CESifo WORKING PAPERS

11027
2024

March 2024

# Testing with Vectors of Statistics: Revisiting Combined Hypothesis Tests with an Application to Specification Testing

*Lena S. Bjerkander, Jonas Dovern, Hans Manner*

CESifo

## Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
https://www.cesifo.org/en/wp
An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: https://www.cesifo.org/en/wp

CESifo *Working Paper No. 11027*

# Testing with Vectors of Statistics: Revisiting Combined Hypothesis Tests with an Application to Specification Testing

## Abstract

We review tests of null hypotheses that consist of many subsidiary null hypotheses, including tests that have not received much attention in the econometrics literature. We study test performance in the context of specification testing for linear regressions based on a Monte Carlo study. Overall, parametric tests that use (transformed) *P*-values corresponding to all subsidiary null hypotheses outperform the well-known minimum *P*-value test and a recently proposed test that relies on the non-parametric estimation of the joint density of all subsidiary test statistics.

JEL-Codes: C120, C150.

Keywords: combined hypothesis, *P*-value, multiple hypothesis testing, Fisher test.

*Lena S. Bjerkander**
*Friedrich-Alexander-University*
*Erlangen-Nürnberg / Germany*
*lena.bjerkander@fau.de*

*Jonas Dovern*
*Friedrich-Alexander-University*
*Erlangen-Nürnberg / Germany*
*jonas.dovern@fau.de*

*Hans Manner*
*University of Graz / Austria*
*hans.manner@uni-graz.at*

*corresponding author

March 21, 2024

# 1 Introduction

Hypothesis testing is widely used in econometrics. Inter alia, it is an important tool to justify modelling choices and, therefore, the results of empirical research. In many situations, one wishes to test a joint null hypothesis consisting of multiple subsidiary null hypotheses. In particular, specification tests typically involve using multiple statistics that are motivated by the need to test a number of assumptions underlying a model and/or estimation procedure. A joint test is generally preferable over multiple individual tests to control the overall probability of type I errors. This paper re-examines existing tests—including tests not widely used in the econometrics literature—and investigates their performance in the context of specification testing.

While it is straightforward to perform hypothesis testing based on a single statistic, it is less clear how to evaluate multiple statistics. In general, a vector that collects multiple statistics can form the basis for hypothesis tests. Each statistic may refer to different aspects of the data generating process (DGP). Consider a $d \times 1$ vector of statistics, $\mathbf{S} = (S_1, S_2, \ldots, S_d)'$, where the $d$ elements of the vector are the individual test statistics that correspond to the considered subsidiary hypotheses. Using critical values for each test individually without adjustment when making a test decision is problematic due to the resulting high overall probability of a type I error. One remedy is to determine adjusted decision rules based on the Bonferroni inequality. However, these approaches tend to be too conservative. A more fruitful approach is to aggregate the individual statistics into a single one.

Most of the approaches that we consider below follow this aggregation approach. At least since Dufour et al. (2015), the most commonly used tests in econometrics are a test that relies on the subsidiary statistic with the lowest $P$-value (dating back to Tippett et al., 1931) and a test that relies on the product of the $P$-values of all subsidiary statistics (dating back to Fisher, 1932). Not widely known in econometrics are tests that look at sums of subsidiary $P$-values, after transforming the latter by the distribution functions of the $\chi^2$ distribution (dating back to Lancaster, 1961), the normal distribution (dating back to Stouffer et al., 1949), or the Gamma distribution (recently proposed by Chen, 2021). Under independence of the subsidiary test statistics, it is often possible to derive a null distribution of the overall test statistic analytically.

In economics, the assumption of independent subsidiary statistics is often not realistic. However, even if the subsidiary tests are not independent, simulations can usually be used to determine the test distribution under the null hypothesis (Godfrey, 2005, 2009; Dufour et al., 2015) and, thus, to derive a properly sized test. We will rely on such simulation techniques when applying the different testing approaches below. We consider one non-parametric approach recently proposed by King et al. (2020). This approach handles dependent subsidiary test statistics, too. The main idea of it is to non-parametrically estimate the joint density of all subsidiary test statistics under the null hypothesis and to estimate the $P$-value of the smallest acceptance region test.

A natural application of multiple hypothesis testing lies in specification testing. In this context, one wishes to detect different forms of model misspecification or testing assumptions underlying a specific estimation or inference procedure. The overall null hypothesis is that the chosen model specification is adequate. Usually, there are multiple potential deviations from that overall null hypothesis, not known a priori. Linear regression models, for instance, frequently assume no residual autocorrelation, homoskedastic error terms and no unmodelled non-linearities. The real DGP might deviate from any of those assumptions. Well-known tests are available that allow assessing these inadequacies separately; they need to be suitably combined to obtain a valid test of the overall null hypothesis. Godfrey (2009) notes that such vector of separate diagnostic checks constitutes a situation where the individual tests are potentially dependent.

We analyze the performance of the tests using Monte Carlo simulations in the context of specification testing for linear regressions. The application setup follows Godfrey (2005) who investigates a similar research question, yet considers only the minimum $P$-value statistic. We find that combined tests outperform individual tests tailored to one particular deviation from the null hypothesis as soon as there are two types of deviation from the overall null hypothesis. Those approaches that combine information from all subsidiary tests (especially via a simple product of the subsidiary $P$-values or a sum of the subsidiary $P$-values transformed by the chi squared distribution function) perform substantially better than the minimum $P$-value test in most settings and equally well if only exactly one deviation from the null hypothesis is present. The non-parametric approach by King et al. (2020) performs well but has smaller power than the best parametric tests.

The remainder of the paper is structured as follows. In Section 2, we describe the general testing framework and revisit existing tests. Section 3 introduces specification tests for regression models that we will focus on in our application. Section 4 presents the Monte Carlo study; it includes information about the simulation setup (Section 4.1), the results regarding the size of the tests (Section 4.2) and their power (Section 4.3). Section 5 concludes.

## 2 The Testing Framework

Consider a testing problem with a joint null hypothesis that consists of $d$ subsidiary null hypotheses $H_{01}, H_{02}, \ldots, H_{0d}$. For each null hypothesis, a statistic $S_i$ is available that has a cumulative distribution function (CDF) $F_i$ if $H_{0i}$ is true.[1] The individual statistics are chosen to have correct size and good power to detect deviations from the respective null hypothesis. Individually, one would reject a subsidiary null hypothesis if the observed value of $S_i$ is very unlikely under $H_{0i}$. The $P$-value can be used to design a natural decision rule for such a situation. It is defined as the probability to observe a value of $S_i$ that is as

[1] A related situation occurs if the same null hypothesis is tested multiple times using different tests that might have different power against different alternatives, so-called induced tests.

extreme or more extreme than the realized value $s_i$ under the condition that the null hypothesis is true. Without loss of generality we assume a right-sided test such that $P(S_i \geq s_i|H_0) = 1 - F_i(s_i) = P_i$. Given significance level $\alpha$, $H_{0i}$ is rejected when $P_i \leq \alpha$. Under the assumption that the $F_i$ are continuous, each $P_i$ is described by a uniform distribution between 0 and 1 when the associated subsidiary null hypothesis is true, i.e., $P_i|H_{0i} \sim U(0,1)$.

The null hypothesis of the joint testing problem is given as the intersection where all individual null hypotheses are true:

$$H_0 : H_{01}, H_{02}, \ldots, H_{0d} \text{ are all true} \tag{1}$$

This overall null hypothesis is tested against the alternative hypothesis that at least one of the subsidiary null hypotheses is not true:

$$H_1 : \text{at least one of } H_{01}, H_{02}, \ldots, H_{0d} \text{ is not true} \tag{2}$$

We now turn to the issue of aggregating the components of $\mathbf{S} = (S_1, S_2, \ldots, S_d)'$ or the corresponding $P$-values, $\mathbf{P} = (P_1, P_2, \ldots, P_d)'$, into one test statistic.

## 2.1 Parametric Tests

Combining information about several statistics/$P$-values associated with multiple subsidiary null hypotheses and generating an overall statistic for testing a joint hypothesis is not a novel idea. Early proposals date back to Tippett et al. (1931), Fisher (1932), and Good (1955). In general terms, the idea is to form a test statistic for the joint hypothesis as a function of either the subsidiary test statistics directly, i.e., $S^* = f(\mathbf{S})$, or the corresponding $P$-values, i.e., $S^* = f(\mathbf{P})$. Many approaches assume a test statistic of the form

$$S^* = f\left(w_1 g(P_1), w_2 g(P_2), \ldots, w_d g(P_d)\right), \tag{3}$$

where the $w_i$ are weights attached to the individual subsidiary $P$-values and $g()$ is some function used to transform the $P$-values before aggregation.

Any feasible approach requires that the distribution of $S^*$ under the null hypothesis is either known or can be approximated by simulation. We can distinguish three cases. First, the null distribution is analytically available for some approaches of the form given by equation (3) under the condition that the subsidiary tests are independent of each other. We mention these null distributions—if available—for the particular combined test discussed below.

Second, if no analytical distribution is available even though the subsidiary tests are independent, estimation of the distribution function of $S^*$ by simulation is straightforward. In particular, the $P$-values in equation (3) follow independent $U(0,1)$ distributions under the joint null hypothesis in that case. Hence, one can simulate $m$ draws of $S^*$ under $H_0$ by simulating $m$ independent draws from uniform

distributions for $P_1, \ldots, P_d$ (potentially determining $P$-value-dependent weights based on these draws, too). This is not computationally expensive and for given $d$, critical values could even be tabulated based on such simulation.

Finally, whenever the subsidiary test statistics are not independent of each other, the results from Dufour et al. (2015) show how one can use bootstrap simulations to estimate the null distributions—and, thus, critical regions—of combined tests. An important requirement is that the joint null distribution of the subsidiary tests does not depend on nuisance parameters, and neither do the weights in equation (3). Essentially, the approach implies bootstrapping data under the assumption that $H_0$ is true to obtain $m$ sets of subsidiary statistics and their $P$-values (with the same joint distribution as $\mathbf{S}$ under $H_0$), which, in turn, can be used to determine an empirical $P$-value by calculating the rank of the combined test statistics obtained for the observed sample in the sequence of bootstrapped combined statistics. The test is exact when $m$ is chosen such that $\alpha(m+1)$ is an integer. In the remainder of this section, we review available testing approaches that implement the general idea of combining subsidiary statistics into one overall statistic.

**Minimum $P$-value test**

In the econometrics literature, the minimum $P$-value test is probably the most commonly used approach. It dates back to Tippett et al. (1931). The decision rule of the minimum $P$-value test is based on the smallest of the subsidiary $P$-values. In terms of the general framework in equation (3) that implies that the weights are given by the indicator function $w_i = \mathbb{1}(P_i < P_j, \forall j \neq i)$. The statistic is then given by

$$S^* = S_{min} = 1 - P_{min} = 1 - \min_{i=1,\ldots,d}\{P_i\}. \tag{4}$$

$H_0$ is rejected whenever $S_{min}$ is large. $S_{min}$ follows a Beta distribution with parameters 1 and $d$ under the null hypothesis when the subsidiary $P$-values are independent. When independence is violated, the bootstrap suggested in Dufour et al. (2015) can be used to estimate the null distribution.

**Product of $P$-values test / Fisher test**

One alternative is a test based on the product of all subsidiary $P$-values or, equivalently, the sum of the logarithms of these $P$-values. Dating back to Fisher (1932), it is also known as the Fisher test. The test statistic is given by

$$S^* = S_F = -2\sum_{i=1}^{d} \ln P_i. \tag{5}$$

$H_0$ is rejected whenever $S_F$ is large. Under independence of the subsidiary $P$-values the null distribution of $S_F$ is $\chi^2_{2d}$. When independence is violated, one can again bootstrap critical values and the $P$-value.

**Weighted product of $P$-value tests**

Instead of weighting all $P$-values equally ($w_i = 1$, for all $i = 1, \ldots, d$), one can also construct a test

statistic based on a weighted product of the subsidiary $P$-values or a weighted sum of their logarithms:

$$S^* = S_{wF} = -2\sum_{i=1}^{d} w_i \ln P_i. \tag{6}$$

Again, $H_0$ is rejected whenever $S_{wF}$ is very large. The weights could reflect prior information (Good, 1955), could be based on the $P$-values themselves (Wilkinson, 1951), or on the standard error of estimated effect sizes (Whitlock, 2005) when, for instance, t-tests from multiple studies are to be combined. Dufour et al. (2015) consider one variant of this idea that we implement below. In particular, subsidiary $P$-values below a significance level $\alpha^*$ (we will choose $\alpha^* = \alpha$) receive a weight of 1 while insignificant $P$-values receive a zero weight. We label the corresponding test statistic $S_{wF(*)}$.[2] Since the weights depend on the subsidiary $P$-values, no analytic null distribution is available. The distribution has to be simulated if the subsidiary tests are independent and can be bootstrapped following Dufour et al. (2015) if they are dependent.

**$\chi^2$-transform test**

We now turn to alternative tests that are not commonly known in the econometrics literature in the context of testing combined hypotheses based on multiple subsidiary statistics. One of those tests applies a different transformation to the subsidiary $P$-values before aggregating them. The $\chi^2$-transform test (Lancaster, 1961) is given by

$$S^* = S_{\chi^2} = \sum_{i=1}^{d} F_{\chi^2}^{-1}\left(1 - P_i\right),$$

where $F_{\chi^2}^{-1}$ denotes the quantile function of the $\chi^2$ distribution with 1 degree of freedom. Obviously, the null distribution of $S_{\chi^2}$ is $\chi^2_d$ if the subsidiary $P$-values are independent. In case they are not, critical values and the $P$-value can be bootstrapped.

In addition, one can consider a weighted version of $S_{\chi^2}$, analogously to the weighted version of the Fisher test $S_{wF}$ described before. We also consider this option in our simulations below and denote it by $S_{\chi^2(*)}$.

**$Z$-transform test**

Originally proposed in Stouffer et al. (1949), this test addresses an asymmetry inherent to the Fisher test (Whitlock, 2005) by using the quantile function of a standard normal distribution, $\Phi^{-1}$, before aggregating the subsidiary $P$-values. The test statistic is given by

$$S^* = S_Z = \sum_{i=1}^{d} \Phi^{-1}\left(1 - P_i\right)/\sqrt{d}. \tag{7}$$

[2]Dufour et al. (2015) consider an alternative variant that includes the $\tilde{d}$ smallest $P$-values with a weight of 1 while discarding all other $P$-values. Since $d$ is small in our application, we do not consider this variant.

$H_0$ is rejected whenever $S_Z$ is large. If the subsidiary $P$-values are independent and $H_0$ is true, $S_Z$ follows a standard normal distribution. Otherwise its distribution and, thus, critical values and the $P$-value need to be bootstrapped.

**Gamma-transform test**

Finally, a recent proposal is to use the Gamma distribution for transforming the subsidiary $P$-values prior to their aggregation. In particular, Chen (2021) mentions an implementation that uses the subsidiary $P$-values to determine the shape parameters of the used Gamma distributions.The statistic is given by

$$S^* = S_G = \sum_{i=1}^{d} F^{-1}_{G(1/P_i;1)} \left(1 - P_i\right), \tag{8}$$

where $F^{-1}_{G(1/P_i;1)}$ is the quantile function of a Gamma distribution with shape parameter $1/P_i$ and rate parameter 1.[3] $H_0$ is rejected whenever $S_G$ is large. The null distribution of $S_G$ is unknown and has to be simulated if the subsidiary tests are independent or otherwise bootstrapped.

## 2.2 A Non-Parametric Test

King et al. (2020, KZA) suggest a testing framework that abstracts from combining the subsidiary $P$-values into one statistic. Their approach is based on a non-parametric estimation of the joint density $f(\mathbf{S})$ of the individual statistics. The region with the highest density determines the acceptance region of the test with probability $1 - \alpha$. Since the joint density is unknown in most situations, the authors recommend to estimate $f(\mathbf{S})$ using a kernel density estimator.[4] Inference is based on densities estimated with bootstrapped samples of $\mathbf{S}$, obtained under the overall null hypothesis. In a procedure called double simulation method (DSM), two independent samples of $m$ values of $\mathbf{S}$ are drawn, one to obtain an estimate of the joint density, $\hat{f}_m(\mathbf{S})$, and the other to evaluate the estimated density at values generated when $H_0$ is true. Alternatively, the single simulation method (SSM) is based on one sample of size $m$ of $\mathbf{S}$ under the null that is combined with a leave-one-out kernel density estimation of $f(\mathbf{S})$. Both methods produce $m$ values of the estimated joint density evaluated at draws of the vector of statistics under the null hypothesis. To calculate the $P$-value of the test, these values are compared to the estimated density evaluated at the observed value of $\mathbf{S}$. In particular, the $P$-value is the fraction of density evaluations under the null hypothesis that are smaller than the density evaluated at the vector of statistics based on the data sample used for testing, i.e., the fraction of those values from the simulation under $H_0$ with a

[3] Chen (2021) discusses that the test is equivalent for any value of the rate parameter and proposes a value of 1 for convenience.

[4] The kernel density estimator depends on choice of the bandwidth matrix $H$. We follow King et al. (2020) and use a diagonal bandwidth matrix, where the $i$th diagonal element is based on the estimated standard deviation of $S_i$.

lower density than the one actually observed:

$$\hat{P}_n^{King} = \frac{\sum_{j=1}^{m} \mathbb{1}\left(\hat{f}_m(\mathbf{S}_{H_0}^{(j)}) \leq \hat{f}_m(\mathbf{s})\right)}{n}, \tag{9}$$

where $\hat{f}_m()$ is evaluated at draws of $\mathbf{S}_{H_0}^{(j)}$ under the null hypothesis and at the observed vector of statistics $\mathbf{s} = (s_1, \ldots, s_d)'$, respectively. The test is exact when for a given significance level $\alpha$ the value $\alpha(m+1)$ is an integer. We consider the double simulation version of this test.

# 3 Application: Specification Tests

An obvious situation in which multiple test statistics have to be combined are diagnostic checks for regression specification. Such specification tests are commonly used to evaluate whether the assumptions, which are made to derive certain properties of estimators, hold true. There are several established individual tests available in many software packages to detect different forms of misspecification.

Our study considers four common deviations from the overall null hypothesis of a linear regression model with independent, homoskedastic, and standard normally distributed error terms: residual autocorrelation, heteroskedasticity, non-normality, and/or non-linearity. To test against each of these four alternatives, we use well-established test statistics by Ljung and Box (1978), Breusch and Pagan (1979), Neyman (1937), and Ramsey (1969), respectively. In our case, the vector of subsidiary statistics is, thus, of dimension $4 \times 1$.

Without loss of generality, we consider the simple linear regression model

$$y_t = \beta x_t + \varepsilon_t, \tag{10}$$

where we draw $x_t$ randomly from the $U(0,1)$ distribution, $\beta = 1$ and the random errors are Gaussian and independent of each other across time, $\varepsilon_t \overset{iid}{\sim} N(0,1)$. We assume throughout that the model's parameters are estimated by the OLS estimator.

We test the subsidiary hypothesis of no residual autocorrelation based on the test by Ljung and Box (1978), thus

$$S_1 = T(T+2)\sum_{k=1}^{L} \frac{\rho(k)^2}{(T-k)}, \tag{11}$$

where $\rho(k)$ is the $k$th order autocorrelation coefficient for k = 1, 2, $\ldots, L$. The null hypothesis of no autocorrelation is rejected for large values of $S_1$. Under the null hypothesis, $S_1$ asymptotically follows a $\chi^2$ distribution with $L$ degrees of freedom.

The second statistic in $\mathbf{S}$ is Breusch and Pagan's (1979) test statistic that allows testing the null hypothesis of no heteroskedasticity. It is based on the auxiliary regression $\hat{\varepsilon}_t^2 = \gamma_0 + \gamma_1 x_t + u_t$, where the

$\hat{\varepsilon}_t$ are the residuals from the regression model in (10). The test statistic is given by

$$S_2 = T\,R^2, \tag{12}$$

where $R^2$ is the coefficient of determination (R-squared) of the auxiliary regression. Under the null hypothesis, $S_2$ is $\chi^2$ distributed with degrees of freedom equal to the number of predictors in the linear model (in our case 1).

The smooth test by Neyman (1937) is commonly used to detect deviations from the null hypothesis of normally distributed errors. Since the test actually tests for uniformity on [0, 1], the residuals from the regression model in (10) need to be transformed first. Specifically, we calculate $Z_t = \Phi(\hat{\varepsilon}_t)$ where $\Phi$ is the cumulative distribution function of the standard normal distribution. Hence, $Z_t \overset{iid}{\sim} U[0,1]$ under $H_0$. Neyman's test statistic is then

$$S_3 = \sum_{j=1}^{4} \left( T^{-1/2} \sum_{t=1}^{T} b_j(Z_t) \right)^2, \tag{13}$$

where $b_1$, $b_2$, $b_3$, and $b_4$ are the normalized Legendre polynomials on [0, 1]. When the hypothesis of normally distributed error terms is true, the test statistic is $\chi^2$ distributed with four degrees of freedom (the number of Legendre polynomials included). The null hypothesis is rejected for large values of $S_3$.

Finally, we use Ramsey's (1969) Regression Equation Specification Error Test (RESET) to identify a potential nonlinear relationship between $y_t$ and $x_t$ in (10). The test compares the explanatory power of the original model with the explanatory power of a model that includes non-linear terms. These additional regressors are added in the form of a polynomial of the fitted values from the original model. The auxiliary regression model with one squared fitted value is given by: $y_t = x_t\alpha + \hat{y}_t^2\gamma + \zeta$, where the $\hat{y}_t$ denotes the fitted value. We use a Wald test to evaluate the null hypothesis that the coefficient $\gamma$ is equal to zero. Thus, the test statistic is

$$S_4 = \hat{\gamma}' \left( \begin{bmatrix} 0 & 1 \end{bmatrix} \Sigma_{\hat{\gamma}} \begin{bmatrix} 0 & 1 \end{bmatrix}' \right)^{-1} \hat{\gamma}, \tag{14}$$

where $\Sigma_{\hat{\gamma}}$ is the covariance estimate of the auxiliary regression model. Under the null hypothesis, $S_4$ follows a $\chi^2$ distribution with one degree of freedom.

The distribution of the four test statistics in $\mathbf{S}$ are available in closed form under the respective subsidiary null hypotheses. Still, we use a bootstrap building on Dufour et al. (2015) to approximate the null distribution of the combined parametric tests because we want to account for potential dependence between the four subsidiary tests. To do so, we simulate $100,000$ data sets according to the DGP assumed under $H_0$ and described by equation (3). For each of these data sets, we then estimate the null model and compute the four tests and their corresponding $P$-values. In a last step, this allows us to compute

100,000 draws of the combined test statistics from Section 2. We can use the empirical distribution function of those draws obtained under the overall null hypothesis to make test decisions in the Monte Carlo simulations described in the next section.

# 4 Monte Carlo Simulations

## 4.1 The Setup

In accordance with the four test statistics in vector $\mathbf{S}$, we examine four departures from the null hypothesis of i.i.d. errors. These departures are autocorrelated errors, heteroskedasticity, non-normal errors, and non-linear associations between the dependent and independent variables in the regression (10). We explore each departure individually as well as combinations of the four deviations. Additionally, we include a randomized simulation design where each departure occurs with some predetermined probability.

To assess the empirical sizes of the testing approaches, we simulate errors from the standard normal distribution, i.e. $\varepsilon_t \sim N(0,1)$. To analyze their power, we use the following four DGPs for $\varepsilon_t$:

1. We model serial correlation as a stationary AR(1) process. The error term is defined as $\varepsilon_t = \rho\varepsilon_{t-1} + u_t$, where $u_t \sim N(0,1)$ and $\rho$ is the autocorrelation coefficient. We set $\rho = 0.15$. When simulating the autocorrelated errors, a burn-in sample of 100 observations is used throughout.

2. We model heteroskedasticity such that $\varepsilon_t$ has a time-varying variance $Var(\varepsilon_t|x_t) = \sigma^2 h_t$. We define $h_t$ such that $Var(\varepsilon_t|x_t)$ increases with $t$. In particular, we assume $\varepsilon_t = (e_{1,i(t)} + e_{2,t})/\sqrt{3}$, where $e_{1,t} \sim N(0,1)$ and $e_{2,t} \sim N(0,2)$ and $i(t)$ is defined according to $i(t) = rank\,(z \cdot \phi(e_{1,t}))$ with $z \sim U[0,1]$ and $\phi$ being the probability density function of the standard normal distribution. The index $i(t)$ ensures that larger absolute values of $e_{1,t}$ tend to occur more frequently for higher $t$. Finally, we sort the $x$ values by size such that the variance of $\varepsilon_t$ increases with $x_t$.

3. To generate non-normal errors, we draw values from Student's $t$-distribution, i.e., $\varepsilon_t \sim t_k$ where $k$ are the degrees of freedom. We set $k = 7$.

4. To induce a nonlinear relationship between $y_t$ and $x_t$ in deviation from the regression model (10), we specify the errors as $\varepsilon_t = x_t^2\delta + u_t$, where $\delta = 1.5$ and $u_t \sim N(0,1)$.

In a first step, we simulate four different DGPs, each exhibiting one of the deviations from the null. The standard tests hold a power advantage in these scenarios, as they are specifically designed to detect a single deviation from the null hypothesis in the data. In the second step, we generate data where two of the aforementioned deviations are present, resulting in a total of six distinct DGPs. The data are generated according to the above description except when simulating the DGP with heteroskedastic and non-normal errors. In this case, we set $e_1 \sim t_7$ and $e_2 \sim t_3$ in order to regulate the strength

of the deviation from the null. We keep this specification for all combined alternatives that include heteroskedasticity and non-normality. Next, we consider situations in which three of the four specified deviations coincide. This results in a total of four different DGPs. The final data is generated such that all four deviations from the null are present simultaneously.

To create an alternative where it is unknown which and how many of the deviations from the overall null hypothesis are behind the DGP, we use a randomized simulation design. The number and type of deviations from the null are chosen at random from the four deviations above. Each one is selected according to a Bernoulli process with some probability. We set this probability to either 0.5, 0.7, or 0.9. We exclude cases where none of the deviations are chosen, i.e., we avoid that the errors are in full accordance with the overall null hypothesis.

All simulations are executed with $n = 20{,}000$ Monte Carlo iterations. We consider sample sizes $T = \{50, 100, 200\}$ and choose a nominal significance level of $\alpha = 0.05$ throughout.

## 4.2 Results: Size

Tables 1 and 2 show the empirical size for each test obtained for different sample sizes. The first table presents sizes for the individual tests (Ljung-Box test, Breusch-Pagan test, Neyman-Smooth test, and the RESET test); the second table presents sizes for the alternative ways of aggregating these four tests into one (as described in Section 2).

Overall and despite minor differences, all tests achieve empirical sizes very close to the nominal value of 0.05. This is true for all sample sizes that we consider. Hence, all test approaches are able to correctly control the probability of type I errors.[5]

Table 1: Sizes of standard tests

| T | Ljung-Box | Breusch-Pagan | Neymann-Smooth | RESET |
|---|---|---|---|---|
| 50 | 0.055 | 0.057 | 0.050 | 0.056 |
| 100 | 0.047 | 0.054 | 0.049 | 0.053 |
| 200 | 0.050 | 0.048 | 0.048 | 0.051 |

*Notes:* Table shows empirical rejection frequencies under the null hypothesis as described in Section 2 for a nominal size of $\alpha = 0.05$. The number of Monte-Carlo iterations is $n = 20,000$. For each sample size, the highest power across all tests is highlighted in bold.

## 4.3 Results: Power

Turning to the power results, we start with the simulations that include only one single type of deviation from the overall null hypothesis. Comparing the rejection frequencies from Table 3 with those in Table 4

[5]The sizes (and power results below) are very similar when we rely on analytical null distributions whenever available. This is due to the fact that the correlations between any of the individual test statistics are small in absolute values under $H_0$ in our setup.

Table 2: Sizes of standard tests

| T | $S_{min}$ | $S_F$ | $S_{wF(*)}$ | $S_{\chi^2}$ | $S_{\chi^2(*)}$ | $S_Z$ | $S_G$ | KZA |
|---|---|---|---|---|---|---|---|---|
| 50 | 0.050 | 0.052 | 0.050 | 0.052 | 0.050 | 0.051 | 0.050 | 0.052 |
| 100 | 0.051 | 0.051 | 0.051 | 0.051 | 0.051 | 0.051 | 0.051 | 0.052 |
| 200 | 0.051 | 0.049 | 0.049 | 0.049 | 0.049 | 0.050 | 0.050 | 0.049 |

*Notes:* Table shows empirical rejection frequencies under the null hypothesis as described in Section 2 for a nominal size of $\alpha = 0.05$. The number of Monte-Carlo iterations is $n = 20,000$. For each sample size, the highest power across all tests is highlighted in bold.

it is evident—and not surprising—that the specialized individual tests outperform the combined tests. Thus, if one knows that, say, residual autocorrelation is the only potential deviation from $H_0$, then relying on the Ljung-Box test is, of course, preferable to using any of the tests that combine test statistics related to all four potential deviations from $H_0$. Among the combined tests, the test by King et al. (2020) has the highest power for all deviations but non-normally distributed error terms. In the latter case, the test from Chen (2021) that transforms the subsidiary $P$-values using the Gamma distribution function before aggregation and the weighted versions of the product of P-values and the $\chi^2$-transform tests, $S_{wF(*)}$ and $S_{\chi^2(*)}$, perform best among the combined tests. Hence, even with only one deviation relying on $S_{min}$ does not seem to be optimal.

Table 3: Powers of standard tests - single deviation

| T | Ljung-Box | Breusch-Pagan | Neyman Smooth | RESET |
|---|---|---|---|---|
| Panel A: Autocorrelation | | | | |
| 50 | **0.134** | 0.051 | 0.050 | 0.062 |
| 100 | **0.276** | 0.051 | 0.055 | 0.058 |
| 200 | **0.528** | 0.051 | 0.054 | 0.060 |
| Panel B: Heteroskedasticity | | | | |
| 50 | 0.056 | **0.197** | 0.047 | 0.072 |
| 100 | 0.056 | **0.380** | 0.050 | 0.071 |
| 200 | 0.057 | **0.679** | 0.050 | 0.067 |
| Panel C: Non-normality | | | | |
| 50 | 0.047 | 0.048 | **0.203** | 0.054 |
| 100 | 0.049 | 0.049 | **0.322** | 0.053 |
| 200 | 0.051 | 0.049 | **0.553** | 0.052 |
| Panel D: Non-linearity | | | | |
| 50 | 0.048 | 0.050 | 0.050 | **0.224** |
| 100 | 0.052 | 0.047 | 0.051 | **0.385** |
| 200 | 0.052 | 0.047 | 0.054 | **0.654** |

*Notes:* This table presents the powers of standard tests when $\alpha = 0.05$. The number of Monte-Carlo iterations is $n = 20,000$. For each sample size, the highest power across all tests is highlighted in bold.

Next, we consider the situation when the DGP has deviations along two dimensions from the null hypothesis. Table 5 shows the power of the individual specialized tests for any combinations of two deviations from $H_0$. In each of the scenarios, the power of tests that are designed to detect one of the two deviations is high. Interestingly, however, the power is smaller (with few exceptions) than in the

Table 4: Powers of combined tests - single deviation

| T | $S_{min}$ | $S_F$ | $S_{wF(*)}$ | $S_{\chi^2}$ | $S_{\chi^2(*)}$ | $S_Z$ | $S_G$ | KZA |
|---|---|---|---|---|---|---|---|---|
| Panel A: Autocorrelation | | | | | | | | |
| 50 | 0.086 | 0.089 | 0.087 | 0.091 | 0.087 | 0.083 | 0.087 | **0.093** |
| 100 | 0.163 | 0.165 | 0.167 | 0.167 | 0.167 | 0.137 | 0.165 | **0.177** |
| 200 | 0.347 | 0.315 | 0.349 | 0.331 | 0.349 | 0.231 | 0.349 | **0.365** |
| Panel B: Heteroskedasticity | | | | | | | | |
| 50 | 0.112 | 0.122 | 0.117 | 0.123 | 0.117 | 0.113 | 0.115 | **0.128** |
| 100 | 0.226 | 0.215 | 0.234 | 0.225 | 0.234 | 0.173 | 0.231 | **0.252** |
| 200 | 0.488 | 0.436 | 0.490 | 0.461 | 0.490 | 0.300 | 0.493 | **0.515** |
| Panel C: Non-normality | | | | | | | | |
| 50 | 0.137 | 0.129 | 0.137 | 0.134 | 0.137 | 0.107 | **0.138** | 0.120 |
| 100 | 0.219 | 0.209 | **0.221** | 0.216 | **0.221** | 0.156 | 0.221 | 0.192 |
| 200 | 0.409 | 0.373 | 0.411 | 0.391 | 0.411 | 0.262 | **0.412** | 0.360 |
| Panel D: Non-linearity | | | | | | | | |
| 50 | 0.131 | 0.129 | 0.132 | **0.133** | 0.132 | 0.107 | **0.133** | 0.132 |
| 100 | 0.234 | 0.218 | 0.236 | 0.226 | 0.236 | 0.164 | 0.237 | **0.243** |
| 200 | 0.464 | 0.413 | 0.467 | 0.439 | 0.467 | 0.274 | 0.469 | **0.477** |

*Notes:* This table presents the powers of combined tests when $\alpha = 0.05$. The number of Monte-Carlo iterations is $n = 20,000$. For each sample size, the highest power across all tests is highlighted in bold.

previous cases with only one deviation from $H_0$. Thus, the probability of type II errors associated with these specialized tests increases in most cases once another deviation is added.

In contrast, the power of all combined tests increases without exception when a second deviation is added (Table 6). Overall, when $T > 50$, the power of these combined tests is now higher than that of the specialized tests in Table 5. In particular, the highest power is obtained by a combined test and most combined tests have higher power than the best specialized test, Neyman's smooth test being the only individual test that is competitive in most scenarios and for small samples. While this is not too surprising, it is interesting to analyze which of the combined testing approaches perform best. Across all combinations of deviations from $H_0$, $S_F$ and $S_{\chi^2}$ exhibit the highest power, the differences in rejection frequencies between the two being only marginal. The next best tests are their weighted versions, $S_{wF(*)}$ and $S_{\chi^2(*)}$, and the non-parametric test by King et al. (2020). $S_G$ also performance at par with those tests except for the setup with residual autocorrelation in combination with heteroskedasticity. $S_Z$ and $S_{min}$ have the lowest power on average. Interestingly, there are no major differences in the relative ranking (according to power) of the combined tests between different sample sizes.

In the next step, we consider alternative DGPs that deviate along three dimensions from the overall null hypothesis. We display the results regarding the power of the combined tests in Panels A to D in Table 7. Note that we put further results for the individual subsidiary test statistics into the appendix since their relative performance only deteriorates further when more deviations from $H_0$ are added; Neyman's Smooth test is the only one that reaches power almost at par with some of the combined tests in some setups. The ranking of the combined tests remains largely unchanged: again, $S_F$ and $F_{\chi^2}$ have the largest power. Naturally, $S_{min}$ falls back because it uses only information from one subsidiary

Table 5: Powers of standard tests - two deviations

| T | Ljung-Box | Breusch-Pagan | Neyman Smooth | RESET |
|---|---|---|---|---|
| Panel A: Autocorrelation and Heteroskedasticity | | | | |
| 50 | 0.127 | 0.118 | **0.206** | 0.110 |
| 100 | **0.261** | 0.255 | 0.201 | 0.108 |
| 200 | 0.512 | **0.532** | 0.176 | 0.112 |
| Panel B: Autocorrelation and Non-normality | | | | |
| 50 | 0.135 | 0.049 | **0.219** | 0.059 |
| 100 | 0.271 | 0.048 | **0.355** | 0.057 |
| 200 | 0.532 | 0.047 | **0.604** | 0.054 |
| Panel C: Autocorrelation and Non-linearity | | | | |
| 50 | 0.127 | 0.070 | 0.185 | **0.292** |
| 100 | 0.252 | 0.076 | 0.324 | **0.495** |
| 200 | 0.487 | 0.099 | 0.575 | **0.785** |
| Panel D: Heteroskedasticity and Non-normality | | | | |
| 50 | 0.058 | **0.266** | 0.146 | 0.089 |
| 100 | 0.061 | **0.394** | 0.221 | 0.083 |
| 200 | 0.062 | **0.526** | 0.376 | 0.084 |
| Panel E: Heteroskedasticity and Non-linearity | | | | |
| 50 | 0.056 | 0.190 | 0.053 | **0.243** |
| 100 | 0.061 | 0.385 | 0.052 | **0.396** |
| 200 | 0.071 | **0.670** | 0.051 | 0.650 |
| Panel F: Non-normality and Non-linearity | | | | |
| 50 | 0.048 | 0.044 | **0.215** | 0.178 |
| 100 | 0.051 | 0.046 | **0.354** | 0.294 |
| 200 | 0.047 | 0.046 | **0.604** | 0.513 |

*Notes:* This table presents the powers of standard tests when $\alpha = 0.05$. The number of Monte-Carlo iterations is $n = 20,000$. For each sample size, the highest power across all tests is highlighted in bold.

$P$-value, failing to incorporate information provided by the other subsidiary statistics.

Finally, we consider the situation in which all four deviations from $H_0$ become relevant. The results in Panel E in Table 7 show the same overall picture. When $T = 200$, the power of $S_F$ and $S_{\chi^2}$ is highest (around 0.93) and roughly 7 percentage points above the power of $S_{min}$.

In addition to the alternatives discussed so far, we consider the (to some extent more realistic) situation where the dimensions along which the DGP deviates from $H_0$ are randomly chosen, i.e., unknown. We assume that each deviation occurs with a probability of 0.5, 0.7, or 0.9 and discard cases where no deviation is chosen. The overall conclusion remains unchanged. Power increases in the probability by which deviations occur (Table 8). $S_F$ and $S_{\chi^2}$ perform best. The non-parametric approach by King et al. (2020) follows at par with $S_{wF(*)}$, $S_{w\chi^2(*)}$, and $S_G$. $S_Z$ has similar power in the case of large samples and a high probability of deviations. The power of $S_{min}$ remains much lower. The differences between the best and worst performance increase slightly in the probability by which deviations occur. Again, the individual subsidiary tests are not competitive and results are in the appendix.

In summary, we draw the following general conclusions from our simulation exercise. First, the

Table 6: Powers of combined tests - two deviations

| T | $S_{min}$ | $S_F$ | $S_{wF(*)}$ | $S_{\chi^2}$ | $S_{\chi^2(*)}$ | $S_Z$ | $S_G$ | KZA |
|---|---|---|---|---|---|---|---|---|
| Panel A: Autocorrelation and Heteroskedasticity | | | | | | | | |
| 50 | 0.198 | **0.250** | 0.214 | 0.245 | 0.214 | 0.239 | 0.212 | 0.206 |
| 100 | 0.302 | **0.395** | 0.339 | 0.391 | 0.339 | 0.378 | 0.329 | 0.344 |
| 200 | 0.563 | **0.667** | 0.612 | 0.666 | 0.612 | 0.621 | 0.605 | 0.624 |
| Panel B: Autocorrelation and Non-normality | | | | | | | | |
| 50 | 0.176 | 0.193 | 0.184 | **0.194** | 0.184 | 0.167 | 0.183 | 0.172 |
| 100 | 0.321 | 0.360 | 0.341 | **0.363** | 0.341 | 0.312 | 0.335 | 0.321 |
| 200 | 0.608 | 0.659 | 0.639 | **0.667** | 0.639 | 0.576 | 0.631 | 0.613 |
| Panel C: Autocorrelation and Non-linearity | | | | | | | | |
| 50 | 0.261 | **0.312** | 0.282 | 0.309 | 0.282 | 0.286 | 0.276 | 0.272 |
| 100 | 0.486 | **0.591** | 0.530 | 0.587 | 0.530 | 0.553 | 0.521 | 0.524 |
| 200 | 0.813 | **0.899** | 0.853 | 0.897 | 0.853 | 0.871 | 0.849 | 0.852 |
| Panel D: Heteroskedasticity and Non-normality | | | | | | | | |
| 50 | 0.208 | 0.234 | 0.223 | **0.237** | 0.223 | 0.202 | 0.220 | 0.227 |
| 100 | 0.345 | 0.369 | 0.363 | **0.376** | 0.363 | 0.309 | 0.359 | 0.363 |
| 200 | 0.529 | 0.559 | 0.552 | **0.570** | 0.552 | 0.463 | 0.550 | 0.540 |
| Panel E: Heteroskedasticity and Non-linearity | | | | | | | | |
| 50 | 0.197 | 0.222 | 0.210 | **0.223** | 0.210 | 0.195 | 0.208 | 0.216 |
| 100 | 0.386 | 0.431 | 0.414 | **0.437** | 0.414 | 0.364 | 0.406 | 0.429 |
| 200 | 0.710 | 0.757 | 0.740 | **0.765** | 0.740 | 0.656 | 0.736 | 0.755 |
| Panel F: Non-normality and Non-linearity | | | | | | | | |
| 50 | 0.198 | 0.207 | 0.204 | **0.210** | 0.204 | 0.174 | 0.203 | 0.185 |
| 100 | 0.335 | 0.367 | 0.353 | **0.372** | 0.353 | 0.312 | 0.347 | 0.329 |
| 200 | 0.615 | 0.657 | 0.640 | **0.663** | 0.640 | 0.563 | 0.638 | 0.614 |

*Notes:* This table presents the powers of combined tests when $\alpha = 0.05$. The number of Monte-Carlo iterations is $n = 20,000$. For each sample size, the highest power across all tests is highlighted in bold.

minimum $P$-value test, $S_{min}$, that is widely used is not an optimal choice in the context of specification testing. Second, instead, the Fisher test, $S_F$, that uses the product of all subsidiary $P$-values and the $\chi^2$-transform test, $S_{\chi^2}$ seem to perform best—independently of the sample size and the exact number of deviations from the overall null. Third, the recently proposed Gamma-transform test, $S_G$, that has not yet been used much in the econometrics literature performs decently, too. Fourth, the best parametric combination tests outperform the non-parametric approach by King et al. (2020). This is fortunate since even though simulation is required to approximate the distribution of the overall test statistic the computational burden is smaller than in case of the non-parametric approach.

# 5 Conclusion

This paper reviews various approaches for testing an overall null hypothesis that consists of multiple subsidiary hypotheses and uses a Monte Carlo simulation to study the tests' performance in the context of specification testing for linear regressions. The review includes tests that are not commonly used in the econometrics literature but might be valuable for future research since they perform well in our simulation. Most approaches that we review imply combining the individual test statistics or corresponding $P$-values

Table 7: Powers of combined tests - three and four deviations

| T | $S_{min}$ | $S_F$ | $S_{wF(*)}$ | $S_{\chi^2}$ | $S_{\chi^2(*)}$ | $S_Z$ | $S_G$ | KZA |
|---|---|---|---|---|---|---|---|---|
| Panel A: Autocorrelation, Non-normality and Heteroskedasticity | | | | | | | | |
| 50 | 0.516 | 0.571 | 0.541 | **0.573** | 0.541 | 0.505 | 0.539 | 0.521 |
| 100 | 0.625 | **0.705** | 0.661 | 0.702 | 0.661 | 0.650 | 0.655 | 0.648 |
| 200 | 0.779 | **0.859** | 0.814 | 0.855 | 0.814 | 0.817 | 0.812 | 0.808 |
| Panel B: Autocorrelation, Non-normality and Non-linearity | | | | | | | | |
| 50 | 0.239 | **0.276** | 0.256 | 0.275 | 0.256 | 0.247 | 0.252 | 0.240 |
| 100 | 0.431 | **0.517** | 0.465 | 0.514 | 0.465 | 0.480 | 0.456 | 0.449 |
| 200 | 0.745 | **0.841** | 0.788 | 0.838 | 0.788 | 0.804 | 0.783 | 0.777 |
| Panel C: Autocorrelation, Heteroskedasticity and Non-linearity | | | | | | | | |
| 50 | 0.270 | **0.340** | 0.293 | 0.337 | 0.293 | 0.327 | 0.290 | 0.287 |
| 100 | 0.444 | **0.570** | 0.492 | 0.562 | 0.492 | 0.556 | 0.484 | 0.497 |
| 200 | 0.748 | **0.862** | 0.803 | 0.857 | 0.803 | 0.844 | 0.796 | 0.810 |
| Panel D: Heteroskedasticity, Non-normality and Non-linearity | | | | | | | | |
| 50 | 0.268 | 0.312 | 0.289 | **0.314** | 0.289 | 0.277 | 0.285 | 0.291 |
| 100 | 0.449 | 0.517 | 0.485 | **0.518** | 0.485 | 0.458 | 0.476 | 0.484 |
| 200 | 0.678 | **0.748** | 0.711 | **0.748** | 0.711 | 0.695 | 0.710 | 0.708 |
| Panel E: Autocorrelation, Heteroskedasticity, Non-normality and Non-linearity | | | | | | | | |
| 50 | 0.553 | 0.619 | 0.582 | **0.619** | 0.582 | 0.557 | 0.578 | 0.559 |
| 100 | 0.698 | **0.784** | 0.738 | 0.781 | 0.738 | 0.742 | 0.731 | 0.727 |
| 200 | 0.856 | **0.928** | 0.892 | 0.925 | 0.892 | 0.913 | 0.887 | 0.886 |

*Notes:* This table presents the powers of combined tests when $\alpha = 0.05$. The number of Monte-Carlo iterations is $n = 20,000$. For each sample size, the highest power across all tests is highlighted in bold.

Table 8: Powers of combined tests - random Alternative

| T | $p$ | $S_{min}$ | $S_F$ | $S_{wF(*)}$ | $S_{\chi^2}$ | $S_{\chi^2(*)}$ | $S_Z$ | $S_G$ | KZA |
|---|---|---|---|---|---|---|---|---|---|
| 50 | 0.5 | 0.261 | 0.285 | 0.274 | **0.288** | 0.274 | 0.254 | 0.271 | 0.263 |
| 100 | 0.5 | 0.410 | 0.456 | 0.431 | **0.457** | 0.431 | 0.408 | 0.429 | 0.426 |
| 200 | 0.5 | 0.652 | 0.693 | 0.678 | **0.697** | 0.678 | 0.619 | 0.676 | 0.676 |
| 50 | 0.7 | 0.345 | 0.385 | 0.363 | **0.386** | 0.363 | 0.344 | 0.360 | 0.347 |
| 100 | 0.7 | 0.503 | **0.580** | 0.535 | 0.578 | 0.535 | 0.534 | 0.531 | 0.529 |
| 200 | 0.7 | 0.733 | **0.799** | 0.766 | 0.798 | 0.766 | 0.748 | 0.764 | 0.764 |
| 50 | 0.9 | 0.477 | 0.530 | 0.502 | **0.531** | 0.502 | 0.480 | 0.497 | 0.482 |
| 100 | 0.9 | 0.616 | **0.704** | 0.652 | 0.700 | 0.652 | 0.664 | 0.648 | 0.642 |
| 200 | 0.9 | 0.811 | **0.894** | 0.850 | 0.891 | 0.850 | 0.866 | 0.844 | 0.848 |

*Notes:* This table presents the powers of combined tests when $\alpha = 0.05$. The number of Monte-Carlo iterations is $n = 20,000$. For each sample size, the highest power across all tests is highlighted in bold.

related to the subsidiary null hypotheses into one aggregate test statistic.

In contrast to many other fields (e.g., bio statistic), where most applications can rely on the assumption of independence of the subsidiary test statistics (e.g., because results from independent experiments are to be aggregated), we take dependencies into account and rely on simulations to calculate overall $P$-values following Godfrey (2005) and Dufour et al. (2015).

We find that the combined tests outperform the individual tests tailored to one particular deviation from the null hypothesis (e.g., the Ljung-Box test for residual autocorrelation) as soon as there are two types of deviation from the overall null hypothesis. Those approaches that combine information from

all subsidiary tests (especially $S_F$ and $S_{\chi^2}$) perform substantially better than the minimum $P$-value test, $S_{min}$, in most cases and equally well if only exactly one deviation is present. This suggest that these alternatives are preferable to the widely used minimum $P$-value test because it is likely that they outperform also in other applications where $d$ is reasonably large and multiple deviations from the overall null hypothesis are likely.

In particular, the version of the Gamma-transform test (Chen, 2021) that we consider performs reasonably well. This approach is particularly promising for other applications with overall null hypotheses formed by a large number of subsidiary hypotheses, i.e., with $d >> 1$. Chen (2021) shows that a modified version of this approach, in which the parameters of the Gamma distribution used to transform the subsidiary $P$-values are optimally chosen using a maximum-likelihood estimation based on the sample of subsidiary $P$-values is universally most powerful under certain assumptions about the distribution of subsidiary $P$-values under the alternative. Note that we cannot implement this test due to the small number of subsidiary $P$-values ($d = 4$), but other applications in econometrics—like multivariate goodness-of-fit tests in the context of, for instance, evaluation of multivariate density forecasts (Diebold et al., 1999; Dovern and Manner, 2020)—potentially imply a much larger $d$.

Another message from our simulations is that a number of parametric tests performs equally well or even better than the non-parametric approach by King et al. (2020) that is computationally more demanding. While this may not be true in all contexts—after all King et al. (2020) demonstrate the merits of their approach, i.a., when testing for residual autocorrelation or heteroskedasticity—it is a potentially very valuable advantage for applications with a large number of subsidiary null hypotheses.

# References


Breusch, T. S. and A. R. Pagan (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica 47*(5), 1287–1294.

Chen, Z. (2021). Optimal tests for combining p-values. *Applied Sciences 12*(1), 322.

Diebold, F. X., J. Hahn, and A. S. Tay (1999). Multivariate density forecast evaluation and calibration in financial risk management: high-frequency returns on foreign exchange. *Review of Economics and Statistics 81*(4), 661–673.

Dovern, J. and H. Manner (2020). Order-invariant tests for proper calibration of multivariate density forecasts. *Journal of Applied Econometrics 35*(4), 440–456.

Dufour, J.-M., L. Khalaf, and M. Voia (2015). Finite-sample resampling-based combined hypothesis tests, with applications to serial correlation and predictability. *Communications in Statistics-Simulation and Computation 44*(9), 2329–2347.

Fisher, R. A. (1932). *Statistical methods for research workers*. Springer.

Godfrey, L. (2009). *Bootstrap tests for regression models*. Springer.

Godfrey, L. G. (2005). Controlling the overall significance level of a battery of least squares diagnostic tests. *Oxford Bulletin of Economics and Statistics 67*(2), 263–279.

Good, I. (1955). On the weighted combination of significance tests. *Journal of the Royal Statistical Society: Series B (Methodological) 17*(2), 264–265.

King, M. L., X. Zhang, and M. Akram (2020). Hypothesis testing based on a vector of statistics. *Journal of Econometrics 219*(2), 425–455.

Lancaster, H. O. (1961). The combination of probabilities: an application of orthonormal functions. *Australian Journal of Statistics 3*(1), 20–33.

Ljung, G. M. and G. E. Box (1978). On a measure of lack of fit in time series models. *Biometrika 65*(2), 297–303.

Neyman, J. (1937). Smooth test for goodness of fit. *Scandinavian Actuarial Journal 1937*(3-4), 149–199.

Ramsey, J. B. (1969). Tests for specification errors in classical linear least-squares regression analysis. *Journal of the Royal Statistical Society: Series B (Methodological) 31*(2), 350–371.

Stouffer, S. A., E. A. Suchman, L. C. Devinney, S. A. Star, and R. M. Williams Jr. (1949). *The American soldier: adjustment during army life. Studies in social psychology in World War II*. Princeton University Press.

Tippett, L. H. C. et al. (1931). *The methods of statistics*. London: Williams & Norgate Ltd.

Whitlock, M. C. (2005). Combining probability from independent tests: the weighted z-method is superior to fisher's approach. *Journal of Evolutionary Biology 18*(5), 1368–1373.

Wilkinson, B. (1951). A statistical consideration in psychological research. *Psychological Bulletin 48*(2), 156–158.

# Appendix - Additional Simulation Results

Table A.1: Powers of standard tests - three and four deviations

| T | Ljung-Box | Breusch-Pagan | Neyman Smooth | RESET |
|---|---|---|---|---|
| Panel A: Autocorrelation, Non-normality and Heteroskedasticity | | | | |
| 50 | 0.118 | 0.255 | **0.573** | 0.133 |
| 100 | 0.253 | 0.398 | **0.604** | 0.132 |
| 200 | 0.502 | 0.541 | **0.661** | 0.131 |
| Panel B: Autocorrelation, Non-normality and Non-linearity | | | | |
| 50 | 0.128 | 0.047 | **0.241** | 0.185 |
| 100 | 0.262 | 0.047 | **0.395** | 0.296 |
| 200 | 0.510 | 0.046 | **0.659** | 0.507 |
| Panel C: Autocorrelation, Heteroskedasticity and Non-linearity | | | | |
| 50 | 0.149 | 0.120 | 0.227 | **0.233** |
| 100 | 0.318 | 0.257 | 0.239 | **0.365** |
| 200 | **0.606** | 0.539 | 0.228 | 0.585 |
| Panel D: Heteroskedasticity, Non-normality and Non-linearity | | | | |
| 50 | 0.059 | **0.266** | 0.154 | 0.215 |
| 100 | 0.062 | **0.396** | 0.239 | 0.329 |
| 200 | 0.067 | **0.534** | 0.395 | 0.520 |
| Panel E: Autocorrelation, Heteroskedasticity, Non-normality and Non-linearity | | | | |
| 50 | 0.132 | 0.251 | **0.594** | 0.213 |
| 100 | 0.304 | 0.398 | **0.645** | 0.309 |
| 200 | 0.570 | 0.541 | **0.710** | 0.474 |

*Notes:* This table presents the powers of standard tests when $\alpha = 0.05$. The number of Monte-Carlo iterations is $n = 20,000$. For each sample size, the highest power across all tests is highlighted in bold.

Table A.2: Powers of standard tests - random alternative

| T | Prob. | Ljung-Box | Breusch-Pagan | Neyman Smooth | RESET |
|---|---|---|---|---|---|
| 50 | 0.5 | 0.099 | 0.138 | **0.241** | 0.157 |
| 100 | 0.5 | 0.171 | 0.210 | **0.313** | 0.235 |
| 200 | 0.5 | 0.312 | 0.334 | **0.423** | 0.349 |
| 50 | 0.7 | 0.108 | 0.173 | **0.339** | 0.184 |
| 100 | 0.7 | 0.217 | 0.276 | **0.411** | 0.267 |
| 200 | 0.7 | 0.397 | 0.402 | **0.516** | 0.413 |
| 50 | 0.9 | 0.125 | 0.225 | **0.489** | 0.210 |
| 100 | 0.9 | 0.268 | 0.354 | **0.546** | 0.295 |
| 200 | 0.9 | 0.513 | 0.496 | **0.632** | 0.460 |

*Notes:* This table presents the powers of individual tests when $\alpha = 0.05$. The number of Monte-Carlo iterations is $n = 20,000$. For each sample size, the highest power across all tests is highlighted in bold.