(When) Can We Detect $p$-Hacking?*

Graham Elliott† Nikolay Kudrin‡ Kaspar Wüthrich§

May 18, 2022

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. 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.

Keywords: $p$-hacking, publication bias, $p$-curve, specification search, selecting instruments, variance bandwidth selection

---

*Some parts of this paper are based on material in the working paper Elliott et al. (2020), which was not included in the final published version (Elliott et al., 2022). K.W. is also affiliated with CESifo and the ifo Institute. The usual disclaimer applies.

†Department of Economics, University of California, San Diego, 9500 Gilman Dr. La Jolla, CA 92093, email: grelliott@ucsd.edu

‡Department of Economics, University of California, San Diego, 9500 Gilman Dr. La Jolla, CA 92093, email: nkudrin@ucsd.edu

§Department of Economics, University of California, San Diego, 9500 Gilman Dr. La Jolla, CA 92093, email: kwuthrich@ucsd.edu
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. (2014a) 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.\footnote{See, for example, Masicampo and Lalande (2012); Simonsohn et al. (2014b); Lakens (2015); Simonsohn et al. (2015); Head et al. (2015); Ulrich and Miller (2015) for early applications and further discussions and Christensen and Miguel (2018) for a review.} To be able to understand how useful tests are in detecting $p$-hacking, 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.\footnote{We focus on detecting $p$-hacking based on the distribution of $p$-values across studies and do not consider tests based on the distribution of $t$-statistics and $z$-scores (e.g., Gerber and Malhotra, 2008a,b; Brodeur et al., 2016, 2020; Bruns et al., 2019; Vivalt, 2019; Adda et al., 2020).}

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. 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 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 conclusions 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 relate power to the costs of $p$-hacking.
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. 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. We then examine in Monte Carlo analyses 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

---

3 Here our focus is on understanding whether we can detect p-hacking, and we do not explicitly model selective publication (e.g., Andrews and Kasy, 2019; Kasy, 2021; Frankel and Kasy, 2022). Publication bias interacting with p-hacking is likely to increase the rejection probability of tests.

4 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.
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.

2 Setup and Testable Restrictions

Elliott et al. (2022) provided a general characterization of the distribution of p-values across studies in the absence of p-hacking for general distributions of true effects. 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 $h \in \mathcal{H}$ indexes parameters of either the exact or asymptotic distribution of the test. Researchers are testing the null hypothesis that $h \in \mathcal{H}_0$ against $h \in \mathcal{H}_1$ with $\mathcal{H}_0 \cap \mathcal{H}_1$ 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 $\Pi$ of effects $h$, which is to say that different researchers testing different hypotheses examine different problems that have different “true” effects. The resulting CDF of p-values across all these studies is then

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

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

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

\footnote{See, for example, Hung et al. (1997), Simonsohn et al. (2014a), and Ulrich and Miller (2018) for numerical and analytical examples of p-curves for specific tests and/or effect distributions.}
Our goal is to evaluate the power of statistical tests based on the different testable implications derived in the literature. Elliott et al. (2022) 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. These conditions hold for many possible distributions $F_h$ that arise in research, for example, normal, folded normal (relevant for two-sided tests), and $\chi^2$ distributions.

When $T$ is normally distributed (for example, tests on means or regression parameters when central limit theorems apply), Elliott et al. (2022) 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.

If researchers do $p$-hack, this affects the distribution of reported $t$-statistics. 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. It may be that humps appear below popular significance levels such as 0.05 violating the monotonicity condition, discontinuities may appear in the $p$-curve, and in the case of tests based on $t$-statistics, the bounds derived in Elliott et al. (2022) may be violated.

Researchers have a number of options in modeling that allow for the possibility of $p$-hacking. To understand the likely violation under the alternative that there is $p$-hacking, and hence the relevant test to use for testing for $p$-hacking, it is useful to consider how different approaches to $p$-hacking impact the $p$-curve. This is also useful in understanding the likely power of tests for $p$-hacking.

We consider two basic approaches to $p$-hacking. The first of these we refer to as 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, 5%). 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.

Our second approach to $p$-hacking is to simply take the smallest $p$-value from a set of correctly specified or overspecified models. We refer to this below as the “minimum” approach. Intuitively, we would expect that for this approach the distribution of $p$-values

\footnote{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., 2014a; Head et al., 2015), are valid in a wide range of empirically relevant settings.}
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 3.3.

The actual form and properties of the *p*-curve differ from situation to situation. We examine four examples in more depth in Section 3.

### 3 Implications of *p*-Hacking

Here we study the shape of the distribution of *p*-values under different forms of *p*-hacking. Appendix A presents the analytical derivations underlying the results in this section.

#### 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.

##### 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 of interest $Y_i$. The data are generated as

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

where $U_i \sim N(0, 1)$. For simplicity, we assume that $X_i$ is non-stochastic. The researchers test the following hypothesis

$$H_0 : \beta = 0 \quad \text{against} \quad H_1 : \beta > 0.$$
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 $N^{-1} \sum_{i=1}^{N} X_i = N^{-1} \sum_{i=1}^{N} Z_{1i}^2 = N^{-1} \sum_{i=1}^{N} Z_{2i}^2 = 1$, that $N^{-1} \sum_{i=1}^{N} Z_{1i}Z_{2i} = \gamma^2$, and that $N^{-1} \sum_{i=1}^{N} Z_{1i} = N^{-1} \sum_{i=1}^{N} X_iZ_{2i} = \gamma$, where $|\gamma| \in (0,1).$ $^8$ Let $h := \sqrt{N}\beta\sqrt{1-\gamma^2}$, where $h$ is drawn from a distribution of effects with support $\mathcal{H} \subseteq [0,\infty)$.

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 $\hat{\beta}_r^1$ to be the OLS estimate from the regression that accords with the chosen $p$-value. The minimum approach takes the minimum of the two $p$-values that can be constructed. 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 A.1, we show that for the thresholding approach the resulting $p$-curve is

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

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

$$\Upsilon_1^t(p; \alpha, h, \rho) = \begin{cases} 1 + \Phi \left( \frac{z_h(p) - \rho z_0(p)}{\sqrt{1-\rho^2}} \right), & \text{if } p \leq \alpha, \\ 2\Phi \left( \frac{z_h(p) \sqrt{1-\rho}}{1+\rho} \right), & \text{if } p > \alpha. \end{cases}$$

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 $\Phi$ 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 $\Phi \left( \frac{z_h(p) - \rho z_0(p)}{\sqrt{1-\rho^2}} \right)$ is decreasing in $h$ that

---

$^7$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.
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 just 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

$$g_1^m(p) = 2 \int_{\mathcal{H}} \exp \left( h z_0(p) - \frac{h^2}{2} \right) \Phi \left( z_h(p) \sqrt{\frac{1-\rho}{1+\rho}} \right) d\Pi(h).$$

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 $\Pi$, so does not have a hump and remains continuous. Because of this, only the tests based on upper bounds have any potential for testing the null hypothesis of no $p$-hacking. If $\Pi$ 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. (2022), the upper end (largest $p$) of which is at $p = 1 - \Phi(h)$.

Figure 1 shows the theoretical $p$-curves for various $h$ and $\gamma$. 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

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

where $\phi$ is the standard normal probability density function (PDF). Note that $\rho$ is always positive and, when all nulls are true (i.e., when $\Pi$ assigns probability one to $h = 0$), $g_1'(p)$ is positive for all $p \in (0, \alpha)$. This can be seen for the dashed line in Figure 1 (left panel). However at $h = 1$ this effect no longer holds, and the $p$-curve is downward sloping. From Figure 1 (right panel) we see that violations for monotonicity are larger for smaller $\gamma$.

Figure 1 indicates that the thresholding approach to $p$-hacking does imply a discontinuity at $p$-values equal to size. Here the size of the discontinuity is larger for larger $h$ and remains for each $\gamma$, 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 5.

---

9For $p > \alpha$, the derivative of $g_1'(p)$ is negative and equal to

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

8
Figure 1: *p*-Curves from covariate selection with thresholding. Left panel: $\gamma = 0.5$. Right panel: $h = 1$.

Figure 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 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: *p*-Curves from covariate selection for threshold and minimum approaches ($h = 1$, $\gamma = 0.5$).
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 \( \rho \). Figure 3 shows the difference between empirical and nominal size. The left panel shows, for nominal size \( \alpha = 0.05 \), how empirical size varies with \( \gamma \). For small \( \gamma \) the tests are highly correlated (\( \rho \) is close to one), leaving little room for effective \( p \)-hacking, and hence there is only a small effect on size. As \( \gamma \) 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 \( \gamma \) (\( \gamma \in \{0.1, 0.5, 0.9\} \)), how the empirical size exceeds nominal size. The lower line is nominal size; the empirical size is larger for each value of \( \gamma \). Essentially, the result is somewhat uniform over this empirical size range, with size coming close to double empirical size for the largest value of \( \gamma \).

![Figure 3: Rejection rate under \( p \)-hacking. Left panel: rejection rate as a function of \( \gamma \) for \( \alpha = 0.05 \). Right panel: rejection rate as a function of \( \alpha \) for \( \gamma \in \{0.1, 0.5, 0.9\} \).](image)

Selectively choosing larger \( t \)-statistics results in selectively choosing larger estimated ef-
fects. The bias for the threshold case is given by

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

and for the minimum approach it is given by

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

The bias as a function of $h$ can be seen in Figure 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 $\gamma$ means a smaller $\rho$, 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 4: Bias from covariate selection for $\gamma \in \{0.1, 0.5, 0.9\}$](image)

### 3.2 Selecting amongst Instruments in IV Regression

#### 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 of interest $Y_i$. 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 following hypothesis

\[
H_0 : \beta = 0 \quad \text{against} \quad H_1 : \beta > 0.
\]

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 \( h := \sqrt{N}|\gamma| \), where \( h \) is drawn from an effect distribution supported on \( \mathcal{H} \subseteq [0, \infty) \).

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 \( \hat{\beta}_t^r \) and \( \hat{\beta}_m^r \), accordingly.

For the threshold approach, the \( p \)-curve (see Appendix A.2 for derivations) is

\[
g_2^t(p) = \int_{\mathcal{H}} \exp \left( h z_0(p) - \frac{h^2}{2} \right) \Upsilon_2^t(p; \alpha, h) d\Pi(h),
\]

where

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

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

$$g_2^m(p) = \int_\mathcal{H} \exp \left( h z_0(p) - \frac{h^2}{2} \right) \Upsilon_2^m(p; \alpha, h) d\Pi(h),$$

where

$$\Upsilon_2^m(p; \alpha, h) = \begin{cases} \phi(z_{z_0(p)}) \zeta(p) + 2\Phi(D_h(p) - z_h(p)), & \text{if } 0 < p \leq 1/2, \\ 2\Phi(z_h(p)), & \text{if } 1/2 < p < 1. \end{cases}$$

The right hand side panel of Figure 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 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 6: $p$-Curves from IV selection for threshold and minimum approaches ($h = 1$).

Figure 6 shows the comparable figure for the IV problem as Figure 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.

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 $\alpha$ 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 7. Size is essentially double the stated size, with the $p$-hacked size at 11% when nominal size is 5%.
Figure 7: Rejection rate under \( p \)-hacking. Left panel: rejection rate as a function of \( \alpha \). Right panel: bias from \( p \)-hacking for different values of \( h \).

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

\[
B^t_2 = B^m_2 - |\gamma|^{-1} \left( \frac{\sqrt{2} - 1}{\sqrt{2}} \right) \phi \left( \frac{h}{\sqrt{2} - \sqrt{2}} - \sqrt{4 - 2\sqrt{2}z_0(\alpha)} \right),
\]

\[
B^m_2 = |\gamma|^{-1} \phi \left( \sqrt{\frac{\sqrt{2} - 1}{\sqrt{2}} h} \right) \Phi \left( \frac{h}{\sqrt{2} - \sqrt{2}} \right) + |\gamma|^{-1} \sqrt{2}\phi(0)(1 - \Phi(\sqrt{2}h)).
\]

The right-hand panel in Figure 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

---

\(^{10}\)Since the first moments of the IV estimators do not exist in just-identified cases (Kinal, 1980), we define \( B^j_2 \) to be the mean of the asymptotic distribution of \( \sqrt{N}(\hat{\beta}^j_r - \beta) \), where \( \hat{\beta}^j_r \) is the \( p \)-hacked estimate and \( j \in \{m, t\} \).
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.

### 3.3 Selecting Across Datasets

Consider a setting where a researcher runs a finite number of \( K > 1 \) independent analyses over which they can choose the best results. In each case the researcher uses a \( t \)-test to test their hypothesis, with \( T_i \sim \mathcal{N}(h, 1) \). 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 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

\[
g^*_3(p) = \int_{\mathcal{H}} \exp \left( h z_0(p) - \frac{h^2}{2} \right) \mathcal{Y}_3^t(p; \alpha, h) d\Pi(h),
\]

where

\[
\mathcal{Y}_3^t(p; \alpha, h) = \begin{cases} 
1 + \Phi \left( z_h(\alpha) \right), & \text{if } p \leq \alpha, \\
2\Phi \left( z_h(p) \right), & \text{if } p > \alpha.
\end{cases}
\]

This is a special case of the results in Section 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 \( \rho \) equal to that correlation and using the results in Section 3.1 would yield the correct distribution of \( p \)-values.

Figure 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 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. For general $K$, the $p$-curve is

$$g_3^m(p; K) = K \int_{\mathcal{H}} \exp \left( h z_0(p) - \frac{h^2}{2} \right) \Phi(z_h(p))^{K-1} d\Pi(h), \quad (1)$$

Equation (1) generalizes the analysis in Ulrich and Miller (2015), who considered the case where all nulls are true, that is $h = 0$. See also Elliott et al. (2022).

Figure 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., 2022). This can be seen in Figure 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 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. (2022).

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.

### 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 \( U_t \stackrel{\text{iid}}{\sim} \mathcal{N}(0, 1) \). 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 \( \kappa \) is the weight in the spectral density estimator.

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.\(^{11}\) In the second approach, the smallest \( p \)-value of the two computed is reported.

In Appendix A.4, we show that the distribution of \( p \)-values has the form

\[
g_4^1(p) = \int_{h} \exp \left( h z_0(p) - \frac{h^2}{2} \right) \Upsilon_4^1(p; \alpha, h, \kappa) d\Pi(h),
\]

\(^{11}\)If \( \hat{\rho} \) is such that \( \hat{\omega}^2 \) is negative, the researcher always reports the initial result.
with \( \Upsilon_4(p; \alpha, h, \kappa) \) taking different forms over different parts of the support of the distribution. Define \( l(p) = (2\kappa)^{-1} \left( \frac{z_0(\alpha)}{z_0(p)} \right)^2 - 1 \) and let \( H_N \) and \( \eta_N \) be the CDF and PDF of \( \hat{\rho} \), respectively. Then we have

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

Figure 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 10 presents the \( p \)-curves for the thresholding case. 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

\[
g^m_4(p) = \int_{\mathcal{H}} \exp \left( hz_0(p) - \frac{h^2}{2} \right) \Upsilon^m_4(p; \alpha, h, \kappa) d\Pi(h),
\]

where

\[
\Upsilon^m_4 = \begin{cases} 
1 - H_N(0) + H_N(-(2\kappa)^{-1}) + \frac{1}{\phi(z_h(p))} \int_{-(2\kappa)^{-1}}^{0} \omega(r) \phi(z_0(p)\omega(r) - h)\eta_N(r)dr, & \text{if } 0 < p \leq 1/2, \\
H_N(0) + \frac{1}{\phi(z_h(p))} \int_{0}^{\infty} \omega(r) \phi(z_0(p)\omega(r) - h)\eta_N(r)dr, & \text{if } 1/2 < p < 1.
\end{cases}
\]

The right-hand side panel in Figure 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 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
\[
G_0(\alpha) = \alpha + (1 - \alpha)(H_N(0) - H_N(-(2\kappa)^{-1})) - \int_{-(2\kappa)^{-1}}^{0} \Phi(z_0(\alpha)\omega(r) - h)\eta_N(r)dr
\]

Figure 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 11: Rejection rate under $p$-hacking from lag length selection with $N = 200$ and $\kappa = 1/2$](image)

Figure 11: Rejection rate under $p$-hacking from lag length selection with $N = 200$ and $\kappa = 1/2$

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 discuss in detail the histogram-based tests,
which are the most powerful and flexible in the simulations, and then provide an overview of alternative tests.

4.1 Histogram-based Tests

Histogram-based tests (Elliott et al., 2022) 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

\[
\pi_j = \int_{x_{j-1}}^{x_j} g(p) \, dp, \quad j = 1, \ldots, J.
\]

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. (2022) show that \(K\)-monotonicity\(^{12}\) restrictions and upper bounds on the \(p\)-curve and its derivatives can be expressed as

\[
H_0: \ A\pi_{-J} \leq b, \quad (2)
\]

for a matrix \(A\) and vector \(b\), where \(\pi_{-J} := (\pi_1, \ldots, \pi_{J-1})'\).\(^{13}\)

To test hypothesis (2), we estimate \(\pi_{-J}\) using the vector of sample proportions \(\hat{\pi}_{-J}\). The estimator \(\hat{\pi}_{-J}\) is asymptotically normal with mean \(\pi_{-J}\), such that the problem of testing (2) becomes 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. (2022), we use the conditional chi-squared test of Cox and Shi (2022), which is easy to implement and remains computationally tractable when the number of bins, \(J\), is moderate or large.

4.2 Other Tests

In addition to the histogram-based tests, we also investigate the performance of several other tests proposed in the literature. These tests can be categorized based on the testable implications they consider.

\(^{12}\)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 \(k\)th derivative of \(g\). By definition, a completely monotone function is \(K\)-monotone for all \(K\).

\(^{13}\)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 \((\pi_1, \ldots, \pi_J)\).
4.2.1 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., 2014a; 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., 2014a).\footnote{An alternative to Fisher’s test is Stouffer’s method (Simonsohn et al., 2015).}

In addition to the classical Binomial test and Fisher’s test, we consider tests based on the least concave majorant (LCM) (Elliott et al., 2022).\footnote{LCM tests have been successfully applied in many different contexts (e.g., Carolan and Tebbs, 2005; Beare and Moon, 2015; Fang, 2019)} 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 $\mathcal{M}\hat{G}$, where $\mathcal{M}$ is the LCM operator. We choose the test statistic $\|\hat{G} - \mathcal{M}\hat{G}\|_\infty$. The uniform distribution is least favorable for the LCM test (Kulikov and Lopuhaä, 2008; Beare, 2021), and critical values can be obtained via simulations.

4.2.2 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. (2022), we use the approach by Cattaneo et al. (2020) with automatic bandwidth selection implemented in the R-package \texttt{rddensity} (Cattaneo et al., 2021).

5 Monte Carlo Simulations

In this section, we investigate the finite sample properties of the tests in Section 4 using a Monte Carlo simulation study. The Monte Carlo study is based on generalizations of the analytical examples of $p$-hacking in Section 3. We do not consider selection across datasets, as this example can be viewed as a special case of covariate selection and IV selection.
5.1 Generalized \( p \)-Hacking Examples

In all examples that we consider, researchers are interested in testing a hypothesis about a scalar parameter \( \beta \):

\[
H_0 : \beta = 0 \quad \text{against} \quad H_1 : \beta > 0. \tag{3}
\]

Researchers may \( p \)-hack their initial results by exploring additional model specifications or estimators and report a different result of their choice. Specifically, we consider the two general approaches to \( p \)-hacking discussed in Section 3: the threshold and the minimum approach. In what follows, we discuss the generalized examples of \( p \)-hacking in more detail.

5.1.1 Selecting Control Variables in Linear Regression

Researchers have access to a random sample with \( N = 200 \) observations generated as

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

where \( X_i \sim \mathcal{N}(0, 1) \) and \( u_i \sim \mathcal{N}(0, 1) \) are independent of each other. Researchers have access to \( K \) additional control variables, \( Z_i := (Z_{1i}, \ldots, Z_{Ki})' \), which are generated as

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

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). 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 16, 17, and 18 show the null and \( p \)-hacked distributions for \( K \in \{3, 5, 7\} \).\(^{16}\) The \( p \)-curves are similar to those in the simple analytical example of Section 3.1. The threshold approach leads to a discontinuity in the \( p \)-curve and may lead to non-increasing \( p \)-curves and

\(^{16}\)To generate these distributions, we run the algorithm one million times and collect \( p \)-hacked and non-\( p \)-hacked results.
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. For example, for the threshold approach, the larger $h$, the higher the probability that they 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.

5.1.2 Selecting amongst Instruments in IV Regression

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

$$Y_i = X_i \beta + U_i,$$
$$X_i = Z_i' \pi + V_i,$$

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

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

where $\xi_i, \epsilon_{Z_{ki}},$ and $\gamma_k$ are independent for all $k$. Also, $\pi_k \sim \text{Uniform}[1, 3], k = 1, \ldots, K$.

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), 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 19 and 20 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 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.

5.1.3 Lag Length Selection in 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. 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 \( Y_i \) on \( X_i \) 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 between 0, 1, 2, 3, and 4
lags). They then use a t-test to test (3) 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 there is no significant result, they try two extra lags etc. If all results are
insignificant, they report the smallest p-value.

Minimum approach. Researchers regress \( Y_i \) on \( X_i \) and calculate the standard error
using Newey-West with 0, 1, 2, 3, and 4 lags and report the specification that yields the
minimum p-value.

The null and p-hacked distributions are displayed in Figure 21. The p-curves exhibit
features similar to those in the simple analytical example in Section 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.

5.2 Simulations

5.2.1 Setup

We model the distribution of p-values as a mixture:
\[
g(p) = \tau \cdot g^p(p) + (1 - \tau) \cdot g^{np}(p)
\]
Here, \( g^p \) 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). In our simulations we vary the fraction of researchers who \( p \)-hack, \( \tau \in [0,1] \). To generate the data, we first simulate the algorithm one million times to obtain samples corresponding to \( g^p \) and \( g^{np} \). Then, to construct samples in every Monte Carlo iteration, we draw with replacement from a mixture of those samples. The overall sample size is \( n = 5000 \). Following Elliott et al. (2022), we apply the test to the subinterval \((0,0.15]\). Therefore, the number of \( p \)-values used in the tests (i.e., effective sample size) depends on the \( p \)-hacking strategy, the distribution of \( h \), and the fraction of \( p \)-hackers \( \tau \).

We compare the finite sample performance of the tests described in Section 4. See Table 1 for more details. 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 the statistical software MATLAB (MATLAB, 2020) and R (R Core Team, 2022).

Table 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 with automatic bandwidth selection (Cattaneo et al., 2021) |

| Testable restriction: upper bounds on \( p \)-curve, first, and second derivative |
|---|
| CSUB | Histogram-based test based on Cox and Shi (2022) with \( J = 15 \) |

| Testable restriction: 2-monotonicity and upper bounds on \( p \)-curve, first, and second derivative |
|---|
| CS2B | Histogram-based test based on Cox and Shi (2022) with \( J = 15 \) |

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 C. The nominal level is 5\% for all tests. All results are based on 1000 simulation draws. Figures 12, 13, and 14 present the
The power for detecting $p$-hacking crucially depends on the type of $p$-hacking, the econometric method, the fraction of $p$-hackers, $\tau$, and the value of $h$. As shown in Section 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 B. 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 (see, for example, the results lag length selection with $h = 0$). Among the tests that exploit the classical monotonicity restriction, the CS1 test exhibits higher power than the LCM and the widely-used Binomial 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 are using a thresholding $p$-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 approach can yield isolated humps right below the cutoff (see Figure 10 and also Figure 21). 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, limiting their usefulness in practice.

Discontinuity tests constitute a useful complement to tests based on monotonicity and upper bounds because $p$-hacking based on thresholding 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 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, 5, and 10). Consistent with the analysis in Section 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 overlap, suggesting that the power of the CS2B test comes from using upper bounds. This finding demonstrates the importance of exploiting upper bounds in addition to monotonicity and continuity restrictions in practice. Figure 13 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 C 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.

Figure 12: Power curves covariate selection with $K = 3$
Figure 13: Power curves IV selection with $K = 3$

Figure 14: 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 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.

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 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 $h$. In Figure 15, 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 15 corresponds to the power of one test under one DGP. We only show results for covariate and IV selection; the bias under lag length selection is always zero. For the thresholding 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 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 15: Power vs. bias
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. A number of different tests have been suggested, and in empirical work based on these tests, evidence for p-hacking is mixed. 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.

**References**

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

Brodeur, A., Lé, M., Sangnier, M., and Zylberberg, Y. (2016). 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*.

Elliott, G., Kudrin, N., and Wüthrich, K. (2020). Detecting p-hacking. arXiv:1906.06711v3.

Elliott, G., Kudrin, N., and Wüthrich, K. (2022). 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.

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

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.

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.

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.

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

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.

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.

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.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Simonsohn, U., Nelson, L. D., and Simmons, J. P. (2014a). P-curve: a key to the file-drawer. *Journal of Experimental Psychology: General*, 143(2):534–547.

Simonsohn, U., Nelson, L. D., and Simmons, J. P. (2014b). p-curve and effect size: Correcting for publication bias using only significant results. *Perspectives on Psychological Science*, 9(6):666–681. PMID: 26186117.

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.

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.

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.
Appendix to “(When) Can We Detect $p$-Hacking?”

Table of Contents

A Detailed Derivations Section 3

A.1 Selecting Control Variables in Linear Regression 

A.2 Selecting amongst Instruments in IV Regression

A.3 Selecting across Datasets

A.4 Variance Bandwidth Selection for Means

B Null and Alternative Distributions MC Study

C Additional Simulation Results

A Detailed Derivations Section 3

A.1 Selecting Control Variables in Linear Regression

A.1.1 $p$-Curve under $p$-Hacking

We denote by $\hat{\sigma}_j$ the standard error of the estimator of $\beta$ 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

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

where

$$
\begin{pmatrix}
W_{xu} \\
W_{zu} \\
W_{zu}
\end{pmatrix} \sim \mathcal{N} 
\begin{pmatrix}
0 \\
0 \\
\gamma^2
\end{pmatrix}, 
\begin{pmatrix}
1 & \gamma & \gamma \\
\gamma & 1 & \gamma^2 \\
\gamma^2 & \gamma & 1
\end{pmatrix}.
$$

Thus, conditional on $h$,

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

where the correlation is $\rho = 1 - \gamma^2$. As the control variable and $X_i$ become more correlated (larger $\gamma$), $\rho$ becomes smaller.
The CDF of \( P_r \) on \((0,1)\) for the threshold case is

\[
G^t_h(p) = \Pr(P_r \leq p) \\
= \Pr(P_1 \leq p \mid P_1 \leq \alpha) \Pr(P_1 \leq \alpha) \\
+ \Pr(\min\{P_1, P_2\} \leq p \mid P_1 > \alpha) \Pr(P_1 > \alpha) \\
= \Pr(P_1 \leq \min\{p, \alpha\}) + (1 - \Pr(P_1 > p, P_2 > p \mid P_1 > \alpha)) \Pr(P_1 > \alpha) \\
= \Pr(T_1 \geq z_0(\min\{p, \alpha\})) + \Pr(T_1 < z_0(\alpha)) - \Pr(T_1 < z_0(\max\{p, \alpha\}), T_2 < z_0(p)) \\
= 1 - \Phi(z_0(\min\{p, \alpha\})) + \Phi(z_0(\alpha)) - \int_{z_0(h)}^{z_0(\max\{p, \alpha\})} \int_{-\infty}^{\min\{p, \alpha\}} f(x, y; \rho)\,dx\,dy,
\]

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 \( G^t_h(p) \) with respect to \( p \) yields:

\[
\frac{dG^t_h(p)}{dp} = \frac{dz_h(p)}{dp} \left[ -\phi(z_h(p)) - \int_{-\infty}^{z_h(\alpha)} f(z_h(p), y; \rho)\,dy \right] \\
= \phi(z_h(p)) \left[ 1 + \Phi\left( \frac{z_h(\alpha) - \rho z_h(p)}{\sqrt{1-\rho^2}} \right) \right].
\]

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

\[
\frac{dG^t_h(p)}{dp} = \frac{2\phi(z_h(p))\Phi\left( \frac{z_h(\alpha) - \rho z_h(p)}{\sqrt{1-\rho^2}} \right)}{\phi(z_0(p))}.
\]

It follows that the PDF of \( p \)-values is

\[
g^t_h(p) = \int_{\mathcal{H}} \frac{dG^t_h(p)}{dp}d\Pi(h) = \int_{\mathcal{H}} \frac{\phi(z_h(p))\Upsilon_1(p; \alpha, h, \rho)}{\phi(z_0(p))}d\Pi(h),
\]

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

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

\[
G^m_h(p) = \Pr(P_r \leq p) \\
= \Pr(P_1 \leq p, P_1 \leq P_2) + \Pr(P_2 \leq p, P_2 < P_1) \\
= \Pr(T_1 \geq z_0(p), T_1 \geq T_2) + \Pr(T_2 \geq z_0(p), T_2 > T_1) \\
= 2 \Pr(\xi_1 \geq z_h(p), \xi_1 \geq \xi_2)
\]
\[
\begin{align*}
= 2 \int_{-\infty}^{z_h(p)} \int_{z_h(p)}^{\infty} f(x, y; \rho) \, dx \, dy + 2 \int_{z_h(p)}^{\infty} \int_{y}^{\infty} f(x, y; \rho) \, dx \, dy,
\end{align*}
\]

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

The derivative of \( G_{\beta}^m(p) \) with respect to \( p \) is
\[
\frac{dG_{\beta}^m(p)}{dp} = 2 \frac{dz_h(p)}{dp} \left[ \int_{z_h(p)}^{\infty} f(x, z_h(p); \rho) \, dx - \int_{-\infty}^{z_h(p)} f(z_h(p), y; \rho) \, dy - \int_{z_h(p)}^{\infty} f(x, z_h(p); \rho) \, dx \right]
\]
\[
= 2 \frac{\phi(z_h(p))}{\phi(z_0(p))} \Phi \left( z_h(p) \sqrt{\frac{1-\rho}{1+\rho}} \right).
\]

Therefore, the PDF of \( p \)-values is
\[
g_{\beta}^m(p) = 2 \int_{\beta}^{\infty} \exp \left( h z_0(p) - \frac{h^2}{2} \right) \Phi \left( z_h(p) \sqrt{\frac{1-\rho}{1+\rho}} \right) \, d\Pi(h).
\]

### A.1.2 Bias of the \( p \)-Hacked Estimator

Fix \( h \) for now. We have \( \hat{\beta}_r = \hat{\beta}_1 + (\hat{\beta}_2 - \hat{\beta}_1)1_{(T_2 > T_1, T_1 < z_0(\alpha))} \). The bias in the \( p \)-hacked estimate is given by
\[
E[\hat{\beta}_r - \beta] = E \left[ \hat{\beta} - \beta + (\hat{\beta}_2 - \hat{\beta}_1)1_{(T_2 > T_1, T_1 < z_0(\alpha))} \right]
\]
\[
= E \left[ (\hat{\beta}_2 - \hat{\beta}_1)1_{(T_2 > T_1, T_1 < z_0(\alpha))} \right]
\]
\[
= \frac{1}{\sqrt{N}} \frac{1}{\sqrt{\rho}} \left[ (\xi_2 - \xi_1)1_{(\xi_2 > \xi_1, \xi_1 < z_h(\alpha))} \right]
\]
\[
= \frac{1}{\sqrt{N}} \frac{1}{\sqrt{\rho}} E \left[ V1_{(V > 0, \xi_1 < z_h(\alpha))} \right],
\]

where \( V = \xi_2 - \xi_1 \sim \mathcal{N}(0, 2(1-\rho)) \) and \( E[V\xi_1] = -(1 - \rho) \). Now
\[
E \left[ V1_{(V > 0, \xi_1 < z_h(\alpha))} \right] = \int_{0}^{\infty} \int_{-\infty}^{z_h(\alpha)} v f_{\xi_1}(v, x) \, dx \, dv
\]
\[
= \int_{0}^{\infty} \int_{-\infty}^{z_h(\alpha)} v f_{\xi_1|V}(x|v) f_V(v) \, dx \, dv
\]
\[
= \int_{0}^{\infty} v f_V(v) \left( \int_{-\infty}^{z_h(\alpha)} f_{\xi_1|V}(x|v) \, dx \right) \, dv,
\]

and
\[
\int_{-\infty}^{z_h(\alpha)} f_{\xi_1|V}(x|v) \, dx = \Pr(\xi_1 < z_h(\alpha) \mid V = v)
\]
\[
= \Pr \left( \frac{\xi_1 + v/2}{\sqrt{1+\rho/2}} < \frac{z_h(\alpha) + v/2}{\sqrt{1+\rho/2}} \right)
= \Phi \left( \frac{z_h(\alpha) + v/2}{\sqrt{1+\rho/2}} \right).
\]

So now we have
\[
E\hat{\beta}_r^t - \beta = 1 \cdot \sqrt{\frac{N}{\sqrt{N}}} \int_0^\infty v \Phi \left( \frac{z_h(\alpha) + v/2}{\sqrt{1+\rho/2}} \right) f_V(v) \, dv,
\]
and the final expression follows by direct integration.

For the minimum approach, \( \hat{\beta}_m^t = \hat{\beta}_1 + (\hat{\beta}_2 - \hat{\beta}_1)1_{\{T_2 > T_1\}} \). The bias in the \( p \)-hacked estimate is given by
\[
E\hat{\beta}_r^m - \beta = E \left[ \hat{\beta}_1 - \beta + (\hat{\beta}_2 - \hat{\beta}_1)1_{\{T_2 > T_1\}} \right]
= \frac{1}{\sqrt{N} N} E \left( V_1(V > 0) \right).
\]

Now \( E \left[ V_1(V > 0) \right] = \sqrt{2(1-\rho)\phi(0)} \) so
\[
E\hat{\beta}_r^m - \beta = \frac{1}{\sqrt{N}} \sqrt{2(1-\rho)} \frac{\phi(0)}{\rho}.
\]

It follows that \( E\hat{\beta}_r^t \leq E\hat{\beta}_r^m \) because
\[
EV1_{\{V > 0, \xi_1 < z_h(\alpha)\}} = \int_0^\infty v f_V(v) \left( \int_{-\infty}^{z_h(\alpha)} f_{\xi_1|V}(x|v) \, dx \right) \, dv 
\leq \int_0^\infty v f_V(v) \, dv.
\]

### A.2 Selecting amongst Instruments in IV Regression

#### A.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
\[
\hat{\beta}_{12} = \beta + \left[ \frac{\sum X_i Z_{1i}^2}{\sum Z_{1i}^2} + \frac{\sum X_i Z_{2i}^2}{\sum Z_{2i}^2} + o_p(1) \right]^{-1} \left[ \sum X_i Z_{1i} \sum U_i Z_{1i} + \sum X_i Z_{2i} \sum U_i Z_{2i} \right]
\]
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,
\[
\hat{\beta}_j = \beta + \frac{\sum Z_j U_i}{\sum X_i Z_{ji}}, \quad j = 1, 2,
\]
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
\[
G_h^t(p) = \Pr(P_r \leq p)
= \Pr(P_{12} \leq p \mid P_{12} \leq \alpha) \Pr(P_{12} \leq \alpha)
+ \Pr(\min\{P_1, P_2, P_{12}\} \leq p \mid P_{12} > \alpha) \Pr(P_{12} > \alpha)
= \Pr(P_{12} \leq \min\{p, \alpha\}) + \Pr(P_{12} > \alpha)
- \Pr(P_1 > p, P_2 > p, P_{12} > p \mid P_{12} > \alpha) \Pr(P_{12} > \alpha)
= \Pr(T_{12} \geq z_0(\min\{p, \alpha\})) + \Pr(T_{12} < z_0(\alpha))
- \Pr(T_1 < z_0(p), T_2 < z_0(p), T_{12} < z_0(\max\{p, \alpha\}))
= 1 - \Phi(z_{\sqrt{2}h}(\min\{p, \alpha\})) + \Phi(z_{\sqrt{2}h}(\alpha)) - \Phi(z_h(p)) \Phi(D_h(\max\{p, \alpha\})) - z_h(p)
- \int_{D_h(\max\{p, \alpha\}) - z_h(p)}^{z_h(p)} \phi(x) \Phi(D_h(\max\{p, \alpha\})) - x \, dx.
\]
The last equality follows because for \(p \leq 1/2\) 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
\[
\Pr(W_1 < a, W_2 < a, W_1 + W_2 < b) = \int_{-\infty}^{a} \int_{-\infty}^{b-a} \phi(x) \phi(y) \, dx \, dy
+ \int_{b-a}^{a} \int_{-\infty}^{b-x} \phi(x) \phi(y) \, dy \, dx
= \Phi(a) \Phi(b - a) + \int_{b-a}^{a} \phi(x) \Phi(b - x) \, dx
\]
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
\[
\frac{dG_t^i(p)}{dp} = \frac{\phi(z_{\sqrt{2}h}(p))(1 - 2\Phi((1 - \sqrt{2})z_0(p))) + 2C(z_h(p), D_h(p))}{\phi(z_0(p))}, \quad p \in (\alpha, 1/2).
\]

For \( p > 1/2 \), we have \( 2z_h(p) < D_h(\max\{p, \alpha\}) \), and similar arguments yield
\[
G_t^i(p) = 1 - \text{Pr}(W_1 < z_h(p), W_2 < z_h(p), W_1 + W_2 < D_h(\max\{p, \alpha\}))
\]
\[
= 1 - \int_{-\infty}^{z_h(p)} \phi(x) \phi(y) dxdy
\]
\[
= 1 - \Phi^2(z_h(p))
\]
and
\[
\frac{dG_t^i(p)}{dp} = \frac{2\phi(z_h(p)\Phi(z_h(p))}{\phi(z_0(p))}, \quad p > 1/2.
\]

Since \( g^t_i(p) = \int_{\mathcal{H}} \frac{G_t^i(w)}{dp} d\Pi(h) \), we have
\[
g_2^t(p) = \int_{\mathcal{H}} \exp \left( hz_0(p) - \frac{h^2}{2} \right) \Upsilon_2^i(p; \alpha, h) d\Pi(h),
\]
where \( \zeta(p) = 1 - 2\Phi((1 - \sqrt{2})z_0(p)) \), \( cv_1(p) = z_0(p) \) and
\[
\Upsilon_2^i(p; \alpha, h) = \begin{cases} 
\frac{\phi(z_{\sqrt{2}h}(p))}{\phi(z_0(p))} + 2\Phi(D_h(\alpha) - z_h(p)), & \text{if } 0 < p \leq \alpha, \\
\frac{\phi(z_{\sqrt{2}h}(p))}{\phi(z_0(p))} \zeta(p) + 2\Phi(D_h(p) - z_h(p)), & \text{if } \alpha < p \leq 1/2, \\
2\Phi(z_h(p)), & \text{if } 1/2 < p < 1.
\end{cases}
\]

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

### A.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
\[
\hat{\beta}_t = \hat{\beta}_{12} + (\hat{\beta}_1 - \hat{\beta}_{12})1_{\{A_N\}} + (\hat{\beta}_1 - \hat{\beta}_{12})1_{\{B_N\}},
\]
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,
\[
\sqrt{N} |\gamma| (\hat{\beta}_t^i - \beta) \xrightarrow{d} \xi^i := \frac{W_1 + W_2}{2} + \frac{W_1 - W_2}{2}1_{\{A\}} + \frac{W_2 - W_1}{2}1_{\{B\}},
\]
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,

\[
E[\xi^t] = E[(W_1 - W_2)1_{\{A\}}] = E[W_11_{\{A\}}] - E[W_21_{\{A\}}].
\]

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

\[
A = \begin{cases} 
\sqrt{2}h + \frac{W_1 + W_2}{\sqrt{2}} < z_0(\alpha) \\
h + W_1 > \sqrt{2}h + \frac{W_1 + W_2}{\sqrt{2}} \\
W_1 > W_2
\end{cases} \iff \begin{cases} 
W_2 < \sqrt{2}z_0(\alpha) - 2h - W_1 \\
W_2 < (\sqrt{2} - 1)(W_1 - \sqrt{2}h) \iff \begin{cases} 
W_1 < \sqrt{2}z_0(\alpha) - 2h - W_2 \\
W_1 > \sqrt{2}h + \frac{W_2}{\sqrt{2} - 1} \iff \begin{cases} 
W_1 > W_2
\end{cases}
\end{cases}
\end{cases}
\]

From equation (\ast\ast), 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 (\ast\ast) is equivalent to

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

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

\[
\Pr(A|W_2) = \Pr \left( W_1 \in \left( W_2, \sqrt{2}z_0(\alpha) - 2h - W_2 \right) | W_2 \right) \cdot 1_{\{W_2 < -h\}} + \Pr \left( W_1 \in \left( 2h + \frac{W_2}{\sqrt{2} - 1}, \sqrt{2}z_0(\alpha) - 2h - W_2 \right) | W_2 \right) \cdot 1_{\{W_2 \geq -h\}}
\]

\[
= \Phi \left( \sqrt{2}z_0(\alpha) - 2h - W_2 \right) - \Phi(W_2) \cdot 1_{\{W_2 < -h\}} + \Phi \left( \sqrt{2}z_0(\alpha) - 2h - W_2 \right) - \Phi \left( \sqrt{2}h + \frac{W_2}{\sqrt{2} - 1} \right) \cdot 1_{\{-h \leq W_2 \leq (\sqrt{2} - 1)z_0(\alpha) - h\}}
\]

\[
= \Phi \left( \sqrt{2}z_0(\alpha) - 2h - W_2 \right) \cdot 1_{\{W_2 \leq (\sqrt{2} - 1)z_0(\alpha) - h\}} - \Phi(W_2) \cdot 1_{\{W_2 < -h\}} - \Phi \left( \sqrt{2}h + \frac{W_2}{\sqrt{2} - 1} \right) \cdot 1_{\{-h \leq W_2 \leq (\sqrt{2} - 1)z_0(\alpha) - h\}}
\]
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 1.**
\[
\int_{L}^{U} w \phi(w) \Phi(aw + b) dw = \Phi(aL + b) - \Phi(aU + b) + \frac{a}{\sqrt{1 + a^2}} \phi \left( \frac{b}{\sqrt{1 + a^2}} \right) \times \left[ \Phi \left( \sqrt{1 + a^2}U + \frac{ab}{\sqrt{1 + a^2}} \right) - \Phi \left( \sqrt{1 + a^2}L + \frac{ab}{\sqrt{1 + a^2}} \right) \right]
\]

**Proof.**
\[
\int_{L}^{U} w \phi(w) \Phi(aw + b) dw = - \int_{L}^{U} \Phi(aw + b) d\phi(w)
\]
\[
= \Phi(aL + b) - \Phi(aU + b) + \int_{L}^{U} \phi(w) d\Phi(aw + b)
\]
\[
= \Phi(aL + b) - \Phi(aU + b) + a \int_{L}^{U} \phi(w) \Phi(aw + b) dw
\]
\[
= \Phi(aL + b) - \Phi(aU + b) + a \phi \left( \frac{b}{\sqrt{1 + a^2}} \right) J,
\]

where
\[
J = \int_{L}^{U} \phi \left( \sqrt{1 + a^2}w + \frac{ab}{\sqrt{1 + a^2}} \right) dw
\]
\[
= \frac{1}{\sqrt{1 + a^2}} \left[ \Phi \left( \sqrt{1 + a^2}U + \frac{ab}{\sqrt{1 + a^2}} \right) - \Phi \left( \sqrt{1 + a^2}L + \frac{ab}{\sqrt{1 + a^2}} \right) \right].
\]

Finally,
\[
E[W_1 1_{\{A\}}] = E[W_1 P(A|W_1)]
\]
\[
= E[W_1 \Phi(\min\{W_1, \sqrt{2}z_0(\alpha) - 2h - W_1, (\sqrt{2} - 1)(W_1 - \sqrt{2}h)\})]
\]
\[
= \int_{-\infty}^{-h} w \phi(w) \Phi(w) dw
\]
\[
+ \int_{-h}^{z_0(\alpha) - h} w \phi(w) \Phi((\sqrt{2} - 1)w - \sqrt{2}(\sqrt{2} - 1)h) dw
\]
\[
+ \int_{z_0(\alpha) - h}^{+\infty} w \phi(w) \Phi(\sqrt{2}z_0(\alpha) - 2h - w) dw
\]

8
where the fourth equality follows from the direct application of Lemma 1.

\[
E[W_{21(A)}] = E[W_2 \Pr(A|W_2)]
\]
\[
= E \left[ W_2 \Phi \left( \sqrt{2}z_0(\alpha) - 2h - W_2 \right) \cdot 1\{W_2 \leq (\sqrt{2} - 1)z_0(\alpha) - h\} \right] - E \left[ W_2 \Phi(W_2) \cdot 1\{W_2 < -h\} \right] - E \left[ W_2 \Phi \left( \sqrt{2}h + \frac{W_2}{\sqrt{2} - 1} \right) \cdot 1\{-h \leq W_2 \leq (\sqrt{2} - 1)z_0(\alpha) - h\} \right]
\]
\[
= \int_{-\infty}^{(\sqrt{2} - 1)z_0(\alpha) - h} w\phi(w)\Phi \left( \sqrt{2}z_0(\alpha) - 2h - w \right)dw - \int_{-\infty}^{-h} w\phi(w)\Phi(w)dw - \int_{-h}^{(\sqrt{2} - 1)z_0(\alpha) - h} w\phi(w)\Phi \left( \sqrt{2}h + \frac{w}{\sqrt{2} - 1} \right)dw
\]
\[
= -\Phi(z_0(\alpha) - h)\phi((\sqrt{2} - 1)z_0(\alpha) - h) - \frac{1}{\sqrt{2}}\phi(z_0(\alpha) - \sqrt{2}h)(1 - \Phi((\sqrt{2} - 1)z_0(\alpha))) + (1 - \Phi(h))\phi(h) - \frac{1}{\sqrt{2}}\phi(0)(1 - \Phi(\sqrt{2}h))
\]
\[
- (1 - \Phi(h))\phi(h) + \Phi(z_0(\alpha) - h)\phi((\sqrt{2} - 1)z_0(\alpha) - h) - \frac{1}{\sqrt{2}}\phi(z_0(\alpha) - \sqrt{2}h)(1 - \Phi((\sqrt{2} - 1)z_0(\alpha)))
\]
\[
- \frac{1}{\sqrt{4 - 2\sqrt{2}}} \phi \left( \frac{\sqrt{2} - 1}{\sqrt{2}}h \right) \left[ \Phi \left( \frac{h}{\sqrt{2} - \sqrt{2}} \right) - \Phi \left( \frac{h}{\sqrt{2} - \sqrt{2}} - \sqrt{4 - 2\sqrt{2}z_0(\alpha)} \right) \right]
\]
\[
= -\frac{1}{\sqrt{2}}\phi(z_0(\alpha) - \sqrt{2}h)(1 - \Phi((\sqrt{2} - 1)z_0(\alpha))) - \frac{1}{\sqrt{2}}\phi(0)(1 - \Phi(\sqrt{2}h)) - \frac{1}{\sqrt{4 - 2\sqrt{2}}} \phi \left( \frac{\sqrt{2} - 1}{\sqrt{2}}h \right) \left[ \Phi \left( \frac{h}{\sqrt{2} - \sqrt{2}} \right) - \Phi \left( \frac{h}{\sqrt{2} - \sqrt{2}} - \sqrt{4 - 2\sqrt{2}z_0(\alpha)} \right) \right].
\]
Combining these results gives us the first-order bias of \( \hat{\beta}_r^t, B_2' = |\gamma|^{-1} E[\xi^t], \) where

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

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

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

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,

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

and

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

Putting these together gives the asymptotic bias of \( \hat{\beta}_r^m, B_2^m = |\gamma|^{-1} E[\xi^m], \) where

\[
E[\xi^m] = \frac{1}{\sqrt{2} - \sqrt{2}} \phi \left( \sqrt{\frac{2}{2} - 1} h \right) \Phi \left( \frac{h}{\sqrt{2} - \sqrt{2}} \right) + \sqrt{2} \phi(0)(1 - \Phi(\sqrt{2} h)).
\]
A.3 Selecting across Datasets

With the assumption that the datasets are independent, we have that the \( K \) \( t \)-statistics are distributed as \( T \sim \mathcal{N}(h \lambda_K, I_K) \) where \( \lambda_K \) 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 A.1 with \( \rho = 0 \). Hence the result for the thresholding case is \( g^t_1 \) evaluated at \( \rho = 0 \) and for the minimum case is \( g^m_1 \) also evaluated at \( \rho = 0 \).

For general \( K \), note that for the minimum case

\[
G^m(p; K) = \Pr(\max(T_1, T_2, \ldots, T_K) \geq z_0(p)) = 1 - \Pr(T_1 \leq z_0(\alpha), T_2 \leq z_0(\alpha), \ldots, T_K \leq z_0(\alpha)) = 1 - (\Phi(z_h(p)))^K
\]

Setting \( p = \alpha \) gives the size after \( p \)-hacking for a nominal value of \( \alpha \). Differentiating with respect to \( p \) generates the \( p \)-curve

\[
g^m_3(p; K) = \frac{d}{dp} \left( 1 - (\Phi(z_h(p)))^K \right) = -K \Phi(z_h(p))^{K-1} \frac{d}{dp} \Phi(z_h(p)) = K \frac{d}{dp} \Phi(z_h(p))^{K-1} \frac{\phi(z_h(p))}{\phi(z_0(p))}
\]

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

A.4 Variance Bandwidth Selection for Means

Note that \( T_0 \sim \mathcal{N}(h, 1) \) and \( \hat{\rho} \) are independent.\(^{17}\) 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 \( \alpha \) or if \( \hat{\omega}^2 < 0 \), otherwise they report the \( p \)-value associated with the largest \( t \)-statistic. Fixing \( h \), we have

\[
G^t_h(p) = \Pr(P_r \leq p) = \Pr(T_0 \geq z_0(p), T_0 \geq z_0(\alpha)) + \Pr(T_0 \geq z_0(p), T_0 < z_0(\alpha), -\infty < \hat{\rho} \leq -2(\kappa)^{-1})
\]

\(^{17}\)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.
+ \Pr(T_0 \geq z_0(p), T_0 \geq 0, T_0 < z_0(\alpha), \hat{\rho} > 0)
+ \Pr(T_1 \geq z_0(p), T_0 \geq 0, T_0 < z_0(\alpha), -(2\kappa)^{-1} < \hat{\rho} \leq 0)
+ \Pr(T_1 \geq z_0(p), T_0 \leq 0, T_0 < z_0(\alpha), \hat{\rho} > 0)
+ \Pr(T_0 \geq z_0(p), T_0 \leq 0, T_0 < z_0(\alpha), -(2\kappa)^{-1} < \hat{\rho} \leq 0)

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

$$g_h^t(p) = \int_{\mathcal{H}} \exp \left( hz_0(p) - \frac{h^2}{2} \right) \Upsilon^t_4(p; \alpha, h, \kappa) d\Pi(h),$$

where

$$\Upsilon^t_4 = \begin{cases} 
1 + \frac{1}{\phi(z_h(p))} \int_{-(2\kappa)^{-1}}^{l(p)} \omega(r)\phi(z_0(p)\omega(r) - h)\eta_N(r)dr, & \text{if } 0 < p \leq \alpha, \\
1 - H_N(0) + H_N(-(2\kappa)^{-1}) + \frac{1}{\phi(z_h(p))} \int_{-(2\kappa)^{-1}}^{0} \omega(r)\phi(z_0(p)\omega(r) - h)\eta_N(r)dr, & \text{if } \alpha < p \leq 1/2, \\
H_N(0) + \frac{1}{\phi(z_h(p))} \int_{0}^{\infty} \omega(r)\phi(z_0(p)\omega(r) - h)\eta_N(r)dr, & \text{if } 1/2 < p < 1.
\end{cases}$$

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

Note that

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

For \( p \leq 1/2 \), we have

\[
G^m_h(p) = H_N(0) - H_N(-(2\kappa)^{-1}) + (1 - \Phi(z_h(p)))(1 - H_N(0) + H_N(-(2\kappa)^{-1}))
- \int_{-(2\kappa)^{-1}}^{0} \Phi(z_0(p)\omega(r) - h)\eta_N(r)dr
\]

and, for \( p > 1/2 \), we have

\[
G^m_h(p) = 1 - H_N(0) + (1 - \Phi(z_h(p)))H_N(0) - \int_{0}^{\infty} \Phi(z_0(p)\omega(r) - h)\eta_N(r)dr.
\]

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

\[
g^m_4(p) = \int_{\mathcal{H}} \exp \left( h z_0(p) - \frac{h^2}{2} \right) \Upsilon^m_4(p; \alpha, h) d\Pi(h),
\]

where

\[
\Upsilon^m_4 = \begin{cases} 
1 - H_N(0) + H_N(-(2\kappa)^{-1}) + \frac{1}{\phi(z_h(p))} \int_{-(2\kappa)^{-1}}^{0} \omega(r)\phi(z_0(p)\omega(r) - h)\eta_N(r)dr, & \text{if } 0 < p \leq 1/2, \\
H_N(0) + \frac{1}{\phi(z_h(p))} \int_{0}^{\infty} \omega(r)\phi(z_0(p)\omega(r) - h)\eta_N(r)dr, & \text{if } 1/2 < p < 1.
\end{cases}
\]
B Null and Alternative Distributions MC Study

Figure 16: Null and $p$-hacked distributions for covariate selection with $K = 3$
Figure 17: Null and p-hacked distributions for covariate selection with $K = 5$
Figure 18: Null and $p$-hacked distributions for covariate selection with $K = 7$
Figure 19: Null and $p$-hacked distributions for IV selection with $K = 3$
Figure 20: Null and $p$-hacked distributions for IV selection with $K = 5$
Figure 21: Null and $p$-hacked distributions for lag length selection
C Additional Simulation Results

Figure 22: Power curves for \( h \sim \chi^2(1) \). Thresholding (left column) and minimum (right column).
Figure 23: Power curves covariate selection with $K = 5$. Thresholding (left column) and minimum (right column).
Figure 24: Power curves covariate selection with $K = 7$. Thresholding (left column) and minimum (right column).
Figure 25: Power curves IV selection with $K = 5$. Thresholding (left column) and minimum (right column).