special case of covariate and IV selection.

2.5.1 Generalized p-Hacking Examples

In all examples that we consider, researchers are interested in testing a hypothesis about a scalar parameter β:

$$H_0 : \beta = 0 \qquad \text{against} \qquad H_1 : \beta > 0. \tag{2.3}$$

The results for two-sided tests of $H_0 : \beta = 0$ against $H_1 : \beta \neq 0$ are similar. See Figure 2.13.

Researchers may p-hack their initial results by exploring additional model specifications

[16] LCM tests have been successfully applied in many different contexts (e.g., Carolan and Tebbs, 2005; Beare and Moon, 2015; Fang, 2019).

or estimators and report a different result of their choice. Specifically, we consider the two general approaches to p-hacking discussed in Section 2.3: the threshold and the minimum approach. In what follows, we discuss the generalized examples of p-hacking in more detail.

2.5.1.1 Selecting Control Variables in Linear Regression

Researchers have access to a random sample with $N = 200$ observations generated as $Y_i = X_i\beta + u_i$, where $X_i \sim \mathcal{N}(0,1)$ and $u_i \sim \mathcal{N}(0,1)$ are independent of each other. There are K additional control variables, $Z_i := (Z_{1i}, \ldots, Z_{Ki})'$, which are generated as

$$Z_{ki} = \gamma_k X_i + \sqrt{1-\gamma_k^2}\varepsilon_{Z_k,i}, \quad \varepsilon_{Z_k,i} \sim \mathcal{N}(0,1), \quad \gamma_k \sim U[-0.8, 0.8], \quad k = 1, \ldots, K.$$

We set $\beta = h/\sqrt{N}$ with $h \in \{0,1,2\}$ and show results for $h \sim \chi^2(1)$ in the Appendix.

Researchers use either a threshold or a minimum approach to p-hacking.

Threshold approach. Researchers regress Y_i on X_i and Z_i and test (3.12). Denote the resulting p-value as P. If $P \leq 0.05$, the researchers report the p-value. If $P > 0.05$, they regress Y_i on X_i, trying all $(K-1) \times 1$ subvectors of Z_i as controls and select the result with the smallest p-value. If the smallest p-value is larger than 0.05, they continue and explore all $(K-2) \times 1$ subvectors of Z_i etc. If all results are insignificant, they report the smallest p-value.

Minimum approach. Researchers run regressions of Y_i on X_i and each possible configuration of covariates Z_i and report the minimum p-value.

Figures C.4, C.5, and C.6 show the null and p-hacked distributions for $K \in \{3,5,7\}$ and figure B.4 shows the null and p-hacked distributions for $K = 3$ when researchers report p-values for two-sided tests.[17] The p-curves are similar to those in the simple analytical example of Section 2.3.1. The threshold approach leads to a discontinuity in the p-curve and may lead

[17] To generate these distributions, we run the algorithm one million times and collect p-hacked and non-p-hacked results.

to non-increasing p-curves and humps below significance thresholds. By contrast, reporting the minimum p-value across all possible specifications generally leads to continuous and non-increasing p-curves. The distribution of h is an important determinant of the shape of the p-curve, especially when researchers use the threshold approach. The larger h, the higher the probability that the researchers find significant results in the initial specification and thus will not engage in further specification search. Finally, as expected, the violations of the testable restrictions are more pronounced when K is large, that is when researchers have many degrees of freedom.

2.5.1.2 Selecting amongst Instruments in IV Regression

Researchers have access to a random sample with $N = 200$ observations generated as

$$\begin{aligned} Y_i &= X_i\beta + U_i, \\ X_i &= Z_i'\pi + V_i, \end{aligned}$$

where $U_i \sim \mathcal{N}(0,1)$ and $V_i \sim \mathcal{N}(0,1)$ with $\text{Cov}(U_i, V_i) = 0.5$. The instruments $Z_i := (Z_{1i}, \ldots, Z_{Ki})'$ are generated as

$$Z_{ki} = \gamma_k \xi_i + \sqrt{1-\gamma_k^2}\varepsilon_{Z_k,i}, \quad \xi_i \sim \mathcal{N}(0,1), \quad \varepsilon_{Z_k,i} \sim \mathcal{N}(0,1), \quad \gamma_k \sim U[-0.8, 0.8], \quad k = 1, \ldots, K,$$

where $\xi_i, \varepsilon_{Z_k,i}$, and γ_k are independent for all k. Also, $\pi_k \overset{iid}{\sim} U[1,3]$, $k = 1, \ldots, K$. We set $\beta = h/(3\sqrt{N})$ with $h \in \{0,1,2\}$ and show results for $h \sim \chi^2(1)$ in the Appendix.

Researchers use either a threshold or a minimum approach to p-hacking.

Threshold approach. Researchers estimate the model using all instruments Z_i, test (3.12), and obtain the p-value P. If $P \leq 0.05$, the researchers report the p-value. If $P > 0.05$, they try all $(K-1) \times 1$ subvectors of Z_i as instruments and select the result corresponding to the smallest p-value. If the smallest p-value is larger than 0.05, they continue and explore all $(K-2) \times 1$ subvectors of Z_i etc. If all results are insignificant, they report the smallest

p-value.

Minimum approach. The researchers run IV regressions of Y_i on X_i using each possible configuration of instruments and report the minimum p-value.

Figures B.5 and B.6 display the null and p-hacked distributions for $K \in \{3,5\}$. We do not show results for $K = 7$ since there is a very high concentration of p-values at zero in this case. The p-curves in the general case here are similar to those in the simple analytical example of Section 2.3.2. As with covariate selection, the threshold approach yields discontinuous p-curves and may lead to non-increasingness and humps, whereas reporting the minimum p-value leads to continuous and decreasing p-curves. The distribution of h and the number of instruments, K, are important determinants of the shape of the p-curve. We note that the distribution of p-values under the null hypothesis of no p-hacking when $h = 0$ is not exactly uniform because of the relatively small sample size.

2.5.1.3 Lag Length Selection in Regression

Researchers have access to a random sample with $N = 200$ observations from $Y_t = X_t\beta + U_t$, where $X_t \sim \mathcal{N}(0,1)$ and $U_t \sim \mathcal{N}(0,1)$ are independent. We set $\beta = h/\sqrt{N}$ with $h \in \{0,1,2\}$ and show results for $h \sim \chi^2(1)$ in the Appendix.

Researchers use either a threshold or a minimum approach to p-hacking.

Threshold approach. Researchers first regress Y_t on X_t and calculate the standard error using the classical Newey-West estimator with the number of lags selected using the Bayesian Information Criterion (researchers only choose up to 4 lags). They then use a t-test to test (3.12) and calculate the p-value P. If $P \leq 0.05$, the researchers report the p-value. If $P > 0.05$, they try Newey-West estimator with one extra lag. If the result is not significant, they try two extra lags etc. If all results are insignificant, they report the smallest p-value.

Minimum approach. Researchers regress Y_t on X_t, calculate the standard error using Newey-West with 0 to 4 lags, and report the minimum p-value.

The null and p-hacked distributions are displayed in Figure B.7. The p-curves exhibit features similar to those in the simple analytical example in Section 2.3.4. The threshold approach induces a sharp spike right below 0.05. The reason is that p-hacking via judicious lag selection does not lead to huge improvements in terms of p-value. Both approaches to p-hacking lead to a discontinuity at 0.5.

2.5.2 Simulations

2.5.2.1 Setup

We model the distribution of reported p-values as a mixture:

$$g^o(p) = \tau \cdot g^d(p) + (1-\tau) \cdot g^{np}(p)$$

Here, g^d is the distribution under the different p-hacking approaches described above; g^{np} is the distribution in the absence of p-hacking (i.e., the distribution of the first p-value that the researchers obtain). The parameter $\tau \in [0,1]$ captures the fraction of researchers who engage in p-hacking.

For our main results, we focus on settings without publication bias and set the publication probability equal to one, irrespective of the reported p-value, $\Pr(S = 1 \mid p) = 1$. To assess the impact of publication bias, we also consider two types of publication bias that differ with respect the publication probability $\Pr(S = 1 \mid p)$.

Sharp publication bias. The publication probability $\Pr(S = 1 \mid p)$ is a step function. We set the probability of publishing a result that is significant at the 5% level to one, $\Pr(S = 1 \mid p) = 1$ for $p \leq 0.05$, and the probability of publishing an insignificant result to 0.1, $\Pr(S = 1 \mid p) = 0.1$ for $p > 0.05$. Hence, significant results are 10 times more likely to be published than insignificant ones.

Smooth publication bias. The publication probability is a smooth function of the reported p-value. We set $\Pr(S = 1 \mid p) = \exp(-A \cdot p)$, where we choose $A = 8.45$ to make results comparable across both types of publication bias.[18]

To generate the data, we first simulate the p-hacking algorithms one million times to obtain samples corresponding to g^d and g^{np}. Then, to construct samples in every Monte Carlo iteration, we draw $n = 5000$ p-values with replacement from a mixture of those samples and keep each p-value, p_i, with probability $\Pr(S = 1 \mid p_i)$. Following Elliott et al. (2022b), we apply the tests to the subinterval $(0, 0.15]$. Therefore, the effective sample size depends on the p-hacking strategy, the distribution of h, the presence and type of publication bias, and the fraction of p-hackers τ.

Table 2.1. Tests for p-hacking

Testable restriction: non-increasingness of p-curve	
CS1	Histogram-based test based on Cox and Shi (2022) with $J = 15$
LCM	LCM test
Binomial	Binomial test with bins $[0.40, 0.45)$ and $[0.45, 0.50]$
Testable restriction: continuity of p-curve	
Discontinuity	Density discontinuity test (Cattaneo et al., 2021)
Testable restriction: upper bounds on p-curve, 1st, and 2nd derivative	
CSUB	Histogram-based test based on Cox and Shi (2022) with $J = 15$
Testable restriction: 2-monotonicity and upper bounds on p-curve, 1st, and 2nd derivative	
CS2B	Histogram-based test based on Cox and Shi (2022) with $J = 15$

We compare the finite sample performance of the tests described in Section 2.4. See

[18] When $A = 8.45$, the ratio between $\int_0^{0.05} \Pr(S = 1 \mid p)dp$ and $\int_{0.05}^1 \Pr(S = 1 \mid p)dp$ is the same for both types of publication bias.

Table 2.1 for more details.[19] We do not show results for Fisher's test since we found that this test has essentially no power for detecting the types of p-hacking we consider. The simulations are implemented using `MATLAB` (MATLAB, 2020) and `R` (R Core Team, 2022).

2.5.2.2 Power Curves

In this section, we present power curves for the different data generating processes (DGPs). For covariate and instrument selection, we focus on the results for $K = 3$ in the main text and present the results for larger values of K in Appendix B.3. The nominal level is 5% for all tests. All results are based on 5000 simulation draws. Figures 2.12–2.15 present the results.

The power for detecting p-hacking crucially depends on the type of p-hacking, the econometric method, the fraction of p-hackers, τ, and the value of h. As shown in Section 2.3, when researchers p-hack using a threshold approach, the p-curves are discontinuous at the threshold, may violate the upper bounds, and may be non-monotonic; see also Appendix C.5. Thus, tests exploiting these testable restrictions may have power when the fraction of p-hackers is large enough.

The CS2B test, which exploits monotonicity restrictions and bounds, has the highest power overall. However, this test may exhibit some small size distortions when the effective sample size is small (e.g., lag length selection with $h = 0$). Among the tests that exploit monotonicity of the entire p-curve, the CS1 test typically exhibits higher power than the LCM test. The LCM test can exhibit non-monotonic power curves because the test statistic converges to zero in probability for strictly decreasing p-curves (Beare and Moon, 2015).

The widely-used Binomial test often exhibits low power. The reason is that the p-hacking approaches we consider do not lead to isolated humps or spikes near 0.05, even if researchers use a threshold p-hacking approach. There is one notable exception. When researchers engage in variance bandwidth selection, our theoretical results show that p-hacking based on the threshold

[19] For CS1, CSUB and CS2B tests, the optimization routine fails to converge for some realizations of the data due to nearly singular covariance matrix estimates. We count these cases as non-rejections of the null in our Monte Carlo simulations.

approach can yield isolated humps right below the cutoff (see Figure 2.10 and also Figure B.7). By construction, the Binomial test is well-suited for detecting this type of p-hacking and is among the most powerful tests in this case. Our results for the Binomial test demonstrate the inherent disadvantage of using tests that only exploit testable implications locally. Such tests only have power against very specific forms of p-hacking, which limits their usefulness in practice.

Discontinuity tests are a useful complement to tests based on monotonicity and upper bounds because p-hacking based on threshold approaches often yields pronounced discontinuities. These tests are particularly powerful for detecting p-hacking based on lag length selection, which leads to spikes and pronounced discontinuities at 0.05 as discussed above.

When researchers always report the minimum p-value, the power of the tests is much lower than when they use a threshold approach. The minimum approach to p-hacking does not lead to violations of monotonicity and continuity over $p \in (0, 0.15]$. Therefore, by construction, tests based on these restrictions have no power, irrespective of the fraction of researchers who are p-hacking.

In Section 2.3, we show theoretically that the minimum approach may yield violations of the upper bounds. The range over which the upper bounds are violated and the extent of these violation depend on h and the econometric method used by the researchers (see Figures 2.2, 2.5, and 2.10). Consistent with the analysis in Section 2.3, the simulations show a moderate amount of power for the tests based on upper bounds (CSUB and CS2B) for IV selection with $h = 0$ and $h = 1$ when a sufficiently large fraction of researchers p-hacks. Our theoretical results show that the violations of the upper bounds may be small, so the moderate power of these tests is not unexpected.

Under the minimum approach, the power curves of the CSUB and CS2B tests are very similar, suggesting that the power of the CS2B test comes mainly from using upper bounds. This finding demonstrates the importance of exploiting upper bounds in addition to monotonicity and continuity restrictions in practice. Figure 2.14 further shows that the power of the CSUB and the CS2B test may not be monotonic in h. On the one hand, for large h, there are more p-values close

to zero, where the upper bounds are more difficult to violate. On the other hand, the effective sample size increases with h, leading to more power. Finally, the results for covariate and IV selection in Appendix B.3 show that the larger K — the more degrees of freedom the researchers have when p-hacking — the higher the power of the CSUB and CS2B test.

Finally, we compare the results in Figure 2.12, which are based on researchers testing the one-sided hypothesis (3.12), to those in Figure 2.13, which are based on researchers engaging in covariate selection based on two-sided tests. While there are some differences between the power curves, the overall rankings of the tests in terms of their power properties are the same. Interestingly, the power of the CSUB and the CS2B test under the minimum approach can be higher when the researchers use two-sided tests.

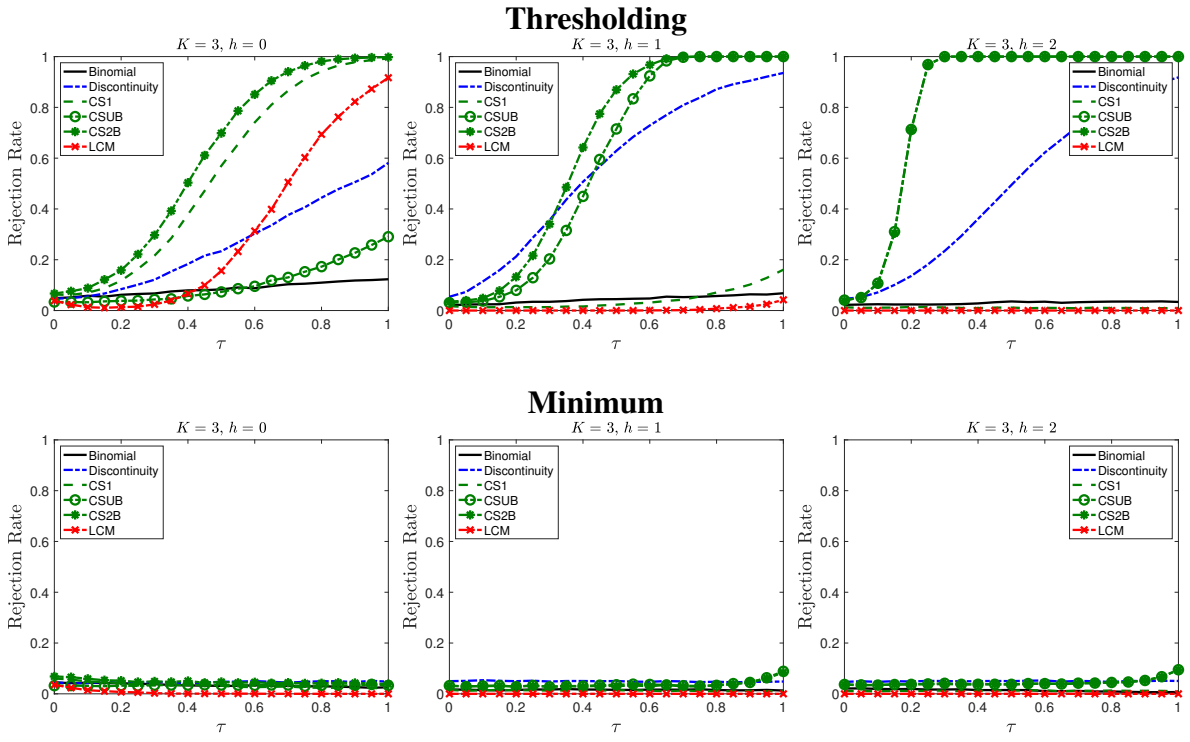

Figure 2.12. Power curves covariate selection with $K = 3$.

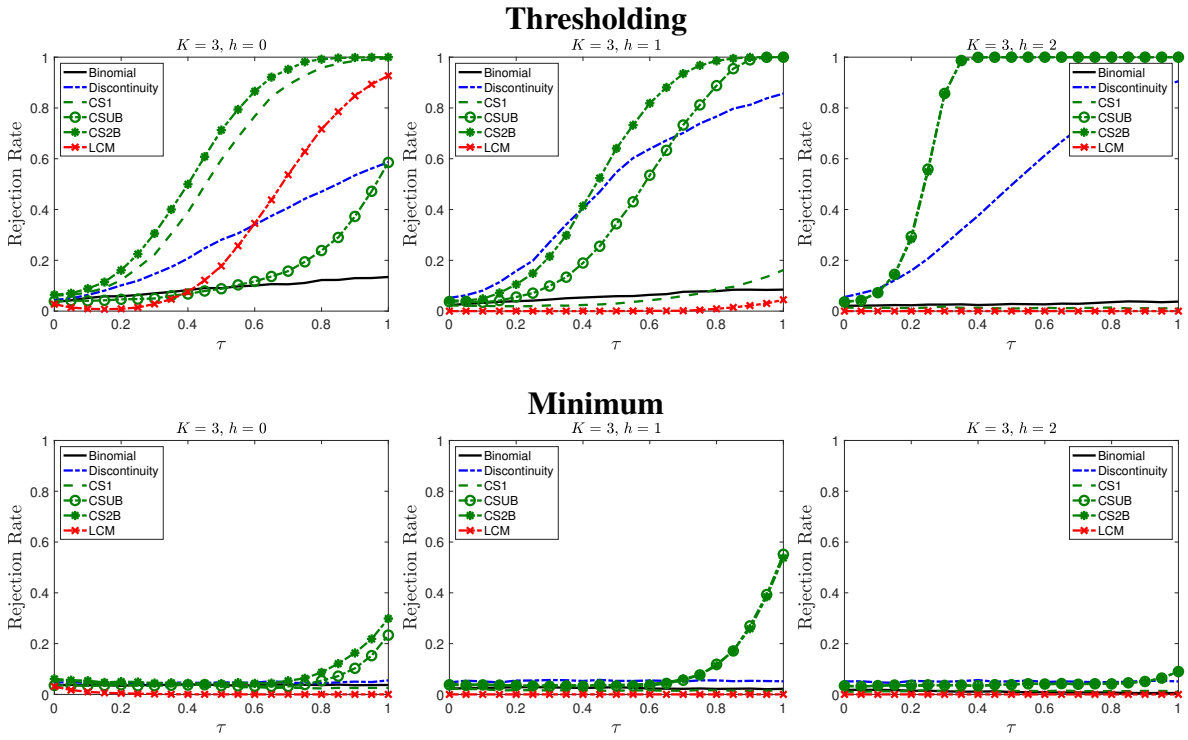

Figure 2.13. Power curves covariate selection with $K = 3$ (2-sided tests).

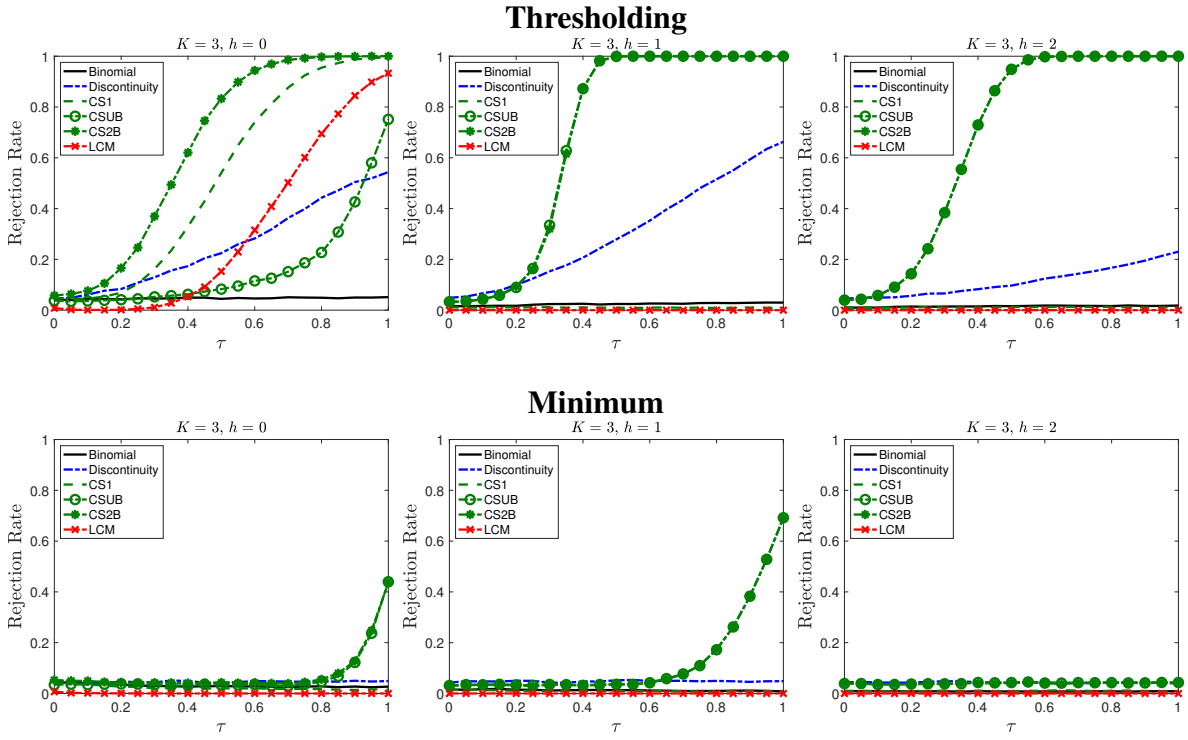

Figure 2.14. Power curves IV selection with $K = 3$.

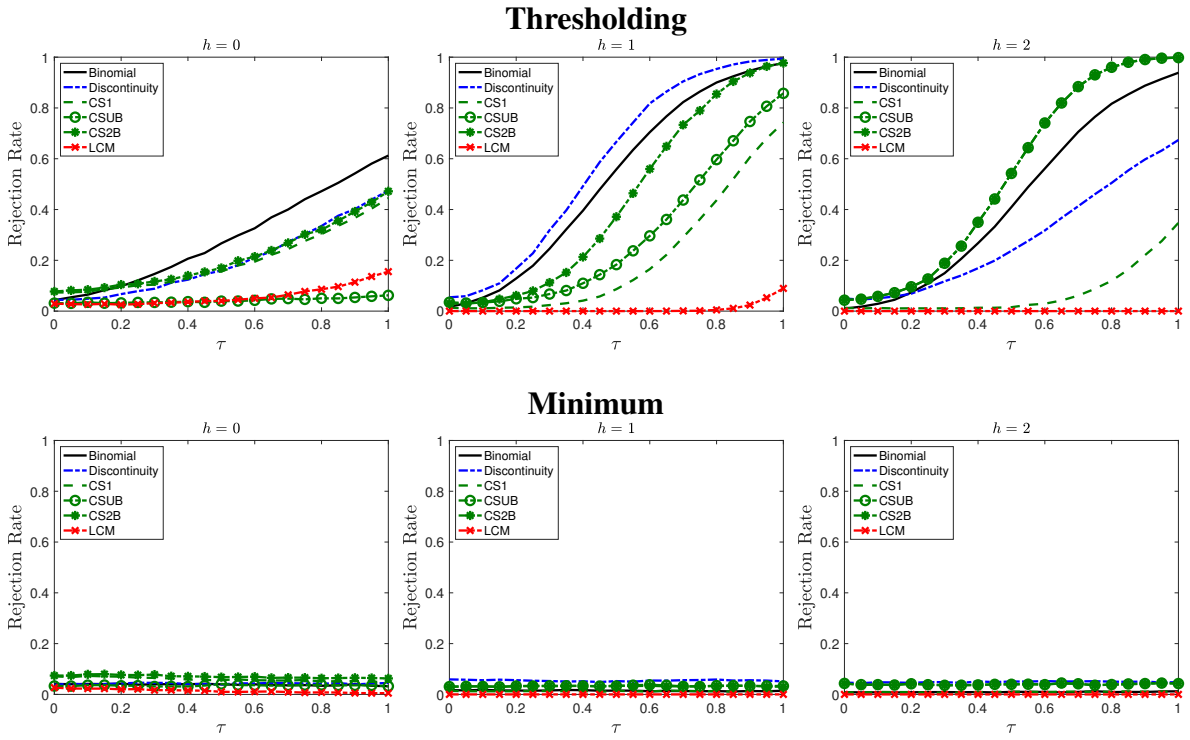

Figure 2.15. Power curves lag length selection.

Overall, the tests' ability to detect *p*-hacking is highly context-specific and can be low in some cases. As we show theoretically in Section 2.3, this is because *p*-hacking may not lead to violations of the testable restrictions used by the statistical tests for *p*-hacking. Moreover, even if *p*-hacking leads to violations of the testable restrictions, these violations may be small and can thus only be detected based on large samples of *p*-values. Regarding the choice of testable restrictions, the simulations demonstrate the importance of exploiting upper bounds in addition to monotonicity and continuity for constructing powerful tests against plausible *p*-hacking alternatives.

2.5.2.3 Power vs. Costs of *p*-Hacking

Our simulation results show that the tests' ability to detect *p*-hacking crucially depends on the shape of the *p*-curve under the alternative of *p*-hacking, which depends on the type of *p*-hacking, the econometric method, the distribution of effects, and the fraction of *p*-hackers.

This result is expected since the alternative space of p-curves under p-hacking is very large. It begs the question: What alternatives are relevant for empirical practice?

To determine which alternatives are relevant, we consider the costs of p-hacking. As discussed in Section 2.3, there are at least two types of costs: size distortions when $h = 0$ and bias in the estimated coefficients. Here we focus on the bias, which is relevant for all values of h. Our theoretical results show that the bias is particularly large for small values of h. This is intuitively clear in that this is where there is more need for p-hacking to get small p-value.

In Figure 2.16, we plot the relationship between bias and power for the threshold and the minimum approach across all DGPs and all K. We compare three tests: the classical Binomial test, the discontinuity test, and the CS2B test (the most powerful test overall). Each dot in Figure 2.16 corresponds to the power of one test under one DGP.[20] We only show results for covariate and IV selection; the bias under lag length selection is always zero. For the threshold approach, we present results for $\tau = 0.25$. For the minimum approach, we set $\tau = 0.75$ since no test has non-trivial power for $\tau = 0.25$.

When researchers p-hack using a threshold approach, there is a positive association between the average bias and the power of all three tests: the higher the costs of p-hacking, the higher the power of the tests on average. The CS2B test has power close to one when the bias is large. This test is able to detect p-hacking with high probability when it is costly, even when only 25% of the researchers are p-hacking. Although less powerful than the CS2B test, the discontinuity test also has high power in settings where p-hacking yields large biases. By contrast, the power of the Binomial test does not exceed 30%, even when p-hacking leads to large biases.

p-Hacking based on the minimum approach is difficult to detect. This type of p-hacking does not lead to violations of continuity and monotonicity. As a result, the discontinuity and the Binomial test, by construction, have no power, irrespective of the magnitude of the bias. Our simulations confirm this. Therefore, we only discuss results for the CS2B test, which may

[20] We only include results for one-sided tests and exclude the results for two-sided tests in Figure 2.13.

have power since the minimum approach may yield p-curves that violate the upper bounds. Our simulation results suggest that the relationship between bias and power crucially depends on the econometric method. The CS2B test does not have meaningful power for covariate selection, irrespective of the bias. By contrast, there is again a positive relationship between bias and power for IV selection.

Overall, our results show that whenever the tests have non-trivial power, there is a positive association between their power and the bias from p-hacking. This is desirable from a meta-analytic perspective. However, we also document cases where the proposed tests do not have non-trivial power, even when p-hacking is very costly.

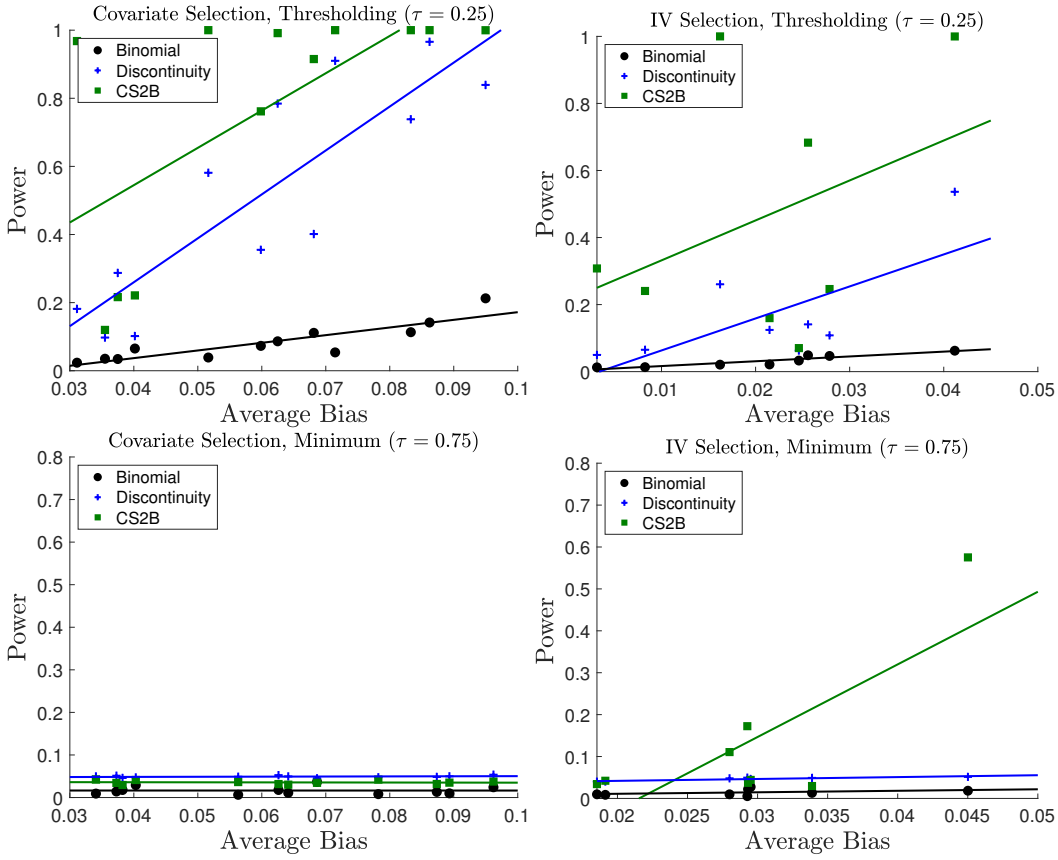

Figure 2.16. Power vs. bias.

2.5.2.4 The Impact of Publication Bias

Here we investigate the impact of publication bias on the power of the tests for testing the joint null hypothesis of no p-hacking and no publication bias. Table 2.2 presents the results for $h = 0$ and $\tau = 0.5$. In Appendix B.3, we also report results for $h \in \{1,2\}$ and $h \sim \chi^2(1)$.

The impact of publication bias on power depends on the testable restrictions that the tests exploit. Both types of publication bias can substantially increase the power of the CSUB and the CS2B test, which exploit upper bounds. This is expected since both forms of publication bias favor small p-values, which leads steeper p-curves that are more likely to violate the upper bounds, as discussed in Section 2.2.4. The difference in power with and without publication bias is particularly stark under the minimum approach to p-hacking: publication bias can lead to nontrivial power even when the CSUB and the CS2B test have very low power for detecting p-hacking.

For the tests based on monotonicity of the entire p-curve (CS1 and LCM), the results depend on the type of publication bias. Sharp publication bias tends to increase power, whereas smooth publication bias can lower power. Due to its local nature, sharp publication bias does not increase the power of the Binomial test. This again demonstrates the advantages of using "global" tests.

Sharp publication bias accentuates existing discontinuities and leads to discontinuities in otherwise smooth p-curves. It is thus not surprising that the discontinuity test is much more powerful under sharp publication bias. By contrast, smooth publication bias can decrease the power of the discontinuity test.

Overall, our results suggest that publication bias, sharp publication bias in particular, can lead to high power, even in settings where p-hacking is difficult to detect. This finding is relevant when interpreting empirical results. Specifically, we have documented several cases where even the best tests exhibit low power for detecting p-hacking. In such cases, rejections are likely due to the presence of publication bias.

Table 2.2. The effect of publication bias: 1-sided tests, $h = 0$, $\tau = 0.5$

	Test					
	Binomial	Discontinuity	CS1	CSUB	CS2B	LCM
Cov Selection ($K = 3$, thresholding)						
No Pub Bias	0.083	0.234	0.57	0.073	0.698	0.156
Sharp Pub Bias	0.083	0.997	0.635	0.998	0.998	0.764
Smooth Pub Bias	0.052	0.155	0.059	1	1	0.006
Cov Selection ($K = 3$, minimum)						
No Pub Bias	0.031	0.047	0.032	0.034	0.047	0.001
Sharp Pub Bias	0.031	0.931	0.081	0.998	0.998	0.001
Smooth Pub Bias	0.024	0.045	0.012	1	1	0
IV Selection ($K = 3$, thresholding)						
No Pub Bias	0.045	0.224	0.541	0.083	0.833	0.153
Sharp Pub Bias	0.045	0.994	0.583	1	1	0.736
Smooth Pub Bias	0.033	0.133	0.046	1	1	0.004
IV Selection ($K = 3$, minimum)						
No Pub Bias	0.029	0.042	0.02	0.033	0.043	0
Sharp Pub Bias	0.029	0.95	0.054	1	1	0
Smooth Pub Bias	0.018	0.047	0.011	1	1	0
Lag Selection (thresholding)						
No Pub Bias	0.267	0.16	0.156	0.039	0.17	0.04
Sharp Pub Bias	0.267	0.994	0.259	0.996	0.996	0.147
Smooth Pub Bias	0.155	0.109	0.028	1	1	0
Lag Selection (minimum)						
No Pub Bias	0.04	0.039	0.057	0.041	0.068	0.01
Sharp Pub Bias	0.04	0.931	0.134	0.995	0.995	0.02
Smooth Pub Bias	0.029	0.047	0.017	1	1	0

2.6 Conclusion

The ability of researchers to choose between possible results to report to put their work in the best possible light (p-hack) has reasonably caused concern within the empirical sciences. General approaches to limit or detect this ability are welcome, for example, replication studies and pre-registration. One strand of detection is to undertake meta-studies examining reported p-values (the p-curve) over many papers. Interpreting empirical work based on these tests requires a careful understanding of their ability to detect p-hacking.

We examine from both a statistical and scientific perspective how well these tests are able to detect p-hacking in practice. To do this, we examine four situations where we might expect researchers to p-hack — search over control variables in linear regressions, search over available instruments in IV regressions, selecting amongst datasets, and selecting bandwidth in variance estimation. In a stylized version of each of these, we show how p-hacking affects the distribution of p-values under the alternative, which tells us which types of tests might have power. These results motivate Monte Carlo experiments in more general settings.

Threshold approaches to p-hacking (where a predetermined significance level is targeted) result in p-curves that typically have discontinuities, p-curves that exceed upper bounds under no p-hacking, and less often violations of monotonicity restrictions. Many tests have some power to find such p-hacking, and the best tests are those exploiting both monotonicity and upper bounds and those based on testing for discontinuities. p-Hacking based on reporting the minimum p-value does not result in p-curves exhibiting discontinuities or monotonicity violations. However, tests based on bound violations have some power. Overall this second approach to p-hacking is much harder to detect.

From a scientific perspective, the relevant question is how hard it is to detect p-hacking when the costs of p-hacking — size distortions when there is no effect to find and biases induced in estimates through reporting the best results — are high. From this perspective, the results are more positive. For the p-hacking strategies we examine, the opportunities to change results

significantly through p-hacking are often limited. Test statistics are often quite highly (positively) correlated when they are based on a single dataset so that the effects can be small. We show that for the threshold case the power of tests that work well is positively correlated with the biases in estimated effects induced by p-hacking. This is less so when the minimum p-value is reported because of the low power of the tests in general.

Of final note is that this study examines situations where the model is correctly specified or over-specified, so estimates are consistent for their true values. For poorly specified models, for example, the omission of important variables that leads to omitted variables (confounding) effects, it is possible to generate a larger variation in p-values. Such problems with empirical studies are well understood and perhaps best found through theory and replication than meta-studies.

Chapter 2, in full, is currently being prepared for submission for publication of the material. It is joint work with Graham Elliott and Kaspar Wüthrich. The dissertation author is a primary author of this material.

Chapter 3

Robust Caliper Tests

Abstract

Caliper tests are widely used to test for the presence of p-hacking and publication bias based on the distribution of the z-statistics across studies. We show that without additional restrictions on the distribution of true effects, Caliper tests may suffer from substantial size distortions. We propose a modification of the existing Caliper test, referred to as the Robust Caliper test, which is shown to control size irrespective of the true effect distribution. We also propose a way of correcting the regression-based version of the Caliper test that allows for the inclusion of additional covariates. The proposed tests are easy to implement and perform well in practice.

3.1 Introduction

Publication bias and p-hacking undermine the credibility of empirical findings reported in the scientific journals. A growing body of literature is concerned with detecting and understanding the magnitude and the impact of these phenomena based on the samples of published statistical results (e.g., Gerber and Malhotra, 2008a,b; Brodeur et al., 2016b, 2020b; Vivalt, 2019; Bruns et al., 2019; Adda et al., 2020; Elliott et al., 2022b,a). One approach to detecting p-hacking

and publication bias is to analyze the distribution of observed z-statistics across studies.[1]

Among the statistical tools routinely used to formally test for the presence of p-hacking and publication bias is the Caliper test. Originally proposed by Gerber and Malhotra (2008a,b) for detecting publication bias in sociology research, this is a local test based on the idea that in the absence of selection or manipulation, the observed fractions of marginally significant and marginally insignificant results in a small neighborhood around the chosen significance cutoff should be the same. Recently in economics, Vivalt (2019) adopted the same methodology to examine how significance inflation has varied across time, methods and disciplines. A prominent paper by Brodeur et al. (2020b) also uses the Caliper test and its regression-based augmentation to examine how extent of p-hacking varies by the econometric method that researchers use. Other applications include Brodeur et al. (2021, 2022b,a).

The logic behind the Caliper test described above does not take into account the underlying distribution of true effects that the researchers are dealing with. In this paper, we show that this standard version of the Caliper test fails to control size for certain distributions of true effects. We propose a corrected version of the Caliper test that we refer to as the *Robust Caliper test*, which compares the proportion difference of marginally significant and insignificant results to the bound implied by the worst-case distribution of true effects.

Our contribution to the literature is two-fold. First, by construction, the resulting test is uniformly valid over the set of all possible distributions of effects. In this way, the paper contributes to the literature on methods for testing for p-hacking and publication bias. Second, we propose a procedure to quantify the extent of p-hacking/publication bias present in the literature. Specifically, given a chosen confidence level, we construct a lower bound on the share of literature affected by p-hacking. To our knowledge, this paper is the first to provide a formal way of measuring the extent of p-hacking without assuming a counterfactual distribution of test statistics, absent p-hacking, contributing to the literature on evaluation of the extent of p-hacking

[1] More classic approaches to analyzing the impact and magnitude of publication bias include, for example, Rosenthal (1979) and Gleser and Olkin (1996). A great review of related methods, their advantages and limitations, can be found in Schmid et al. (2020)

and publication bias (Head et al., 2015; Andrews and Kasy, 2018; Brodeur et al., 2016b, 2020b).

When the econometrician has additional information regarding the distribution of true effects, such as its parametric form, we derive a version of the Caliper test that uses this information to construct an alternative bound on the proportion difference that allows for the test to be less conservative and consequently improve power.

We also consider a regression-based modification of the Caliper test introduced by Brodeur et al. (2020b). We show that the regression-based test suffers a similar size control problem and admits a correction under parametric first stage estimation of the distribution of true effects. We propose a bootstrap version of the corrected test that does not require analytical derivation of the variance matrix components.

The Caliper tests can be equivalently formulated and applied to the distributions of reported p-values (p-curves). In this case, the tests will compare the number of p-values just below a significance threshold (typically, 0.05) to the number just above it. Given the non-increasingness of the p-curve absent p-hacking (Elliott et al., 2022b), the standard Caliper test is only valid when the p-curve is flat, i.e. when the distribution of true effects is a point mass at zero so that all the null hypotheses considered by researchers in the literature are true. This condition is unlikely to hold in reality. On the other hand, the Robust Caliper test takes into account the maximum absolute slope of the p-curve in the neighborhood of the target threshold for constructing the bound on proportion difference. Therefore, this paper complements existing methods for detecting p-hacking that use shape restrictions on the p-curve (see, e.g., Simonsohn et al. (2014); Elliott et al. (2022b,a)).

The rest of the paper is organized as follows. Section 3.2 provides the general setup, introduces the standard Caliper test and shows its invalidity. Section 3.3 derives the Robust Caliper test and shows how it can be used to evaluate the extent of p-hacking. Section 3.4 discusses how to improve Caliper test and its regression-based version given a parametric estimate of the distribution of true effects. Section 3.5 conducts a simulation experiment that shows validity of proposed tests and examines their power. Section 3.6 applies Robust Caliper

tests to the dataset of Brodeur et al. (2020b). Section 3.7 concludes. Appendix contains all derivations, proofs and additional results.

3.2 Setup

3.2.1 The Distribution of z-statistics

Consider the problem of testing a two-sided hypothesis about a scalar parameter θ

$$H_0 : \theta = \theta_0 \quad \text{against} \quad H_1 : \theta \neq \theta_0. \tag{3.1}$$

Suppose that researchers have access to an estimator $\hat{\theta}$ of θ based on a sample of n observations. The estimator is assumed to be asymptotically normal such that

$$\sqrt{n}\left(\hat{\theta} - \theta\right) \xrightarrow{d} \mathcal{N}(0, \sigma^2),$$

where σ^2 is the asymptotic variance of $\hat{\theta}$ that can be consistently estimated by $\hat{\sigma}^2$.

Define the usual z-ratio as

$$z := \frac{\sqrt{n}(\hat{\theta} - \theta_0)}{\hat{\sigma}}.$$

For a given $h \in \mathbb{R}$, asymptotic normality of $\hat{\theta}$ implies that $z \xrightarrow{d} \mathcal{N}(h, 1)$, where $h := \sqrt{n}\left(\theta - \theta_0\right)/\sigma$. The absolute value of the z-ratio, $|z|$, is asymptotically distributed according to a folded normal distribution with location parameter h and scale parameter 1. In most studies, the sample sizes are large enough to rely on asymptotic approximations. Therefore, in what follows, we are going to assume that, conditional on h for a given study, the z-statistic comes from $\mathcal{N}(h, 1)$ distribution.

We are interested in the distribution of the absolute value of z across studies, where we compute z given values of h, which are drawn from a probability distribution Π with support

$\mathscr{H} \subset \mathbb{R}$. We refer to Π as the *distribution of true effects*. The density of $|z|$ is given by

$$g_{|z|}(x) = \int_{\mathscr{H}} [\phi(x+h) + \phi(x-h)]\, d\Pi(h), \tag{3.2}$$

where ϕ is the density of a standard normal distribution. Note that the derivative of $g_{|z|}(x)$ is

$$g'_{|z|}(x) = \int_{\mathscr{H}} [(h-x)\phi(x-h) - (x+h)\phi(x+h)]\, d\Pi(h).$$

and its sign will generally depend on x and Π. In practice, scientists may well be focusing on hypotheses that are difficult to distinguish from the null, which corresponds to Π's that assign a lot of mass to z-statistics near two. As a result, the distribution of z-statistics, while not being p-hacked, may still look suggestive of p-hacking due to the hump near the significance threshold. See Section **??** and Figure 3.1 for a concrete example.

Illustrative Example. *If $\Pi(h)$ is the normal distribution with mean μ_h and variance σ_h^2, then*

$$g_{|z|}(x) = \frac{\phi((x-\mu_h)/\sqrt{\sigma_h^2+1}) + \phi((x+\mu_h)/\sqrt{\sigma_h^2+1})}{\sqrt{\sigma_h^2+1}}.$$

Remark 3.1 (One-sided tests)**.** *When researchers are concerned with a one-sided version of (3.1) with alternative hypothesis being $\theta > \theta_0$, one may also consider the density of signed z-statistics $g_z(x) = \int_{\mathscr{H}} \phi(x+h) d\Pi(h)$. All results of this paper remain valid even if some fraction of observations are generated by one-sided testing problems. Therefore, in what follows, we focus on the case of two-sided hypotheses and absolute z-values.*

3.2.2 Caliper Tests

Caliper tests are based on the distribution of absolute z-values. The idea underlying the caliper test is to compare the fraction of absolute z-statistics right above and below a particular threshold t. If the fraction right above t exceeds 0.5, it is interpreted as evidence for p-hacking

and/or publication bias, where the value of t is often chosen to be 1.96 or some other standard significance cutoff, in practice.

Let b denote a bandwidth chosen by the researcher. Define the fraction of absolute z-statistics right below and right above t as

$$p_L(b) := \Pr(t-b \leq |z| \leq t) = \int_{t-b}^{t} g_{|z|}(x)dx = \int_{\mathscr{H}} K_L(h;t,b)d\Pi(h) \tag{3.3}$$

and

$$p_U(b) := \Pr(t \leq |z| \leq t+b) = \int_{t}^{t+b} g_{|z|}(x)dx = \int_{\mathscr{H}} K_U(h;t,b)d\Pi(h) \tag{3.4}$$

respectively, where $K_L(h;t,b) = \Phi(t+h)+\Phi(t-h)-\Phi(t-b+h)-\Phi(t-b-h)$, $K_U(h;t,b) = \Phi(t+b+h)+\Phi(t+b-h)-\Phi(t+h)-\Phi(t-h)$ and $\Phi(\cdot)$ is the CDF of a standard normal distribution. Finally, define

$$\Delta_b := p_U(b) - p_L(b) = \int_{\mathscr{H}} K(h;t,b)d\Pi, \tag{3.5}$$

where $K(h;t,b) = K_U(h;t,b) - K_L(h;t,b)$.

The formal hypothesis testing problem behind the Caliper test is

$$H_0 : \Delta_b \leq 0 \quad \text{against} \quad H_1 : \Delta_b > 0. \tag{3.6}$$

The hypothesis (3.6) can be tested using (exact) Binomial test.[2]

Remark 3.2. *Brodeur et al. (2020b) consider a regression-based implementation of Caliper tests. In particular, they restrict their sample to $[t-b,t+b]$ and model the conditional probability of a*

[2]The Binomial test examines that $\frac{p_U(b)}{p_L(b)+p_U(b)} \leq 0.5$ by comparing N_L,the number of observed $|z|$-values in $[t-b,t)$, to N_U, the number of observed $|z|$-values in $[t,t+b]$. The p-value for the Binomial test is calculated as $1-F_{Bin}(N_U-1;N_L+N_U,0.5)$, where $F_{Bin}(x;N,p)$ is the CDF of the Binomial distribution with success probability p and N number of trials. Note that the exact Binomial test requires independent observations.

significant result as a function of covariates. A simplified version of their model is

$$\Pr\left(|z| \geq t \mid |z| \in [t-b, t+b], X\right) = \Lambda(X'\delta),$$

where X is a vector of covariates of interest and Λ is a suitable link function such as the Probit or Logit link function. We analyze this approach in more detail in Section 3.4.2.

3.2.3 Size Distortions of Caliper Tests

Previous papers (e.g., Gerber and Malhotra, 2008a,b; Brodeur et al., 2020b; Vivalt, 2019; Bruns et al., 2019; Adda et al., 2020) have interpreted the existence of humps, i.e. excess mass, just above the t threshold as evidence of p-hacking. However, this interpretation partly relies on the implicit monotonicity assumption on the shape of the density of $|z|$-statistics. As pointed out by Elliott et al. (2020), this monotonicity assumption is only satisfied under additional restrictions on the allowable set of distributions of true effect.[3] Consequently, in general, humps generated by the distribution of local alternatives result in size violations for the tests attributing humps in the z-curve to p-hacking.

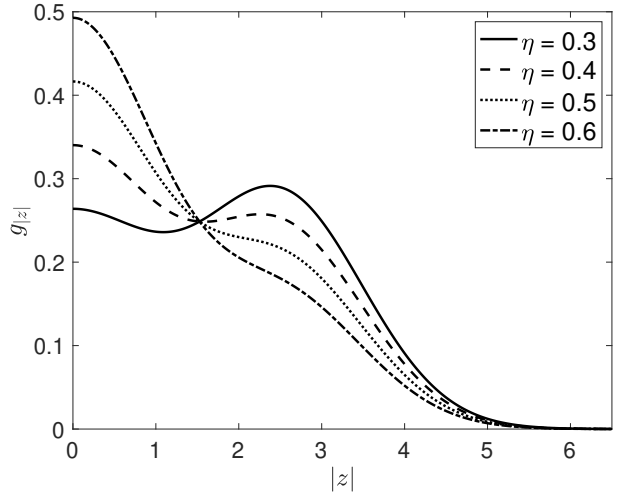

Figure 3.1. z-curves

Consider the following simple example. Let Π be a two-point distribution that assigns

[3] Specifically, the density of absolute z-statistics is monotone if Π admits a unimodal continuous density function.

probability $\eta \in (0,1)$ to $h = 0$ and probability $1-\eta$ to $h = 2.5$. Thus, the observed results are coming from a mixture of studies in which researchers examine true null hypotheses and studies in which researchers focus on hypotheses with alternatives that are difficult to distinguish from the null. Figure 3.1 plots $|z|$-curves for $\eta \in \{0.3, 0.4, 0.5, 0.6\}$. The shape of $g_{|z|}$ is highly dependent on η, and may result in humps around $t = 1.96$ even in the absence of p-hacking.

The shape of $g_{|z|}$ directly translates to the rejection rates of the Caliper test. Figure 3.2 displays the empirical rejection rate of a Caliper test based on exact Binomial tests for $t = 1.96$ and $b \in \{0.05, 0.1, 0.2\}$, where the nominal level is 5%. The sample size is $N = 5000$ corresponding to the empirical application of Section 3.6.[4] In our setting, the average local sample size ranges from 58 ($b = 0.05,\ \eta = 1$) to 686 ($b = 0.2,\ \eta = 0$). Our simulations demonstrate that Caliper tests can suffer from substantial size distortions when τ is small. When this is the case, the z-curve exhibits humps induced by the distribution of alternatives (cf. Figure 3.1).

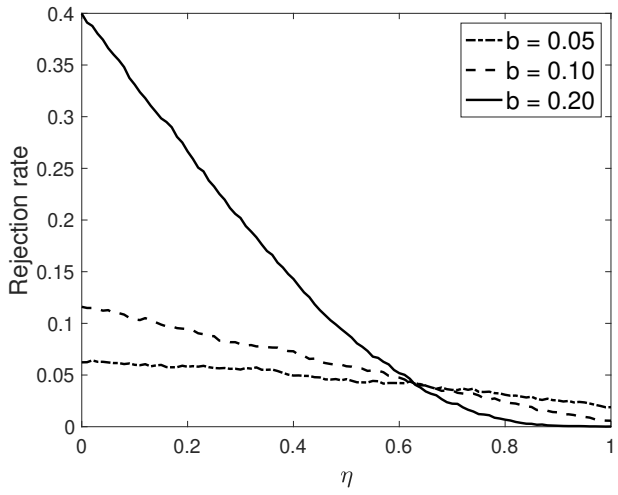

Figure 3.2. Size distortions caliper test. Nominal level: 5%. Based on simulations with 10000 repetitions.

Proposition 3.1 provides a formal explanation for the findings of the above Monte Carlo experiment. The asymptotic rejection probability of the Binomial test can be made arbitrarily

[4]Note that the relevant sample size for the Caliper test is not the overall sample size, but the local sample size in the $[t^* - b, t^* + b]$ interval, which depends on the distribution of alternatives and on the choice of b.

large by choosing a suitable distribution of true effects.

Proposition 3.1. *Suppose $Z_1, ..., Z_N$ are i.i.d. random observations of z-statistics as in Section 3.2.1. Then, for any $0 < b \leq t$,*

$$\sup_{\Pi} \limsup_{N \to \infty} \Pr(1 - F_{Bin}(N_U - 1; N_L + N_U, 0.5) < \alpha) = 1,$$

where $\alpha \in (0,1)$, $N_U = \sum_{i=1}^{N} 1\{t < |Z_i| \leq t + b\}$, $N_L = \sum_{i=1}^{N} 1\{t - b \leq |Z_i| \leq t\}$ and the supremum is taken over all probability distributions on $\mathbb{R}$.

3.3 Robust caliper tests

Let $\Delta_b(\Pi)$ denote the value of Δ_b that is implied by the distribution of true effects Π. If Π is known to the econometrician, the testing problem (3.6) can be adjusted in a straightforward way by utilizing this knowledge. Indeed, the null hypothesis in testing problem $H_0 : \Delta_b \leq \Delta_b(\Pi)$ vs. $H_1 : \Delta_b > \Delta_b(\Pi)$ holds under no *p*-hacking, and the Binomial test with appropriately adjusted success probability controls size.[5]

Illustrative Example (continued). *When the distribution of true effects is $\mathcal{N}(\mu_h, \sigma_h^2)$, the true value of Δ_b in the absence of p-hacking becomes*

$$\Delta_b(\mathcal{N}(\mu_h, \sigma_h^2)) = \sum_{(i,j) \in \{-1,1\}^2} \Phi\left(\frac{t + ib + j\mu_h}{\sqrt{\sigma_h^2 + 1}}\right) - \Phi\left(\frac{t + j\mu_h}{\sqrt{\sigma_h^2 + 1}}\right).$$

3.3.1 Worst-Case Correction

In practice, the distribution of true effects is never known to the econometrician. This motivates the development of the Caliper test that is agnostic to the distribution true effects. Here we propose a modified version of the Caliper test, which we refer to as the Robust Caliper Test.

[5] In this case, the success probability parameter for the Binomial test under the null needs to be calculated as the value of $\frac{p_U(b)}{p_L(b)+p_U(b)}$ implied by Π.

The idea is the following. Given the explicit form of $g_{|z|}$ in equation (3.2), we can compute the maximum value of Δ_b that can be achieved in the absence of p-hacking. We denote this value as $\overline{\Delta}_b$. Then we consider the modified testing problem

$$H_0 : \Delta_b \leq \overline{\Delta}_b \quad \text{against} \quad H_1 : \Delta_b > \overline{\Delta}_b. \tag{3.7}$$

To determine the value of $\overline{\Delta}_b$, we use the definition of Δ_b (equation (3.5)). In the absence of p-hacking, we have

$$\overline{\Delta}_b := \Delta(\Pi_b^*) = \sum_{(i,j)\in\{-1,1\}^2} \Phi(t + ib + jh^*) - \Phi(t + jh^*),$$

where Π_b^* is a probability measure that assigns all mass to an element $h^* \in \arg\max_{h\in\mathscr{H}} K(h;t,b)$.

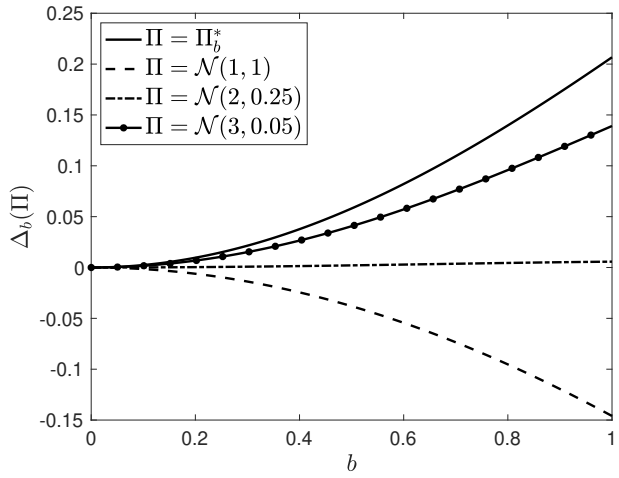

Figure 3.3. The bound on the proportion differences as a function of b ($t = 1.96$).

For given values of b and t, the value of h^* can be found numerically. Figure 3.3 shows the graph $\overline{\Delta}_b$ for $t = 1.96$ as a function of b and Δ_b for three different choices of Π. As we can see, depending of the distribution of true effects, Δ_b can take large negative values ($\Pi = \mathcal{N}(1,1)$), be close to zero ($\Pi = \mathcal{N}(2,0.25)$) or take large positive values close to the upper bound $\overline{\Delta}_b$ ($\Pi = \mathcal{N}(3,0.05)$).

The hypothesis (3.7) can be tested using a simple one-sided t-test. The estimate of Δ_b is

$\hat{\Delta}_b = \hat{p}_U(b) - \hat{p}_L(b)$, where $\hat{p}_U(b)$ and $\hat{p}_L(b)$ are the sample proportions of marginally significant and marginally insignificant results respectively. The variance of $\hat{\Delta}_b$ is $\Omega_b = p_U(b)(1-p_U(b)+p_L(b))+p_L(b)(1-p_L(b)+p_U(b))$ and can be consistently estimated by plugging in sample proportions. We denote this estimator by $\hat{\Omega}_b$. Thus, as Theorem 3.1 states formally, under the null hypothesis in (3.7), the asymptotic distribution of the statistic

$$T = \frac{\sqrt{N}(\hat{\Delta}_b - \overline{\Delta}_b)}{\sqrt{\hat{\Omega}_b}}$$

is stochastically dominated by the standard normal distribution and we can test the null hypothesis of no p-hacking by comparing T to quantiles of the standard Normal distribution.

Theorem 3.1. *Suppose $Z_1, ..., Z_N$ are i.i.d. random observations of z-statistics as in Section 3.2.1. Then under the null hypothesis of no p-hacking,*

$$\limsup_{N\to\infty} \Pr(T > z_{1-\alpha}) \leq \alpha,$$

for all Π other than point masses at $\pm\infty$. In addition, for any $\gamma \in (0,1)$ and sufficiently large $A > 0$,

$$\limsup_{N\to\infty} \sup_{\Pi\in\mathscr{F}_{A,\gamma}} \Pr(T > z_{1-\alpha}) = \alpha,$$

where $\alpha \in (0,1)$, $z_{1-\alpha} := \Phi^{-1}(1-\alpha)$ and $\mathscr{F}_{A,\gamma} = \{\Pi : \Pi(A) - \Pi(-A) > \gamma\}$.

Remark 3.3 (Testing multiple thresholds jointly)**.** *While p-hacking, different researchers may target different significance thresholds. For instance, if different groups of researchers aim to achieve significant results at 1%, 5% and 10% significance levels, then we should expect the testable restriction to be violated in the neighborhoods of $t_1 = 2.576$, $t_2 = 1.96$ and $t_3 = 1.645$ respectively. One can use the Robust Caliper test to examine each threshold separately. However, doing so would require a certain adjustment for multiple testing in order for this procedure to control size. Instead, one can construct a joint test that combines all thresholds. For bandwidth*

values b_1, b_2 and b_3, the vector of proportion differences, $(\Delta_{b_1}^{t_1}, \Delta_{b_2}^{t_2}, \Delta_{b_3}^{t_3})'$ can be upper-bounded by a vector of corresponding worst-case differences, $(\overline{\Delta}_{b_1}^{t_1}, \overline{\Delta}_{b_2}^{t_2}, \overline{\Delta}_{b_3}^{t_3})'$, component-wise. Moment inequality tests can be used to test that these upper bounds hold given asample of z-statistics. See Appendix C.3 for more details.

Remark 3.4 (Correlated observations)**.** *In practice, we observe multiple results reported in a single paper. As a consequence, the independence assumption of Theorem 1 is likely to be violated when we apply the robust test to the full sample of observed results. In that case, one needs to replace $\hat{\Omega}_b$ with a cluster-robust version of the variance estimator.*

Remark 3.5 (z-curve vs. p-curve)**.** *Caliper tests can be constructed analogously on the basis of observed p-values instead of $|z|$-values. An advantage of using the p-curve for testing p-hacking is that testable restrictions under no p-hacking have a more natural interpretation in terms of shape constraints (monotonicity) on allowable p-curves (see Elliott et al. (2022b)). In general, since there is a one-to-one correspondence between $|z|$-values and p-values, one can always construct a Caliper test based on the p-curve that is equivalent to a given Caliper test based of the z-curve. However, this is not the case if we require a symmetric partition of the local testing interval (see Appendix C.4 for the comparison between two approaches in this case). In this paper, we focus on Caliper tests based on the distribution of absolute z-values.*

3.3.2 Evaluating the Extent of p-hacking

In practice, the observed sample of z-statistics combines a mixture of both p-hacked and non-p-hacked results. Let $\tau \in [0,1]$ be a population probability that the study is p-hacked, we will refer to it as the extent of p-hacking in the literature. Then the density of observed $|z|$-values is a mixture of $g_{|z|}(x)$ and $g_{|z|}^{ph}(x)$ with mixture weights $1-\tau$ and τ respectively:

$$g_{|z|}^{obs}(x) = (1-\tau)g_{|z|}(x) + \tau g_{|z|}^{ph}(x), \tag{3.8}$$

where $g_{|z|}(x)$ is the density of $|z|$ in the absence of p-hacking defined in equation (3.2) and $g^{ph}_{|z|}(x)$ is the density of p-hacked results. The shape of $g^{ph}_{|z|}(x)$ depends on various factors such as the underlying distribution of data in observed studies, the estimation methods used by researchers and the p-hacking strategy employed by researchers. For example, Elliott et al. (2022a) provide analytical and numerical examples of p-hacked distributions for the cases of covariate selection in linear regression, judicious instrument selection in IV regression and judicious lag length selection for the estimation of variance. Section 3.5.1 and Appendix C.5 provide concrete examples of $g^{ph}_{|z|}(x)$ in the context of control variables selection in linear regression. For this section, we do not make any assumptions about the form of $g^{ph}_{|z|}(x)$.

When $\tau = 0$ there is no p-hacking in the literature and $|z|$ follows $g_{|z|}$. Therefore, the hypothesis of no p-hacking can be formulated as

$$H_0 : \tau = 0 \quad \text{against} \quad H_1 : \tau > 0. \tag{3.9}$$

As we have shown in 3.3.1, Robust Caliper test can be used to test (3.9). Suppose that instead of the extreme case $\tau = 0$, we want to test a weaker hypothesis that the extent of p-hacking in the literature is at most $\bar{\tau}$

$$H_0 : \tau \leq \bar{\tau} \quad \text{against} \quad H_1 : \tau > \bar{\tau}. \tag{3.10}$$

Let $\Delta_{b,\bar{\tau}}$ be the proportion difference implied by (3.8) when $\tau = \bar{\tau}$ Under the null hypothesis in (3.10) with $\bar{\tau} > 0$, $\overline{\Delta}_b$ is not a valid upper bound on $\Delta_{b,\bar{\tau}}$. However, the valid upper bound can be easily constructed given the structure of equation (3.8). Specifically, we can show that

$$\overline{\Delta}_{b,\bar{\tau}} := \sup_{\Pi} \Delta_{b,\bar{\tau}} = (1 - \bar{\tau})\overline{\Delta}_b + \bar{\tau}.$$

The first term in the above expression is the worst-case bound on Δ_b scaled by the minimum fraction of non-p-hacked results under the null. The second term is the maximum extent of p-hacking under the null and comes from the fact that the value of the proportion difference

under p-hacking cannot exceed 1. Therefore, analogously to the result of Theorem 3.1, the null hypothesis in (3.10) can be tested by a simple one-sided t-test that compares the test statistic

$$T_{\bar{\tau}} = \frac{\sqrt{N}(\hat{\Delta}_b - \overline{\Delta}_{b,\bar{\tau}})}{\sqrt{\hat{\Omega}_b}}$$

to quantiles of a standard Normal distribution. The confidence interval for τ can be constructed by inverting this test. The $(1-\alpha)100\%$ confidence interval for τ is therefore given by $CI_{1-\alpha}(\tau) := [\underline{\tau}_\alpha, 1]$, where

$$\underline{\tau}_\alpha = \max\left\{0, \frac{\sqrt{N}(\hat{\Delta}_b - \overline{\Delta}_b) - z_{1-\alpha}\sqrt{\hat{\Omega}_b}}{\sqrt{N}(1-\overline{\Delta}_b)}\right\}$$

is the largest extent of p-hacking rejected by the test at level α.

3.4 Estimating Π under the Null

In testing problem $H_0 : \Delta_b \leq \Delta_b(\Pi)$ vs. $H_1 : \Delta_b > \Delta_b(\Pi)$, the distribution of effects, Π, plays the role of a nuisance parameter. In section 3.3, we dealt with it by considering the worst-case value of the nuisance parameter to ensure the uniform validity of the test. In this section, we consider an alternative approach by estimating the nuisance parameter under the null for the test. This approach requires additional assumptions on Π.

3.4.1 Caliper Test with Parametric Π

The robust caliper test proposed in Section 3.3.1 is valid uniformly over the universe of distributions of true effects. However, as it can be deduced from Figure 3.3, Robust Caliper tests can be conservative due to the worst-case construction. The power of the test can be improved in the presence of additional restrictions on the set of allowable distributions of true effects. For instance, it can be done if we have information about the parametric form of Π.[6] The key observation is that under the null of no p-hacking, one can construct an estimate $\hat{\Pi}$ of Π and use

[6]Note that even in the presence of such information, if the parametric family of allowable Π's is rich enough, it is not possible to tighten the upper bound threshold $\overline{\Delta}_b$. For example, if $\mathscr{F}_\Pi = \{\mathcal{N}(\mu_h, \sigma_h^2) : \mu_h \in \mathbb{R}, \sigma_h^2 > 0\}$, then $\sup_{\Pi \in \mathscr{F}_\Pi} \Delta_b(\Pi) = \Delta(\Pi_b^*) = \overline{\Delta}_b$.

it to construct an alternative estimate of $\Delta_b(\Pi)$ as $\tilde{\Delta}_b := \Delta_b(\hat{\Pi})$. The test then can be based on the realized difference between $\hat{\Delta}_b$ and $\tilde{\Delta}_b$. We require parametric models for Π to satisfy the following assumption.

Assumption 3.1 (MLE Regularity)**.** *(i) The model $\{\Pi_\gamma : \gamma \in \Gamma\}$ is differentiable at an inner point γ_0 of $\Gamma \subset \mathbb{R}^k$; (ii) $\Pi_\gamma \neq \Pi_{\gamma_0}$ for every $\gamma \neq \gamma_0$; (iii) the derivative of the PDF (or PMF) $\pi(h;\gamma)$ with respect to γ is integrable on $\mathscr{H}$ for all $\gamma \in \Gamma$; (iv) information matrix of the model at γ_0 is nonsingular; (v) there exists a measurable function $s(\cdot)$ with $E_{\Pi_{\gamma_0}}[s^2] < \infty$ such that, for every γ_1 and γ_2 in a neighborhood of γ_0, $|\log \pi(h;\gamma_1) - \log \pi(h;\gamma_2) \leq s(h)||\gamma_1 - \gamma_2||$.*

Assumption 3.1 contains standard conditions that ensure that the maximum likelihood estimator of Π under no p-hacking is consistent and asymptotically normal. They are satisfied in many parametric settings including the setting of the Illustrative example. Under Assumption 1 the stochastic behavior of $\hat{\Delta}_b - \tilde{\Delta}_b$ is characterized by the following theorem.

Theorem 3.2. *Suppose the model for the distribution of true effects is $\{\Pi_\gamma : \gamma \in \Gamma \subset \mathbb{R}^k\}$ and Assumption 3.1 is satisfied. Let $\hat{\gamma}$ and $\hat{\Pi} = \Pi_{\hat{\gamma}}$ be the MLE estimators of γ and Π respectively. Then under no p-hacking*

$$\sqrt{N}(\hat{\Delta}_b - \tilde{\Delta}_b) \xrightarrow{d} \xi_1 - \xi_2,$$

where

$$\begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} \sim \mathscr{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Omega_b & C'_{\Delta\gamma} D(\Pi) \\ C'_{\Delta\gamma} D(\Pi) & D(\Pi)' \mathscr{I}_\gamma^{-1} D(\Pi) \end{pmatrix} \right),$$

$\mathscr{I}_\gamma = \int_{-\infty}^{\infty} \frac{(\int_{\mathscr{H}} \phi(z-h)\partial\pi(h;\gamma)/\partial\gamma dh)(\int_{\mathscr{H}} \phi(z-h)\partial\pi(h;\gamma)/\partial\gamma dh)'}{\int_{\mathscr{H}} \phi(z-h)\pi(h;\gamma)dh} dz$, *$C_{\Delta\gamma}$ is the asymptotic covariance between $\hat{\Delta}_b$ and $\hat{\gamma}$ and $D(\Pi) = \int_{\mathscr{H}} K(h;t,b)\partial\pi_\gamma(h)/\partial\gamma dh$.*

Every element in the variance matrix of $(\xi_1, \xi_2)'$ can be consistently estimated under standard conditions. Therefore, the test statistic

$$\tilde{T} = \frac{\sqrt{N}(\hat{\Delta}_b - \tilde{\Delta}_b)}{\sqrt{\hat{\Omega}_b - 2\hat{C}'_{\Delta\gamma} D(\hat{\Pi}) + D(\hat{\Pi})' \hat{\mathscr{I}}_\gamma^{-1} D(\hat{\Pi})}}$$

is asymptotically normally distributed and we can test the null hypothesis of no p-hacking by comparing it to quantiles of the standard normal distribution.

Illustrative Example (continued). *Under normality assumption on the distribution of true effects,* $\gamma = (\mu_h, \sigma_h^2)'$ *and* $\hat{\gamma} = \left(\bar{Z}, \max\left\{0, \frac{1}{N}\sum_{i=1}^{N}(Z_i - \bar{Z})^2 - 1\right\}\right)'$ *with* $\mathscr{I}_\gamma = diag\{\sigma_h^2 + 1, 2(\sigma_h^2 + 1)^2\}$. *Define* $\alpha_1^+ := \frac{t+b-\mu_h}{\sqrt{\sigma_h^2+1}}$, $\alpha_2^+ := \frac{t-\mu_h}{\sqrt{\sigma_h^2+1}}$, $\alpha_3^+ := \frac{t-b-\mu_h}{\sqrt{\sigma_h^2+1}}$, $\alpha_1^- := \frac{-t-b-\mu_h}{\sqrt{\sigma_h^2+1}}$, $\alpha_2^- := \frac{-t-\mu_h}{\sqrt{\sigma_h^2+1}}$ *and* $\alpha_3^- := \frac{-t+b-\mu_h}{\sqrt{\sigma_h^2+1}}$. *Then* $C_{\Delta\gamma} = (C_{\Delta\mu_h}, C_{\Delta\sigma_h^2})'$, *where*

$$C_{\Delta\mu_h} = \sqrt{\sigma_h^2+1}\left(2(\phi\left(\alpha_2^+\right) - \phi(\alpha_2^-)) + \phi\left(\alpha_1^-\right) + \phi\left(\alpha_3^-\right) - \phi\left(\alpha_1^+\right) - \phi\left(\alpha_3^+\right)\right),$$

$$C_{\Delta\sigma_h^2} = (\sigma_h^2+1)\left(2(\alpha_2^+\phi\left(\alpha_2^+\right) - \alpha_2^-\phi(\alpha_2^-)) + \alpha_1^-\phi\left(\alpha_1^-\right) + \alpha_3^-\phi\left(\alpha_3^-\right) - \alpha_1^+\phi\left(\alpha_1^+\right) - \alpha_3^+\phi\left(\alpha_3^+\right)\right)$$

and

$$D(\Pi) = \int_{\mathbb{R}} K(h;t,b)\pi(h)\left(\frac{h-\mu_h}{\sigma_h^2}, \frac{(h-\mu_h)^2 - \sigma_h^2}{2\sigma_h^4}\right)' dh.$$

For more complicated families of distributions, the calculations of the expressions for the elements in the variance matrix of $(\xi_1, \xi_2)'$ can be cumbersome and often infeasible analytically. These derivations can be avoided by using Efron's bootstrap that is described in the Algorithm 3.1.

Algorithm 3.1. *Efron's Bootstrap for* $\hat{\Delta}_b - \tilde{\Delta}_b$

Step 1. Given a sample of $\mathscr{Z} = \{Z_1, ..., Z_N\}$ *calculate the estimates* $\hat{\Delta}_b, \hat{\Pi}$ *and* $\tilde{\Delta}_b = \Delta(\hat{\Pi})$

Step 2. Calculate the test statistic $S = \hat{\Delta}_b - \tilde{\Delta}_b$

Step 3. Construct a bootstrap sample $\mathscr{Z}^* = \{Z_1^*, ..., Z_N^*\}$ *by drawing with replacement from* $\mathscr{Z}$ *and use* $\mathscr{Z}^*$ *to calculate* $\hat{\Delta}_b^*, \hat{\Pi}^*$ *and* $\tilde{\Delta}_b^* = \Delta(\hat{\Pi}^*)$

Step 4. Calculate the bootstrap version of the test statistic as $S^* = \hat{\Delta}_b^* - \tilde{\Delta}_b^* - S$

Repeat steps 3 and 4 B times and obtain $\mathscr{S} = \{S_1^*, ..., S_B^*\}$

Step 5. Reject the null hypothesis of no p-hacking if $S > q_{1-\alpha}^*(\mathscr{S})$, *where* $q_{1-\alpha}^*(\mathscr{S})$ *is the* $100(1-\alpha)\%$ *quantile of* $\mathscr{S}$

Remark 3.6. *(Non-parametric estimation of* Π*) Under the set up of Section 3.2 the distribution of true effects can be estimated non-parametrically. Note that asymptotically* $z = h + \xi$*,* $\xi \sim \mathcal{N}(0,1)$*, and thus, given a random sample of* z'*s, the distribution of* h *can be recovered using kernel deconvolution techniques (see Delaigle (2021) for an excellent review). The derivation of the asymptotic distribution of* $\sqrt{N}(\hat{\Delta}_b - \tilde{\Delta}_b)$ *in this case is more involved and left for future work. Nevertheless, the test with a non-parametric first stage can be conducted using the bootstrap procedure described in Algorithm 3.1.*

3.4.2 Regression-Based Test

Brodeur et al. (2020b) propose a regression-based implementation of the Caliper test. In particular, they restrict the sample to include only marginally significant and insignificant results, $|z| \in [t-b, t+b]$, and model the conditional probability of a significant result as a function of binary variables of interest and additional covariates such as author and journal characteristics.[7] More specifically, a version of their model can be written as

$$\Pr\left(|z| \geq t \mid |z| \in [t-b, t+b], M_0, M_1, ..., M_K, X\right) = \Lambda\left(\beta_0 + \sum_{k=1}^{K} \beta_k M_k + X'\delta\right), \tag{3.11}$$

where $M_k \in \{0,1\}, k = 0, ..., K$ represent mutually exclusive categories, $X_i \in \mathcal{X}$ is a vector of covariates of interest and Λ is a suitable link function such as the Probit or Logit link function. For example, M_k's can represent different statistical estimation methods (RCT (baseline), RDD, IV, DID) used by researchers as in Brodeur et al. (2020b). If the results in the baseline category are not p-hacked, then significantly positive values of β_k may be interpreted as evidence of p-hacking in category $k > 0$. Implicitly, this method uses the baseline category as a subset of data on which the actual proportions of marginally significant and marginally insignificant results (implied by the distribution of true effects) can be estimated and compares the observed values in other categories to this benchmark. As Proposition 3.2 formally states, this logic works if the

[7] Brodeur et al. (2020b) argue that including these additional covariates may help to distinguish p-hacking from publication bias.

distribution of true effects is the same among categories and X is independent of the categories. However, in the absence of homogeneity in the distributions of true effects, this procedure is invalid for testing for no p-hacking (Proposition 3.3).

Proposition 3.2. *Suppose that (i) the baseline category ($M_0 = 1$) is not p-hacked; (ii) X is independent of $M_k, k = 0, 1, ..., K$; and (iii) the distributions of true effects for all categories coincide $\Pi_k = \Pi$ for $k = 0, ..., K$. Then under no p-hacking $\beta_k = 0, k = 1, ..., K$ in model (3.11).*

Proposition 3.3. *Suppose X is independent of $M_k, k = 0, 1, ..., K$, $E[X^2] < \infty$ and $\Lambda(\cdot)$ is strictly increasing. Let T_k be the t-statistic for testing $\beta_k = 0$ against $\beta_k > 0$ in model (3.11). Then*

$$\sup_{\Pi_0, ..., \Pi_K} \limsup_{N \to \infty} \Pr(T_k > z_{1-\alpha}) = 1.$$

Corrected Test

The worst-case correction similar to Section 3.3.1 is not available in the context of the regression-based test. However, a valid test can still be constructed if we have estimates of $\Pi_k, k = 0, ..., K$ similar to Section 3.4.1.

Let $F_{X|k}$ be the distribution function of X conditional of $M_k = 1$. Define

$$\eta_k(a) := \eta(a; \delta, F_{X|k}) = \int_{\mathscr{X}} \Lambda(a + x\delta) dF_{X|k}(x)$$

and let the estimate of the inverse of η be defined as

$$\widehat{\eta_k^{-1}}(u) := \eta^{-1}(u; \hat{\delta}, \hat{F}_{X|k}).$$

In addition, define conditional fraction of marginally significant results for category k, $\omega_k := \omega_k(\Pi_k) := \Pr(|z| \geq t \mid |z| \in [t-b, t+b], M_k = 1) = \frac{p_{U,k}}{p_{U,k} + p_{L,k}}$. Note that $\eta_0^{-1}(\omega_0) = \beta_0$ and $\eta_k^{-1}(\omega_k) =$

$\beta_0 + \beta_k$ for $k > 0$. Using this observation, it can be shown that, in the absence of p-hacking,

$$\beta = A\eta^{-1},$$

where $\beta = (\beta_0, ..., \beta_K)'$, $\eta^{-1} = (\eta_0^{-1}(\omega_0), ..., \eta_K^{-1}(\omega_K))'$ and $A = \begin{pmatrix} -1 & 0_{1\times K} \\ -\iota_K & I_K \end{pmatrix}$. The left-hand side, β, can be estimated from the binary regression and the right-hand side, η^{-1}, can be estimated by combining the estimates of $\Pi_0, ..., \Pi_K$ (and implied estimates of $\omega_0, ..., \omega_K$) with the estimate of η. The following theorem shows that a valid test for p-hacking can be constructed based on the difference between these two estimates.

Theorem 3.3. *Suppose the model for the distribution of true effects for subgroup $M_k = 1$ is $\{\Pi_{\gamma_k} : \gamma_k \in \Gamma_k \subset \mathbb{R}^{d_k}\}$ and Assumption 1 is satisfied for every $k = 0, ..., K$. In addition, assume that $E[X^2] < \infty$. Let $\hat{\beta}$ be the MLE estimate of β in model (3.11), $\widehat{\eta^{-1}} = (\widehat{\eta_0^{-1}}(\hat{\omega}_0), ..., \widehat{\eta_K^{-1}}(\hat{\omega}_K))'$ and $\hat{\omega}_k = \omega_k(\Pi_{\hat{\gamma}_k})$ for $k = 0, ..., K$. Then under no p-hacking*

$$\sqrt{N}(\hat{\beta} - A\widehat{\eta^{-1}}) \xrightarrow{d} \zeta_1 - A\zeta_2,$$

where

$$\begin{pmatrix} \zeta_1' & \zeta_2' \end{pmatrix}' \sim \mathcal{N}\left(0_{2(K+1)\times 1}, V\right),$$

and the blocks of V are defined in Appendix C.1.6.

Given the result of Theorem 3.3, the null hypothesis of no p-hacking for category k can be tested by a one-sided t-test constructed on the basis of the kth element of $\sqrt{N}(\hat{\beta} - A\widehat{\eta^{-1}})$. One can also test the null jointly for several categories by using a χ^2-test on a subset of elements in $\sqrt{N}(\hat{\beta} - A\widehat{\eta^{-1}})$.

In practice, the expression for V can be very hard to calculate even when the distributions of true effects are normal. Similar to Section 3.3.1, these derivations can be avoided by using Efron's bootstrap that is described in the Algorithm 3.2.

Algorithm 3.2. *Efron's Bootstrap for* $\hat{\beta} - A\widehat{\eta^{-1}}$

Step 1. Given a sample of $\mathscr{Z} = \{Z_1, ..., Z_N\}$ *calculate the estimates* $\hat{\beta}$ *and* $\widehat{\eta^{-1}}$

Step 2. Calculate the test statistic $S = \hat{\beta} - A\widehat{\eta^{-1}}$

Step 3. Construct a bootstrap sample $\mathscr{Z}^* = \{Z_1^*, ..., Z_N^*\}$ *by drawing with replacement from* $\mathscr{Z}$ *and use* $\mathscr{Z}^*$ *to calculate* $\hat{\beta}^*$ *and* $\widehat{\eta^{-1}}^*$

Step 4. Calculate the bootstrap version of the test statistic as $S^* = \hat{\beta}^* - A\widehat{\eta^{-1}}^* - S$

Repeat steps 3 and 4 B times and obtain $\mathscr{S} = \{S_1^*, ..., S_B^*\}$

Step 5. Reject the null hypothesis of no p-hacking if $S > q_{1-\alpha}^*(\mathscr{S})$*, where* $q_{1-\alpha}^*(\mathscr{S})$ *is the* $100(1-\alpha)\%$ *quantile of* $\mathscr{S}$

3.5 Monte Carlo Simulations

In this section, we investigate the finite sample properties of the tests in Section 3.3 using a Monte Carlo simulation study.

3.5.1 Covariate Selection in Linear Regression

We adopt the Monte Carlo design of Elliott et al. (2022a). Specifically, we consider researchers who have access to a random sample with $N = 200$ observations generated as

$$Y_i = X_i\beta + u_i, \quad i = 1, \ldots, N,$$

where $(X_i, u_i)' \sim i.i.d.\,\mathscr{N}(0, I_2)$, $\beta = h/\sqrt{N}$ and h is drawn from the distribution of true effects Π. For this Monte Carlo experiment, we use the three choices of Π that we considered in Section 3.3.1: $\Pi = \mathscr{N}(1,1)$, $\Pi = \mathscr{N}(2, 0.25)$ and $\Pi = \mathscr{N}(3, 0.05)$. Researchers have access to K additional control variables, $W_i := (W_{1i}, \ldots, W_{Ki})'$, which are generated as

$$W_{ki} = \gamma_k X_i + \sqrt{1-\gamma_k^2}\varepsilon_{W_k,i}, \quad \varepsilon_{W_k,i} \sim \mathscr{N}(0,1), \quad \gamma_k \sim U[-0.8, 0.8], \quad k = 1, \ldots, K.$$

The researchers are interested in testing a hypothesis about β:

$$H_0 : \beta = 0 \qquad \text{against} \qquad H_1 : \beta \neq 0. \tag{3.12}$$

Following Elliott et al. (2022a), we consider two approaches to p-hacking: a thresholding and a maximum approach.

For the threshold approach, researchers first regress Y_i on X_i and W_i and test (3.12). Denote the resulting z-value as Z. If $|Z| \geq t$, the researchers report the absolute z-value. If $|Z| < t$, they regress Y_i on X_i trying all $(K-1) \times 1$ subvectors of W_i as controls and select the result with the largest absolute z-value. If the largest absolute z-value is larger than t, they continue and explore all $(K-2) \times 1$ subvectors of W_i etc. If all results are insignificant, they report the largest absolute z-value.

For the maximum approach, researchers run regressions of Y_i on X_i and each possible configuration of covariates W_i and report the maximum absolute z-value.

Figures C.4, C.5, and C.6 show the null and p-hacked distributions for $K \in \{3,5,7\}$. To generate these distributions, we run the algorithm one million times and collect p-hacked and non-p-hacked results. The threshold approach leads to a discontinuity in the z-curve and may lead to humps just above significance thresholds. On the other hand, when using the maximum approach, the z-curve is generated by maximums across realizations of normal random variables and hence continuous. The distribution of h is an important driver of the shape of the z-curve. The larger h, the higher the probability that researchers find significant results in the initial specification and terminate the specification search when using the thresholding approach. Since larger amount of controls gives researcher more room to p-hack, the hump right next to the significance threshold is more pronounced for larger values of K. Similarly, for the maximum approach larger K makes the distribution more shifted to the right.

3.5.2 *P*-hacking at $t = 1.96$

In this section we present the power curves of our tests when the p-hacking occurs at a single significance threshold ($t = 1.96$) and the tests are using this threshold as a target.

We draw our Monte Carlo samples from the distribution $g^{obs}_{|z|}$, that is a mixture of p-hacked ($g^{ph}_{|z|}$) and non-p-hacked ($g_{|z|}$) distributions

$$g^{obs}_{|z|} = (1-\tau)g_{|z|} + \tau g^{ph}_{|z|},$$

where $\tau \in [0,1]$. We use sample sizes $N \in \{1000, 5000\}$. All results are based on 5,000 Monte Carlo repetitions. The nominal level of all tests is 5%.

Figures 3.4 and 3.5 display the power of the Robust Caliper test and Caliper tests based on a parametric (normal) estimate of Π for the case $K = 5$ (results for $K = 3$ and $K = 7$ are reported in Appendix C.5). The parametric tests are implemented using both the asymptotic distribution and bootstrap where we use 1,000 bootstrap repetitions in all cases. We show the results for two choices of b: $b = 0.1$ and $b = 0.5$. Figure 3.4 shows the power for the overall sample size of $N = 1000$. It can be seen that all tests control size for every choice of Π. For the thresholding p-hacking approach the power of the Robust Caliper test varies significantly with Π and b. For $b = 0.5$ and $\Pi \in \{\mathcal{N}(1,1), \mathcal{N}(2,0.25)\}$ the test has no power. The reason for that is the fact that the worst-case bound is very weak for $b = 0.5$ and far above the true value of the proportion difference implied by these distributions of true effects (cf. Figure 3.3). When $\Pi = \mathcal{N}(3,0.05)$ the test with $b = 0.5$ starts detecting p-hacking after the extent of p-hacking exceeds 40%. For $b = 0.1$ the power of the Robust Caliper test is much higher, however, the most amount of power is again achieved in case $\Pi = \mathcal{N}(3,0.05)$ that generates the highest value of the proportion difference under the null.

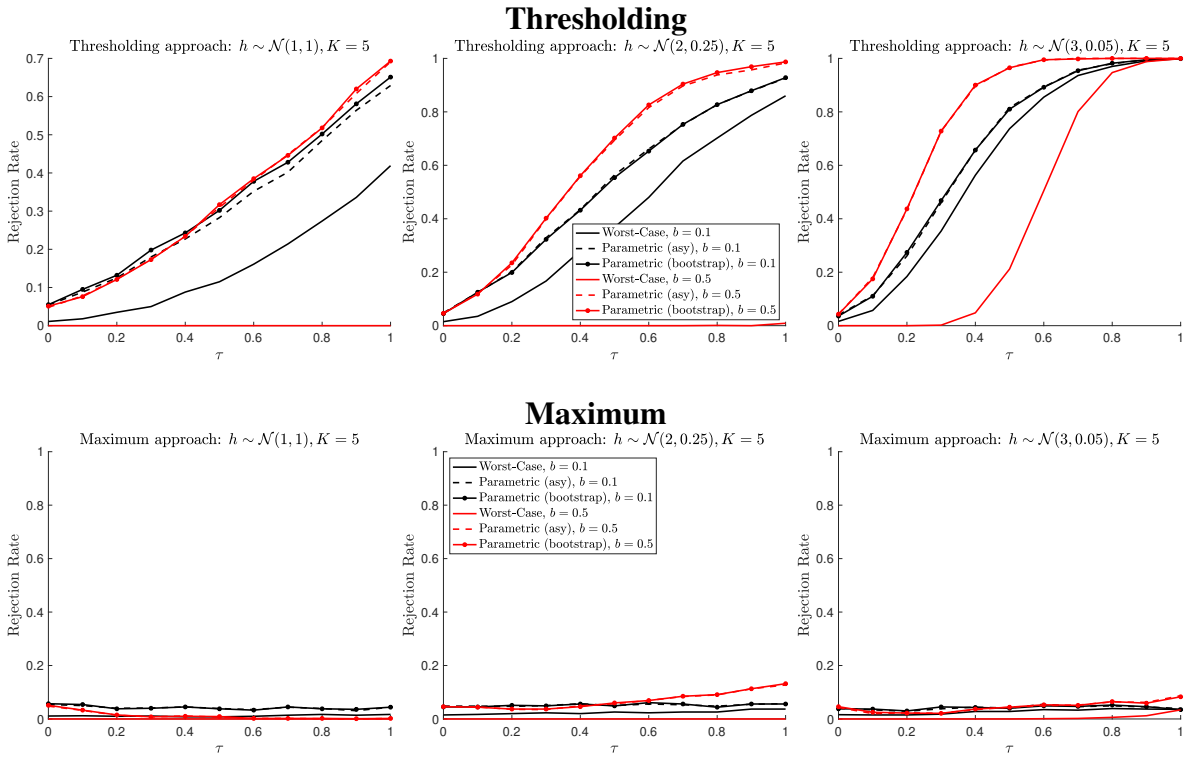

Figure 3.4. Power curves covariate selection with $K = 5$ and $t = 1.96$. Sample size is 1000.

The Caliper tests based on the parametric estimate of Π are more powerful for both $b = 0.1$ and $b = 0.5$. This is expected because this tests use additional information about the shape of the underlying distribution of true effects. As we can see, the results for the asymptotic and bootstrap versions of the tests are almost identical demonstrating the validity of the bootstrap procedure. In contrast to the case of the Robust Caliper test, the power of the parametric versions of the test is higher for larger value of b. The explanation for this is that larger b allows us to use more information about the shape of the distribution of true effects under the null.

Finally, when researchers use the maximum approach to p-hacking, non of the tests have power to detect it. As it was pointed out in Elliott et al. (2022a), the maximum approach does not lead to significant violations of the testable restrictions and is hard to detect.

Figure 3.5 displays the results when the sample size is taken to be $N = 5000$. The larger sample size naturally leads to higher power. At the same time, the figures are qualitatively similar to the ones in case $N = 1000$.

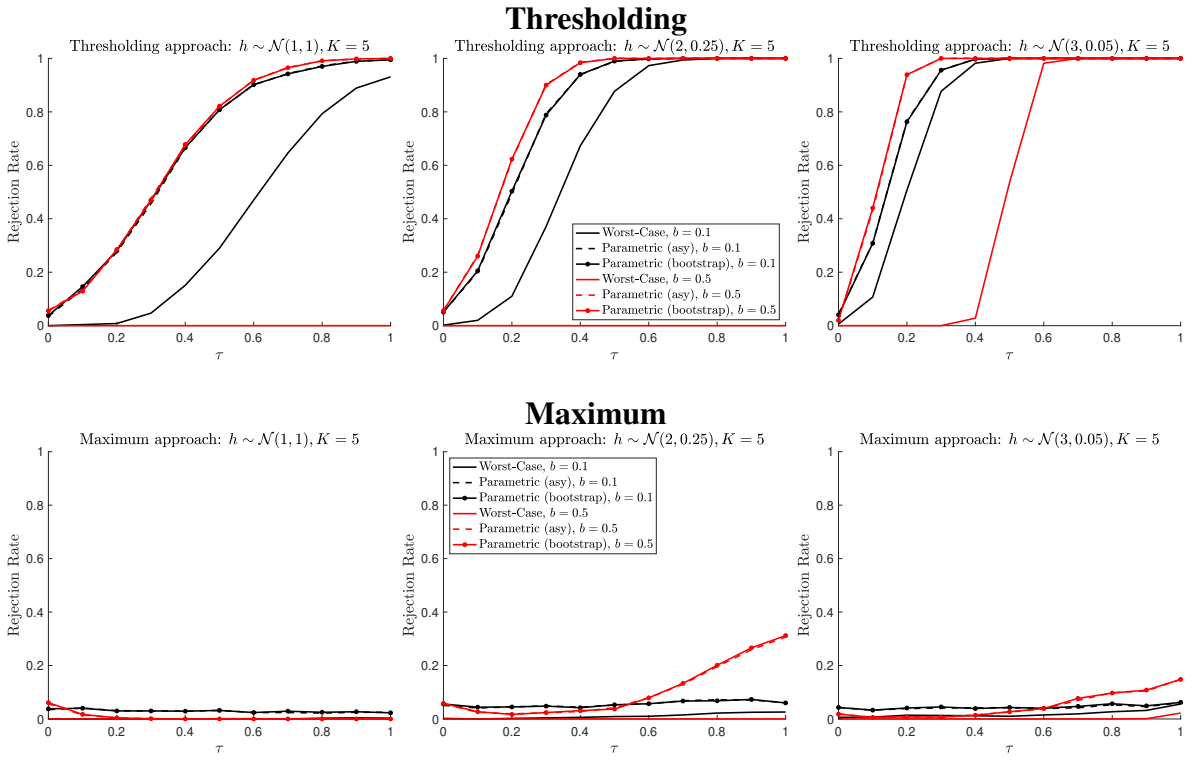

Figure 3.5. Power curves covariate selection with $K = 5$ and $t = 1.96$. Sample size is 5000.

3.5.3 *P*-hacking at Multiple Thresholds

In this section, we consider the scenario when researchers p-hack at three different significance levels: 1%, 5% and 10% (p-hacking occurs with equal probability at each threshold). The data-generating process is the same as in Section 3.5.2 but the p-hacked distribution is the mixture of p-hacked distributions for different thresholds. Figure C.7 in Appendix C.5 displays these distributions for different Π. We focus on the case $K = 5$, thresholding approach and use sample size $N = 1000$. To test the null hypothesis of no p-hacking we use joint Caliper tests. As in the previous Monte Carlo experiment, we use both Robust and parametric[8] versions of the test.

Figure 3.6 shows the results. All tests control size. As before, the power of the Robust test is very small for $b = 0.5$ due to the weakness of the upper bounds. At the same time, the

[8]Since the distribution of the moment inequalities test of Cox and Shi (2022) is non-standard, the bootstrap version of the parametric test is not available

power of the parametric test is higher in all cases and depends negatively on the bandwidth choice.

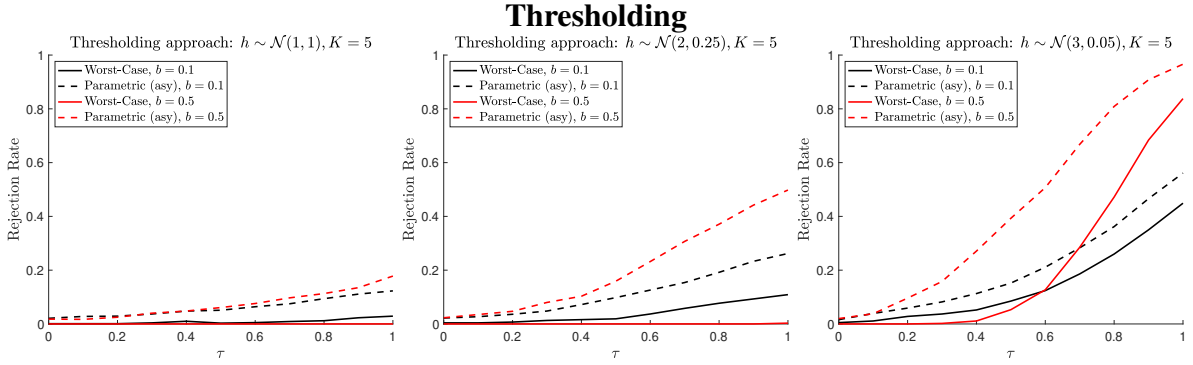

Figure 3.6. Power curves covariate selection with $K = 5$ and multiple thresholds. Sample size is 1000.

3.6 Application

Here we reanalyze the data collected by Brodeur et al. (2020b), which contain information about 21,740 t-tests from 684 articles published in Top 25 economic journals. After excluding observations with missing information, there are 21,156 tests from 680 papers. The data are divided into categories based on the estimation method used by researchers: Difference-in-Difference (DID), Instrumental Variables (IV), Randomized Control Trials (RCT) and Regression Discontinuity design (RDD).

We apply the standard Binomial test and the Robust Caliper test developed in this paper to these data and compare the results. Because the z-values may be correlated within papers, we use cluster-robust estimators of the variance for the Robust test. Following Brodeur et al. (2020b) we examine three most common significance thresholds: 2.576 (1% level), 1.96 (5% level) and 1.645 (10% level). We try seven values of b ($b = 0.05, 0.075, 0.1, 0.2, 0.3, 0.4, 0.5$). For every case we report $\underline{\tau}_{5\%}$, the lower end of the 95% confidence interval for the extent of p-hacking. We apply these tests to subgroups DID, IV, RCT and RDD separately. Table 3.1 replicates Table 3 from Brodeur et al. (2020b) and shows the results (p-values) of testing p-hacking at 5% threshold. The results for the 1% and 10% thresholds can be found in Appendix C.6.

As we can see from Table 3.1, the p-values of the Robust Caliper Test are always smaller than the ones coming from the Binomial test. This is expected because the Robust test uses the worst-case bound to test against in order to preserve size. For the values of b greater than or equal to 0.2, the Robust test does not reject the null of no p-hacking on any significance level. The reason for this is that for such large values of b, the fluctuation of the z-curve generated by the distribution of true effects can be quite large and cannot be distinguished from p-hacking. At the same time, for $b = 0.1$, the null hypothesis is rejected at 5% level for DID and IV and for $b = 0.075$ the null is additionally rejected at 10%. level for RCT. As for the extent of p-hacking, for $b = 0.075$ we can conclude with 95% confidence that at least 35% of DID results and at least 17% of IV results are p-hacked. The results for $b = 0.05$ are not very different from the results for $b = 0.075$.

Table 3.1. Binomial and Robust Caliper Tests, 5% Significance threshold (p-values)

	DID	IV	RCT	RDD
Proportion Significant in 1.96 ± 0.5	0.5341	0.5404	0.478	0.475
Binomial Test	0.0119	0.0026	0.9593	0.8777
Robust Caliper Test ($\underline{\tau}_{5\%}$)	1 (0)	0.9999 (0)	1 (0)	1 (0)
# Tests in 1.96 ± 0.5	1071	1162	1613	579
Proportion Significant in 1.96 ± 0.4	0.5353	0.5337	0.4909	0.4774
Binomial Test	0.0161	0.0168	0.7363	0.8232
Robust Caliper Test ($\underline{\tau}_{5\%}$)	0.9974 (0)	0.998 (0)	1 (0)	1 (0)
# Tests in 1.96 ± 0.4	893	965	1324	465
Proportion Significant in 1.96 ± 0.3	0.5264	0.5237	0.4946	0.4737
Binomial Test	0.073	0.0895	0.6227	0.8283
Robust Caliper Test ($\underline{\tau}_{5\%}$)	0.9484 (0)	0.9662 (0)	1 (0)	0.9998 (0)
# Tests in 1.96 ± 0.3	720	758	1023	361

Table 3.1. (cont.) Binomial and Robust Caliper Tests, 5% Significance threshold (p-values)

Proportion Significant in 1.96 ± 0.2	0.5506	0.5379	0.5043	0.4846
Binomial Test	0.0108	0.0389	0.396	0.6547
Robust Caliper Test ($\underline{\tau}_{5\%}$)	0.4373 (0)	0.5228 (0)	0.9779 (0)	0.9812 (0)
# Tests in 1.96 ± 0.2	494	515	704	227
Proportion Significant in 1.96 ± 0.1	0.6232	0.566	0.5656	0.5505
Binomial Test	0	0.0134	0.0052	0.1251
Robust Caliper Test ($\underline{\tau}_{5\%}$)	0.005 (0.36)	0.0441 (0.02)	0.1245 (0)	0.3907 (0)
# Tests in 1.96 ± 0.1	284	265	366	109
Proportion Significant in 1.96 ± 0.075	0.6491	0.595	0.578	0.5465
Binomial Test	0	0.0028	0.0036	0.1659
Robust Caliper Test ($\underline{\tau}_{5\%}$)	0.0055 (0.35)	0.0086 (0.17)	0.0666 (0)	0.2707 (0)
# Tests in 1.96 ± 0.075	228	200	282	86
Proportion Significant in 1.96 ± 0.05	0.6648	0.5929	0.6649	0.6207
Binomial Test	0	0.0111	0	0.024
Robust Caliper Test ($\underline{\tau}_{5\%}$)	0.0066 (0.3)	0.0127 (0.11)	0.0007 (0.34)	0.0753 (0)
# Tests in 1.96 ± 0.05	182	140	194	58
Total obs	5780	5158	7101	3117

Note: $[\underline{\tau}_{5\%},1]$ is the 95% confidence interval for the extent of p-hacking.

The results show that the conclusion of the standard Caliper test and the Robust Caliper test can be quite different especially when b is large (0.2 or above). The observed difference can be due to the fact that the standard Caliper test does not control size and is prone to false rejections when b is large.

3.6.1 Joint Test

In this section we apply the Robust Caliper test to testing for no p-hacking jointly at three significance thresholds that we consider in our analysis. To test moment inequalities for the

joint test we use the method of Cox and Shi (2022). As in the previous subsection, to account for within-paper dependence, we use cluster-robust estimators of the covariance matrix of sample proportions.

Table 3.2. Joint Robust Caliper Tests, {1%, 5%, 10%} Significance thresholds (p-values)

	DID	IV	RCT	RDD
Robust Caliper Test ($\underline{\tau}_{5\%}$)	1 (0)	1 (0)	1 (0)	1 (0)
# Tests in $t_i \pm 0.5, i = 1,2,3$	1785	1896	2747	1023
Robust Caliper Test ($\underline{\tau}_{5\%}$)	0.7096 (0)	1 (0)	1 (0)	1 (0)
# Tests in $t_i \pm 0.4, i = 1,2,3$	1600	1753	2461	917
Robust Caliper Test ($\underline{\tau}_{5\%}$)	0.9977 (0)	1 (0)	1 (0)	1 (0)
# Tests in $t_i \pm 0.3, i = 1,2,3$	1423	1586	2185	808
Robust Caliper Test ($\underline{\tau}_{5\%}$)	0.8925 (0)	0.3648 (0)	1 (0)	1 (0)
# Tests in $t_i \pm 0.2, i = 1,2,3$	1091	1190	1666	611
Robust Caliper Test ($\underline{\tau}_{5\%}$)	0.0569 (0)	0.2388 (0)	0.7222 (0)	0.8359 (0)
# Tests in $t_i \pm 0.1, i = 1,2,3$	635	656	914	330
Robust Caliper Test ($\underline{\tau}_{5\%}$)	0.0161 (0.01)	0.0905 (0)	0.4753 (0)	0.4038 (0)
# Tests in $t_i \pm 0.075, i = 1,2,3$	484	496	664	252
Robust Caliper Test ($\underline{\tau}_{5\%}$)	0.1036 (0)	0.0486 (0.01)	0.0172 (0.01)	0.2906 (0)
# Tests in $t_i \pm 0.05, i = 1,2,3$	362	348	457	170
Total obs	5780	5158	7101	3117

Note: [$\underline{\tau}_{5\%}$,1] is the 95% confidence interval for the extent of p-hacking.

Table 3.2 reports the results (p-values) of the tests for $b = 0.05, 0.075, 0.1, 0.2, 0.3, 0.4, 0.5$. As we can see, the joint test is able to reject the null of no p-hacking at 5% significance level for DID with $b = 0.075$ and for IV and RCT when $b = 0.05$. For $b = 0.1$, the tests reject on a subsample of DID at 10% level. The reported value of $\underline{\tau}_{5\%}$ do not exceed 1% in all of the cases. Overall, it can be observed that the amount of rejections by the joint test is much smaller than by

the test that is focused exclusively on the 5% threshold. One possible reason for that is that there is not as much p-hacking occurring at 1% and 10% thresholds and thus the joint test loses power by taking them into consideration.

3.7 Conclusion

This paper analyses the Caliper test that is one of the most commonly used statistical methods for testing p-hacking and publication bias given the sample distribution of reported z-statistics. We demonstrate that the version of the test routinely used by researchers in practice fails to control size and may lead to a significant overrejection of the null hypothesis of no p-hacking for certain distributions of true effects. We develop a Robust version of the Caliper test that is agnostic to the distribution of true effects and always controls size. We show how the proposed test can be used to draw inference on the extent of p-hacking in the literature. In addition, we propose a way of incorporating distributional assumptions regarding underlying true effects to construct more powerful Caliper tests.

We confirm validity of the proposed tests and examine their power using Monte Carlo experiments. We apply the Robust Caliper test to reanalyze a dataset from the literature and compare it to the standard test. Given the validity of the Robust test, we recommend researchers to use it in practice instead of the non-robust version.

Chapter 3, in part, is currently being prepared for submission for publication of the material. The dissertation author is the sole author of this material.

Appendix

A Additional results and proofs for Chapter 1

A.1 Additional details Section 1.4.3

A.1.1 Bounds on proportions and their differences

The bounds on the proportions and their differences implied by hypothesis (1.13) are not sharp in general. Here we derive sharp bounds by directly extremizing the proportions and their differences.

For the one-sided t-tests, the population proportion, π_j, can be written as

$$\begin{aligned}
\pi_j = \int_{x_{j-1}}^{x_j} g_1(p)dp &= \int_{x_{j-1}}^{x_j} \int_{[0,\infty)} e^{-h^2/2} e^{hcv_1(p)} d\Pi(h)dp \\
&= \int_{[0,\infty)} \left(\int_{x_{j-1}}^{x_j} e^{-h^2/2} e^{hcv_1(p)} dp \right) d\Pi(h) \\
&= \int_{[0,\infty)} \left(\int_{cv_1(x_j)}^{cv_1(x_{j-1})} \phi(t-h)dt \right) d\Pi(h) \\
&= \int_{[0,\infty)} \lambda_{1,j}(cv_1,h) d\Pi(h),
\end{aligned}$$

where $\lambda_{1,j}(cv,h) := \Phi(cv(x_{j-1})-h) - \Phi(cv(x_j)-h)$. For the two-sided t-tests, $\pi_j = \int_{x_{j-1}}^{x_j} g_2(p)dp = \int_{\mathbb{R}} \lambda_{2,j}(cv_2,h)d\Pi(h)$, where $\lambda_{2,j}(cv,h) := \lambda_{1,j}(cv,h) + \lambda_{1,j}(cv,-h)$.

Since $\lambda_{1,j}(cv_1,h)$, as a function of h, attains its maximum at $h_j^* = \frac{cv_1(x_{j-1})+cv_1(x_j)}{2}$, for the one-sided t-tests $\pi_j \leq 2\Phi\left(\frac{cv_1(x_{j-1})-cv_1(x_j)}{2}\right) - 1 := \vartheta_{1,j}^{(0)}$. In case of the two-sided t-tests, the bound, $\vartheta_{2,j}^{(0)} := \max_{h\in\mathbb{R}} \lambda_{2,j}(cv_2,h)$, can be calculated numerically.

For the bounds on the kth differences of π's, note that, for $j = 1,\ldots,J-k$, $\Delta_j^k =$

$\sum_{i=0}^{k}(-1)^i\binom{k}{i}\pi_{k+j-i}$ and therefore

$$|\Delta_j^k| \leq \vartheta_{s,j}^{(k)} := \max_{h\in\mathscr{H}_{(s)}}\left\{\sum_{i=0}^{k}(-1)^{i+k}\binom{k}{i}\lambda_{s,k+j-i}(cv_s,h)\right\}, \quad j=1,\ldots,J-k,$$

where $\mathscr{H}_{(1)}=[0,\infty)$, $\mathscr{H}_{(2)}=\mathbb{R}$, and $s=1$ and $s=2$ for the one- and two-sided t-tests, respectively. These bounds can be computed numerically.

A.1.2 Null hypothesis

The null hypothesis formulated in terms of the proportions is

$$H_0:\ 0\leq(-1)^k\Delta^k\leq\vartheta_s^{(k)},\ \sum_{j=1}^{J}\pi_j=1,\ \text{ for all } k=0,\ldots,K, \tag{13}$$

where Δ^k is a $(J-k)\times 1$ vector of kth differences of π's, $\Delta^0=\pi$, $\vartheta_s^{(k)}:=(\vartheta_{s,1}^{(k)},\ldots,\vartheta_{s,J-k}^{(k)})'$ is the vector of upper bounds on $|\Delta^k|$ (cf. Appendix A.1.1), $s=1$ for one-sided tests, and $s=2$ for two-sided tests. The inequalities in (13) are interpreted element-wise.

Let D_m be $(m-1)\times m$ differencing matrix of the following form:

$$D_m := \begin{pmatrix} -1 & 1 & 0 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & -1 & 1 \end{pmatrix}.$$

In addition, define the $J\times 1$ vector $e_J:=(0,\ldots,1)'$, $(J-1)\times 1$ vector $i_{J-1}:=(1,\ldots,1)'$, and matrix $F:=[-I_{J-1},i_{J-1}]'$. Using this notation, we can write $(-1)^k\Delta^k=D^k\pi,\ k=1,\ldots,K$, where $D^k:=(-1)^kD_{J-k+1}\times\cdots\times D_J$. Note that the restrictions under the null are equivalent to $\mathscr{D}_K\pi\geq c$ and $\pi=e_J-F\pi_{-J}$, where $\mathscr{D}_K=[-1,1]'\otimes[I_J,D^{1'},\ldots,D^{K'}]'$ and $c=[\vartheta_s^{(0)'},\ldots,\vartheta_s^{(K)'},0'_{(K+1)(J-K/2)\times 1}]'$ The symbol $\otimes$ denotes the Kronecker product. We can thus express the null hypothesis (13) as $H_0:\ A\pi_{-J}\leq b$, where $A:=\mathscr{D}_KF$ and $b:=\mathscr{D}_Ke_J-c$.

When testing on a subinterval $(0,\alpha]$, the bounds need to be re-scaled. We use a consistent (under the null) estimator of $G(\alpha)$ to re-scale the bounds. In particular, we use bounds $\vartheta_{s,j}^{(k)}=$

$\vartheta_{s,j}^{(k)}/\hat{G}(\alpha)$, where $\hat{G}(\alpha)$ is the fraction of p-values below α.

A.2 Proofs

A.2.1 Proof of Lemma 1.1

Note that for claim (i) $\{cv(p) : p \in (0,1)\} = \mathbb{R}$ and for claims (ii) and (iii) $\{cv(p) : p \in (0,1)\} = (0,\infty)$.

Claim (i): In this case $f(x) = \phi(x)$ and $f_h(x) = \phi(x-h)$. It follows that, for all $h \geq 0$, $f_h'(x)f(x) - f'(x)f_h(x) = h\phi(x)\phi(x-h) \geq 0$.

Claim (ii): In this case $f(x) = 2\phi(x)$ and $f_h(x) = \phi(x-h) + \phi(x+h)$, where $x \geq 0$. After taking derivatives and collecting terms we get

$$f_h'(x)f(x) - f'(x)f_h(x) = 2\phi(x)h(\phi(x-h) - \phi(x+h)) = 2\phi(x)\phi(x+h)h(e^{2xh} - 1) \geq 0,$$

because $h(e^{2xh} - 1) \geq 0$ for any h.

Claim (iii): In this case $f(x) := f(x;d) = \frac{1}{2^{d/2}\Gamma(d/2)}x^{d/2-1}e^{-x/2}$ and $f_h(x) = \sum_{j=0}^{\infty}\frac{e^{-h/2}(h/2)^j}{j!}f(x;d+2j)$, where $x > 0$. Note that $f'(x;d) = f(x;d)\left((d-2)x^{-1} - 1\right)/2$. After taking derivatives and collecting terms we get

$$\begin{aligned} f_h'(x)f(x) - f'(x)f_h(x) &= \sum_{j=0}^{\infty}\frac{e^{-h/2}(h/2)^j}{2j!}f(x;d+2j)f(x;d)\left[((d+2j-2)x^{-1} - 1) - ((d-2)x^{-1} - 1)\right] \\ &= \sum_{j=0}^{\infty}\frac{e^{-h/2}(h/2)^j}{j!}f(x;d+2j)f(x;d)jx^{-1} \geq 0, \end{aligned}$$

since every term in the last sum is non-negative. □

A.2.2 Proof of Theorem 1.1

Recall that $\beta(p,h) = 1 - F_h(cv(p))$, where $cv(p) = F^{-1}(1-p)$. Under Assumption 1.1,

$$\begin{aligned}\frac{\partial^2 \beta(p,h)}{\partial p^2} &= \frac{f_h'(cv(p))cv'(p)f(cv(p)) - f'(cv(p))cv'(p)f_h(cv(p))}{f(cv(p))^2} \\ &= \frac{cv'(p)}{f(cv(p))^2}\left[f_h'(cv(p))f(cv(p)) - f'(cv(p))f_h(cv(p))\right].\end{aligned}$$

Non-increasingness of g now follows by Assumption 1.2 and because $cv'(p)/f(cv(p))^2 \leq 0$. Continuous differentiability is implied by Assumption 1.1. □

A.2.3 Proofs of Theorems 1.2 and 1.3

Note that the p-curves for the one-sided and two-sided t-tests are given by

$$g_1(p) = \int_{[0,\infty)} \Psi(cv_1(p),h)\exp\{-h^2/2\}d\Pi(h), \tag{14}$$

$$g_2(p) = \frac{1}{2}\int_{\mathbb{R}} (\Psi(cv_2(p),h) + \Psi(cv_2(p),-h))\exp\{-h^2/2\}d\Pi(h) \tag{15}$$

where $\Psi(x,y) := \exp\{xy\}$. We start by proving an auxiliary lemma about $\Psi(x,y)$.

Lemma A.1. *For $k \geq 1$, the k^{th} derivative of $\Psi(cv_s(p),h)$ is*

$$\Psi^{(k)}(cv_s(p),h) = (-1)^k \frac{h\sum_{j=0}^{k-1} A_j^k(cv_s(p))[cv_s(p)+h]^j}{s^k(\phi(cv_s(p)))^k}\Psi(cv_s(p),h),$$

where coefficients $A_j^k(cv_s(p))$ are polynomials in $cv_s(p)$ with non-negative coefficients and $s = 1$ for one-sided and $s = 2$ for two-sided t-tests.

Proof. By direct computation, the first derivative of $\Psi(cv_s(p),h)$ with respect to p is $\Psi^{(1)}(cv_s(p),h) = -\frac{h}{s\phi(cv_s(p))}\Psi(cv_s(p),h)$. We use induction to derive the kth derivative of $\Psi(cv_s(p),h)$. Suppose

that for $k > 1$

$$\Psi^{(k)}(cv_s(p),h) = (-1)^k \frac{h\sum_{j=0}^{k-1} A_j^k(cv_s(p))[cv_s(p)+h]^j}{s^k(\phi(cv_s(p)))^k}\Psi(cv_s(p),h),$$

where coefficients $A_j^k(cv_s(p))$ are polynomials in $cv_s(p)$ with non-negative coefficients. Define $B_0^k = (k-1)cv_s(p)A_0^k(cv_s(p))$, $B_j^k = (k-1)cv_s(p)A_j^k(cv_s(p)) + A_{j-1}^k(cv_s(p))$ for $j = 1,\ldots,k-1$, and $B_k^k = A_{k-1}^k(cv_s(p))$; $C_j^k = \partial A_j^k(cv_s(p))/\partial cv_s(p) + (j+1)A_{j+1}^k(cv_s(p))$ for $j = 0,\ldots,k-2$, $C_{k-1}^k = \partial A_{k-1}^k(cv_s(p))/\partial cv_s(p)$, and $C_k^k = 0$. Now differentiate $\Psi^{(k)}(cv_s(p),h)$ with respect to p to get

$$\begin{aligned}
\Psi^{(k+1)}(cv_s(p),h) &= (-1)^{k+1}\frac{h^2\sum_{j=0}^{k-1} A_j^k(cv_s(p))[cv_s(p)+h]^j}{s^{k+1}(\phi(cv_s(p)))^{k+1}}\Psi(cv_s(p),h) \\
&\quad +(-1)^{k+1}\frac{(hcv_s(p)k)\sum_{j=0}^{k-1} A_j^k(cv_s(p))[cv_s(p)+h]^j}{s^{k+1}(\phi(cv_s(p)))^{k+1}}\Psi(cv_s(p),h) \\
&\quad +(-1)^{k+1}\frac{h\sum_{j=0}^{k-1}(\partial A_j^k(cv_s(p))/\partial cv_s(p))[cv_s(p)+h]^j}{s^{k+1}(\phi(cv_s(p)))^{k+1}}\Psi(cv_s(p),h) \\
&\quad +(-1)^{k+1}\frac{h\sum_{j=1}^{k-1} jA_j^k(cv_s(p))[cv_s(p)+h]^{j-1}}{s^{k+1}(\phi(cv_s(p)))^{k+1}}\Psi(cv_s(p),h) \\
&= (-1)^{k+1}\frac{\Psi(cv_s(p),h)}{s^{k+1}(\phi(cv_s(p)))^{k+1}}\left\{h\sum_{j=0}^{k}(B_j^k + C_j^k)[cv_s(p)+h]^j\right\}.
\end{aligned}$$

Since $A_j^k(cv_s(p)), j = 0,\ldots,k-1$ are polynomials with non-negative coefficients, B_j^k and C_j^k are also polynomials with non-negative coefficients for every $j = 0,\ldots,k$. It follows that

$$\Psi^{(k+1)}(cv_s(p),h) = (-1)^{k+1}\frac{h\sum_{j=0}^{k} A_j^{k+1}(cv_s(p))[cv_s(p)+h]^j}{s^{k+1}(\phi(cv_s(p)))^{k+1}}\Psi(cv_s(p),h),$$

where $A_j^{k+1}(cv_s(p)) = B_j^k + C_j^k, j = 0,\ldots,k$. This completes the induction step. □

Using Lemma A.1, we now proof Theorem 1.2 and Theorem 1.3.

Proof of Theorem 1.2. Lemma A.1 and equations (14)–(15) directly imply that $0 \le (-1)^k g_1^{(k)}(p)$, for $p \in (0,1/2]$ and $0 \le (-1)^k g_2^{(k)}(p)$, for $p \in (0,1)$ for $k = 1,2,\ldots$. The result for the two-sided

case follows from the fact that $h\{[cv_2(p)+h]^j\Psi(cv_2(p),h)-[cv_2(p)-h]^j\Psi(cv_2(p),-h)\}\geq 0$ for every $j\in\mathbb{N}$ and every $h\in\mathbb{R}$. □

Proof of Theorem 1.3. Consider first the one-sided t-test. Lemma A.1 implies that

$$(-1)^k g_1^{(k)}(p)\leq \mathscr{B}_1^{(k)}(p):=\max_{h\geq 0}\left\{|\Psi^{(k)}(cv_1(p),h)|\exp\{-h^2/2\}\right\},$$

where the inequality holds for every $p\in(0,1)$ and the maximum is finite for every $p\in(0,1)$ since $|\Psi^{(k)}(cv_1(p),h)|\exp\{-h^2/2\}$ is finite for every $h\geq 0$ and converges to zero as h goes to infinity. For the upper bound on $g_1(p)$, note that for $p\in(0,1/2]$, $\max_{h\geq 0}\left\{|\Psi(cv_1(p),h)|\exp\{-h^2/2\}\right\}=$ $\Psi(cv_1(p),cv_1(p))\exp\{-cv_1^2(p)/2\}=\exp\{cv_1^2(p)/2\}$. For $p>1/2$ and $h\geq 0$, $hcv_1(p)-cv_1^2(p)/2<$ 0 and hence $g_1(p)\leq 1$.

For two-sided tests, by the above arguments and symmetry, we have

$$(-1)^k g_2^{(k)}(p)\leq \mathscr{B}_2^{(k)}(p):=\max_{h\in\mathbb{R}}\left\{|\Psi^{(k)}(cv_2(p),h)+\Psi^{(k)}(cv_2(p),-h)|\exp\{-h^2/2\}/2\right\},$$

where the upper bound is finite for every $p\in(0,1)$.

For the upper bound on $g_2(p)$, one can show that for $p\geq 2(1-\Phi(1))$, the first-order condition for maximizing $(\Psi(cv_2(p),h)+\Psi(cv_2(p),-h))\exp\{-h^2/2\}/2$ has only one solution, $h_o=0$. By checking second-order conditions we can verify that 0 is the maximum. For $p<2(1-\Phi(1))$, 0 becomes local minimum, and there are two additional non-zero symmetric solutions to the first-order condition that satisfy the second-order condition for a maximum and result in identical values of the objective function.

□

B Additional results and derivation for Chapter 2

B.1 Detailed Derivations Section 2.3

B.1.1 Selecting Control Variables in Linear Regression

B.1.1.1 p-Curve under p-Hacking

We denote by $\hat{\sigma}_j$ the standard error of the estimator of β when using Z_j as the control variable ($j = 1,2$). Under our assumptions, because the variance of U is known, we have

$$\hat{\sigma}_j^2 = \frac{1}{1-\gamma^2}, \quad j = 1,2.$$

Therefore, the t-statistic for testing $H_0 : \beta = 0$ is distributed as follows

$$T_j = \frac{\sqrt{N}\hat{\beta}_j}{\hat{\sigma}_j} \overset{d}{=} h + \frac{W_{xu} - \gamma W_{z_j u}}{\sqrt{1-\gamma^2}}, \quad j = 1,2,$$

where

$$\begin{pmatrix} W_{xu} \\ W_{z_1 u} \\ W_{z_2 u} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \gamma & \gamma \\ \gamma & 1 & \gamma^2 \\ \gamma & \gamma^2 & 1 \end{pmatrix} \right).$$

Thus, conditional on h,

$$\begin{pmatrix} T_1 \\ T_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} h \\ h \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right),$$

where the correlation is $\rho = 1 - \gamma^2$. As the control variables and X_i become more correlated (larger γ), ρ becomes smaller.

The CDF of P_r on $(0,1)$ for the threshold case is

$$\begin{aligned}
G_h^t(p) &= \Pr(P_r \le p) \\
&= \Pr(P_1 \le p \mid P_1 \le \alpha)\Pr(P_1 \le \alpha) \\
&\quad + \Pr(\min\{P_1, P_2\} \le p \mid P_1 > \alpha)\Pr(P_1 > \alpha) \\
&= \Pr(P_1 \le \min\{p,\alpha\}) + (1 - \Pr(P_1 > p, P_2 > p \mid P_1 > \alpha))\Pr(P_1 > \alpha) \\
&= \Pr(T_1 \ge z_0(\min\{p,\alpha\})) + \Pr(T_1 < z_0(\alpha)) - \Pr(T_1 < z_0(\max\{p,\alpha\}), T_2 < z_0(p)) \\
&= 1 - \Phi(z_h(\min\{p,\alpha\})) + \Phi(z_h(\alpha)) - \int_{-\infty}^{z_h(p)} \int_{-\infty}^{z_h(\max\{p,\alpha\})} f(x,y;\rho)dxdy,
\end{aligned}$$

where $f(x,y;\rho) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\{-\frac{x^2-2\rho xy+y^2}{2(1-\rho^2)}\}$.

For $p \in (0,\alpha)$, differentiating $G_h^t(p)$ with respect to p yields:

$$\begin{aligned}
\frac{dG_h^t(p)}{dp} &= \frac{dz_h(p)}{dp}\left[-\phi(z_h(p)) - \int_{-\infty}^{z_h(\alpha)} f(z_h(p), y;\rho)dy\right] \\
&= \frac{\phi(z_h(p))\left[1 + \Phi\left(\frac{z_h(\alpha)-\rho z_h(p)}{\sqrt{1-\rho^2}}\right)\right]}{\phi(z_0(p))}.
\end{aligned}$$

For $p \in (\alpha,1)$, the derivative is

$$\frac{dG_h^t(p)}{dp} = \frac{2\phi(z_h(p))\Phi\left(\frac{z_h(p)-\rho z_h(p)}{\sqrt{1-\rho^2}}\right)}{\phi(z_0(p))}.$$

It follows that the PDF of p-values is

$$g_1^t(p) = \int_{\mathscr{H}} \frac{dG_h^t(p)}{dp} d\Pi(h) = \int_{\mathscr{H}} \frac{\phi(z_h(p))\Upsilon_1^t(p;\alpha,h,\rho)}{\phi(z_0(p))} d\Pi(h),$$

where $\Upsilon_1^t(p;\alpha,h,\rho) = 1_{\{p\le\alpha\}}\left[1 + \Phi\left(\frac{z_h(\alpha)-\rho z_h(p)}{\sqrt{1-\rho^2}}\right)\right] + 1_{\{p>\alpha\}} 2\Phi\left(\frac{z_h(p)-\rho z_h(p)}{\sqrt{1-\rho^2}}\right)$. The final expression follows because $\phi(z_h(p))/\phi(z_0(p)) = \exp\left(hz_0(p) - \frac{h^2}{2}\right)$.

For the case when the researchers report the minimum of two p-values, $P_r = \min\{P_1, P_2\}$, we have

$$\begin{aligned} G_h^m(p) &= \Pr(P_r \le p) \\ &= \Pr(P_1 \le p, P_1 \le P_2) + \Pr(P_2 \le p, P_2 < P_1) \\ &= \Pr(T_1 \ge z_0(p), T_1 \ge T_2) + \Pr(T_2 \ge z_0(p), T_2 > T_1) \\ &= 2\Pr(\xi_1 \ge z_h(p), \xi_1 \ge \xi_2) \\ &= 2\int_{-\infty}^{z_h(p)} \int_{z_h(p)}^{\infty} f(x,y;\rho)dxdy + 2\int_{z_h(p)}^{\infty} \int_{y}^{\infty} f(x,y;\rho)dxdy, \end{aligned}$$

where $\xi_j = T_j - h$, $j = 1,2$.

The derivative of $G_h^m(p)$ with respect to p is

$$\begin{aligned} \frac{dG_h^m(p)}{dp} &= 2\frac{dz_h(p)}{dp}\left[\int_{z_h(p)}^{\infty} f(x, z_h(p);\rho)dx - \int_{-\infty}^{z_h(p)} f(z_h(p), y;\rho)dy - \int_{z_h(p)}^{\infty} f(x, z_h(p);\rho)dx\right] \\ &= 2\frac{\phi(z_h(p))}{\phi(z_0(p))}\Phi\left(z_h(p)\sqrt{\frac{1-\rho}{1+\rho}}\right). \end{aligned}$$

Therefore, the PDF of p-values is

$$g_1^m(p) = 2\int_{\mathscr{H}} \exp\left(hz_0(p) - \frac{h^2}{2}\right)\Phi\left(z_h(p)\sqrt{\frac{1-\rho}{1+\rho}}\right)d\Pi(h).$$

B.1.1.2 Bias of the p-Hacked Estimator

Fix h for now. We have $\hat{\beta}_r^t = \hat{\beta}_1 + (\hat{\beta}_2 - \hat{\beta}_1)1_{\{T_2>T_1, T_1<z_0(\alpha)\}}$. The bias in the p-hacked estimate is given by

$$\begin{aligned} E\hat{\beta}_r^t - \beta &= E\left[\hat{\beta}_1 - \beta + (\hat{\beta}_2 - \hat{\beta}_1)1_{\{T_2>T_1, T_1<z_0(\alpha)\}}\right] \\ &= E\left[(\hat{\beta}_2 - \hat{\beta}_1)1_{\{T_2>T_1, T_1<z_0(\alpha)\}}\right] \\ &= \frac{1}{\sqrt{N}\sqrt{\rho}} E\left[(\xi_2 - \xi_1)1_{\{\xi_2>\xi_1, \xi_1<z_h(\alpha)\}}\right] \\ &= \frac{1}{\sqrt{N}\sqrt{\rho}} E\left[V 1_{\{V>0, \xi_1<z_h(\alpha)\}}\right], \end{aligned}$$

where $V = \xi_2 - \xi_1 \sim \mathcal{N}(0, 2(1-\rho))$ and $E[V\xi_1] = -(1-\rho)$. Now

$$\begin{aligned} E\left[V 1_{\{V>0_1, \xi_1<z_h(\alpha)\}}\right] &= \int_0^\infty \int_{-\infty}^{z_h(\alpha)} v f_{V,\xi_1}(v,x)dxdv \\ &= \int_0^\infty \int_{-\infty}^{z_h(\alpha)} v f_{\xi_1|V}(x|v) f_V(v)dxdv \\ &= \int_0^\infty v f_V(v) \left(\int_{-\infty}^{z_h(\alpha)} f_{\xi_1|V}(x|v)dx\right) dv, \end{aligned}$$

and

$$\begin{aligned} \int_{-\infty}^{z_h(\alpha)} f_{\xi_1|V}(x|v)dx &= \Pr\left(\xi_1 < z_h(\alpha) \mid V = v\right) \\ &= \Pr\left(\frac{\xi_1 + v/2}{\sqrt{\frac{1+\rho}{2}}} < \frac{z_h(\alpha) + v/2}{\sqrt{\frac{1+\rho}{2}}}\right) \\ &= \Phi\left(\frac{z_h(\alpha) + v/2}{\sqrt{\frac{1+\rho}{2}}}\right). \end{aligned}$$

So now we have

$$E\hat{\beta}_r^t - \beta = \frac{1}{\sqrt{N}\sqrt{\rho}} \int_0^\infty v\Phi\left(\frac{z_h(\alpha) + v/2}{\sqrt{\frac{1+\rho}{2}}}\right) f_V(v)dv,$$

and the final expression follows by direct integration.

For the minimum approach, $\hat{\beta}_r^m = \hat{\beta}_1 + (\hat{\beta}_2 - \hat{\beta}_1)1_{\{T_2>T_1\}}$. The bias in the p-hacked estimate is given by

$$\begin{aligned} E\hat{\beta}_r^m - \beta &= E\left[\hat{\beta}_1 - \beta + (\hat{\beta}_2 - \hat{\beta}_1)1_{\{T_2>T_1\}}\right] \\ &= \frac{1}{\sqrt{N}\sqrt{\rho}} E\left(V 1_{\{V>0\}}\right). \end{aligned}$$

Now $E\left[V 1_{\{V>0\}}\right] = \sqrt{2(1-\rho)}\phi(0)$ so

$$E\hat{\beta}_r^m - \beta = \frac{1}{\sqrt{N}}\sqrt{\frac{2(1-\rho)}{\rho}}\phi(0).$$

It follows that $E\hat{\beta}_r^t \leq E\hat{\beta}_r^m$ because

$$\begin{aligned} EV1_{\{V>0,\xi_1<z_h(\alpha)\}} &= \int_0^\infty v f_V(v)\left(\int_{-\infty}^{z_h(\alpha)} f_{\xi_1|V}(x|v)dx\right)dv \\ &\leq \int_0^\infty v f_V(v)dv. \end{aligned}$$

B.1.2 Selecting amongst Instruments in IV Regression

B.1.2.1 p-Curve under p-Hacking

Since Z_1 and Z_2 are assumed to be uncorrelated, the IV estimator with 2 instruments is

$$\hat{\beta}_{12} = \beta + \left[\frac{(\sum X_i Z_{1i})^2}{\sum Z_{1i}^2} + \frac{(\sum X_i Z_{2i})^2}{\sum Z_{2i}^2} + o_P(1)\right]^{-1}\left[\frac{\sum X_i Z_{1i}\sum U_i Z_{1i}}{\sum Z_{1i}^2} + \frac{\sum X_i Z_{2i}\sum U_i Z_{2i}}{\sum Z_{2i}^2}\right]$$

with asymptotic variance $1/2\gamma^2$. Therefore, the t-statistic is

$$T_{12} = \sqrt{N}\hat{\beta}_{12}\sqrt{2}|\gamma| \xrightarrow{d} h + \frac{W_1 + W_2}{2},$$

where $(W_1, W_2)' \sim \mathcal{N}(0, I_2)$. With one instrument,

$$\hat{\beta}_j = \beta + \frac{\sum Z_{ji} U_i}{\sum X_i Z_{ji}}, \quad j = 1, 2,$$

and the asymptotic variance is $1/\gamma^2$ and

$$T_j = \sqrt{N}\hat{\beta}_j |\gamma| \xrightarrow{d} h + W_j.$$

Note that T_{12} is asymptotically equivalent to $\frac{T_1+T_2}{\sqrt{2}}$.

For now, we fix h. Define $z_h(p) = z_0(p) - h$ and $D_h(p) = \sqrt{2} z_{\sqrt{2}h}(p)$, where $z_0(p) = \Phi^{-1}(1-p)$. The (asymptotic) CDF of P_r on $(0, 1/2]$ is

$$\begin{aligned}
G_h^t(p) &= \Pr(P_r \le p) \\
&= \Pr(P_{12} \le p \mid P_{12} \le \alpha) \Pr(P_{12} \le \alpha) \\
&\quad + \Pr(\min\{P_1, P_2, P_{12}\} \le p \mid P_{12} > \alpha) \Pr(P_{12} > \alpha) \\
&= \Pr(P_{12} \le \min\{p, \alpha\}) + \Pr(P_{12} > \alpha) \\
&\quad - \Pr(P_1 > p, P_2 > p, P_{12} > p | P_{12} > \alpha) \Pr(P_{12} > \alpha) \\
&= \Pr(T_{12} \ge z_0(\min\{p, \alpha\})) + \Pr(T_{12} < z_0(\alpha)) \\
&\quad - \Pr(T_1 < z_0(p), T_2 < z_0(p), T_{12} < z_0(\max\{p, \alpha\})) \\
&= 1 - \Phi(z_{\sqrt{2}h}(\min\{p, \alpha\})) + \Phi(z_{\sqrt{2}h}(\alpha)) - \Phi(z_h(p))\Phi(D_h(\max\{p, \alpha\}) - z_h(p)) \\
&\quad - \int_{D_h(\max\{p,\alpha\}) - z_h(p)}^{z_h(p)} \phi(x)\Phi(D_h(\max\{p, \alpha\}) - x) dx.
\end{aligned}$$

The last equality follows because for $p \le 1/2$ we have $2z_h(p) > D_h(\max\{p, \alpha\})$, $\Pr(T_1 < z_0(p), T_2 < z_0(p), T_{12} < z_0(\max\{p, \alpha\})) = \Pr(W_1 < z_h(p), W_2 < z_h(p), W_1 + W_2 < D_h(\max\{p, \alpha\}))$∎

and

$$\begin{aligned}\Pr(W_1 < a, W_2 < a, W_1 + W_2 < b) &= \int_{-\infty}^{a}\int_{-\infty}^{b-a} \phi(x)\phi(y)dxdy \\ &\quad + \int_{b-a}^{a}\int_{-\infty}^{b-x} \phi(x)\phi(y)dydx \\ &= \Phi(a)\Phi(b-a) + \int_{b-a}^{a} \phi(x)\Phi(b-x)dx\end{aligned}$$

for $2a > b$.

The derivative of $G_h^t(p)$ with respect to p on $(0,\alpha)$ is

$$\frac{dG_h^t(p)}{dp} = \frac{\phi(z_{\sqrt{2}h}(p)) + 2C(z_h(p), D_h(\alpha))}{\phi(z_0(p))}, \quad p \in (0,\alpha),$$

where $C(x,y) := \phi(x)\Phi(y-x)$.

For $p \in (\alpha, 1/2)$ the derivative is

$$\frac{dG_h^t(p)}{dp} = \frac{\phi(z_{\sqrt{2}h}(p))(1 - 2\Phi((1-\sqrt{2})z_0(p))) + 2C(z_h(p), D_h(p))}{\phi(z_0(p))}, \quad p \in (\alpha, 1/2).$$

For $p > 1/2$, we have $2z_h(p) < D_h(\max\{p,\alpha\})$, and similar arguments yield

$$\begin{aligned}G_h^t(p) &= 1 - \Pr(W_1 < z_h(p), W_2 < z_h(p), W_1 + W_2 < D_h(\max\{p,\alpha\})) \\ &= 1 - \int_{-\infty}^{z_h(p)}\int_{-\infty}^{z_h(p)} \phi(x)\phi(y)dxdy \\ &= 1 - \Phi^2(z_h(p))\end{aligned}$$

and

$$\frac{dG^t(p)}{dp} = \frac{2\phi(z_h(p)\Phi(z_h(p))}{\phi(z_0(p)}, \quad p > 1/2.$$

Since $g^t(p) = \int_{\mathscr{H}} \frac{G_h^t(p)}{dp} d\Pi(h)$, we have

$$g_2^t(p) = \int_{\mathscr{H}} \exp\left(hz_0(p) - \frac{h^2}{2}\right) \Upsilon_2^t(p;\alpha,h) d\Pi(h),$$

where $\zeta(p) = 1 - 2\Phi((1-\sqrt{2})z_0(p))$, $cv_1(p) = z_0(p)$ and

$$\Upsilon_2^t(p;\alpha,h) = \begin{cases} \frac{\phi(z_{\sqrt{2}h}(p))}{\phi(z_h(p))} + 2\Phi(D_h(\alpha) - z_h(p)), & \text{if } 0 < p \leq \alpha, \\ \frac{\phi(z_{\sqrt{2}h}(p))}{\phi(z_h(p))}\zeta(p) + 2\Phi(D_h(p) - z_h(p)), & \text{if } \alpha < p \leq 1/2, \\ 2\Phi(z_h(p)), & \text{if } 1/2 < p < 1. \end{cases}$$

The p-curve for the minimum approach arises as a corollary of the above results.

B.1.2.2 Bias of the p-Hacked Estimator

For the bias, consider the estimator for the causal effect given by the threshold approach. The p-hacked estimator is given by

$$\hat{\beta}_r^t = \hat{\beta}_{12} + (\hat{\beta}_1 - \hat{\beta}_{12})1_{\{A_N\}} + (\hat{\beta}_1 - \hat{\beta}_{12})1_{\{B_N\}},$$

where we define sets $A_N = \{T_{12} < z_0(\alpha), T_1 > T_{12}, T_1 > T_2\}$ and $B_N = \{T_{12} < z_0(\alpha), T_2 > T_{12}, T_2 > T_1\}$. Using the same standard 2SLS results used above to generate the approximate distributions of the t-statistics and by the continuous mapping theorem,

$$\sqrt{N}|\gamma|(\hat{\beta}_r^t - \beta) \xrightarrow{d} \xi^t := \frac{W_1 + W_2}{2} + \frac{W_1 - W_2}{2}1_{\{A\}} + \frac{W_2 - W_1}{2}1_{\{B\}},$$

where $A = \{\sqrt{2}h + \frac{W_1+W_2}{\sqrt{2}} < z_0(\alpha), h + W_1 > \sqrt{2}h + \frac{W_1+W_2}{\sqrt{2}}, W_1 > W_2\}$ and $B = \{\sqrt{2}h + \frac{W_1+W_2}{\sqrt{2}} < z_0(\alpha), h + W_2 > \sqrt{2}h + \frac{W_1+W_2}{\sqrt{2}}, W_2 > W_1\}$. By the symmetry of the problem,

$$E[\xi^t] = E[(W_1 - W_2)1_{\{A\}}] = E[W_1 1_{\{A\}}] - E[W_2 1_{\{A\}}].$$

To compute the expectations, we need to calculate $P(A|W_1)$ and $P(A|W_2)$. Note that

$$A = \begin{cases} \sqrt{2}h + \frac{W_1+W_2}{\sqrt{2}} < z_0(\alpha) \\ h + W_1 > \sqrt{2}h + \frac{W_1+W_2}{\sqrt{2}} \\ W_1 > W_2 \end{cases} \Leftrightarrow \begin{cases} W_2 < \sqrt{2}z_0(\alpha) - 2h - W_1 \\ W_2 < (\sqrt{2}-1)(W_1 - \sqrt{2}h) \\ W_2 < W_1 \end{cases} \Leftrightarrow \begin{cases} W_1 < \sqrt{2}z_0(\alpha) - 2h - W_2 \\ W_1 > \sqrt{2}h + \frac{W_2}{\sqrt{2}-1} \\ W_1 > W_2 \end{cases} \quad (*)$$

From equation $(*)$, we have

$$\Pr(A|W_1) = \Phi(\min\{W_1, \sqrt{2}z_0(\alpha) - 2h - W_1, (\sqrt{2}-1)(W_1 - \sqrt{2}h)\}),$$

where

$$\min\{W_1, \sqrt{2}z_0(\alpha) - 2h - W_1, (\sqrt{2}-1)(W_1 - \sqrt{2}h)\} = \begin{cases} W_1, \text{ if } W_1 < -h, \\ \sqrt{2}z_0(\alpha) - 2h - W_1, \text{ if } W_1 > z_0(\alpha) - h, \\ (\sqrt{2}-1)(W_1 - \sqrt{2}h), \text{ if } W_1 \in (-h, z_0(\alpha) - h). \end{cases}$$

Also, the last system of inequalities in equation $(*)$ is equivalent to

$$W_1 \in \begin{cases} \left(W_2, \sqrt{2}z_0(\alpha) - 2h - W_2\right), \text{ if } W_2 < -h, \\ \left(\sqrt{2}h + \frac{W_2}{\sqrt{2}-1}, \sqrt{2}z_0(\alpha) - 2h - W_2\right), \text{ if } W_2 \geq -h. \end{cases}$$

Note that $\left(\sqrt{2}h + \frac{W_2}{\sqrt{2}-1}, \sqrt{2}z_0(\alpha) - 2h - W_2\right)$ is non-empty when $W_2 \leq (\sqrt{2}-1)z_0(\alpha) - h$, and

$\left(W_2, \sqrt{2}z_0(\alpha) - 2h - W_2\right)$ is non-empty for all values of $W_2 < -h$. Therefore,

$$\begin{aligned}
\Pr(A|W_2) &= \Pr\left(W_1 \in \left(W_2, \sqrt{2}z_0(\alpha) - 2h - W_2\right)|W_2\right) \cdot 1_{\{W_2 < -h\}} \\
&\quad + \Pr\left(W_1 \in \left(\sqrt{2}h + \frac{W_2}{\sqrt{2}-1}, \sqrt{2}z_0(\alpha) - 2h - W_2\right)|W_2\right) \cdot 1_{\{W_2 \geq -h\}} \\
&= \left[\Phi\left(\sqrt{2}z_0(\alpha) - 2h - W_2\right) - \Phi(W_2)\right] \cdot 1_{\{W_2 < -h\}} \\
&\quad + \left[\Phi\left(\sqrt{2}z_0(\alpha) - 2h - W_2\right) - \Phi\left(\sqrt{2}h + \frac{W_2}{\sqrt{2}-1}\right)\right] \cdot 1_{\left\{-h \leq W_2 \leq (\sqrt{2}-1)z_0(\alpha) - h\right\}} \\
&= \Phi\left(\sqrt{2}z_0(\alpha) - 2h - W_2\right) \cdot 1_{\left\{W_2 \leq (\sqrt{2}-1)z_0(\alpha) - h\right\}} - \Phi(W_2) \cdot 1_{\{W_2 < -h\}} \\
&\quad - \Phi\left(\sqrt{2}h + \frac{W_2}{\sqrt{2}-1}\right) \cdot 1_{\left\{-h \leq W_2 \leq (\sqrt{2}-1)z_0(\alpha) - h\right\}}
\end{aligned}$$

To finish the calculation of expectations, we will need to calculate several integrals of the form

$$\int_L^U w\phi(w)\Phi(aw+b)dw.$$

The following result is therefore useful.

Lemma B.1.

$$\begin{aligned}
\int_L^U w\phi(w)\Phi(aw+b)dw &= \Phi(aL+b)\phi(L) - \Phi(aU+b)\phi(U) + \frac{a}{\sqrt{1+a^2}}\phi\left(\frac{b}{\sqrt{1+a^2}}\right) \\
&\quad \times \left[\Phi\left(\sqrt{1+a^2}U + \frac{ab}{\sqrt{1+a^2}}\right) - \Phi\left(\sqrt{1+a^2}L + \frac{ab}{\sqrt{1+a^2}}\right)\right]
\end{aligned}$$

Proof.

$$\begin{aligned}
\int_L^U w\phi(w)\Phi(aw+b)dw &= -\int_L^U \Phi(aw+b)d\phi(w) \\
&= \Phi(aL+b)\phi(L) - \Phi(aU+b)\phi(U) + \int_L^U \phi(w)d\Phi(aw+b) \\
&= \Phi(aL+b)\phi(L) - \Phi(aU+b)\phi(U) + a\int_L^U \phi(w)\phi(aw+b)dw \\
&= \Phi(aL+b)\phi(L) - \Phi(aU+b)\phi(U) + a\phi\left(\frac{b}{\sqrt{1+a^2}}\right)J,
\end{aligned}$$

where

$$\begin{aligned} J: &= \int_L^U \phi\left(\sqrt{1+a^2}w+\frac{ab}{\sqrt{1+a^2}}\right) \\ &= \frac{1}{\sqrt{1+a^2}}\left[\Phi\left(\sqrt{1+a^2}U+\frac{ab}{\sqrt{1+a^2}}\right)-\Phi\left(\sqrt{1+a^2}L+\frac{ab}{\sqrt{1+a^2}}\right)\right]. \end{aligned}$$

□

Finally,

$$\begin{aligned} E[W_1 1_{\{A\}}] &= E[W_1 P(A|W_1)] \\ &= E[W_1\Phi(\min\{W_1,\sqrt{2}z_0(\alpha)-2h-W_1,(\sqrt{2}-1)(W_1-\sqrt{2}h)\})] \\ &= \int_{-\infty}^{-h} w\phi(w)\Phi(w)dw \\ &\quad+\int_{-h}^{z_0(\alpha)-h} w\phi(w)\Phi((\sqrt{2}-1)w-\sqrt{2}(\sqrt{2}-1)h)dw \\ &\quad+\int_{z_0(\alpha)-h}^{+\infty} w\phi(w)\Phi(\sqrt{2}z_0(\alpha)-2h-w)dw \\ &= -(1-\Phi(h))\phi(h)+\frac{1}{\sqrt{2}}\phi(0)(1-\Phi(\sqrt{2}h)) \\ &\quad+(1-\Phi(h))\phi(h)-\Phi((\sqrt{2}-1)z_0(\alpha)-h)\phi(z_0(\alpha)-h) \\ &\quad+\frac{\sqrt{2}-1}{\sqrt{4-2\sqrt{2}}}\phi\left(\frac{\sqrt{2}-1}{\sqrt{2-\sqrt{2}}}h\right)\left[\Phi\left(\frac{h}{\sqrt{2-\sqrt{2}}}\right)-\Phi\left(\frac{h}{\sqrt{2-\sqrt{2}}}-\sqrt{4-2\sqrt{2}}z_0(\alpha)\right)\right] \\ &\quad+\Phi((\sqrt{2}-1)z_0(\alpha)-h)\phi(z_0(\alpha)-h)-\frac{1}{\sqrt{2}}\phi(z_0(\alpha)-\sqrt{2}h)[1-\Phi((\sqrt{2}-1)z_0(\alpha))] \\ &= \frac{1}{\sqrt{2}}\phi(0)(1-\Phi(\sqrt{2}h))-\frac{1}{\sqrt{2}}\phi(z_0(\alpha)-\sqrt{2}h)[1-\Phi((\sqrt{2}-1)z_0(\alpha))] \\ &\quad+\frac{\sqrt{2}-1}{\sqrt{4-2\sqrt{2}}}\phi\left(\frac{\sqrt{2}-1}{\sqrt{2-\sqrt{2}}}h\right)\left[\Phi\left(\frac{h}{\sqrt{2-\sqrt{2}}}\right)-\Phi\left(\frac{h}{\sqrt{2-\sqrt{2}}}-\sqrt{4-2\sqrt{2}}z_0(\alpha)\right)\right], \end{aligned}$$

where the fourth equality follows from the direct application of Lemma B.1.

$$
\begin{aligned}
E[W_2 1_{\{A\}}] &= E[W_2 \Pr(A|W_2)] \\
&= E\left[W_2\Phi\left(\sqrt{2}z_0(\alpha)-2h-W_2\right)\cdot 1_{\left\{W_2\le(\sqrt{2}-1)z_0(\alpha)-h\right\}}\right] \\
&\quad -E\left[W_2\Phi(W_2)\cdot 1_{\{W_2<-h\}}\right] \\
&\quad -E\left[W_2\Phi\left(\sqrt{2}h+\frac{W_2}{\sqrt{2}-1}\right)\cdot 1_{\left\{-h\le W_2\le(\sqrt{2}-1)z_0(\alpha)-h\right\}}\right] \\
&= \int_{-\infty}^{(\sqrt{2}-1)z_0(\alpha)-h} w\phi(w)\Phi\left(\sqrt{2}z_0(\alpha)-2h-w\right)dw \\
&\quad -\int_{-\infty}^{-h} w\phi(w)\Phi(w)dw \\
&\quad -\int_{-h}^{(\sqrt{2}-1)z_0(\alpha)-h} w\phi(w)\Phi\left(\sqrt{2}h+\frac{w}{\sqrt{2}-1}\right)dw \\
&= -\Phi(z_0(\alpha)-h)\phi((\sqrt{2}-1)z_0(\alpha)-h)-\frac{1}{\sqrt{2}}\phi(z_0(\alpha)-\sqrt{2}h)(1-\Phi((\sqrt{2}-1)z_0(\alpha))) \\
&\quad +(1-\Phi(h))\,\phi(h)-\frac{1}{\sqrt{2}}\phi(0)\left(1-\Phi\left(\sqrt{2}h\right)\right) \\
&\quad -(1-\Phi(h))\phi(h)+\Phi(z_0(\alpha)-h)\phi((\sqrt{2}-1)z_0(\alpha)-h) \\
&\quad -\frac{1}{\sqrt{4-2\sqrt{2}}}\phi\left(\sqrt{\frac{\sqrt{2}-1}{\sqrt{2}}}h\right)\left[\Phi\left(\frac{h}{\sqrt{2-\sqrt{2}}}\right)-\Phi\left(\frac{h}{\sqrt{2-\sqrt{2}}}-\sqrt{4-2\sqrt{2}}z_0(\alpha)\right)\right] \\
&= -\frac{1}{\sqrt{2}}\phi(z_0(\alpha)-\sqrt{2}h)(1-\Phi((\sqrt{2}-1)z_0(\alpha)))-\frac{1}{\sqrt{2}}\phi(0)\left(1-\Phi\left(\sqrt{2}h\right)\right) \\
&\quad -\frac{1}{\sqrt{4-2\sqrt{2}}}\phi\left(\sqrt{\frac{\sqrt{2}-1}{\sqrt{2}}}h\right)\left[\Phi\left(\frac{h}{\sqrt{2-\sqrt{2}}}\right)-\Phi\left(\frac{h}{\sqrt{2-\sqrt{2}}}-\sqrt{4-2\sqrt{2}}z_0(\alpha)\right)\right].
\end{aligned}
$$

Combining these results gives us the first-order bias of $\hat{\beta}_r^t$, $\mathscr{B}_2^t = |\gamma|^{-1}E[\xi^t]$, where

$$
\begin{aligned}
E[\xi^t] &= \frac{1}{\sqrt{2-\sqrt{2}}}\phi\left(\sqrt{\frac{\sqrt{2}-1}{\sqrt{2}}}h\right)\left(\Phi\left(\frac{h}{\sqrt{2-\sqrt{2}}}\right)-\Phi\left(\frac{h}{\sqrt{2-\sqrt{2}}}-\sqrt{4-2\sqrt{2}}z_0(\alpha)\right)\right) \\
&\quad +\sqrt{2}\phi(0)\left(1-\Phi\left(\sqrt{2}h\right)\right).
\end{aligned}
$$

Finally, for the minimum approach, the p-hacked estimator is given by

$$\hat{\beta}_r^m = \hat{\beta}_{12} + (\hat{\beta}_1 - \hat{\beta}_{12})1_{\{A_N\}} + (\hat{\beta}_1 - \hat{\beta}_{12})1_{\{B_N\}},$$

where now we define sets $A_N = \{T_1 > T_{12}, T_1 > T_2\}$ and $B_N = \{T_2 > T_{12}, T_2 > T_1\}$.

Thus

$$\sqrt{N}|\gamma|(\hat{\beta}_r^m - \beta) \xrightarrow{d} \xi^m := \frac{W_1 + W_2}{2} + \frac{W_1 - W_2}{2}1_{\{A\}} + \frac{W_2 - W_1}{2}1_{\{B\}},$$

where $A = \{W_1 > \max\{W_2, \sqrt{2}h + W_2/(\sqrt{2}-1)\}\}$ and $B = \{W_2 > \max\{W_1, \sqrt{2}h + W_1/(\sqrt{2}-1)\}\}$.

By the symmetry of the problem,

$$E[\xi^m] = E[(W_1 - W_2)1_{\{A\}}] = E[W_1 1_{\{A\}}] - E[W_2 1_{\{A\}}].$$

Note that $\Pr(A|W_2) = 1 - \Phi(\max\{W_2, \sqrt{2}h + W_2/(\sqrt{2}-1)\})$ and $\Pr(A|W_1) = \Phi(\min\{W_1, W_1(\sqrt{2}-1) - \sqrt{2}(\sqrt{2}-1)h\})$. Therefore,

$$\begin{aligned}
E[W_1 1_{\{A\}}] &= E[W_1 \Pr(A|W_1)] \\
&= E\left[W_1 \Phi(\min\{W_1, W_1(\sqrt{2}-1) - \sqrt{2}(\sqrt{2}-1)h\})\right] \\
&= \frac{1}{\sqrt{2}}\phi(0)(1 - \Phi(\sqrt{2}h)) + \frac{\sqrt{2}-1}{\sqrt{4-2\sqrt{2}}}\phi\left(\sqrt{\frac{\sqrt{2}-1}{\sqrt{2}}}h\right)\Phi\left(\frac{h}{\sqrt{2-\sqrt{2}}}\right)
\end{aligned}$$

and

$$\begin{aligned} E[W_2 1_{\{A\}}] &= E[W_2 \Pr(A|W_2)] \\ &= -E\left[W_2 \Phi(\max\{W_2, \sqrt{2}h + W_2/(\sqrt{2}-1)\})\right] \\ &= -\frac{1}{\sqrt{2}}\phi(0)(1-\Phi(\sqrt{2}h)) - \frac{1}{\sqrt{4-2\sqrt{2}}}\phi\left(\sqrt{\frac{\sqrt{2}-1}{\sqrt{2}}}h\right)\Phi\left(\frac{h}{\sqrt{2-\sqrt{2}}}\right) \end{aligned}$$

Putting these together gives the asymptotic bias of $\hat{\beta}_r^m$, $\mathscr{B}_2^m = |\gamma|^{-1}E[\xi^m]$, where

$$E[\xi^m] = \frac{1}{\sqrt{2-\sqrt{2}}}\phi\left(\sqrt{\frac{\sqrt{2}-1}{\sqrt{2}}}h\right)\Phi\left(\frac{h}{\sqrt{2-\sqrt{2}}}\right) + \sqrt{2}\phi(0)(1-\Phi(\sqrt{2}h)).$$

B.1.3 Selecting across Datasets

With the assumption that the datasets are independent, we have that the K t-statistics are distributed as $T \sim \mathscr{N}(h\iota_K, I_K)$ where ι_K is a $K \times 1$ vector of ones. The assumption that each dataset tests for the same effect mathematically appears as h being the mean for all t-statistics. For $K = 2$, the definition for P_r in this example is the same as that in Appendix B.1.1 with $\rho = 0$. Hence the result for the thresholding case is g_1^t evaluated at $\rho = 0$ and for the minimum case is g_1^m also evaluated at $\rho = 0$.

For general K, note that for the minimum case

$$\begin{aligned} G^m(p;K) &= \Pr(\max(T_1, T_2, \ldots, T_K) \geq z_0(p) \\ &= 1 - \Pr(T_1 \leq z_0(\alpha), T_2 \leq z_0(\alpha), \ldots, T_K \leq z_0(\alpha)) \\ &= 1 - (\Phi(z_h(p))^K \end{aligned}$$

Setting $p = \alpha$ gives the size after p-hacking for a nominal value of α. Differentiating

with respect to p generates the p-curve

$$\begin{aligned} g_3^m(p;K) &= \frac{d}{dp}\left(1-(\Phi(z_h(p))^K\right) \\ &= -K\Phi(z_h(p))^{K-1}\frac{d}{dp}\Phi(z_h(p)) \\ &= K\Phi(z_h(p))^{K-1}\frac{\phi(z_h(p))}{\phi(z_0(p))} \end{aligned}$$

The expression in the text follows directly from integrating over h.

B.1.4 Variance Bandwidth Selection for Means

Note that $T_0 \sim \mathcal{N}(h,1)$ and $\hat{\rho}$ are independent.[9] Also note that $T_1 \geq T_0$ happens in the following cases: (i) $T_0 \geq 0$ and $0 < \omega^2(\hat{\rho}) \leq 1$, equivalent to $-(2\kappa)^{-1} < \hat{\rho} \leq 0$; (ii) $T_0 < 0$ and $\hat{\rho} > 0$. The researchers report the p-value corresponding to T_0 if the result is significant at level α or if $\hat{\omega}^2 < 0$, otherwise they report the p-value associated with the largest t-statistic. Fixing h, we have

$$\begin{aligned} G_h^t(p) &= \Pr(P_r \leq p) \\ &= \Pr(T_0 \geq z_0(p), T_0 \geq z_0(\alpha)) \\ &\quad + \Pr(T_0 \geq z_0(p), T_0 < z_0(\alpha), -\infty < \hat{\rho} \leq -(2\kappa)^{-1}) \\ &\quad + \Pr(T_0 \geq z_0(p), T_0 \geq 0, T_0 < z_0(\alpha), \hat{\rho} > 0) \\ &\quad + \Pr(T_1 \geq z_0(p), T_0 \geq 0, T_0 < z_0(\alpha), -(2\kappa)^{-1} < \hat{\rho} \leq 0) \\ &\quad + \Pr(T_1 \geq z_0(p), T_0 \leq 0, T_0 < z_0(\alpha), \hat{\rho} > 0) \\ &\quad + \Pr(T_0 \geq z_0(p), T_0 \leq 0, T_0 < z_0(\alpha), -(2\kappa)^{-1} < \hat{\rho} \leq 0) \end{aligned}$$

We can rewrite these expressions using the independence of T_0 and $\hat{\rho}$. For $p \leq \alpha \leq 1/2$,

[9] The independence follows from the fact that $\hat{\rho}$ is a function of $V := (U_2 - \bar{U}, \ldots, U_N - \bar{U})'$, $T_0 = h + \sqrt{N}\bar{U}$ and that V and $\bar{U}$ are independent.

this is

$$G_h^t(p) \;=\; 1-\Phi(z_h(p))+\int_{-(2\kappa)^{-1}}^{l(p)}(\Phi(z_h(\alpha))-\Phi(z_0(p)\omega(r)-h))\eta_N(r)dr.$$

The last term follows since $\Pr(T_1 \geq z_0(p), 0 \leq T_0 \leq z_0(\alpha), -(2\kappa)^{-1} < \hat{\rho} \leq 0)$ can be written as $\Pr(z_0(p)\omega(\hat{\rho}) \leq T_0 \leq z_0(\alpha), -(2\kappa)^{-1} < \hat{\rho} \leq l(p))$ and

$$l(p) = \frac{1}{2\kappa}\left(\left(\frac{z_0(\alpha)}{z_0(p)}\right)^2 - 1\right),$$

which is the largest value for $\hat{\rho}$ at each $p \leq \alpha$ for which the interval for T_0 is nonempty.

For $\alpha < p \leq 1/2$, we have

$$G_h^t(p) \;=\; 1-\Phi(z_h(p))(1-H_N(0)+H_N(-(2\kappa)^{-1}))-\int_{-(2\kappa)^{-1}}^{0}\Phi(z_0(p)\omega(r)-h)\eta_N(r)dr$$

For $\alpha < p$ and $p > 1/2$, we obtain

$$G_h^t(p) \;=\; 1-\Phi(z_h(p))H_N(0)-\int_{0}^{\infty}\Phi(z_0(p)\omega(r)-h)\eta_N(r)dr$$

Differentiating with respect to p and integrating over the distribution of h gives the density

$$g_4^t(p) = \int_{\mathscr{H}} \exp\left(hz_0(p)-\frac{h^2}{2}\right)\Upsilon_4^t(p;\alpha,h,\kappa)d\Pi(h),$$

where

$$\Upsilon_4^t = \begin{cases} 1+\frac{1}{\phi(z_h(p))}\int_{-(2\kappa)^{-1}}^{l(p)}\omega(r)\phi(z_0(p)\omega(r)-h)\eta_N(r)dr, & \text{if } 0 < p \leq \alpha, \\ 1-H_N(0)+H_N(-(2\kappa)^{-1})+\frac{1}{\phi(z_h(p))}\int_{-(2\kappa)^{-1}}^{0}\omega(r)\phi(z_0(p)\omega(r)-h)\eta_N(r)dr, & \text{if } \alpha < p \leq 1/2, \\ H_N(0)+\frac{1}{\phi(z_h(p))}\int_{0}^{\infty}\omega(r)\phi(z_0(p)\omega(r)-h)\eta_N(r)dr, & \text{if } 1/2 < p < 1. \end{cases}$$

The derivations for the minimum p-value approach are analogous and presented below. Note that

$$\begin{aligned} G_h^m(p) &= \Pr(P_r \le p) \\ &= \Pr(T_0 \ge z_0(p), -\infty < \hat{\rho} \le -(2\kappa)^{-1}) \\ &\quad + \Pr(T_0 \ge z_0(p), T_0 \ge 0, \hat{\rho} > 0) \\ &\quad + \Pr(T_1 \ge z_0(p), T_0 \ge 0, -(2\kappa)^{-1} < \hat{\rho} \le 0) \\ &\quad + \Pr(T_1 \ge z_0(p), T_0 \le 0, \hat{\rho} > 0) \\ &\quad + \Pr(T_0 \ge z_0(p), T_0 \le 0, -(2\kappa)^{-1} < \hat{\rho} \le 0) \end{aligned}$$

For $p \le 1/2$, we have

$$\begin{aligned} G_h^m(p) &= H_N(0) - H_N(-(2\kappa)^{-1}) + (1-\Phi(z_h(p)))(1 - H_N(0) + H_N(-(2\kappa)^{-1})) \\ &\quad - \int_{-(2\kappa)^{-1}}^{0} \Phi(z_0(p)\omega(r) - h)\eta_N(r)dr \end{aligned}$$

and, for $p > 1/2$, we have

$$G_h^m(p) = 1 - H_N(0) + (1-\Phi(z_h(p)))H_N(0) - \int_0^{\infty} \Phi(z_0(p)\omega(r) - h)\eta_N(r)dr.$$

Differentiating with respect to p and integrating over the distribution of h gives the density

$$g_4^m(p) = \int_{\mathscr{H}} \exp\left(hz_0(p) - \frac{h^2}{2}\right) \Upsilon_4^m(p;\alpha,h)d\Pi(h),$$

where

$$\Upsilon_4^m = \begin{cases} 1 - H_N(0) + H_N(-(2\kappa)^{-1}) + \frac{1}{\phi(z_h(p))} \int_{-(2\kappa)^{-1}}^{0} \omega(r)\phi(z_0(p)\omega(r) - h)\eta_N(r)dr, & \text{if } 0 < p \le 1/2, \\ H_N(0) + \frac{1}{\phi(z_h(p))} \int_0^{\infty} \omega(r)\phi(z_0(p)\omega(r) - h)\eta_N(r)dr, & \text{if } 1/2 < p < 1. \end{cases}$$

B.2 Null and Alternative Distributions MC Study

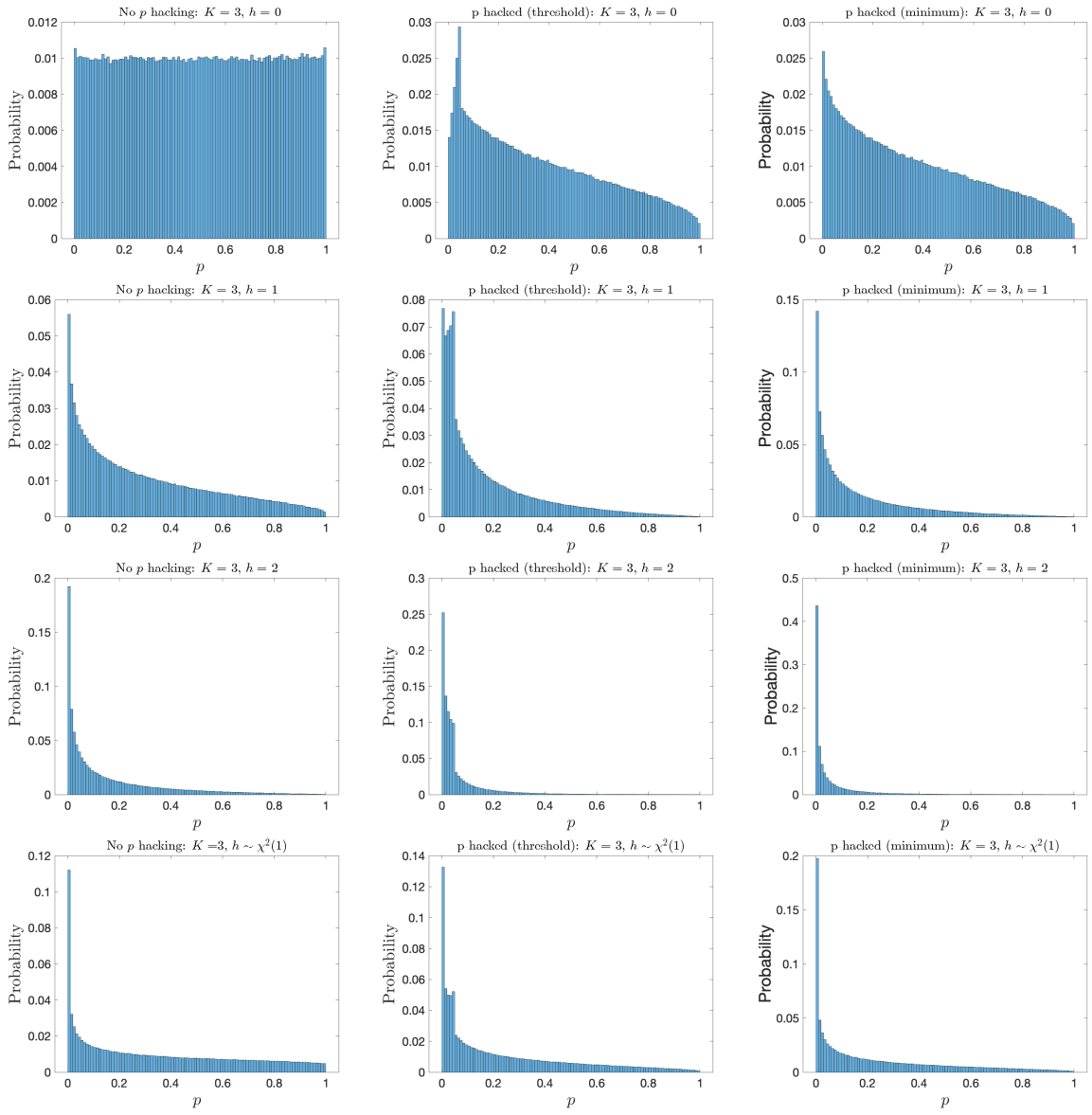

Figure B.1. Null and p-hacked distributions for covariate selection with $K = 3$.

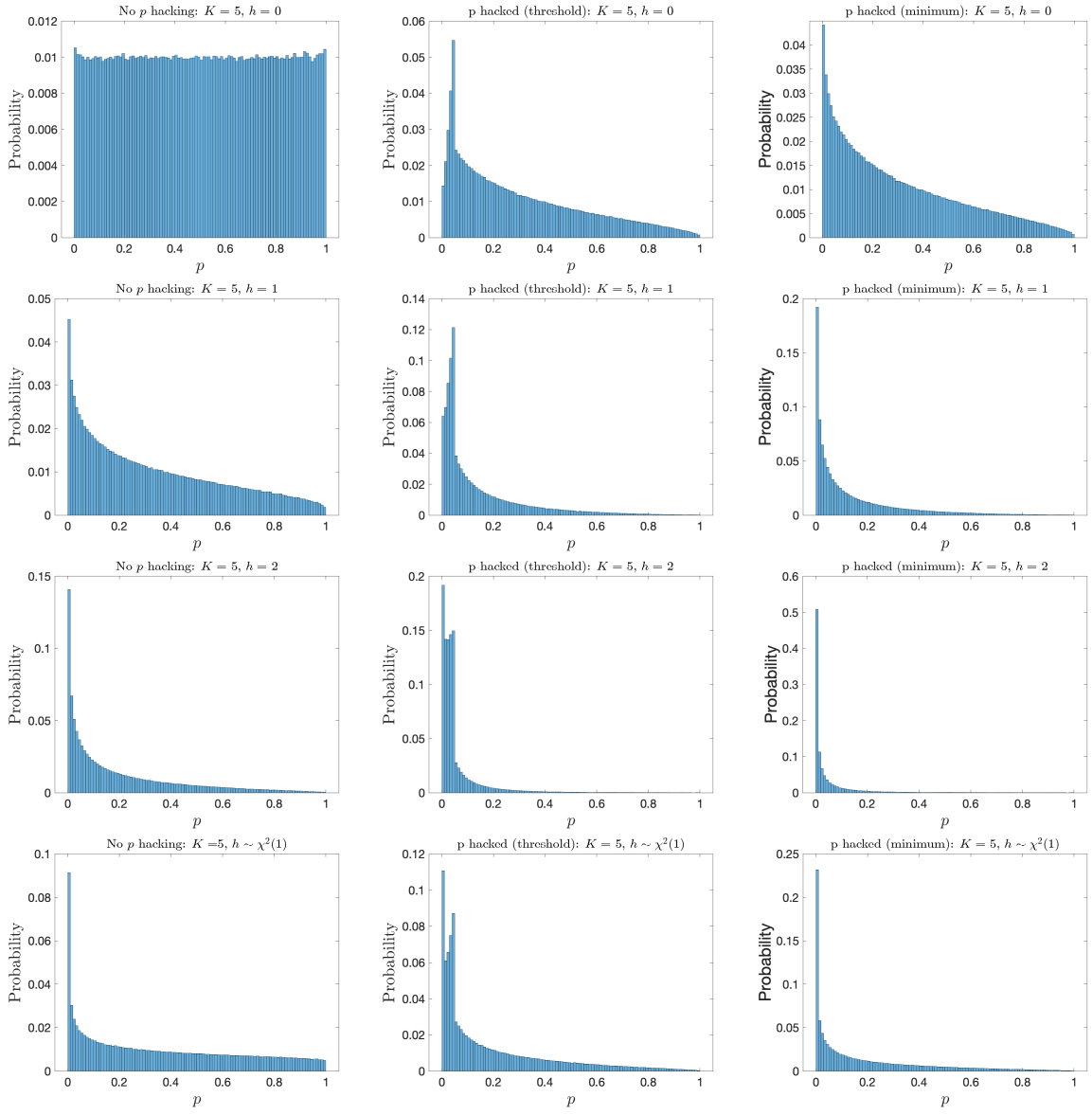

Figure B.2. Null and p-hacked distributions for covariate selection with $K = 5$.

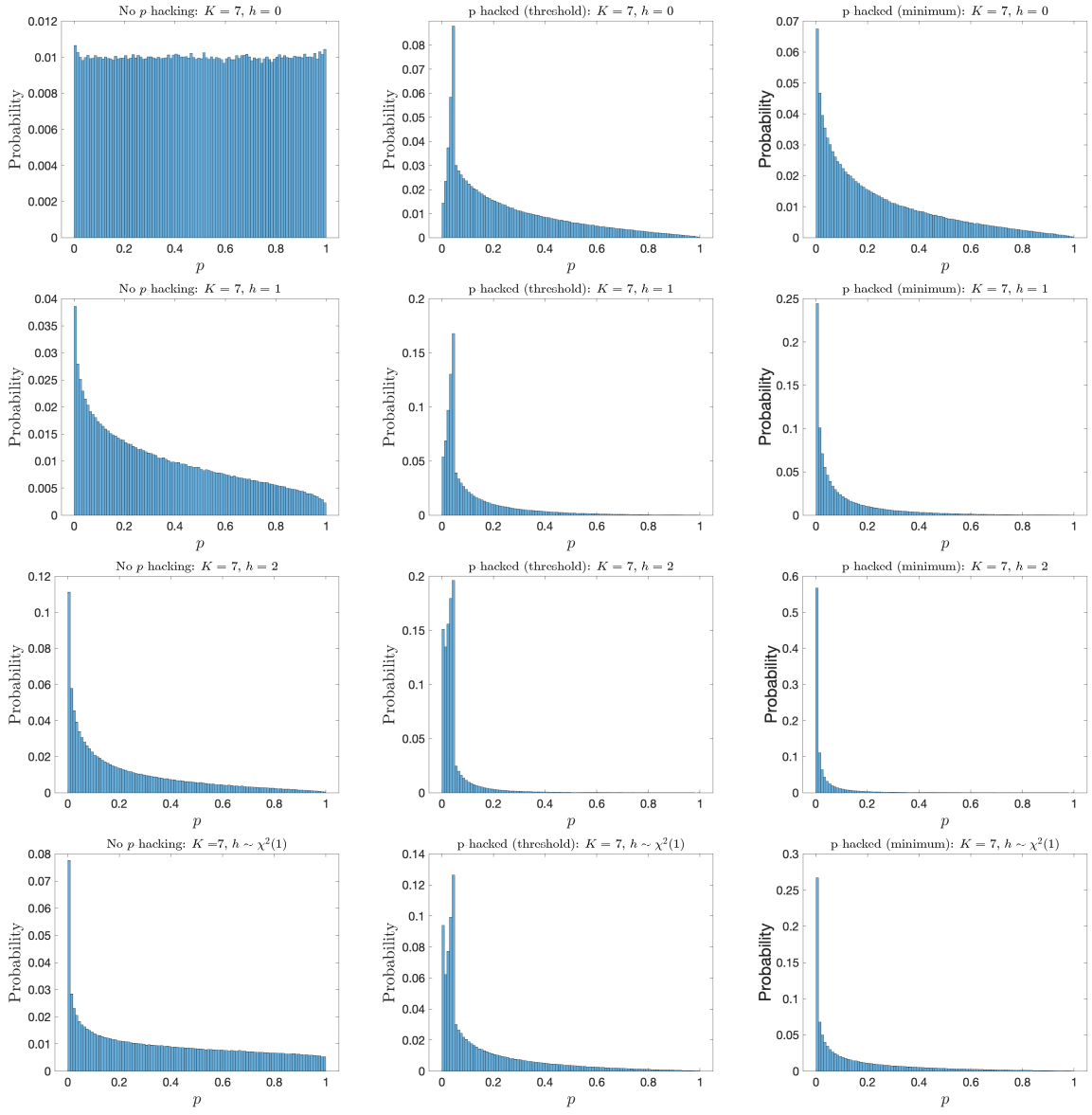

Figure B.3. Null and p-hacked distributions for covariate selection with $K = 7$.

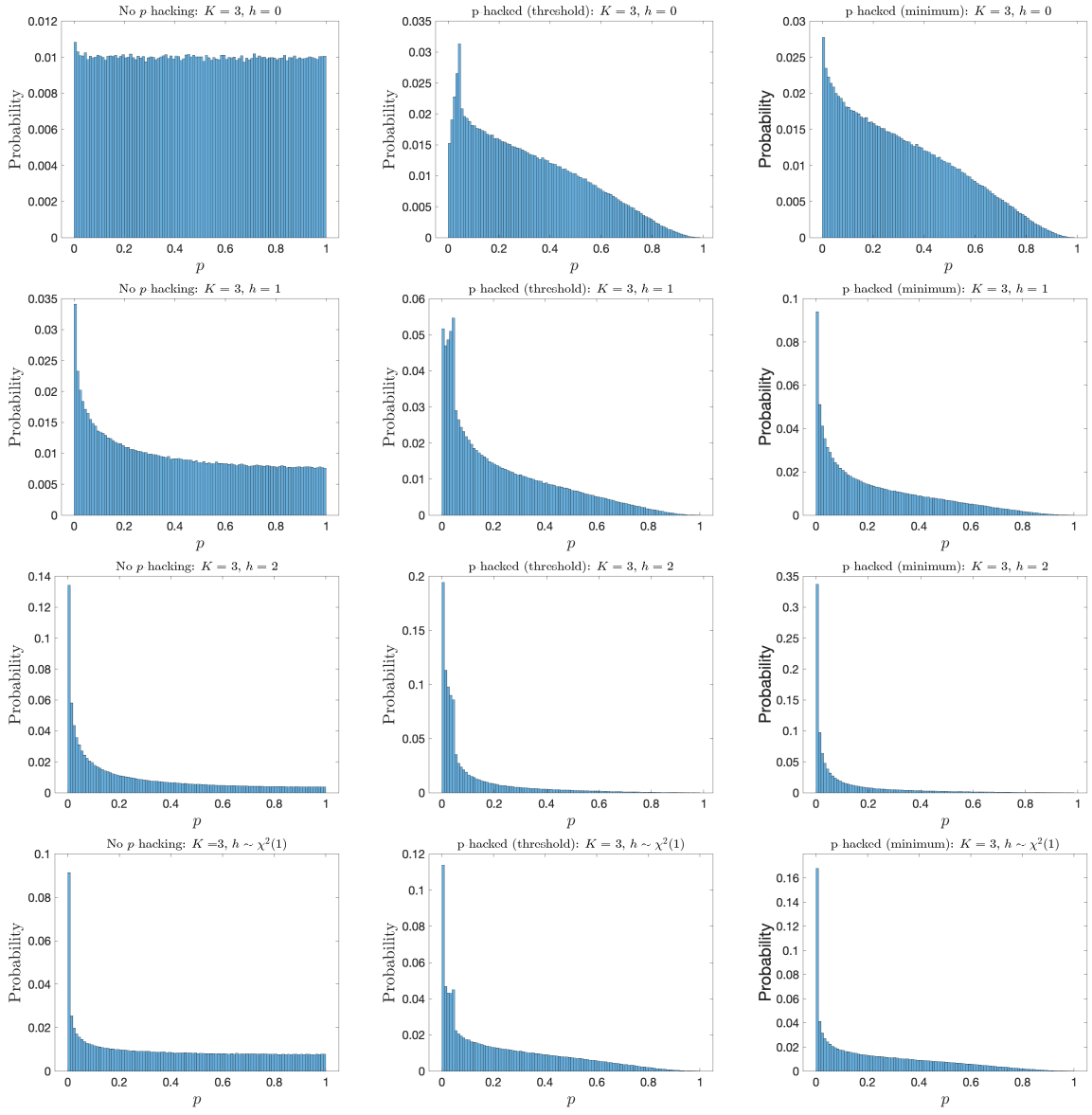

Figure B.4. Null and p-hacked distributions for covariate selection with $K = 3$ and researchers using two-sided test.

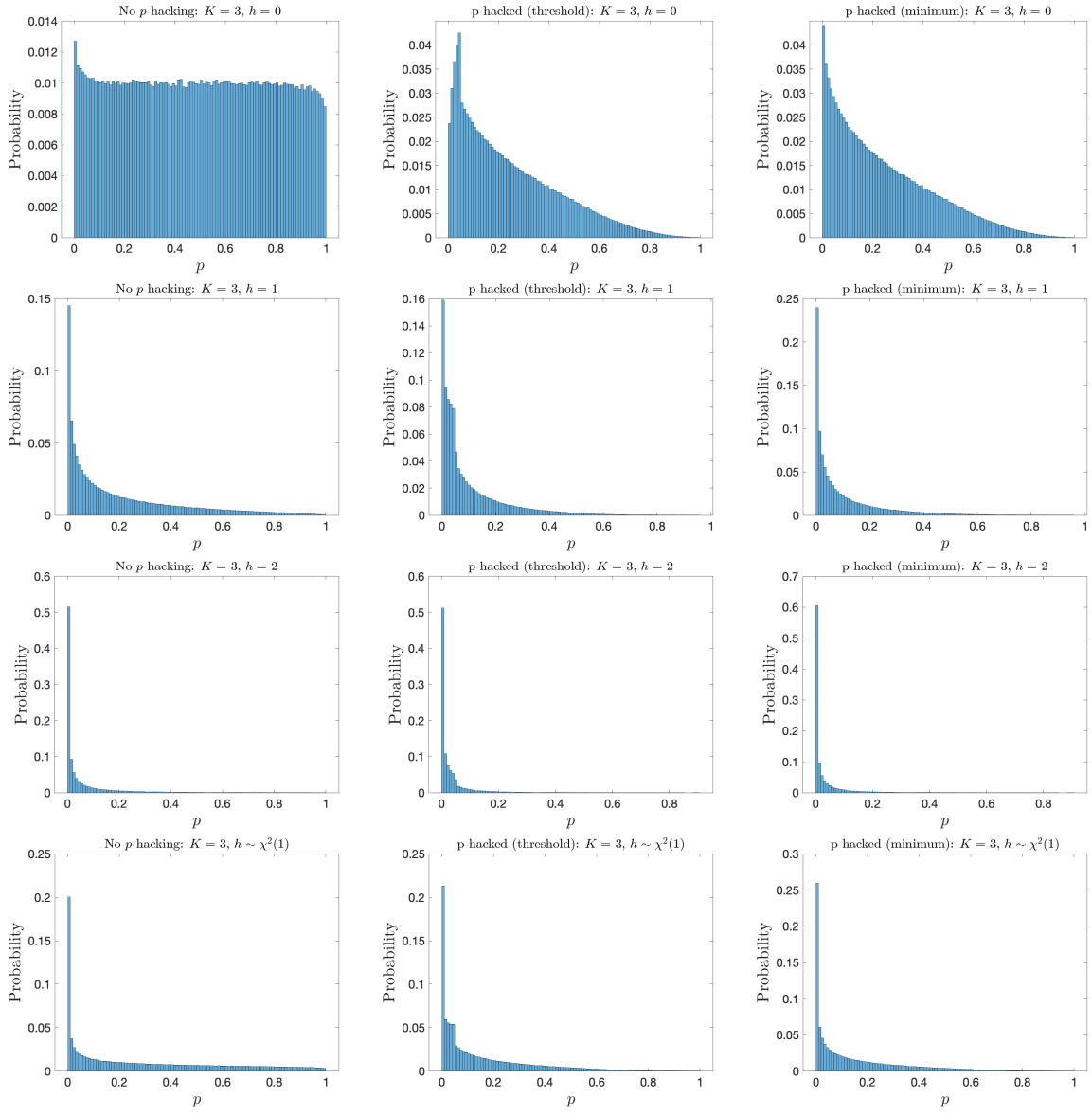

Figure B.5. Null and p-hacked distributions for IV selection with $K = 3$.

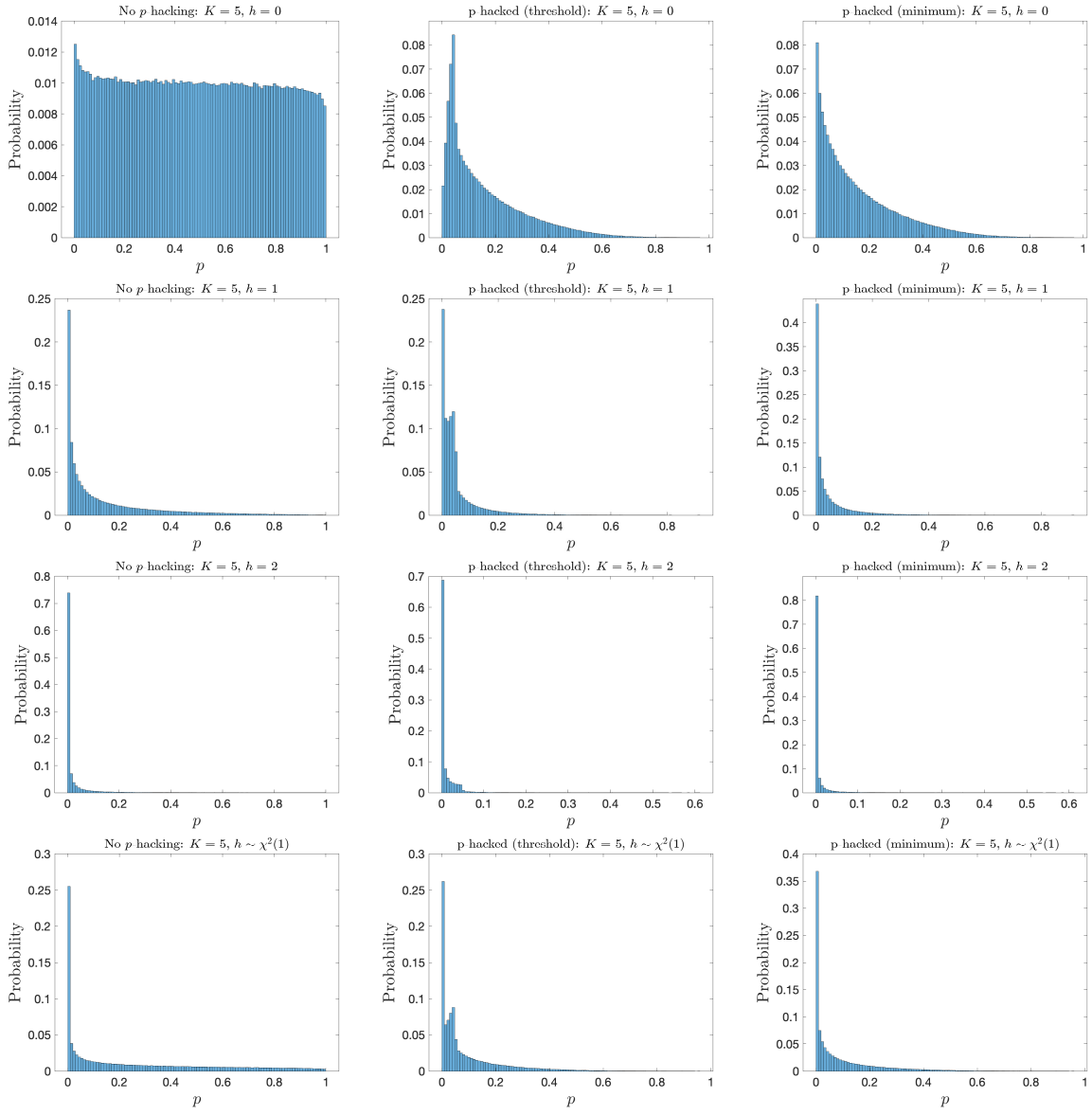

Figure B.6. Null and p-hacked distributions for IV selection with $K = 5$.

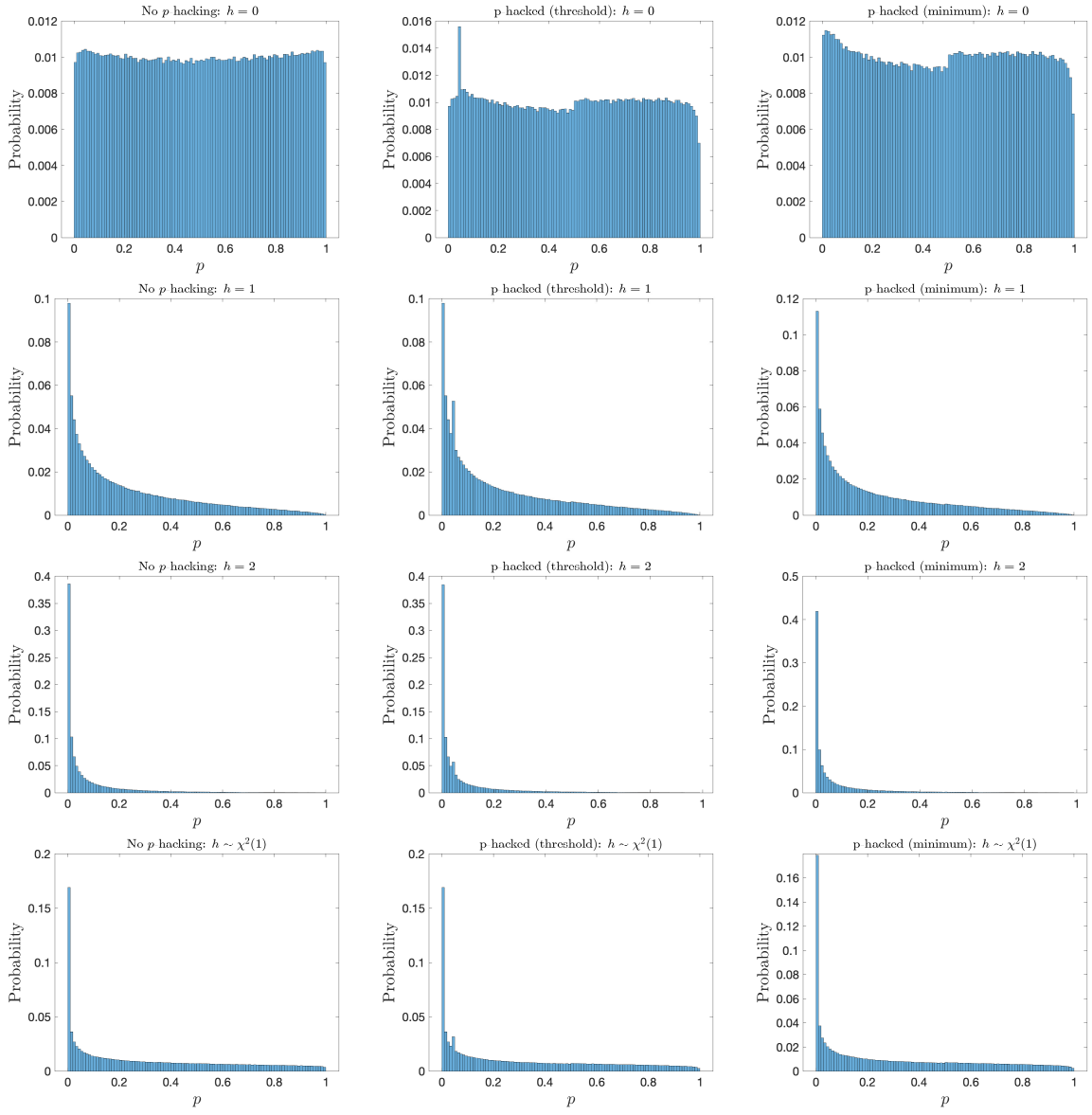

Figure B.7. Null and p-hacked distributions for lag length selection.

B.3 Additional Simulation Results

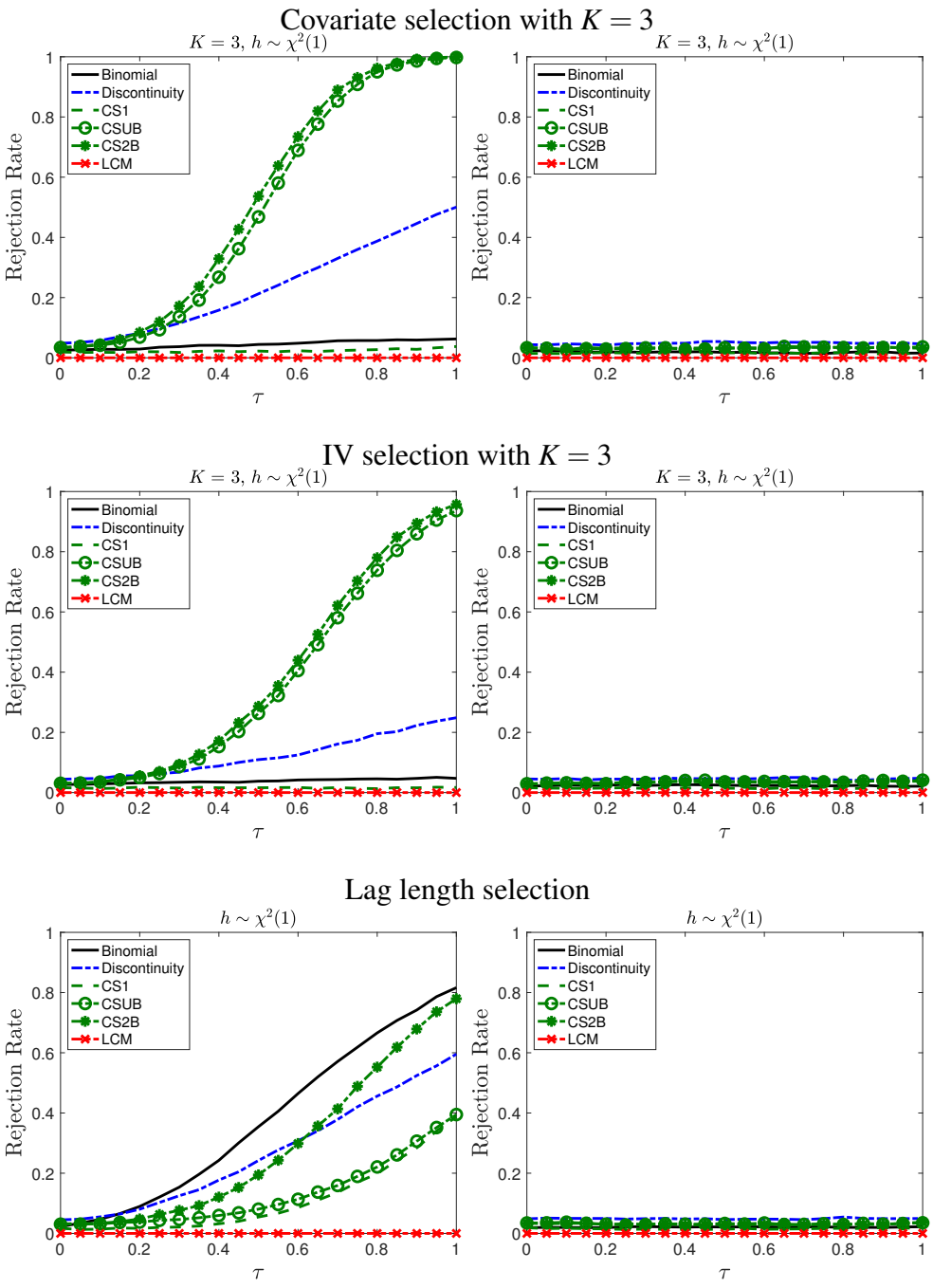

Figure B.8. Power curves for $h \sim \chi^2(1)$. Thresholding (left column) and minimum (right column).

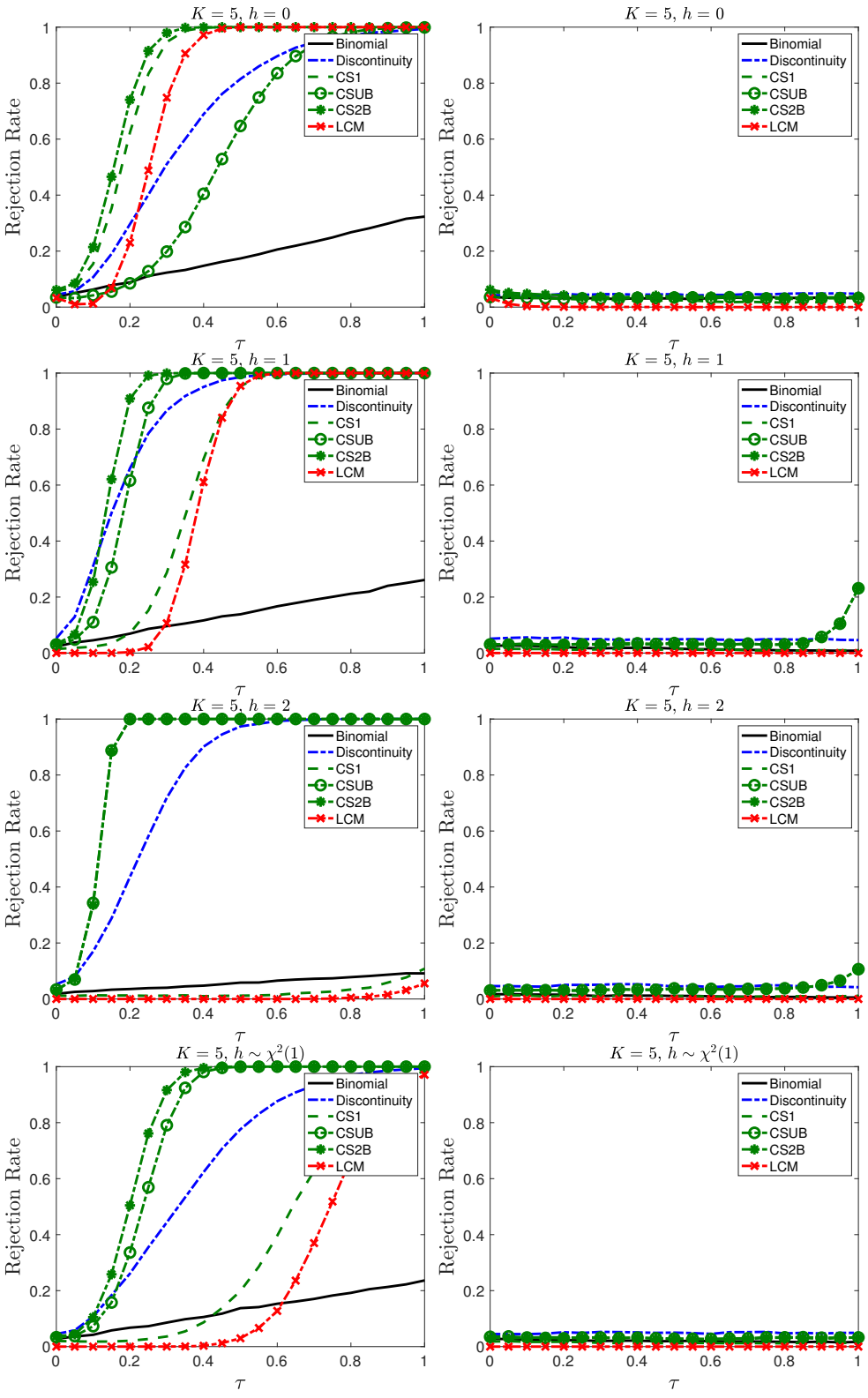

Figure B.9. Power curves covariate selection with $K = 5$. Thresholding (left column) and minimum (right column).

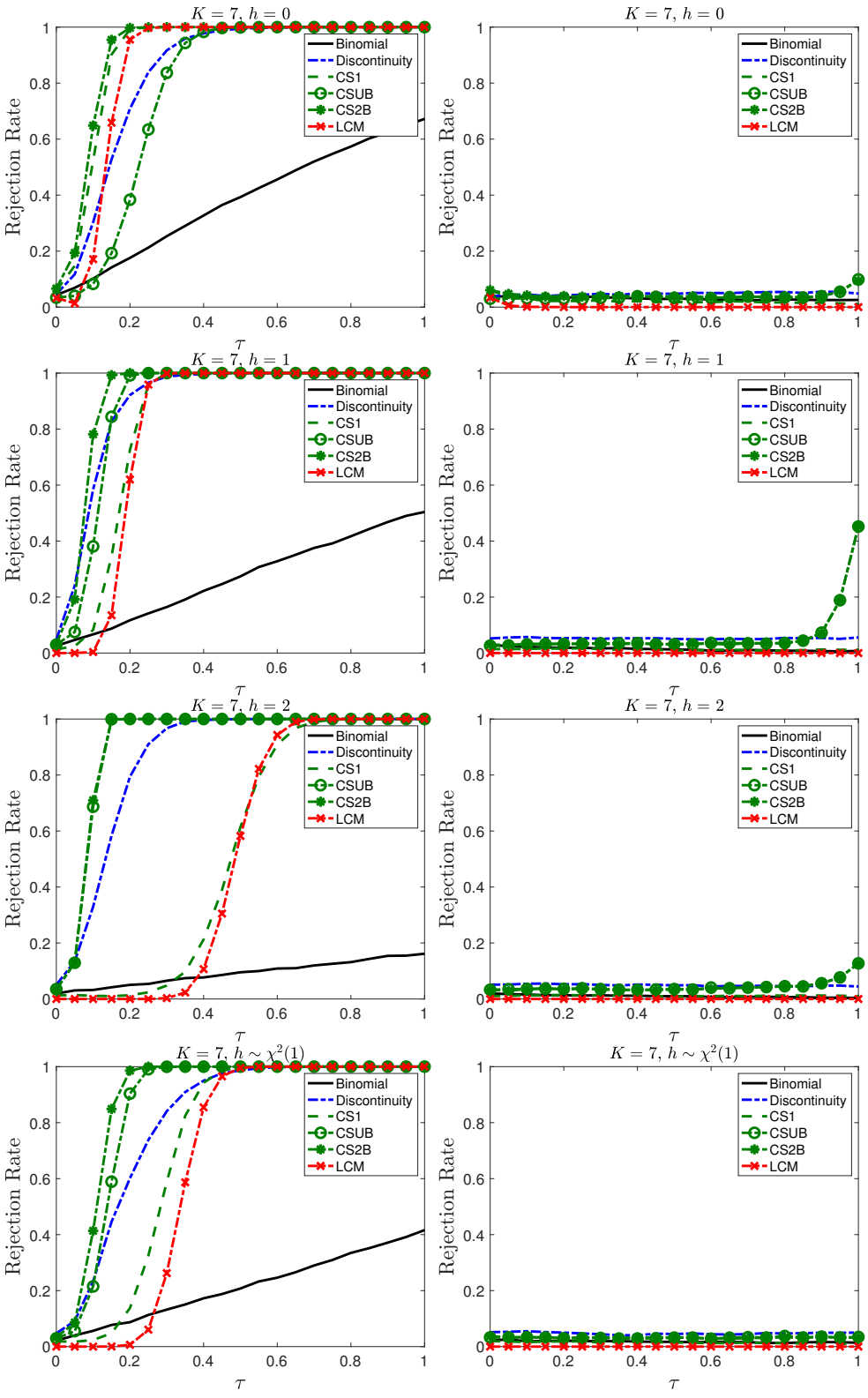

Figure B.10. Power curves covariate selection with $K = 7$. Thresholding (left column) and minimum (right column).

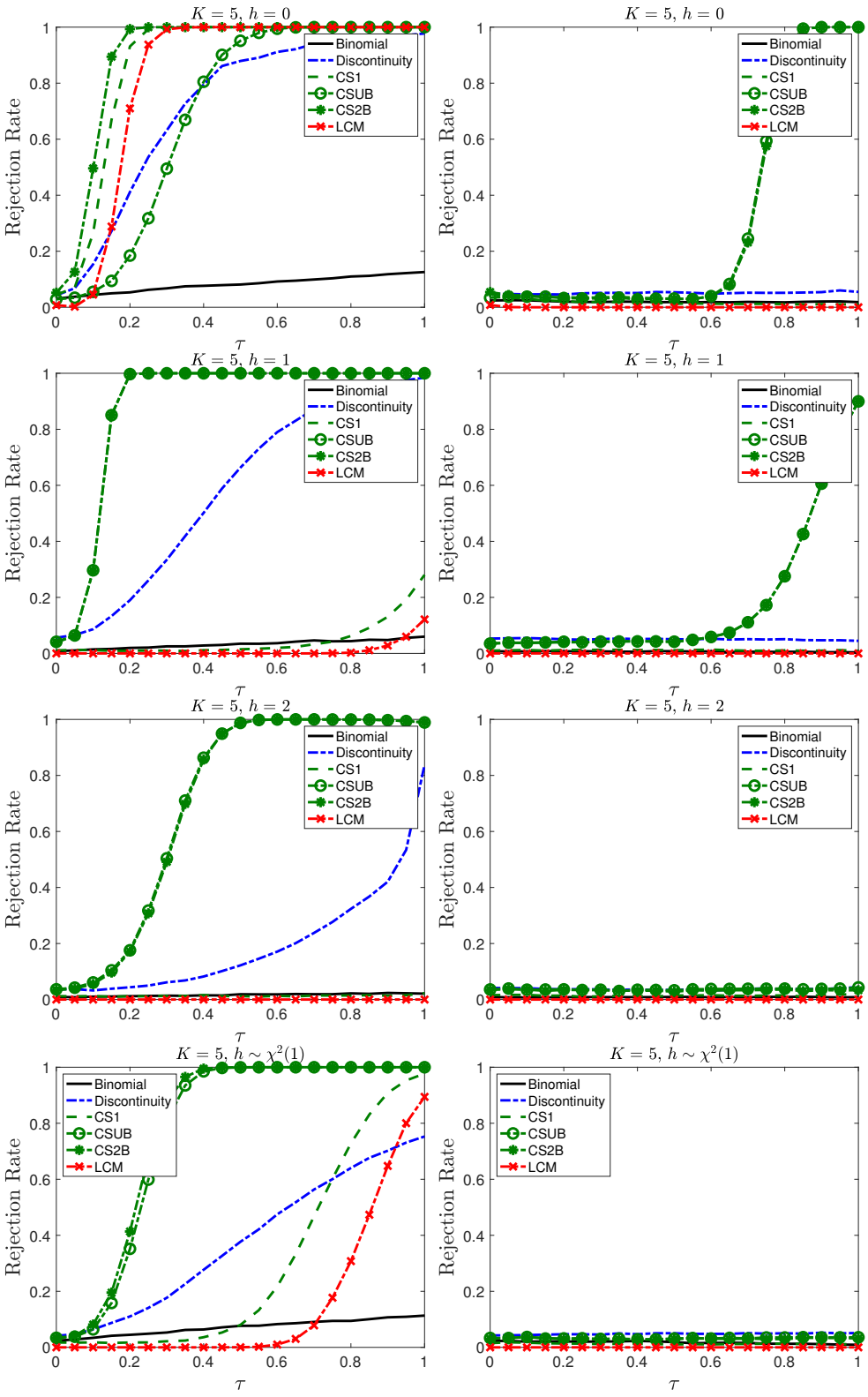

Figure B.11. Power curves IV selection with $K = 5$. Thresholding (left column) and minimum (right column).

Table B.1. The effect of publication bias: 1-sided tests, $h = 1$, $\tau = 0.5$

	Test					
	Binomial	Discontinuity	CS1	CSUB	CS2B	LCM
Cov Selection ($K = 3$, thresholding)						
No Pub Bias	0.045	0.629	0.023	0.715	0.869	0
Sharp Pub Bias	0.045	1	0.044	1	1	0.004
Smooth Pub Bias	0.021	0.377	0.009	1	1	0
Cov Selection ($K = 3$, minimum)						
No Pub Bias	0.016	0.05	0.014	0.036	0.03	0
Sharp Pub Bias	0.016	0.93	0.034	0.999	0.999	0
Smooth Pub Bias	0.008	0.05	0.011	1	1	0
IV Selection ($K = 3$, thresholding)						
No Pub Bias	0.024	0.279	0.01	0.999	0.999	0
Sharp Pub Bias	0.024	0.992	0.024	0.999	0.999	0
Smooth Pub Bias	0.012	0.157	0.009	1	1	0
IV Selection ($K = 3$, minimum)						
No Pub Bias	0.012	0.052	0.011	0.036	0.036	0
Sharp Pub Bias	0.012	0.914	0.028	0.999	0.999	0
Smooth Pub Bias	0.01	0.045	0.008	1	1	0
Lag Selection (thresholding)						
No Pub Bias	0.558	0.667	0.086	0.183	0.372	0
Sharp Pub Bias	0.558	1	0.093	1	1	0.002
Smooth Pub Bias	0.325	0.406	0.023	1	1	0
Lag Selection (minimum)						
No Pub Bias	0.015	0.052	0.014	0.033	0.034	0
Sharp Pub Bias	0.015	0.928	0.036	1	1	0
Smooth Pub Bias	0.009	0.053	0.011	1	1	0

Table B.2. The effect of publication bias: 1-sided tests, $h = 2$, $\tau = 0.5$

	Test					
	Binomial	Discontinuity	CS1	CSUB	CS2B	LCM
Cov Selection ($K = 3$, thresholding)						
No Pub Bias	0.036	0.494	0.012	1	1	0
Sharp Pub Bias	0.036	0.998	0.027	0.999	0.999	0
Smooth Pub Bias	0.018	0.28	0.01	1	1	0
Cov Selection ($K = 3$, minimum)						
No Pub Bias	0.015	0.05	0.013	0.037	0.037	0
Sharp Pub Bias	0.015	0.925	0.029	0.998	0.998	0
Smooth Pub Bias	0.008	0.046	0.01	1	1	0
IV Selection ($K = 3$, thresholding)						
No Pub Bias	0.016	0.097	0.013	0.949	0.948	0
Sharp Pub Bias	0.016	0.994	0.036	0.969	0.969	0
Smooth Pub Bias	0.01	0.08	0.012	0.997	0.997	0
IV Selection ($K = 3$, minimum)						
No Pub Bias	0.008	0.041	0.009	0.043	0.042	0
Sharp Pub Bias	0.008	0.951	0.031	0.966	0.966	0
Smooth Pub Bias	0.004	0.042	0.012	0.12	0.119	0
Lag Selection (thresholding)						
No Pub Bias	0.41	0.237	0.015	0.542	0.543	0
Sharp Pub Bias	0.41	0.998	0.026	0.997	0.997	0
Smooth Pub Bias	0.22	0.128	0.011	0.999	0.999	0
Lag Selection (minimum)						
No Pub Bias	0.01	0.051	0.011	0.04	0.04	0
Sharp Pub Bias	0.01	0.918	0.03	0.996	0.996	0
Smooth Pub Bias	0.006	0.042	0.011	0.953	0.953	0

Table B.3. The effect of publication bias: 1-sided tests, $h \sim \chi^2(1)$, $\tau = 0.5$

	Test					
	Binomial	Discontinuity	CS1	CSUB	CS2B	LCM
Cov Selection ($K = 3$, thresholding)						
No Pub Bias	0.044	0.213	0.022	0.468	0.536	0
Sharp Pub Bias	0.044	0.994	0.043	0.999	0.999	0
Smooth Pub Bias	0.027	0.125	0.012	1	1	0
Cov Selection ($K = 3$, minimum)						
No Pub Bias	0.018	0.054	0.016	0.034	0.031	0
Sharp Pub Bias	0.018	0.901	0.043	0.999	0.999	0
Smooth Pub Bias	0.014	0.045	0.013	0.133	0.133	0
IV Selection ($K = 3$, thresholding)						
No Pub Bias	0.038	0.109	0.016	0.263	0.287	0
Sharp Pub Bias	0.038	0.988	0.041	0.999	0.999	0
Smooth Pub Bias	0.024	0.065	0.013	0.931	0.931	0
IV Selection ($K = 3$, minimum)						
No Pub Bias	0.025	0.047	0.014	0.036	0.036	0
Sharp Pub Bias	0.025	0.926	0.041	0.999	0.999	0
Smooth Pub Bias	0.013	0.043	0.013	0.078	0.077	0
Lag Selection (thresholding)						
No Pub Bias	0.354	0.242	0.054	0.081	0.194	0
Sharp Pub Bias	0.354	0.997	0.097	0.998	0.998	0
Smooth Pub Bias	0.202	0.109	0.019	0.232	0.228	0
Lag Selection (minimum)						
No Pub Bias	0.021	0.046	0.014	0.033	0.03	0
Sharp Pub Bias	0.021	0.923	0.054	0.997	0.997	0
Smooth Pub Bias	0.014	0.043	0.014	0.064	0.062	0

C Additional results and proofs for Chapter 3

C.1 Proofs

C.1.1 Proof of Proposition 3.1

Let $N_b = N_U + N_L$ and N be the total number of observations. Note that $N_U/N \xrightarrow{p} p_U(b)$ and $N_L/N \xrightarrow{p} p_L(b)$ and for any $0 < b \leq t$, there is a distribution $\Pi_{t,b}$ such that $0 < p_L(b) < p_U(b)$. Then, for this choice of $\Pi = \Pi_{t,b}$, we have

$$
\begin{aligned}
1 &\geq \Pr(1 - F_{Bin}(N_U - 1; n, 0.5) < \alpha) \\
&= \Pr\left(1 - \Phi\left(\frac{N_U - N_L - 2}{\sqrt{N_b}}\right) + \Phi\left(\frac{N_U - N_L - 2}{\sqrt{n}}\right) - F_{Bin}(N_U - 1; N_b, 0.5) < \alpha\right) \\
&\geq \Pr\left(1 - \Phi\left(\frac{N_U - N_L - 2}{\sqrt{N_b}}\right) + \left|\Phi\left(\frac{N_U - N_L - 2}{\sqrt{N_b}}\right) - F_{Bin}(N_U - 1; N_b, 0.5)\right| < \alpha\right) \\
&= \Pr\left(1 - \Phi\left(\sqrt{N}\frac{N_U/N - N_L/N - 2/N}{\sqrt{N_b/N}}\right) + \left|\Phi\left(\frac{N_U - N_L - 2}{\sqrt{N_b}}\right) - F_{Bin}(N_U - 1; N_b, 0.5)\right| < \alpha\right) \\
&\to 1,
\end{aligned}
$$

where the last step follows due to the fact that $\left|\Phi\left(\frac{N_U - N_L - 2}{\sqrt{N_b}}\right) - F_{Bin}(N_U - 1; N_b, 0.5)\right| \xrightarrow{p} 0$ due to the Central Limit Theorem and $\sqrt{N}\frac{N_U/N - N_L/N - 2/N}{\sqrt{N_b/N}} \xrightarrow{p} \infty$ as $N \to \infty$. □

C.1.2 Proof of Theorem 3.1

In the absence of p-hacking, $\hat{\Omega}_b \xrightarrow{p} \Omega_b$ and $\frac{\sqrt{N}(\hat{\Delta}_b - \Delta(\Pi))}{\sqrt{\hat{\Omega}_b}} \xrightarrow{d} \mathcal{N}(0,1)$. We have,

$$
\begin{aligned}
\Pr(T > z_{1-\alpha}) &= \Pr\left(\frac{\sqrt{N}(\hat{\Delta}_b - \overline{\Delta}_b)}{\sqrt{\hat{\Omega}_b}} > z_{1-\alpha}\right) \\
&= \Pr\left(\frac{\sqrt{N}(\hat{\Delta}_b - \Delta(\Pi))}{\sqrt{\hat{\Omega}_b}} - \frac{\sqrt{N}(\overline{\Delta}_b - \Delta(\Pi))}{\sqrt{\hat{\Omega}_b}} > z_{1-\alpha}\right) \\
&\leq \Pr\left(\frac{\sqrt{N}(\hat{\Delta}_b - \Delta(\Pi))}{\sqrt{\hat{\Omega}_b}} > z_{1-\alpha}\right)
\end{aligned}
$$

where the weak inequality holds as equality for $\Pi = \Pi_b^*$.

Now note that $\frac{\sqrt{N}(\hat{\Delta}_b - \Delta(\Pi))}{\sqrt{\hat{\Omega}_b}}$ is a t-statistic for testing that $E[Y] = 0$, where $Y = 1\{p \in [t+b, t)\} - 1\{p \in (t, t-b]\} - \Delta(\Pi)$. Clearly, $|Y - E[Y]| \leq 1 + \overline{\Delta}_b$ for all Π. Also, note that $\Omega_b = Var(Y) = p_U(b)(1 - p_U(b) + p_L(b)) + p_L(b)(1 - p_L(b) + p_U(b)) \geq p_U(b)p_L(b)$.

$$\begin{aligned} \inf_{\Pi \in \mathscr{F}_{A,\gamma}} p_u(b) &= \inf_{\Pi \in \mathscr{F}_{A,\gamma}} \int_{\mathscr{H}} K_U(h;t,b) d\Pi(h) \\ &= \gamma \underline{p}_1(A), \end{aligned}$$

where $p_1(A) > 0$ since $K_U(h;t,b)$ is zero only when Π put all its mass to either $+\infty$ or $-\infty$. Similarly,

$$\begin{aligned} \inf_{\Pi \in \mathscr{F}_{A,\gamma}} p_l(b) &= \inf_{\Pi \in \mathscr{F}_{A,\gamma}} \int_{\mathscr{H}} K_L(h;t,b) d\Pi(h) \\ &= \gamma \underline{p}_2(A), \end{aligned}$$

where $p_2(A) > 0$ since $K_L(h;t,b)$ is zero only when Π put all its mass to either $+\infty$ or $-\infty$.

It follow than the the condition 8 form Romano (2004) is satisfied for $\mathscr{F}_{A,\gamma}$ since

$$E\left[\frac{|Y - E[Y]|^{2+\varepsilon}}{\Omega_b^{2+\varepsilon}}\right] \leq \frac{(1 + \overline{\Delta}_b)^{2+\varepsilon}}{(\gamma^2 p_1(A) p_2(A))^{2+\varepsilon}}$$

. Therefore, Theorem 5(i) from Romano (2004) implies that

$$\left| \sup_{\Pi \in \mathscr{F}_{A,\gamma}} \Pr(T > z_{1-\alpha}) - \alpha \right| \to 0 \text{ as } n \to \infty.$$

□

C.1.3 Proof of Theorem 3.2

Under Assumption 1 the maximum likelihood estimator $\hat{\gamma}$ is consistent and asymptotically normal (Van der Vaart (2000), Theorem 5.39)

$$\sqrt{N}(\hat{\gamma}-\gamma) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}_{\gamma}^{-1})$$

Note that

$$\sqrt{N}(\hat{\Delta}_b - \tilde{\Delta}_b) = \sqrt{N}(\hat{\Delta}_b - \Delta_b(\Pi)) - \sqrt{N}(\Delta_b(\hat{\Pi}) - \Delta_b(\Pi)).$$

Now

$$\sqrt{N}(\hat{\Delta}_b - \Delta_b(\Pi)) \xrightarrow{d} \xi_1 \sim \mathcal{N}(0, \Omega_b)$$

and

$$\begin{aligned}
\sqrt{N}(\Delta_b(\hat{\Pi}) - \Delta_b(\Pi)) &= \sqrt{N}\left(\int_{\mathcal{H}} K(h;t,b)\pi(h;\hat{\gamma})dh - \int_{\mathcal{H}} K(h;t,b)\pi(h;\gamma)dh\right) \\
&= \int_{\mathcal{H}} K(h;t,b)\sqrt{N}(\pi(h;\hat{\gamma}) - \pi(h;\gamma))dh \\
&= \sqrt{N}(\hat{\gamma}-\gamma)' \int_{\mathcal{H}} K(h;t,b)\partial\pi(h;\bar{\gamma})/\partial\gamma dh \\
&= \sqrt{N}(\hat{\gamma}-\gamma)'D(\Pi) + o_P(1),
\end{aligned}$$

where $\bar{\gamma}_i \in (\gamma_i, \hat{\gamma}_i)$ for $i = 1, ..., k$. Thus,

$$\sqrt{N}(\Delta_b(\hat{\Pi}) - \Delta_b(\Pi)) \xrightarrow{d} \xi_2 \sim \mathcal{N}(0, D(\Pi)'\mathcal{I}_{\gamma}^{-1}D(\Pi))$$

and

$$Cov(\xi_1, \xi_2) = \lim_{N\to\infty} Cov(\sqrt{N}(\hat{\Delta}_b - \Delta_b(\Pi)), \sqrt{N}(\hat{\gamma}-\gamma)'D(\Pi)) = C'_{\Delta\gamma}D(\Pi).$$

□

C.1.4 Proof of Proposition 3.2

Let F_X be the distribution function of X and define

$$\eta(a) = \eta(a;\delta,F_X) = \int_{\mathscr{X}} \Lambda(a+x'\delta)dF_X(x).$$

If $\Pi_0 = \Pi_k$, then $\omega_0 = \omega_k$. On the other hand,

$$\begin{aligned}\omega_k &= \int_{\mathscr{X}} \Pr(|z| > t||z| \in N_b(t), M_k = 1, X)dF_X(x) \\ &= \int_{\mathscr{X}} \Lambda(\beta_0 + \beta_k + x\delta)dF_X(x).\end{aligned}$$

Therefore,

$$\begin{aligned}\beta_k &= \eta^{-1}(\omega_k) - \beta_0 \\ &= \eta^{-1}(\omega_k) - \eta^{-1}(\omega_0) \\ &= 0.\end{aligned}$$

□

C.1.5 Proof of Proposition 3.3

Let

$$\mathscr{H}^* = \arg\max_{h\in\mathscr{H}} \left\{ \frac{K_U(h;t,b) - K_L(h;t,b)}{K(h;t,b)} \right\}.$$

and $h_0^* \notin \mathscr{H}^*$ and $h_k^* \in \mathscr{H}^*$. It follows that if Π_0 is a point mass at h_0^* and Π_k is a point mass at h_k^*, then $\omega_0 < \omega_k$. Note that $|h_k^*| < \infty$.

Not let $\hat{\beta}_k$ be the estimate of β_k and $\hat{V}_k$ be the estimate of its variance. Since $|h_k^*| < \infty$, $\hat{V}_k \xrightarrow{p} V_k < \infty$ and from the proof of Proposition 2 we know that

$$\hat{\beta}_k \xrightarrow{p} \beta_k = \eta^{-1}(\omega_k) - \eta^{-1}(\omega_0).$$

$\eta^{-1}(a)$ is an increasing function. To see this, note that $\eta(a)$ is an increasing function because $\Lambda(\cdot)$ is a CDF and the inverse of increasing function is increasing. It follows then that

$$T_k := \frac{\sqrt{N}(\hat{\beta}_k - \beta_k)}{\sqrt{\hat{V}_k}} \xrightarrow{p} \infty.$$

□

C.1.6 Proof of Theorem 3.3

First, we establish the asymptotic expansions for local proportions for each category:

$$\begin{aligned}
\sqrt{N}\begin{pmatrix} \hat{p}_{U,k} - p_{U,k} \\ \hat{p}_{L,k} - p_{L,k} \end{pmatrix} &= \begin{pmatrix} \int_{\mathscr{H}} K_U(h)\sqrt{N}(\pi_k(h;\hat{\gamma}_k) - \pi_k(h;\gamma_k))dh \\ \int_{\mathscr{H}} K_L(h)\sqrt{N}(\pi_k(h;\hat{\gamma}_k) - \pi_k(h;\gamma_k))dh \end{pmatrix} \\
&= \begin{pmatrix} \sqrt{N}(\hat{\gamma}_k - \gamma_k)' D_U(\Pi_k) \\ \sqrt{N}(\hat{\gamma}_k - \gamma_k)' D_L(\Pi_k) \end{pmatrix} + o_P(1),
\end{aligned}$$

where $D_L(\Pi)$ and $D_U(\Pi)$ defined analogously to $D(\Pi)$ for lower and upper proportions respectively. Let $\omega_k = \frac{p_{U,k}}{p_{U,k}+p_{L,k}}$, then using the first-order Taylor expansion we get

$$\begin{aligned}
\sqrt{N}(\hat{\omega}_k - \omega_k) &= \frac{p_{L,k}\sqrt{N}(\hat{p}_{U,k} - p_{U,k}) - \sqrt{N}(\hat{p}_{L,k} - p_{L,k})}{(p_{U,k} + p_{L,k})^2} + o_P(1) \\
&= \sqrt{N}(\hat{\gamma}_k - \gamma_k)' \Upsilon_{\omega,k} + o_P(1),
\end{aligned}$$

where $\Upsilon_{\omega,k} = \frac{p_{L,k} D_U(\Pi_k) - D_L(\Pi_k)}{(p_{U,k}+p_{L,k})^2}$.

Let $Y_{U,i} = 1\{t < |Z_i| < t+b\}$ and $Y_{L,i} = 1\{t-b < |Z_i| \leq t\}$. The log-likelihood function of the problem can be written as

$$L = \frac{1}{N}\sum_{i=1}^{N} Y_{U,i} \log \Lambda(M_i'\beta + X_i'\delta) + Y_{L,i} \log(1 - \Lambda(M_i'\beta + X_i'\delta))$$

The first-order condition for the maximum likelihood problem is then

$$\frac{1}{N}\sum_{i=1}^{N}\Psi(Y_{U,i},Y_{L,i},M_i,X_i;\hat{\beta},\hat{\delta})=0,$$

where $\Psi(Y_{U,i},Y_{L,i},M_i,X_i;\beta,\delta)=\begin{pmatrix}\Psi_{\beta,i}\\ \Psi_{\delta,i}\end{pmatrix}=\frac{(Y_{U,i}(1-\Lambda(M_i'\beta+X_i'\delta))-Y_{L,i}\Lambda(M_i'\beta+X_i'\delta))\Lambda'(M_i'\beta+X_i'\delta)}{\Lambda(M_i'\beta+X_i'\delta)(1-\Lambda(M_i'\beta+X_i'\delta))}\begin{pmatrix}M_i\\ X_i\end{pmatrix}$

It follows using standard Taylor expansion argument that

$$\sqrt{N}\begin{pmatrix}\hat{\beta}-\beta\\ \hat{\delta}-\delta\end{pmatrix}=-H^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\Psi(Y_{U,i},Y_{L,i},M_i,X_i;\beta,\delta)+o_P(1),$$

where $H:=E\left[\left(\frac{Y_{U,i}(\Lambda_i''\Lambda_i-(\Lambda_i')^2)}{\Lambda_i^2}+\frac{Y_{L,i}(\Lambda_i''(1-\Lambda_i)-(\Lambda_i')^2)}{(1-\Lambda_i)^2}\right)\begin{pmatrix}M_i\\ X_i\end{pmatrix}'\begin{pmatrix}M_i\\ X_i\end{pmatrix}\right]:=\begin{pmatrix}H_{\beta\beta} & H_{\beta\delta}\\ H_{\beta\delta}' & H_{\delta\delta}\end{pmatrix}$ and for what follows we will define the partition of H^{-1} as $H^{-1}=\begin{pmatrix}A_{\beta\beta} & A_{\beta\delta}\\ A_{\beta\delta}' & A_{\delta\delta}\end{pmatrix}$.

Define

$$\eta_k(a)=\eta(a;\delta,F_{X|k})=\int_{\mathscr{X}}\Lambda(a+x'\delta)dF_{X|k}(x)$$

and let

$$\widehat{\eta_k^{-1}}(u)=\eta^{-1}(u;\hat{\delta},\hat{F}_{X|k})$$

Note that

$$\begin{aligned}\widehat{\eta_k^{-1}}(u) &= \eta^{-1}(u;\delta,\hat{F}_{X|k})+\Upsilon_{\delta}(u;\bar{\delta}_k,\hat{F}_{X|k})(\hat{\delta}-\delta)\\ &= \eta^{-1}(u;\delta,F_{X|k})+\Upsilon_{\Lambda,k}(u)\int_{\mathscr{X}}\Lambda(\eta^{-1}(u)+x\delta)d(\hat{F}_{X|k}-F_{X|k})\\ &+ \Upsilon_{\delta,k}(u)(\hat{\delta}-\delta)+o_P(1),\end{aligned}$$

where $\Upsilon_{\delta,k}(u) = \Upsilon_\delta(u;\delta,F_{X|k}) = -\frac{\int_{\mathscr{X}} x'\Lambda'(\eta^{-1}(u;\delta,F_{X|k})+x\delta)dF_{X|k}(x)}{\int_{\mathscr{X}} \Lambda(\eta^{-1}(u;\delta,F_{X|k})+x'\delta)dF_{X|k}(x)}$ is the derivative of $\eta^{-1}(u;\delta,F_{X|k})$ with respect to δ and the second equality follows from the fact that $\eta^{-1}(u;\delta,F_{X|k})$ as a functional of $F_{X|k}$ is Hadamard differentiable with derivative $\eta^{-1}_{F_{X|k}}(H) = \Upsilon_{\Lambda,k}(u)\int_{\mathscr{X}} \Lambda(\eta_k^{-1}(u)+x'\delta)dH$, where $\Upsilon_{\Lambda,k}(u) = -\frac{1}{\int_{\mathscr{X}} \Lambda(\eta^{-1}(u;\delta,F_{X|k})+x'\delta)dF_{X|k}(x)}$. It follows then

$$\begin{aligned}
\sqrt{N}(\widehat{\eta_k^{-1}}(\hat{\omega}_k) - \eta_k^{-1}(\omega_k)) &= \sqrt{N}(\widehat{\eta_k^{-1}}(\hat{\omega}_k) - \eta_k^{-1}(\hat{\omega}_k)) + \sqrt{N}(\eta_k^{-1}(\hat{\omega}_k) - \eta_k^{-1}(\omega_k)) \\
&= \Upsilon_{\Lambda,k}(\hat{\omega}_k)\int_{\mathscr{X}} \Lambda(\eta_k^{-1}(\hat{\omega}_k)+x'\delta)d\sqrt{N}(\hat{F}_{X|k} - F_{X|k}) \\
&+ \Upsilon_{\delta,k}(\hat{\omega}_k)\sqrt{N}(\hat{\delta}-\delta) \\
&- \Upsilon_{\Lambda,k}(\bar{\omega}_k)\sqrt{N}(\hat{\omega}_k - \omega_k) + o_P(1) \\
&= \Upsilon_{\Lambda,k}(\omega_k)\int_{\mathscr{X}} \Lambda(\eta_k^{-1}(\omega_k)+x'\delta)d\sqrt{N}(\hat{F}_{X|k} - F_{X|k}) \\
&+ \Upsilon_{\delta,k}(\omega_k)\sqrt{N}(\hat{\delta}-\delta) \\
&- \Upsilon_{\Lambda,k}(\omega_k)\sqrt{N}(\hat{\gamma}_k - \gamma_k)'\Upsilon_{\omega,k} + o_P(1)
\end{aligned}$$

Now it follows that

$$\sqrt{N}(\widehat{\eta^{-1}} - \eta^{-1}) \xrightarrow{d} \mathscr{N}(0, V_{\eta\eta}),$$

where the diagonal elements of $V_{\eta\eta}$ are given by

$$\begin{aligned}
[V_{\eta\eta}]_{kk} &= \Upsilon^2_{\Lambda,k}(\omega_k)(V_{\Lambda_k}(\omega_k) - 2C_{\Lambda_k\gamma_k}\Upsilon_{\omega,k} + \Upsilon'_{\omega,k}\mathscr{I}^{-1}_{\gamma_k}\Upsilon_{\omega,k}) + \Upsilon^2_{\delta,k}(\omega_k)A_{\delta\delta} \\
&+ 2\Upsilon_{\Lambda,k}(\omega_k)\Upsilon_{\delta,k}(\omega_k)(C_{\Lambda_k\delta} - C'_{\delta\gamma_k}\Upsilon_{\omega,k})
\end{aligned}$$

and the off diagonal elements are given by

$$\begin{aligned}[V_{\eta\eta}]_{kk'} &= \Upsilon_{\delta,k}(\omega_k)\Upsilon_{\delta,k'}(\omega_{k'})A_{\delta\delta} - \Upsilon_{\delta,k}(\omega_k)\Upsilon_{\Lambda,k'}(\omega_{k'})C'_{\delta,\gamma_{k'}}\Upsilon_{\omega,k'} \\ &- \Upsilon_{\delta,k'}(\omega_{k'})\Upsilon_{\Lambda,k}(\omega_k)C'_{\delta,\gamma_k}\Upsilon_{\omega,k},\end{aligned}$$

where $C_{\Lambda_k\delta}$ is the asymptotic covariance between $\int_{\mathscr{X}}\Lambda(\eta_k^{-1}(\hat{\omega}_k)+x'\delta)d\sqrt{N}(\hat{F}_{X|k}-F_{X|k})$ and $\hat{\delta}$, $C_{\delta\gamma_k}$ is the asymptotic covariance between $\hat{\delta}$ and $\hat{\gamma}_k$ and V_{Λ_k} is the asymptotic variance of $\int_{\mathscr{X}}\Lambda(\eta_k^{-1}(\hat{\omega}_k)+x\delta)d\sqrt{N}(\hat{F}_{X|k}-F_{X|k})$.

Finally,

$$\begin{aligned}\sqrt{N}(\hat{\beta}-A\widehat{\eta^{-1}}) &= \sqrt{N}(\hat{\beta}-\beta)+A\sqrt{N}(\widehat{\eta^{-1}}-\eta^{-1}) \\ &\xrightarrow{d} \zeta_1+A\zeta_2,\end{aligned}$$

where

$$\begin{pmatrix}\zeta_1\\ \zeta_2\end{pmatrix} \sim \mathscr{N}\left(\begin{pmatrix}0\\ 0\end{pmatrix}, \begin{pmatrix}A_{\beta\beta} & V_{\beta\eta}\\ V'_{\beta\eta} & V_{\eta\eta}\end{pmatrix}\right)$$

and the kth row of matrix $V_{\beta\eta}$ is given by

$$\begin{aligned}\Upsilon_{\delta,k}(\omega_k)A_{\beta\delta} &+ A_{\beta\beta}(\Upsilon_{\Lambda,k}C_{\Lambda_k\Psi_\beta} - \Upsilon_{\Lambda,k}(\omega_k)C'_{\Psi_\beta\gamma_k}\Upsilon_{\omega,k}) \\ &+ A_{\beta\delta}(\Upsilon_{\Lambda,k}C_{\Lambda_k\Psi_\delta} - \Upsilon_{\Lambda,k}(\omega_k)C'_{\Psi_\delta\gamma_k}\Upsilon_{\omega,k}),\end{aligned}$$

where $C_{\Lambda_k\Psi_\beta}$ is the asymptotic covariance between $\int_{\mathscr{X}}\Lambda(\eta_k^{-1}(\hat{\omega}_k)+x'\delta)d\sqrt{N}(\hat{F}_{X|k}-F_{X|k})$ and $N^{-1}\sum_{i=1}^N\Psi_{\beta,i}$ and $C_{\Psi_\beta\gamma_k}, C_{\Lambda_k\Psi_\delta}$ and $C_{\Psi_\delta\gamma_k}$ are defined analogously. □

C.2 Illustrative Example Derivations

The expressions for $g_{|z|}(x)$ and $\Delta_b(\mathscr{N}(\mu_h,\sigma_h^2))$ can be obtained by direct integration. For the rest, define $\xi_{U,i} := 1\{t < Z_i < t+b\} + 1\{-t-b < Z_i < -t\}$, $\xi_{L,i} := 1\{t-b < Z_i <$

$t\}+1\{-t<Z_i<-t+b\}$ and $\xi_i := \xi_{U,i}-\xi_{L,i}$. Note that $\hat{\Delta}_b = \frac{1}{N}\sum_{i=1}^N \xi_i$. Also define $Q(a,b) := \Phi(b)-\Phi(a)$. Using the properties of truncated normal distribution we can obtain

$$E[Z_i 1\{t-b<Z_i<t\}] = \mu_h Q(\alpha_1^+,\alpha_2^+)+\sqrt{\sigma_h^2+1}\left(\phi\left(\alpha_1^+\right)-\phi\left(\alpha_2^+\right)\right)$$

and similarly

$$E[Z_i 1\{-t<Z_i<-t+b\}] = \mu_h Q(\alpha_2^-,\alpha_1^-)+\sqrt{\sigma_h^2+1}\left(\phi\left(\alpha_2^-\right)-\phi\left(\alpha_1^-\right)\right),$$

$$E[Z_i 1\{t<Z_i<t+b\}] = \mu_h Q(\alpha_2^+,\alpha_3^+)+\sqrt{\sigma_h^2+1}\left(\phi\left(\alpha_2^+\right)-\phi\left(\alpha_3^+\right)\right),$$

$$E[Z_i 1\{-t-b<Z_i<-t\}] = \mu_h Q(\alpha_3^-,\alpha_2^-)+\sqrt{\sigma_h^2+1}\left(\phi\left(\alpha_3^-\right)-\phi\left(\alpha_2^-\right)\right).$$

It follows that

$$\begin{aligned} Cov(\sqrt{N}(\hat{\Delta}_b-\Delta_b),\sqrt{N}(\hat{\mu}_h-\mu_h)) &= E[Z_i\xi_i]-E[Z_i]E[\xi_i] \\ &= E[Z_i\xi_{U,i}]-E[Z_i\xi_{L,i}]-E[Z_i](E[\xi_{U,i}]-E[\xi_{L,i}]) \\ &= \sqrt{\sigma_h^2+1}\left(2\phi\left(\alpha_2^+\right)-\phi\left(\alpha_1^+\right)-\phi\left(\alpha_3^+\right)\right) \\ &- \sqrt{\sigma_h^2+1}\left(2\phi\left(\alpha_2^-\right)-\phi\left(\alpha_1^-\right)-\phi\left(\alpha_3^-\right)\right) \end{aligned}$$

Note that, since $\Pr(\frac{1}{N}\sum_{i=1}^N (Z_i-\bar{Z})^2 \le 1)\to 0$ as $N\to\infty$, $\hat{\sigma}_h^2$ is asymptotically equivalent to $\frac{1}{N-1}\sum_{i=1}^N (Z_i-\bar{Z})^2 = \frac{1}{N}\sum_{i=1}^N Z_i^2 - \frac{1}{N(N-1)}\sum_{i\neq j} Z_iZ_j - 1$. Therefore, the covariance between

$\hat{\Delta}_b$ and $\hat{\sigma}_h^2$ is

$$\begin{aligned} Cov(\hat{\Delta}_b, \hat{\sigma}_h^2) &= \frac{1}{N} Cov(\xi_i, Z_i^2) - \frac{1}{N} Cov(\xi_i, \frac{1}{N-1} \sum_{i \neq j} Z_i Z_j) \\ &= \frac{1}{N} Cov(\xi_i, Z_i^2) - \frac{2\sum_{i \neq j} E[Z_j]}{N} Cov(\xi_i, Z_i) \\ &= \frac{1}{N} Cov(\xi_i, Z_i^2) - \frac{2\mu_h}{N} Cov(\xi_i, Z_i) \end{aligned}$$

Again, using the properties of truncated normal distribution, we have

$$\begin{aligned} Cov(1\{t-b < Z_i < t\}, Z_i^2) &= E[1\{t-b < Z_i < t\} Z_i^2] - E[1\{t-b < Z_i < t\}] E[Z_i^2] \\ &= Q(\alpha_1^+, \alpha_2^+)(\sigma_h^2 + 1) \left(1 + \frac{\alpha_1^+ \phi(\alpha_1^+) - \alpha_2^+ \phi(\alpha_2^+)}{Q(\alpha_1^+, \alpha_2^+)} - \left(\frac{\phi(\alpha_1^+) - \phi(\alpha_2^+)}{Q(\alpha_1^+, \alpha_2^+)} \right)^2 \right) \\ &+ Q(\alpha_1^+, \alpha_2^+) \left(\mu_h + \frac{\phi(\alpha_1^+) - \phi(\alpha_2^+)}{Q(\alpha_1^+, \alpha_2^+)} \sqrt{\sigma_h^2 + 1} \right)^2 \\ &- Q(\alpha_1^+, \alpha_2^+)(\sigma_h^2 + 1 + \mu_h^2) \\ &= (\sigma_h^2 + 1) \left(\alpha_1^+ \phi(\alpha_1^+) - \alpha_2^+ \phi(\alpha_2^+) \right) + 2\mu_h \sqrt{\sigma_h^2 + 1} (\phi(\alpha_1^+) - \phi(\alpha_2^+)) \end{aligned}$$

and similarly

$$Cov(1\{t < Z_i < t+b\}, Z_i^2) = (\sigma_h^2 + 1) \left(\alpha_2^+ \phi(\alpha_2^+) - \alpha_3^+ \phi(\alpha_3^+) \right) + 2\mu_h \sqrt{\sigma_h^2 + 1} (\phi(\alpha_2^+) - \phi(\alpha_3^+)).$$

$$Cov(1\{-t-b < Z_i < -t\}, Z_i^2) = (\sigma_h^2 + 1) \left(\alpha_3^- \phi(\alpha_3^-) - \alpha_2^- \phi(\alpha_2^-) \right) + 2\mu_h \sqrt{\sigma_h^2 + 1} (\phi(\alpha_3^-) - \phi(\alpha_2^-)).$$

$$Cov(1\{-t < Z_i < -t+b\}, Z_i^2) = (\sigma_h^2 + 1) \left(\alpha_2^- \phi(\alpha_2^-) - \alpha_1^- \phi(\alpha_1^-) \right) + 2\mu_h \sqrt{\sigma_h^2 + 1} (\phi(\alpha_2^-) - \phi(\alpha_1^-)).$$

Combining these results together gives

$$\begin{aligned} Cov(\sqrt{N}(\hat{\Delta}_b - \Delta_b), \sqrt{N}(\hat{\sigma}_h^2 - \sigma_h^2)) &= Cov(\xi_{U,i}, Z_i^2) - Cov(\xi_{L,i}, Z_i^2) - 2\mu_h Cov(\xi_i, Z_i) \\ &= (\sigma_h^2+1)(2\alpha_2^+\phi(\alpha_2^+) - \alpha_1^+\phi(\alpha_1^+) - \alpha_3^+\phi(\alpha_3^+)) \\ &- (\sigma_h^2+1)(2\alpha_2^-\phi(\alpha_2^-) - \alpha_1^-\phi(\alpha_1^-) - \alpha_3^-\phi(\alpha_3^-)) \end{aligned}$$

Finally, since $\pi(h) = \phi((h-\mu_h)/\sigma_h)/\sigma_h$, we have

$$\begin{aligned} \frac{\partial \pi}{\partial \gamma} &= \frac{\partial \phi((h-\mu_h)/\sigma_h)/\sigma_h}{\partial(\mu_h, \sigma_h^2)'} \\ &= \begin{pmatrix} \frac{h-\mu_h}{\sigma_h^2} \\ \frac{(h-\mu_h)^2 - \sigma_h^2}{2\sigma_h^4} \end{pmatrix} \pi(h) \end{aligned}$$

and the expression for $D(\Pi)$ follows.

C.3 Testing Multiple Thresholds

We can use standard moment inequality tests to combine multiple thresholds in a single test. Let $\mathscr{T} = \{t_1, ..., t_J\}$ be a set of thresholds and let $\mathscr{B} = \{b_1, ..., b_J\}$ be the set of corresponding bandwidth values with $0 < b_j \leq t_j$ (we allow them to differ across thresholds). Since the null hypothesis 3.7 holds for any value of b, we can consider the following testing problem as

$$H_0 : \Delta_{b_j}^{t_j} \leq \overline{\Delta}_{b_j}^{t_j} \text{ for all } t_j \in \mathscr{T} \quad \text{against} \quad H_1 : \Delta_{b_j}^{t_j} > \overline{\Delta}_{b_j}^{t_j} \text{ for some } t_j \in \mathscr{T}, \tag{16}$$

where $\overline{\Delta}_{b_j}^{t_j}$ is the upper bound for threshold t_j and bandwidth b_j.

This hypothesis can be tested by testing moment inequalities with, for instance, Cox and Shi (2022) test. Define $\Delta^{\mathscr{T}} := (\Delta_{b_1}^{t_1}, ..., \Delta_{b_J}^{t_J})'$ and $\overline{\Delta}^{\mathscr{T}} := (\overline{\Delta}_{b_1}^{t_1}, ..., \overline{\Delta}_{b_J}^{t_J})'$. Following Cox and Shi (2022), we test the null by comparing $\chi = \inf_{q: q \leq \overline{\Delta}^{\mathscr{T}}} N(\hat{\Delta}^{\mathscr{T}} - q)'\hat{\Omega}_b^{-1}(\hat{\Delta}^{\mathscr{T}} - q)$ to the critical value from a χ^2 distribution with the number of degrees of freedom equal to the number of active

inequalities. Here $\hat{\Omega}$ is a consistent estimate of the variance matrix of $\hat{\Delta}^{\mathscr{T}}$.

C.4 *P*-values vs *z*-values

In this section we compare Caliper tests based on the distribution of z-values and Caliper tests based on the distribution of p-values. To examine the difference between using z-values and p-values for caliper tests, we exploit the Monte Carlo design of Section 3.5.1. We concentrate on the case $K = 5$ and sample sizes of $N = 1000$.

Clearly, since there is a one-to-one correspondence between $|z|$-values and p-values, the test based on z-values that compares proportions in $|z| \in [t-b,t)$ versus $|z| \in [t,t+b]$ is numerically equivalent to the test based on p-values that compares proportions in $p \in [2(1-\Phi(|t+b|)), 2(1-\Phi(|t|))]$ versus $p \in (2(1-\Phi(|t|)), 2(1-\Phi(|t-b|))]$. In this case, the test based on p-values divides the interval $[2(1-\Phi(|t+b|)), 2(1-\Phi(|t-b|))]$ asymmetrically. In practice, researchers tend to use a symmetric partition of the testing interval. If we restrict attention to the tests that divide a chosen subinterval symmetrically, then the two approaches are hard to compare because $|z| \in [t-b,t+b]$ and $p \in [p-b,p+b]$ define very different subsamples of data and so it is hard to understand at what values of b the comparison is meaningful.

C.4.1 Using relative bandwidth

One possible way to make two approaches comparable is to consider relative bandwidth, that is use subsamples of $|z|$-values in $[t(1-b), t(1+b)]$ and subsamples of p-values in $[p(1-b), p(1-b)]$, where $b \in (0,1)$. Unfortunately, even in this case the comparison is difficult because intervals are still quite different due to Jensen's inequality. As simulations indicate, for some values of b, test based of $|z|$-values will generate more power and for other values tests based on p-values will be more powerful.

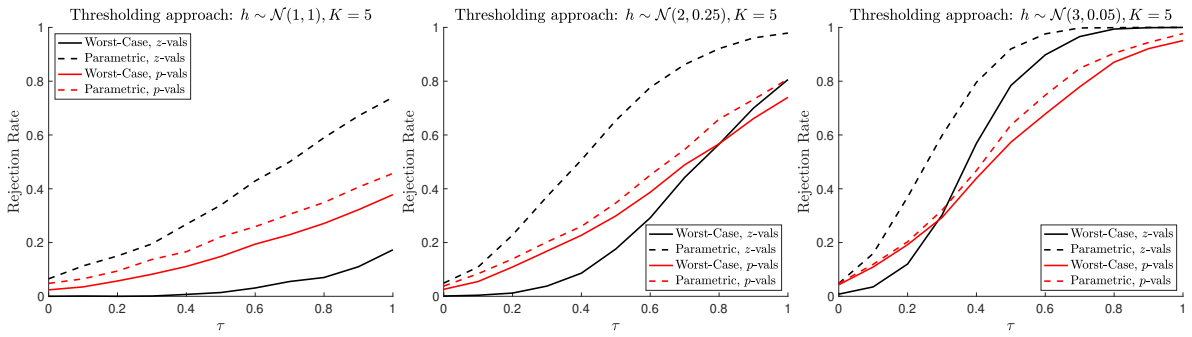

Figure C.1. Power curves, $b = 0.1$.

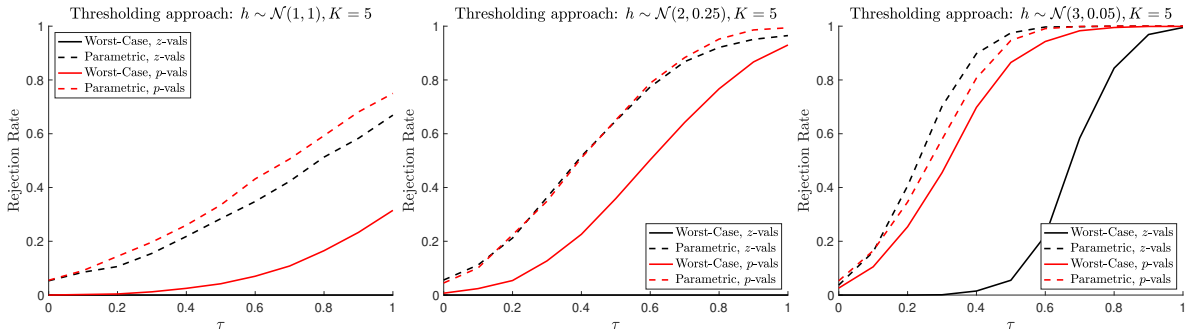

Figure C.2. Power curves, $b = 0.3$.

Figures C.1 and C.2 show the comparison between tests based on $|z|$-values on $[1.96(1-b), 1.96(1+b)]$ and and tests based on p-values on $[0.05(1-b), 0.05(1-b)]$ for $b = 0.1$ and $b = 0.3$. We compare both Robust Caliper tests and parametric versions of Caliper test. As we can see, for $b = 0.1$ the Robust Caliper test based on p-value performs better in term of power when $\Pi \in \{\mathcal{N}(1,1), \mathcal{N}(2,0.25)\}$, but for $\Pi = \mathcal{N}(3,0.05)$ the comparison depends on the amount of p-hacking. The parametric tests based on $|z|$-values exhibit higher power than parametric tests based in p-values in all cases. When $b = 0.3$, parametric tests based on p-values have slightly more power (relative to power for $|z|$-values) for the first two choices of Π and slightly less power in the last case. The robust test based on p-values has much more power when $b = 0.3$.

C.4.2 Using a fixed subsample size for the test

In this subsection we try to make two approaches comparable by fixing the local subsample size for testing. For both types of tests define

$$b_k(x) = \arg\min_b \{ \sum_{i=1}^{n} 1\{X_i \in [x-b, x+b] = k\},$$

where X_i is observed $|z|$- or p-value and k is the number of observations we want to include for testing. For our simulations we set $k = 100$. As a result, we compare tests based on $|z|$-values on $[1.96(1-b_{100}(1.96)), 1.96(1+b_{100}(1.96))]$ and and tests based on p-values on $[0.05(1-b_{100}(0.05)), 0.05(1-b_{100}(0.05))]$, where the value of b_{100} depends on the realization of the sample.

Figure C.3 shows the power of the tests based on $|z|$-values and p-values. We can see, that tests based on p-values demonstrate slightly higher power in all cases and for both versions of the test than tests based on $|z|$-values.

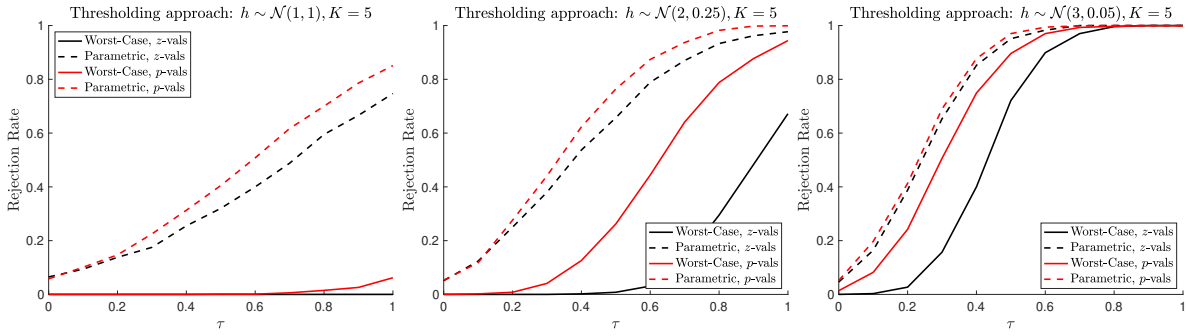

Figure C.3. Power curves, $k = 100$.

C.5 Null and Alternative Distributions MC Study

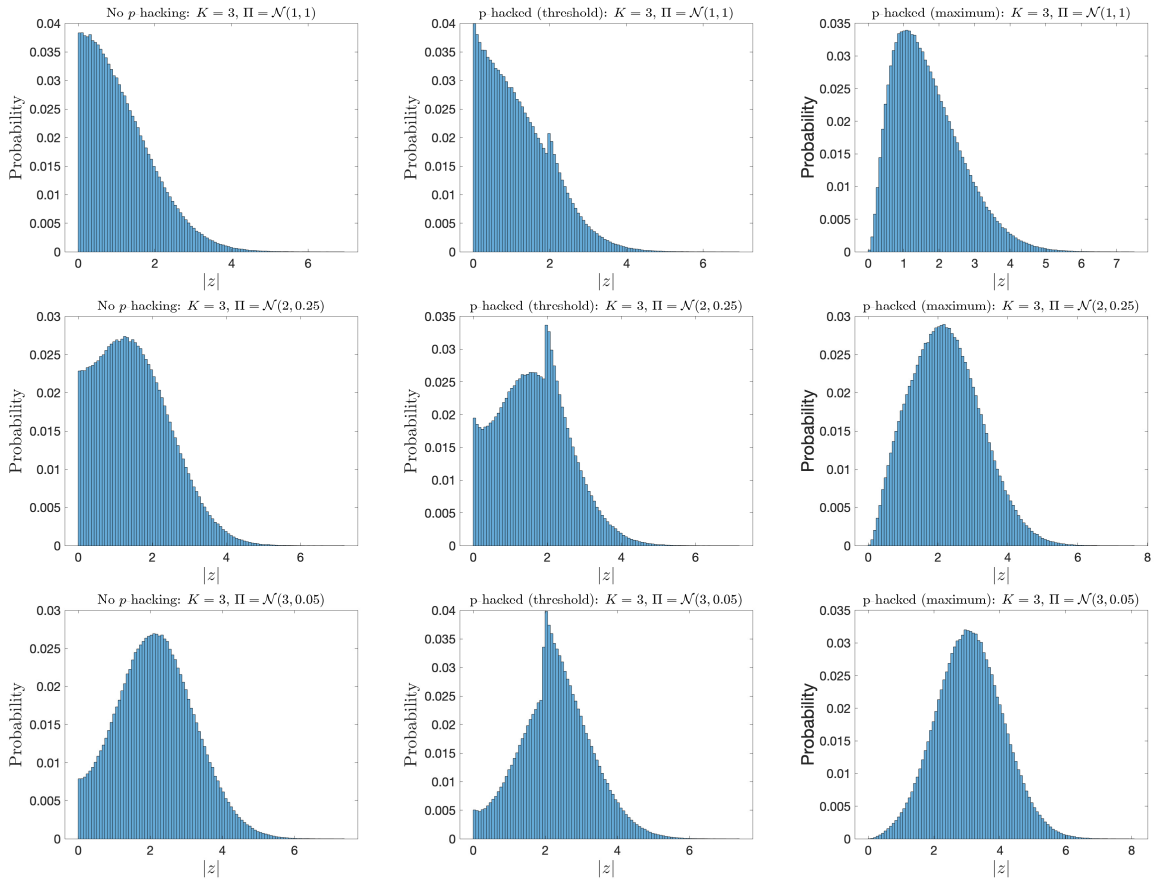

Figure C.4. Null and p-hacked distributions for $K = 3$

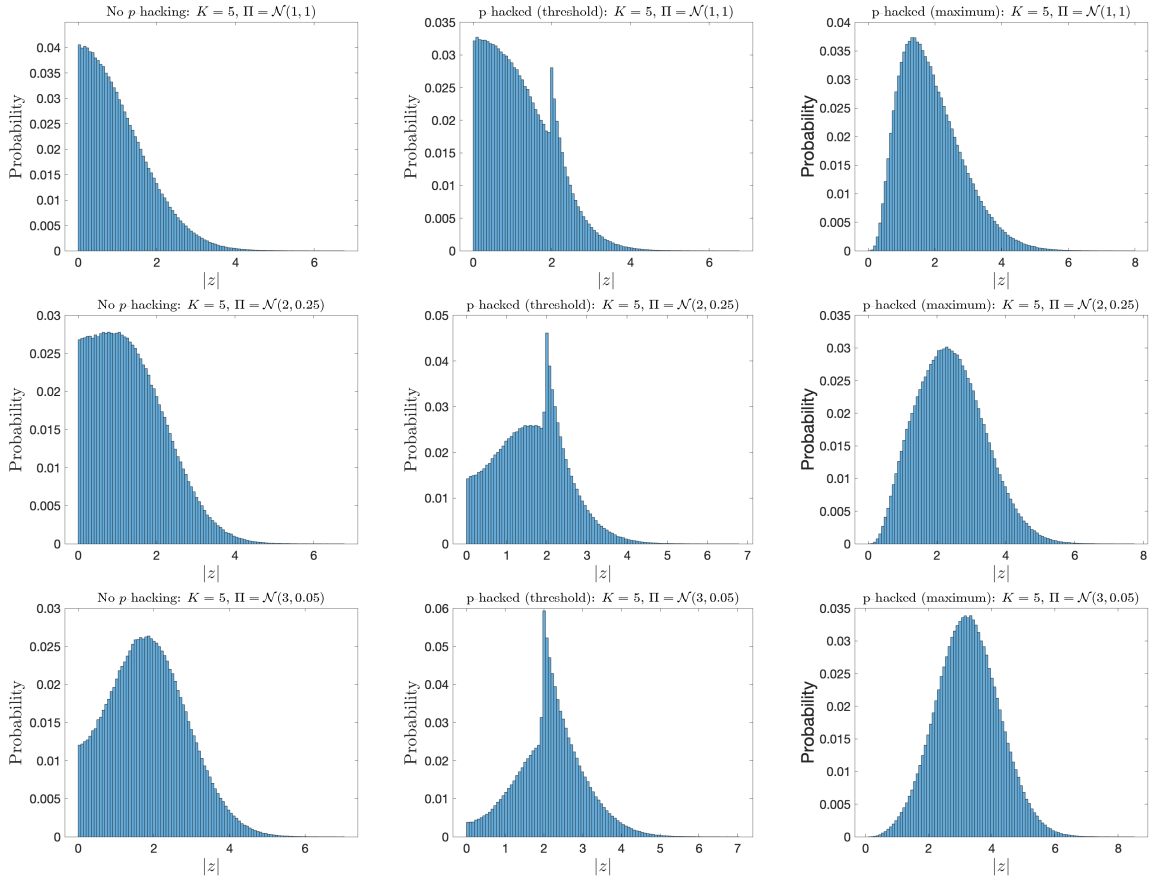

Figure C.5. Null and p-hacked distributions for $K = 5$

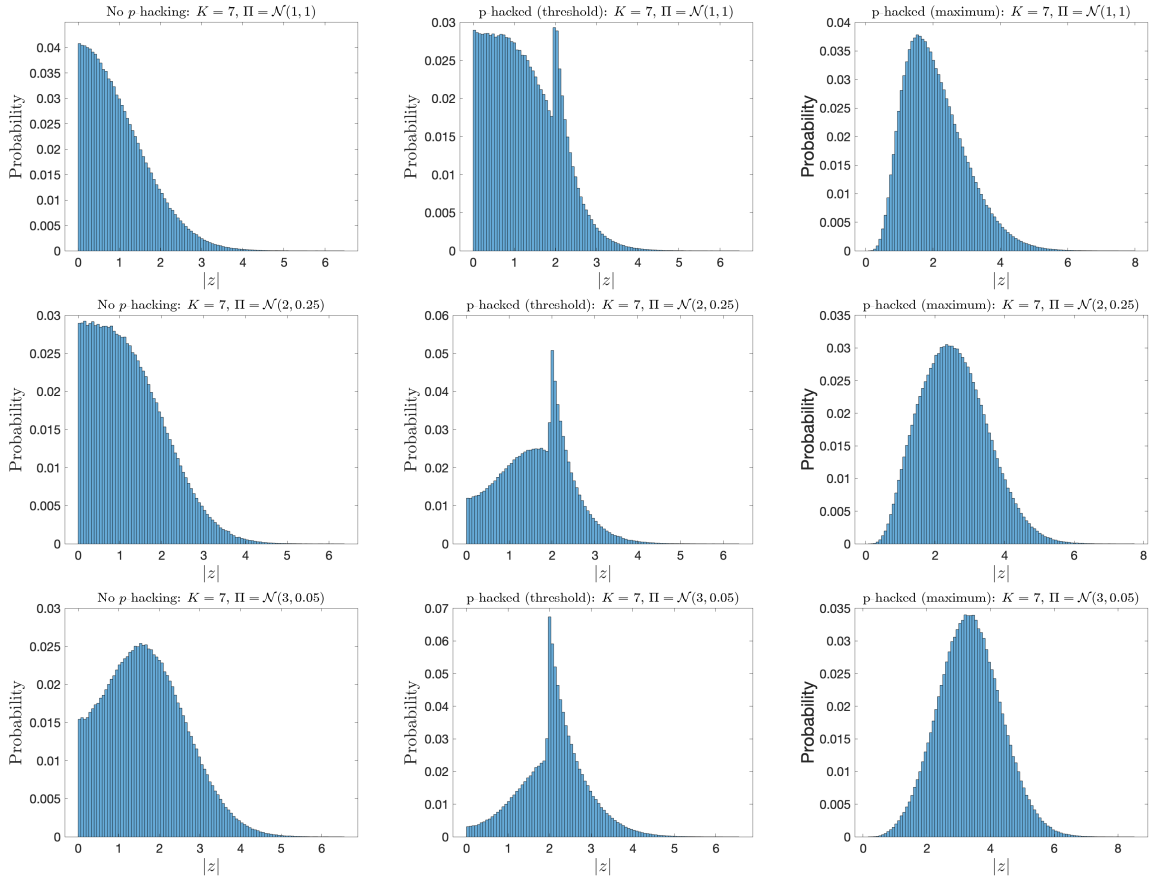

Figure C.6. Null and p-hacked distributions for $K = 7$

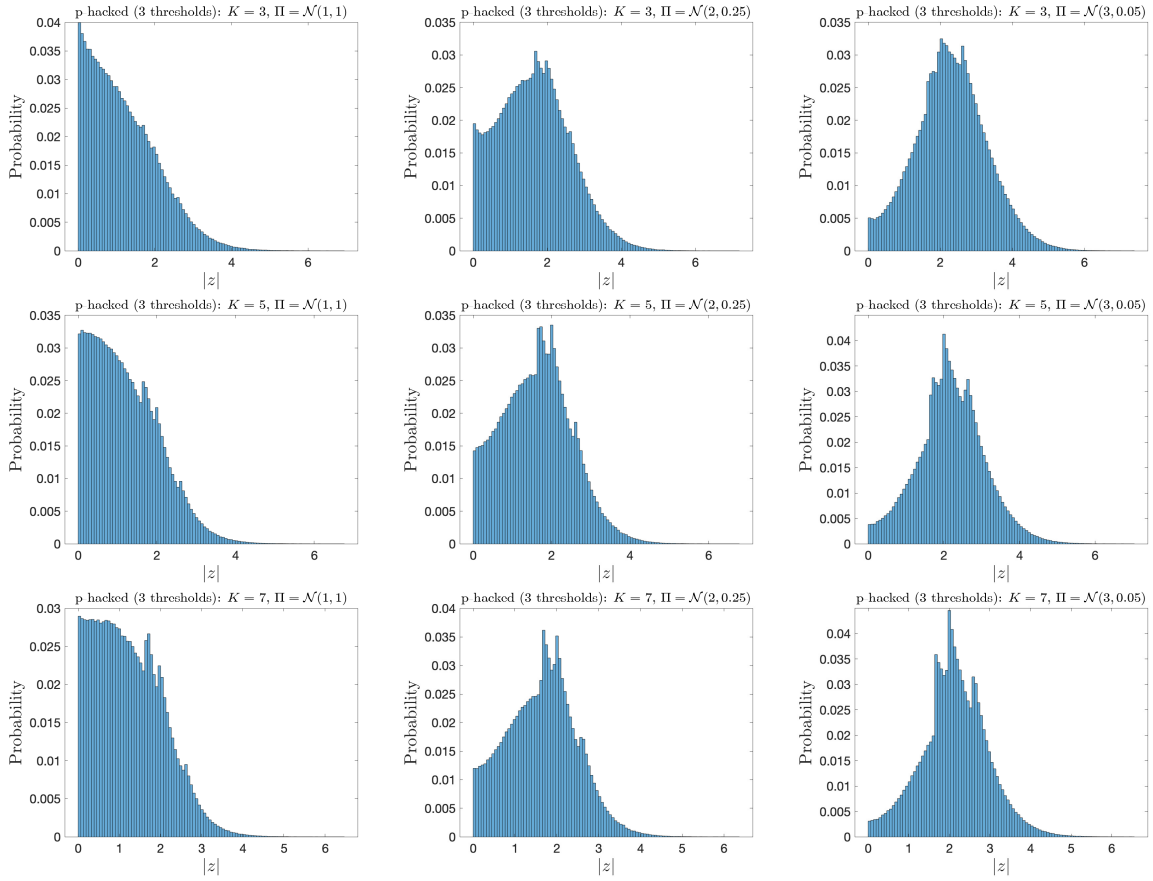

Figure C.7. *p*-hacked distributions: *p*-hacking at multiple thresholds

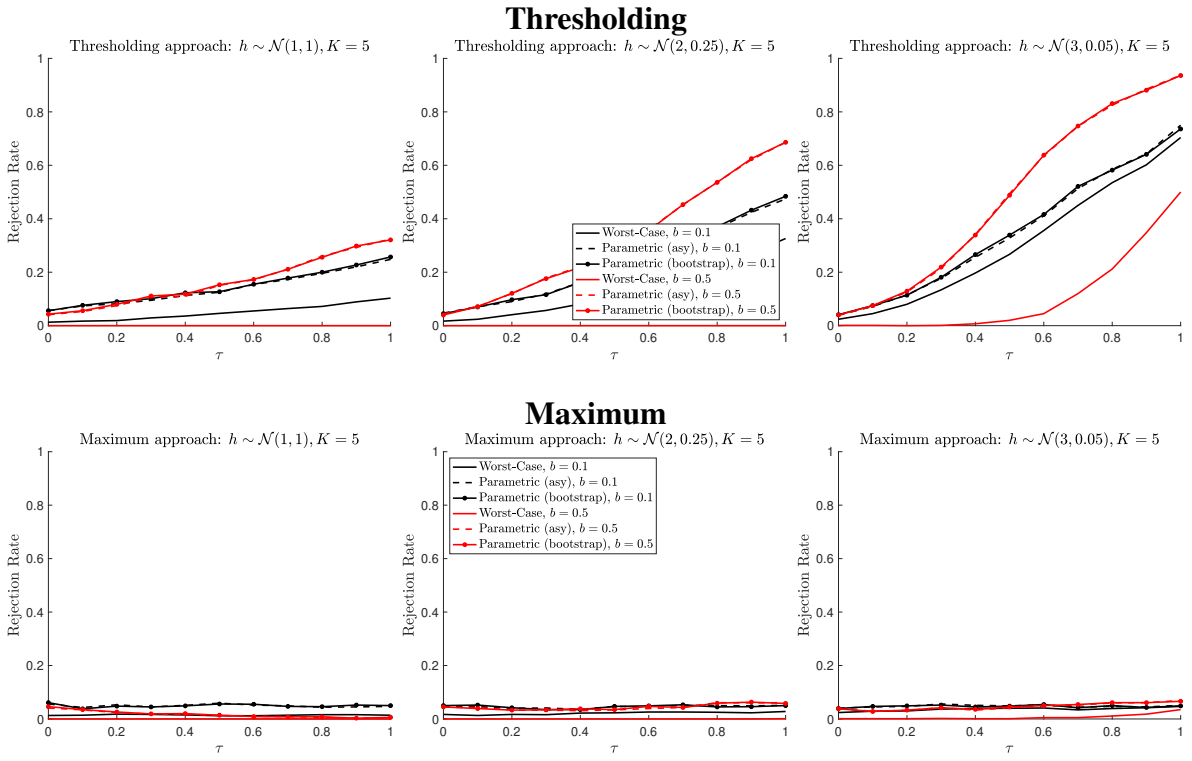

Figure C.8. Power curves covariate selection with $K = 3$. Sample size is 1000.

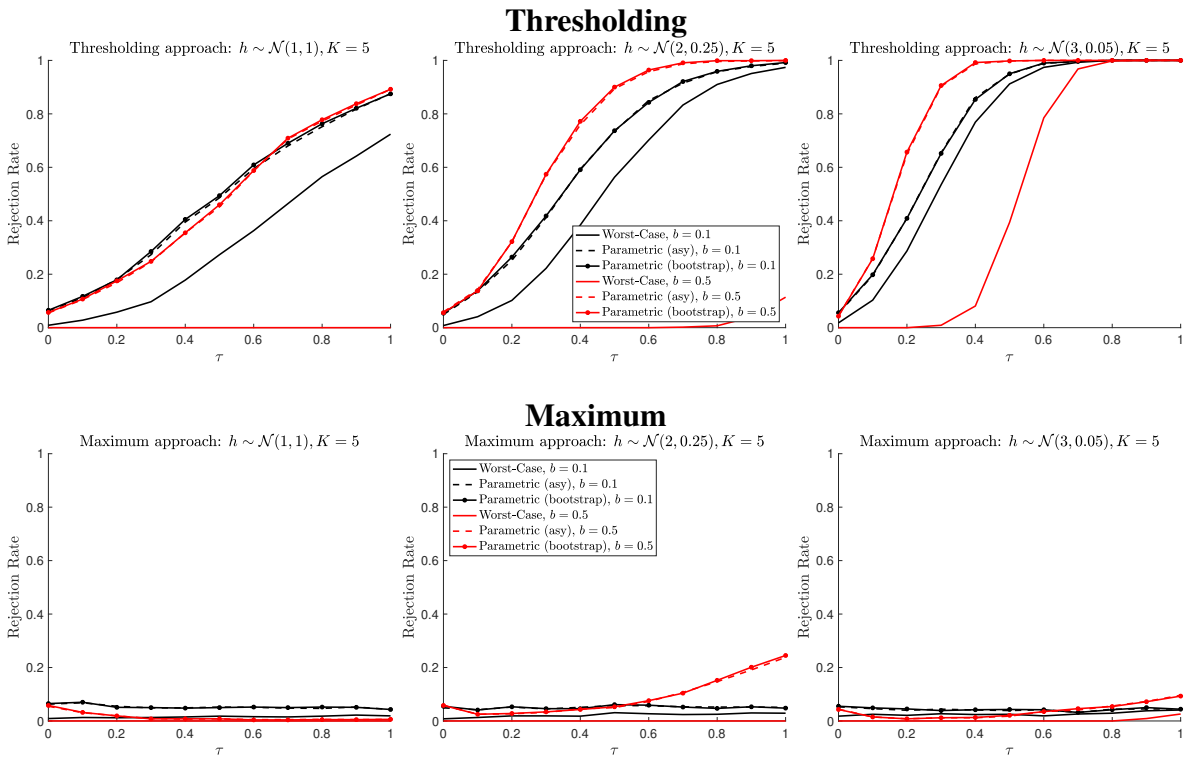

Figure C.9. Power curves covariate selection with $K = 7$. Sample size is 1000.

C.6 Application: Additional Results

Table C.1. Binomial and Robust Caliper Tests, 1% Significance threshold (p-values)

	DID	IV	RCT	RDD
Proportion Significant in 2.576 ± 0.5	0.4134	0.3807	0.3942	0.4111
Binomial Test	1	1	1	0.9999
Robust Caliper Test ($\underline{\tau}_{5\%}$)	1 (0)	1 (0)	1 (0)	1 (0)
# Tests in 2.576 ± 0.5	837	943	1111	467
Proportion Significant in 2.576 ± 0.4	0.3898	0.3962	0.3979	0.4093
Binomial Test	1	1	1	0.9997
Robust Caliper Test ($\underline{\tau}_{5\%}$)	1 (0)	1 (0)	1 (0)	1 (0)
# Tests in 2.576 ± 0.4	644	737	862	364
Proportion Significant in 2.576 ± 0.3	0.3835	0.404	0.4344	0.4126
Binomial Test	1	1	0.9996	0.9975
Robust Caliper Test ($\underline{\tau}_{5\%}$)	1 (0)	1 (0)	1 (0)	1 (0)
# Tests in 2.576 ± 0.3	472	552	663	269
Proportion Significant in 2.576 ± 0.2	0.3839	0.4329	0.4717	0.4032
Binomial Test	1	0.9941	0.8735	0.995
Robust Caliper Test ($\underline{\tau}_{5\%}$)	1 (0)	1 (0)	0.9999 (0)	0.9999 (0)
# Tests in 2.576 ± 0.2	310	365	441	186
Proportion Significant in 2.576 ± 0.1	0.3697	0.4759	0.5063	0.4343
Binomial Test	0.9995	0.7207	0.398	0.8862
Robust Caliper Test ($\underline{\tau}_{5\%}$)	0.9996 (0)	0.8557 (0)	0.8012 (0)	0.9872 (0)
# Tests in 2.576 ± 0.1	165	187	239	99

Table C.1. (cont.) Binomial and Robust Caliper Tests, 10% Significance threshold (p-values)

Proportion Significant in 2.576 ± 0.075	0.38	0.507	0.5548	0.4478
Binomial Test	0.9895	0.4007	0.0796	0.7681
Robust Caliper Test ($\underline{\tau}_{5\%}$)	0.9828 (0)	0.3556 (0)	0.3154 (0)	0.8983 (0)
# Tests in 2.576 ± 0.075	100	142	146	67
Proportion Significant in 2.576 ± 0.05	0.375	0.5049	0.5463	0.3696
Binomial Test	0.9778	0.4219	0.1449	0.9481
Robust Caliper Test ($\underline{\tau}_{5\%}$)	0.9911 (0)	0.2759 (0)	0.304 (0)	0.9674 (0)
# Tests in 2.576 ± 0.05	72	103	108	46
Total obs	5780	5158	7101	3117

Note: [$\underline{\tau}_{5\%}$,1] is the 95% confidence interval for the extent of p-hacking.

Table C.2. Binomial and Robust Caliper Tests, 10% Significance threshold (p-values)

	DID	IV	RCT	RDD
Proportion Significant in 1.645 ± 0.5	0.6026	0.6057	0.5006	0.5042
Binomial Test	0	0	0.4713	0.4025
Robust Caliper Test ($\underline{\tau}_{5\%}$)	0.9141 (0)	0.8749 (0)	1 (0)	1 (0)
# Tests in 1.645 ± 0.5	1014	1050	1738	591
Proportion Significant in 1.645 ± 0.4	0.6183	0.6009	0.5168	0.5198
Binomial Test	0	0	0.0998	0.1804
Robust Caliper Test ($\underline{\tau}_{5\%}$)	0.3548 (0)	0.5635 (0)	1 (0)	0.9983 (0)
# Tests in 1.645 ± 0.4	820	847	1401	479

Table C.2. (cont.) Binomial and Robust Caliper Tests, 10% Significance threshold (p-values)

Proportion Significant in 1.645 ± 0.3	0.5996	0.5712	0.5216	0.5357
Binomial Test	0	0.0002	0.0817	0.0785
Robust Caliper Test ($\underline{\tau}_{5\%}$)	0.4989 (0)	0.5941 (0)	0.9991 (0)	0.898 (0)
# Tests in 1.645 ± 0.3	557	611	997	364
Proportion Significant in 1.645 ± 0.2	0.5822	0.5851	0.5147	0.5397
Binomial Test	0.0005	0.0002	0.2107	0.0929
Robust Caliper Test ($\underline{\tau}_{5\%}$)	0.333 (0)	0.0787 (0)	0.9479 (0)	0.6137 (0)
# Tests in 1.645 ± 0.2	383	417	682	252
Proportion Significant in 1.645 ± 0.1	0.586	0.5637	0.5113	0.5902
Binomial Test	0.0077	0.0292	0.3246	0.0184
Robust Caliper Test ($\underline{\tau}_{5\%}$)	0.1706 (0)	0.0805 (0)	0.7014 (0)	0.1951 (0)
# Tests in 1.645 ± 0.1	186	204	309	122
Proportion Significant in 1.645 ± 0.075	0.5897	0.5584	0.4915	0.6061
Binomial Test	0.01	0.0627	0.5774	0.0133
Robust Caliper Test ($\underline{\tau}_{5\%}$)	0.1195 (0)	0.1477 (0)	0.794 (0)	0.1215 (0)
# Tests in 1.645 ± 0.075	156	154	236	99
Proportion Significant in 1.645 ± 0.05	0.5833	0.5905	0.4968	0.6061
Binomial Test	0.0335	0.0252	0.5	0.032
Robust Caliper Test ($\underline{\tau}_{5\%}$)	0.3898 (0)	0.0274 (0.05)	0.6515 (0)	0.0949 (0)
# Tests in 1.645 ± 0.05	108	105	155	66
Total obs	5780	5158	7101	3117

Note: [$\underline{\tau}_{5\%}$,1] is the 95% confidence interval for the extent of p-hacking.

Bibliography

Adda, J., Decker, C., and Ottaviani, M. (2020). P-hacking in clinical trials and how incentives shape the distribution of results across phases. *Proceedings of the National Academy of Sciences*, 117(24):13386–13392.

Andrews, I. and Kasy, M. (2018). Identification of and correction for publication bias. *forthcoming American Economic Review*.

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

Beare, B. K. (2021). Least favorability of the uniform distribution for tests of the concavity of a distribution function. *Stat*, page e376. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/sta4.376.

Beare, B. K. and Moon, J.-M. (2015). Nonparametric tests of density ratio ordering. *Econometric Theory*, 31(3):471–492.

Brodeur, A., Carrell, S., Figlio, D., and Lusher, L. (2021). Unpacking p-hacking and publication bias. Technical report, Technical Report, Tech. rep.

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

Brodeur, A., Cook, N., and Heyes, A. (2020b). Methods matter: P-hacking and publication bias in causal analysis in economics. forthcoming American Economic Review.

Brodeur, A., Cook, N., and Heyes, A. (2022a). We need to talk about mechanical turk: What 22,989 hypothesis tests tell us about p-hacking and publication bias in online experiments. Technical report, I4R Discussion Paper Series.

Brodeur, A., Cook, N., and Neisser, C. (2022b). P-hacking, data type and data-sharing policy.

Brodeur, A., Lé, M., Sangnier, M., and Zylberberg, Y. (2016a). Replication data for: Star wars: The empirics strike back. Nashville, TN: American Economic Association [publisher], 2016.

Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2019-10-12.

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

Bruns, S. B. (2017). Meta-regression models and observational research. *Oxford Bulletin of Economics and Statistics*, 79(5):637–653.

Bruns, S. B., Asanov, I., Bode, R., Dunger, M., Funk, C., Hassan, S. M., Hauschildt, J., Heinisch, D., Kempa, K., König, J., Lips, J., Verbeck, M., Wolfschütz, E., and Buenstorf, G. (2019). Reporting errors and biases in published empirical findings: Evidence from innovation research. *Research Policy*, 48(9):103796.

Bruns, S. B. and Ioannidis, J. P. A. (2016). p-curve and p-hacking in observational research. *PLOS ONE*, 11(2):1–13.

Carolan, C. A. and Tebbs, J. M. (2005). Nonparametric tests for and against likelihood ratio ordering in the two-sample problem. *Biometrika*, 92(1):159–171.

Cattaneo, M. D., Jansson, M., and Ma, X. (2020). Simple local polynomial density estimators. *Journal of the American Statistical Association*, 115(531):1449–1455.

Cattaneo, M. D., Jansson, M., and Ma, X. (2021). *rddensity: Manipulation Testing Based on Density Discontinuity*. R package version 2.2.

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

Cox, G. and Shi, X. (2022). Simple Adaptive Size-Exact Testing for Full-Vector and Subvector Inference in Moment Inequality Models. *The Review of Economic Studies*, 90(1):201–228.

de Winter, J. C. and Dodou, D. (2015). A surge of *p*-values between 0.041 and 0.049 in recent decades (but negative results are increasing rapidly too). *PeerJ*, 3:e733.

Decker, C. and Ottaviani, M. (2023). Preregistration and credibility of clinical trials. *medRxiv*, pages 2023–05.

Delaigle, A. (2021). Deconvolution kernel density estimation. In *Handbook of Measurement Error Models*, pages 185–220. Chapman and Hall/CRC.

Elliott, G., Kudrin, N., and Wüthrich, K. (2020). Detecting p-hacking. arXiv:1906.06711v3.

Elliott, G., Kudrin, N., and Wüthrich, K. (2022a). (when) can we detect *p*-hacking? *arXiv*

preprint arXiv:2205.07950.

Elliott, G., Kudrin, N., and Wüthrich, K. (2022b). Detecting p-hacking. *Econometrica*, 90(2):887–906.

Fang, Z. (2019). Refinements of the Kiefer-Wolfowitz theorem and a test of concavity. *Electron. J. Statist.*, 13(2):4596–4645.

Gerber, A. and Malhotra, N. (2008a). Do statistical reporting standards affect what is published? publication bias in two leading political science journals. *Quarterly Journal of Political Science*, 3(3):313–326.

Gerber, A. S. and Malhotra, N. (2008b). Publication bias in empirical sociological research: Do arbitrary significance levels distort published results? *Sociological Methods & Research*, 37(1):3–30.

Gleser, L. J. and Olkin, I. (1996). Models for estimating the number of unpublished studies. *Statistics in medicine*, 15(23):2493–2507.

Havranek, T., Kolcunova, D., and Bajzik, J. (2021). When does monetary policy sway house prices? a meta-analysis.

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

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., and Jennions, M. D. (2016). Data from: The extent and consequences of p-hacking in science. Dryad, Dataset.

Hendry, D. F. (1980). Econometrics-alchemy or science? *Economica*, 47(188):387–406.

Hung, H. M. J., O'Neill, R. T., Bauer, P., and Kohne, K. (1997). The behavior of the p-value when the alternative hypothesis is true. *Biometrics*, 53(1):11–22.

Imbens, G. W. (2021). Statistical significance, p-values, and the reporting of uncertainty. *Journal of Economic Perspectives*, 35(3):157–74.

Karunamuni, R. and Alberts, T. (2005). On boundary correction in kernel density estimation. *Statistical Methodology*, 2(3):191 – 212.

Kinal, T. W. (1980). The existence of moments of k-class estimators. *Econometrica: Journal of the Econometric Society*, pages 241–249.

Kudo, A. (1963). A multivariate analogue of the one-sided test. *Biometrika*, 50(3/4):403–418.

Kudrin, N. (2022). Robust caliper tests. Working Paper URL: https://drive.google.com/file/d/1OLNh06fVi2kfffMgMtbm5_ilNisUHzNF/view.

Kulikov, V. N. and Lopuhaä, H. P. (2008). Distribution of global measures of deviation between the empirical distribution function and its concave majorant. *Journal of Theoretical Probability*, 21(2):356–377.

Lakens, D. (2015). What p-hacking really looks like: A comment on Masicampo and LaLande (2012). *The Quarterly Journal of Experimental Psychology*, 68(4):829–832. PMID: 25484109.

Leamer, E. E. (1983). Let's take the con out of econometrics. *The American Economic Review*, 73(1):31–43.

Leggett, N. C., Thomas, N. A., Loetscher, T., and Nicholls, M. E. R. (2013). The life of p: "just significant" results are on the rise. *The Quarterly Journal of Experimental Psychology*, 66(12):2303–2309.

Malovaná, S., Hodula, M., Gric, Z., and Bajzík, J. (2022). Borrower-based macroprudential measures and credit growth: How biased is the existing literature?

Masicampo, E. J. and Lalande, D. R. (2012). A peculiar prevalence of p values just below .05. *The Quarterly Journal of Experimental Psychology*, 65(11):2271–2279.

MATLAB (2020). *version 9.9.0 (R2020b)*. The MathWorks Inc., Natick, Massachusetts.

McCloskey, A. and Michaillat, P. (2022). Incentive-compatible critical values. Working Paper 29702, National Bureau of Economic Research.

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of econometrics*, 142(2):698–714.

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

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Romano, J. P. (2004). On non-parametric testing, the uniform behaviour of the t-test, and related problems. *Scandinavian Journal of Statistics*, 31(4):567–584.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological*

bulletin, 86(3):638.

Schmid, C. H., Stijnen, T., and White, I. (2020). *Handbook of meta-analysis*. CRC Press.

Simonsohn, U. (2020). [91] p-hacking fast and slow: Evaluating a forthcoming aer paper deeming some econ literatures less trustworthy. Data colada: http://datacolada.org/91 (last accessed: August 29, 2022).

Simonsohn, U., Nelson, L. D., and Simmons, J. P. (2014). P-curve: a key to the file-drawer. *Journal of Experimental Psychology: General*, 143(2):534–547.

Simonsohn, U., Simmons, J. P., and Nelson, L. D. (2015). Better p-curves: Making p-curve analysis more robust to errors, fraud, and ambitious p-hacking, a reply to Ulrich and Miller (2015). *Journal of Experimental Psychology: General*, 144(6):1146–1152.

Snyder, C. and Zhuo, R. (2018). Sniff tests in economics: Aggregate distribution of their probability values and implications for publication bias. NBER WP 25058.

StataCorp. (2019). *Stata Statistical Software: Release 16.* College Station, TX.

Ulrich, R. and Miller, J. (2015). p-hacking by post hoc selection with multiple opportunities: Detectability by skewness test?: Comment on Simonsohn, Nelson, and Simmons (2014). *Journal of Experimental Psychology: General*, 144:1137–1145.

Ulrich, R. and Miller, J. (2018). Some properties of p-curves, with an application to gradual publication bias. *Psychological Methods*, 23(3):546–560.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.

Vivalt, E. (2019). Specification searching and significance inflation across time, methods and disciplines. *Oxford Bulletin of Economics and Statistics*, 81(4):797–816.

Wasserstein, R. L. and Lazar, N. A. (2016). The ASA statement on p-values: Context, process, and purpose. *The American Statistician*, 70(2):129–133.

Wolak, F. A. (1987). An exact test for multiple inequality and equality constraints in the linear regression model. *Journal of the American Statistical Association*, 82(399):782–793.

Yang, F., Havranek, T., Irsova, Z., and Novak, J. (2022). Hedge fund performance: A quantitative survey. *Available at SSRN 4151821*.