UNIVERSITY OF CALIFORNIA, SAN DIEGO

Limit Theory for Overfit Models

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

in

Economics

by

Grayson Ford Calhoun

Committee in charge:

Professor Graham Elliott, Co-Chair
Professor Allan Timmermann, Co-Chair
Professor Dimitris Politis
Professor Yixiao Sun
Professor Rossen Valkanov

2009

The dissertation of Grayson Ford Calhoun is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____
Co-Chair

_____
Co-Chair

University of California, San Diego

2009

DEDICATION

To Jess and Blair.

EPIGRAPH

A foolish consistency is the hobgoblin of little minds, . . .

—Ralph Waldo Emerson

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Thank you everyone, especially Jess.

VITA

2001	B. S. in Mathematics, Tufts University
2006	M. S. in Statistics, University of California, San Diego
2009	Ph. D. in Economics, University of California, San Diego

ABSTRACT OF THE DISSERTATION

Limit Theory for Overfit Models

by

Grayson Ford Calhoun
Doctor of Philosophy in Economics
University of California San Diego, 2009
Professor Graham Elliott, Co-Chair
Professor Allan Timmermann, Co-Chair

This dissertation consists of three independent papers. Collectively they attempt to formalize a notion of model "overfit" – the idea that a large econometric model can appear to fit a particular dataset well simply because it is large. This behavior is modeled by using asymptotic approximations that allow the number of regressors in a linear regression model to increase with the number of total observations. The first chapter looks at the behavior of the F-test under this asymptotic theory, shows that the F-test is generally invalid for these overfit models, and derives a correction that gives a valid test statistic. The second chapter looks at the behavior of pseudo out-of-sample comparisons of forecasting models under this asymptotic theory, shows that this asymptotic theory resolves some technical issues that lead to nonstandard test statistics, and gives conditions under which standard procedures remain valid for overfit models. The third chapter conducts an empirical comparison of several methods for comparing forecasting models out-of-sample.

Chapter 1

Hypothesis testing in linear regression when k/n is large

1.1 Introduction

Consider the linear regression model

$$y_t = x_t'\beta + \varepsilon_t \quad t = 1, \ldots, n \tag{1.1.1}$$

with x_t and ε_t uncorrelated. Under standard assumptions, the OLS estimator, $\hat{\beta}$, is consistent and asymptotically normal as n increases to infinity. This asymptotic distribution is the basis for most of the empirical research in economics, but, as Huber (1973) has shown, this approximation is unreliable unless k/n is close to zero; k is the number of regressors in (1.1.1). Huber proves that the OLS coefficient estimator is consistent and asymptotically normal when k increases with n, but only if $k/n \to 0$. In practice, k/n will always be positive and is sometimes large, so it is unclear whether the classic tests that exploit asymptotic normality are themselves reliable. This paper derives the asymptotic distribution of the F-test for the joint significance of a subset of the coefficients in Equation (1.1.1) under a more general limit theory that allows k/n to remain uniformly positive. The conventional F-test is asymptotically invalid under this limit theory, but, despite this theoretical tendency to over-reject, will usually have close to its nominal size in practice. Moreover, this paper derives a modification of the F-test that is

asymptotically valid and demonstrates that this new test performs better than the unmodified F-test in practice.

This paper is not the first to study the asymptotic distribution of estimators like $\hat{\beta}$ as both n and k increase. Previous research has looked at the behavior of M-estimators of (1.1.1) as k increases, of Analysis of Variance (ANOVA) as the number of groups increases, and of instrumental variables estimators as the number of instruments increases. This research has followed two distinct paths. The first looks for the fastest growth rate of k that is compatible with the usual consistency and asymptotic normality results. Typically, $k = o(n)$ is a necessary but insufficient condition for these results to hold. The second approach looks for alternative asymptotic distributions of the coefficient estimators that are correct when k/n remains positive.

These increasing-k asymptotics were first introduced in the context of M-estimation; Huber (1973) argued that the assumption that k is fixed is unrealistic in practice. After proving that $k = o(n)$ is necessary for the OLS estimators to be consistent and asymptotically normal, Huber argues that this condition is likely to be required for any tractable asymptotic theory, and proves normality of the M-estimator of the coefficients of the linear regression model under the stronger condition that $k^3/n \to 0$. This rate was improved by Yohai and Maronna (1979), Portnoy (1984) and Portnoy (1985) . In particular, Portnoy (1984) proves consistency under the condition $k \log k/n \to 0$, and Portnoy (1985) proves asymptotic normality for $(k \log k)^{1.5}/n \to 0$. Further research has extended these results to other estimating functions (Welsh 1989), nonlinear models (He and Shao 2000), and estimation of the distribution of the errors (Portnoy 1986, Mammen 1996, Chen and Lockhart 2001).

In econometrics, interest has focused instead on the properties of IV estimators with a fixed number of coefficients but an increasing number of instruments, l. Bekker (1994) studies the asymptotic behavior of Two-Stage Least Squares and Limited Information Maximum Likelihood (LIML) for normal errors as l/n converges to a positive constant. He finds that LIML is both consistent and asymptotically normal but that 2SLS is not. Bekker's results are extended to non-Gaussian

errors by Hansen, Hausman, and Newey (2008) and Chao, Hausman, Newey, Swanson, and Woutersen (2008). Koenker and Machado (1999) prove the consistency and asymptotic normality of GMM estimators with $l^3/n \to 0$. Stock and Yogo (2005), Chao and Swanson (2005), and Andrews and Stock (2007), among others, combine the many-instruments and the weak instruments literatures and argue that the relationship between the concentration parameter and l is more important than that between the number of observations and l. Han and Phillips (2006) study the limit distributions of nonlinear GMM estimators with many weak instruments, and their approach allows for the estimators to converge to non-normal distributions.

Previous work on the F-test under increasing-k asymptotics has focused largely on ANOVA; this literature finds that the usual F-test is asymptotically invalid unless the design matrix is perfectly balanced (requiring an equal number of observations for each group) and propose a new Gaussian approximation for the statistic that gives an asymptotically valid test. Boos and Brownie (1995) started this research, and it was extended to two-way fixed-effects and mixed models (Akritas and Arnold 2000); to allow for heteroskedasticity (Akritas and Papadatos 2004, Bathke 2004); and to allow for additional covariates (Wang and Akritas 2006, Orme and Yamagata 2006, Orme and Yamagata 2007). Anatolyev (2008) proposes an extension that allows for linear regression under conditions similar to this paper's, but imposes a strong restriction on the matrix of regressors that rules out, among other design matrices, the unbalanced ANOVA examined by the previous papers.

These extensions to the F-test all suggest that the standard test should behave poorly in finite samples unless the number of predictors is quite small. However, the F-test is known to have extremely good performance as a comparison of means, even when the errors are not normal. Scheffé (1959) for example, presents analytic and computational evidence that supports using the F-test even with asymmetric and fat tailed errors. Moreover, the simulations presented in some of the ANOVA papers themselves support using the naïve F statistic instead of their proposed replacements. Akritas and Papadatos (2004), for example, simulate a

5% test with lognormal errors and find that the conventional F-test has size 0.04, while their proposed statistics have size 0.73 and 0.78, a moderate over-rejection.

These corrections have other undesirable features. The approximations do not hold under conventional, fixed-k asymptotics, forcing applied researchers to choose between two incompatible asymptotic approximations before testing. This concern on its own is not inherently problematic, and researchers are often forced to make a similar choice in their empirical work. Since k/n is always positive in practice, it would be reasonable to use the increasing-k limit theory by default, but the simulation evidence favoring the standard F-test suggests that there is little merit to these asymptotics even if they are intuitively compelling. Moreover these results only apply under strong restrictions on the matrix of regressors — either assuming an ANOVA structure or other inhibitive conditions — and so are not relevant for applied economic research.

This paper instead proposes a simple correction to the usual F statistic that gives a valid test under either the conventional fixed-k limit theory or under increasing-k asymptotics. When k is fixed, the correction disappears in the limit and our proposed statistic is asymptotically equivalent to the F-test. When k/n remains positive, the correction does not vanish and improves the size of the test statistic. The simulations presented in this paper indicate that this new statistic performs better than the conventional F-test and also outperforms a Gaussian test that is similar to those proposed for ANOVA.

Since this statistic nests both the standard and nonstandard asymptotic theories, careful study of the correction can explain the F-test's strong performance in simulations. The magnitude of the correction depends on the excess kurtosis of the regression errors, ε_t, and on a particular feature of the matrix of regressors. When the excess kurtosis is zero, no correction is necessary and the F-test is valid. If the excess kurtosis is not zero, the magnitude of the correction depends on the diagonal elements of the projection matrices for the unrestricted and restricted regression models — the restricted model imposes the null hypothesis. In practice, it is likely that the correction will be quite small, and the naïve F-test performs reassuringly well, even if it is technically not asymptotically valid. When the F

statistic returns a value near the critical value for a specific test size, though, the correction can affect whether the test rejects or fails to reject the null hypothesis.

Finally, the usefulness of this statistic is demonstrated through two applications — one for time series macroeconomic data and one for cross-sectional data. The first re-examines Olivei and Tenreyro's (2007) study, "The Timing of Monetary Policy Shocks," and finds further support for their conclusion that monetary policy has a different impact on output in different quarters. The second re-examines Sala-i-Martin's (1997) cross-country economic growth analysis and finds supporting evidence that additional regressors beyond the initial levels of primary school education, GDP per capita, and life expectancy are correlated with a country's economic growth. These variables were singled out by Levine and Renelt (1992) and Sala-i-Martin (1997) as having broad support as determinants of economic growth. The first example studies four different equations, and for each equation tests 51 restrictions on the OLS coefficients with 144 observations; the second tests 64 restrictions with 88 observations.

To reiterate, this paper derives a new statistic that can replace the F statistic in tests and works well for regression models with many regressors. The paper also explains the original F-test's strong performance in simulations and illustrates where it is likely to do poorly in applications. Section 1.2 discusses the new test statistic and studies its asymptotic distributions under the null and alternative hypotheses. Section 1.3 presents monte carlo evidence in favor of the statistic. Section 1.4 presents the empirical exercises. Section 1.5 concludes. The proofs are presented in the appendix.

1.2 Asymptotic Theory and Main Results

This section derives the asymptotic distribution of the F-test of the null hypothesis $R\beta = r$ for the linear equation

$$y_t = x_t'\beta + \varepsilon_t \tag{1.2.1}$$

as $q \to \infty$, $n \to \infty$ and q/n remains uniformly positive; q is the number of restrictions imposed by the null hypothesis. This limiting distribution implies that

the F-test is not valid, and we use this asymptotic theory to find a new statistic, $\hat{G}$, that should be used instead of the F statistic. Comparing $\hat{G}$ to the quantiles from the $F(q, n-k)$ distribution yields an asymptotically valid test. Section 1.2.1 discusses the paper's notation and assumptions, Section 1.2.2 presents asymptotic theory and the new test statistic, and 1.2.3 studies the differences between the uncorrected and corrected statistics in more detail. Since the number of estimated coefficients is assumed to vary with n, a triangular array structure underlies all of this paper's theory. Dependence on n will be suppressed in the body of the text, but will be made explicit in assumptions, theorems, and proofs. Unless otherwise indicated, all limits are taken as $n \to \infty$.

1.2.1 Assumptions

Since the number of restrictions imposed by the null hypothesis and the total number of predictors both increase with n, some assumptions take an unfamiliar form. They are, however, analogous to the usual assumptions that ensure the validity of the F-test under classical (fixed-k) asymptotic theory.[1] The observations are required to be independent, the errors are required to be uncorrelated with the regressors and be homoskedastic, and the matrix $\mathbf{X}'\mathbf{X}$ is required to be uniformly positive definite.

The first assumption defines the behavior of the regressors and errors.

Assumption 1. Define the random array $\{x_{n,t}, \varepsilon_{n,t}; t = 1, \ldots, n\}$ and assume that $\{x_{n,t}, \varepsilon_{n,t}\}$ is uniformly integrable. The elements $x_{n,t}$ are random k_n-vectors of regressors with bounded second moments, and each element $\varepsilon_{n,t}$ is a random scalar error term. For each n, the elements of the series $\{(x_{n,t}, \varepsilon_{n,t}); t = 1, \ldots, n\}$ are independent. There are constants $r > 4$ and $B > 0$ such that $\mathrm{E}|\varepsilon_{n,t}|^r < B$ for all t and n. Moreover,

$$\mathrm{E}(\varepsilon_{n,t} \mid \mathbf{X}_n) = 0.$$

[1] Illustrated by White (2000), for example.

and $\mathrm{E}(\varepsilon_{n,t}^2 \mid \mathbf{X}_n) = \sigma^2 > 0$ for all t and n. The matrices $\mathbf{X}_n$ and $\boldsymbol{\varepsilon}_n$ are defined as

$$\mathbf{X}_n = (x_{n,1}, \ldots, x_{n,n})'$$
$$\boldsymbol{\varepsilon}_n = (\varepsilon_{n,1}, \ldots, \varepsilon_{n,n})'.$$

■

Assumption 1 restricts the errors to be strictly exogenous and conditionally homoskedastic, ruling out time series applications that use lagged dependent variables as predictors. The other details of this assumption could be relaxed. It would be straightforward, for example, to allow the array $\{x_{n,t}, \varepsilon_{n,t}\}$ to satisfy a less restrictive weak dependence condition than full independence, but the requirement that $\mathrm{E}(\varepsilon_{n,t} \mid \mathbf{X}_n) = 0$ is crucial.

The next assumption defines the relationship between $(\varepsilon_{n,t}, x_{n,t})$ and the dependent variable, $y_{n,t}$. The operator $|\cdot|_2$ denotes the Euclidean norm of an arbitrary vector in $\mathbb{R}^p$.

Assumption 2. The dependent and independent variables are related through the equation

$$y_{n,t} = x_{n,t}'\beta_n + \varepsilon_{n,t} \qquad t = 1, \ldots, n \tag{1.2.2}$$

with $|\beta_n|_2 = O(1)$. Also, $\lambda_{\max}(n^{-1}\mathbf{X}_n'\mathbf{X}_n) = O_p(1)$ and $\lambda_{\min}(n^{-1}\mathbf{X}_n'\mathbf{X}_n)^{-1} = O_p(1)$; the functions $\lambda_{\max}$ and $\lambda_{\min}$ return the largest and smallest eigenvalues of their arguments, respectively. ■

The assumption that $|\beta|_2 = O(1)$ ensures that the model does not asymptotically crowd out the error. If $|\beta|_2 \to \infty$ instead, the variance of y_t would also increase to infinity, and in the limit (1.2.2) would behave as though there were no error. Such an asymptotic theory would obviously be of little practical value; in real data, there is a substantial error term.

This assumption that restricts the eigenvalues of $\mathbf{X}'\mathbf{X}$ replaces the standard assumption that $n^{-1}\mathbf{X}'\mathbf{X}$ converges in probability to $\mathrm{E}\, x_t x_t'$, a deterministic and positive definite limit. That standard assumption is inappropriate here, because $n^{-1}\mathbf{X}'\mathbf{X}$ does not converge to any deterministic limit; $\mathbf{X}'\mathbf{X}$ is a $k \times k$ matrix, so its dimension grows with n when k does, and the matrix does not converge at all. For

sequences of matrices like this, it is more natural to look at the convergence of the matrices' eigenvalues rather than of the matrices themselves. When k/n remains positive, though, the eigenvalues of $n^{-1}\mathbf{X}'\mathbf{X}$ do not converge to the eigenvalues of $\mathrm{E}\, x_t x_t'$ either. The limiting behavior of these eigenvalues has been worked out for special cases; Yin (1986), for example, shows that, if the elements of $\mathbf{X}$ are i.i.d. $(0,1)$, the empirical distribution function of the eigenvalues of $n^{-1}\mathbf{X}'\mathbf{X}$ converges in probability to the distribution function $F(x) \equiv \int_0^x f(y)dy$, with

$$f(y) = \begin{cases} \frac{\sqrt{\left(\left(1+\sqrt{c}\right)^2 - y\right)\left(y - \left(1-\sqrt{c}\right)^2\right)}}{2\pi y} & \text{if } \left(1-\sqrt{c}\right)^2 \leq y \leq \left(1+\sqrt{c}\right)^2, \\ 0 & \text{otherwise} \end{cases}$$

and $c = \lim_{n\to\infty} k/n$. Note that the eigenvalue conditions of Assumption 2 are satisfied in this case. In general, Assumption 2 ensures that $n^{-1}\mathbf{X}'\mathbf{X}$ is uniformly positive definite in probability.

Also define the following notation. The OLS coefficient estimators are denoted $\hat{\beta}$ and the residuals are $\hat{\varepsilon}_t$. The null hypothesis of interest is

$$H_o: \qquad R\beta = r. \tag{1.2.3}$$

The next assumption controls the asymptotic behavior of this hypothesis.

Assumption 3. $\{R_n\}$ is a sequence of $q_n \times k_n$ matrices of deterministic restrictions, and $\{r_n\}$ is a sequence of $q_n \times 1$ deterministic vectors. There is a constant B_R such that $\max_{i,j} |R_n^{(ij)}| \leq B_R$, with $R_n^{(ij)}$ the (i,j) element of R_n. Moreover,

$$\lambda_{\max}(R_n(n^{-1}\mathbf{X}_n'\mathbf{X}_n)^{-1}R_n)^{-1} = O_p(1), \quad \lambda_{\min}(R_n(n^{-1}\mathbf{X}_n'\mathbf{X}_n)^{-1}R_n')^{-1} = O_p(1), \tag{1.2.4}$$

and $|r_n|_2 = O(1)$. ■

Assumption 3 is a technical condition on the nature of the theoretical sequence of null hypothesis, and does not need to be (and can not be) checked in practice. It arises only because this paper's asymptotic theory embeds the particular null hypothesis of interest in a sequence of similar hypotheses, and this assumption ensures that the limiting behavior of that sequence is reasonable. It

rules out sequences like $R_1 = I_{k_1}, R_2 = 2 \cdot I_{k_2}, \ldots, R_n = n \cdot I_{k_n}, \ldots$; and guarantees that the restricted model is well-behaved in the limit. Three concrete examples can help illustrate Assumption 3.

Example 1. Suppose that one wants to test that the first coefficient, $\beta_{n,1}$, is zero. Then $R_n = (1, 0, \ldots, 0)$, $r_n = 0$, and $q_n = 1$. Assumption 3 requires that the $(1,1)$ element of $(n^{-1}\mathbf{X}_n'\mathbf{X}_n)^{-1}$ be finite and bounded away from zero in probability, which is also required by Assumption 2.

Example 2. Let $k_n = \lfloor n/2 \rfloor$ and suppose that one wants to test that $\beta_{n,1}$ through $\beta_{\lfloor k_n/2 \rfloor}$ are all zero. Then

$$R_n = \left(I_{\lfloor k_n/2 \rfloor}, \mathbf{0}_{\lfloor k_n/2 \rfloor \times \lfloor k_n/2 \rfloor} \right)$$

and $r_n = \mathbf{0}$. Now Assumption 3 requires that the $\lfloor k_n/2 \rfloor \times \lfloor k_n/2 \rfloor$ top left submatrix of $(n^{-1}\mathbf{X}_n'\mathbf{X}_n)^{-1}$ be uniformly positive definite in probability, which is also implied by Assumption 2.

Example 3. Suppose that one wants to test that $\beta_{n,1} + \cdots + \beta_{n,n} = 0$. For this hypothesis to be consistent with Assumption 3, it should be expressed as $R_n = (n^{-1/2}, \ldots, n^{-1/2})$, $r_n = 0$. There are other equivalent ways to express the hypothesis as well.

To minimize confusion between the F statistic, the F distribution, and the F-test, we denote the conventional F statistic for the hypothesis (1.2.3) as $\hat{F}$,

$$\hat{F} \equiv \frac{\sum_{t=1}^{n} \left(\hat{\varepsilon}_{0,t}^2 - \hat{\varepsilon}_t^2 \right) / q}{\sum_{t=1}^{n} \hat{\varepsilon}_t^2 / (n-k)}$$

with $\hat{\varepsilon}_{0,t}$ denoting the residuals from the restricted model.

1.2.2 Distribution of $\hat{F}$ and Asymptotic Correction

The theory proceeds in two steps. We first find the asymptotic distribution of the F-statistic as both q and n increase together. This distribution gives an unsatisfactory test statistic, but motivates an asymptotically equivalent test that performs better. Lemma 1.2.1 shows that $\hat{F}$ is approximately normal under the null hypothesis.

Lemma 1.2.1. *Suppose that Assumptions 1, 2, and 3 hold, that* $q_n \to \infty$ *and* $k_n \to \infty$ *with* $\lim k_n/n < 1$*, and that the null hypothesis (1.2.3) holds. Then*

$$\frac{\sqrt{q_n}}{\eta_n}\left(\hat{F}_n - 1\right) \xrightarrow{d} N(0,1),$$

with

$$\begin{aligned}
\eta_n^2 &= 2\left(1 + c_n\right) + q_n^{-1}\textstyle\sum_{t=1}^{n} \kappa_{n,t} D_{n,t}, \\
D_{n,t} &= \left(P^*_{n,tt} + c_n P_{n,tt} - c_n\right)^2, \\
P^*_n &= \mathbf{X}_n(\mathbf{X}'_n\mathbf{X}_n)^{-1}R'_n\left(R_n(\mathbf{X}'_n\mathbf{X}_n)^{-1}R'_n\right)^{-1}R_n(\mathbf{X}'_n\mathbf{X}_n)^{-1}\mathbf{X}'_n, \\
P_n &= \mathbf{X}_n(\mathbf{X}'_n\mathbf{X}_n)^{-1}\mathbf{X}'_n,
\end{aligned}$$

$c_n = q_n/(n-k_n)$*, and* $\kappa_{n,t} = \mathrm{E}(\varepsilon_{n,t}^4 \mid \mathbf{X}_n)/\sigma^4 - 3$. ■

Under the null hypothesis,

$$\sqrt{q}(\hat{F} - 1) = \frac{q^{-1/2}\left(\boldsymbol{\varepsilon}'\left(P^* + cP - cI\right)\boldsymbol{\varepsilon}\right)}{(n-k)^{-1}\boldsymbol{\varepsilon}'\left(I - P\right)\boldsymbol{\varepsilon}},$$

so Lemma 1.2.1 follows from the asymptotic normality of the numerator. The denominator converges in probability to σ^2. The numerator can be shown to be asymptotically normal by an existing central limit theorem for quadratic forms, derived by Hall (1984) and de Jong (1987); the convergence in probability of the denominator to σ^2 follows from the same theorem. The details of the proof are presented in the appendix.

Lemma 1.2.1 implies that the standard F-test is an asymptotically valid test statistic only if

$$q^{-1}\sum_{t=1}^{n} \kappa_t D_t \to 0 \quad \text{in probability,}$$

otherwise the asymptotic distribution of $\sqrt{q}\hat{F}$ is not pivotal. For example, observe that if $\varepsilon_t \sim N(0,\sigma^2)$, then $\hat{F} \sim F(q, n-k)$ and additionally $\sqrt{q}(\hat{F}-1) \xrightarrow{d} N(0, 2(1+c))$ (since $\kappa_t = 0$).[2] This convergence implies that the critical values from the $F(q, n-k)$ converge to those of the normal$(0, 2(1+c))$ distribution as both q and $n-k$ increase together. Whenever

$$\sqrt{q}(\hat{F} - 1) \xrightarrow{d} N(0, 2(1+c)),$$

[2] In an abuse of notation, we use c to designate $\lim_{n\to\infty} c_n$.

then, critical values from either the Gaussian or the $F(q, n-k)$ distribution can be used to test, implying that $q^{-1}\sum_{t=1}^{n}\kappa_t D_t \to 0$ is necessary for the naïve F-test to be valid.

This sum converges to zero in three cases. First, if the excess kurtosis of the errors is zero the summation is identically zero, as the example with Normal errors illustrates. Second, if the design matrix, $\mathbf{X}$, is balanced, so that $P_{ss} = P_{tt}$ and $P^*_{ss} = P^*_{tt}$ for all s and t, then all of the elements D_t are equal. In that case, since both P^* and P are idempotent matrices,

$$P_{tt} = n^{-1}\,\mathrm{trace}(P) = k/n \qquad a.s.$$

and

$$P^*_{tt} = n^{-1}\,\mathrm{trace}(P^*) = q/n \qquad a.s.,$$

so each $D_t = 0$ almost surely and the sum is again identically zero. Finally, if $q/n \to 0$ then each of the elements P^*_{tt} converges to zero in probability[3] and $c_n \to 0$, so again, each element of D_t converges to zero in probability. If none of those conditions are met, the sum generally remains positive.

This Gaussian approximation suffers from some limitations. First, and most importantly, using Gaussian critical values for the F-test as the basis for a test statistic performs worse than using the naïve critical values from the F distribution. Section 1.3 presents simulations that illustrate this point, and previous research is consistent with this claim. Akritas and Papadatos (2004), for example, run simulations for the F-test in a similar ANOVA application, and the naïve F-test has lower Type-I error than their Gaussian alternatives.

The second limitation is that this approximation forces researchers to choose between asymptotic approximations. If k is fixed, the approximation implied by Lemma 1.2.1 does not hold because $q\hat{F}$ is asymptotically chi-square. Ideally, Lemma 1.2.1's approximation should contain the fixed-k result as a special case. If this new approximation were much more accurate than the usual approximation, researchers might be convinced to abandon the standard F-test to use the new test. But, again, the Gaussian test statistic seems to perform worse.

[3]The eigenvalue restrictions on $\mathbf{X}'\mathbf{X}$ and $R(\mathbf{X}'\mathbf{X})^{-1}R'$ ensure that the individual P^*_{tt}-s do not deviate too far from their average value of q/n, which is now assumed to converge to zero.

In light of these concerns, we propose rescaling the F statistic and then comparing that new statistic to the $F(q, n-k)$ critical values. Observe that Lemma 1.2.1 implies

$$\frac{\sqrt{2(1+c)q}}{\eta}(\hat{F}-1) \xrightarrow{d} N(0, 2(1+c))$$

under Assumptions 1 – 3. As discussed, the normal$(0, 2(1+c))$ distribution and the $F(q, n-k)$ distribution are related: any sequence of random variables G_n that satisfies $\sqrt{q_n}(G_n - 1) \xrightarrow{d} N(0, 2(1+c))$ is approximately $F(q, n-k)$ as well when both q and $n-k$ are large. Theorem 1.2.2 defines G_n as

$$G_n = \eta^{-1}\sqrt{2(1+c)}(\hat{F}-1)+1,$$

and exploits this relationship. This random variable could form the basis of an infeasible test statistic instead of $\hat{F}$ — if G exceeds the $(1-\alpha)$ quantile of the $F(q, n-k)$ distribution, the test would reject.

Theorem 1.2.2. *Suppose that the conditions of Lemma 1.2.1 hold but q_n and k_n may be bounded. Define the random variable*

$$G_n = v_n\hat{F}_n + (1-v_n), \qquad v_n = \frac{\sqrt{2(1+c_n)}}{\eta_n}.$$

Under (1.2.3), $\mathbf{P}[G_n > z_{n,\alpha}] \to \alpha$; each $z_{n,\alpha}$ is the $(1-\alpha)$ critical value for the F distribution with q_n and $n-k_n$ degrees of freedom. ■

This asymptotic approximation suffers from none of the drawbacks of the asymptotically-normal approximation. The simulations presented in Section 1.3 demonstrate that the feasible test statistic based on G performs at least as well as the F-test. Moreover, this asymptotic result holds whether q and k increase or not, so researchers do not have to choose between asymptotic theories. When q/n remains positive, the discussion preceding the theorem applies, and when q is bounded, $\eta^2 \to 2(1+c)$ in probability so the correction vanishes. In the second case, the new random variable, G, and the F statistic, $\hat{F}$, are asymptotically equivalent.

The correction term v should be viewed as a variance correction. When the innovations have positive excess kurtosis, the variance of the F statistic is larger than predicted by the $F(q, n-k)$ distribution. For small values of q, the variance

is only slightly larger, but for large values of q relative to $n-k$, this discrepancy can invalidate the F-test. Applying the proposed correction, v, simply re-scales the F statistic so that its true variance matches that of the F distribution.

This correction must be estimated to make testing feasible; in particular, η^2 is an unknown random variable and must be estimated. Such an estimate is complicated slightly by the necessary degree-of-freedom corrections, but is straightforward to calculate. The next lemma gives an estimator for η^2.

Lemma 1.2.3. *Suppose that the conditions of Theorem 1.2.2 hold. Then*

$$2(1+c_n) + q_n^{-1} \sum_{t=1}^{n} \left(\frac{\hat{\psi}_{n,t}}{\hat{\sigma}_n^4} - 3 \right) D_{n,t} = \eta_n^2 + o_p(1) \tag{1.2.5}$$

if the smallest eigenvalue of $\mathbf{\Gamma}_n + \mathbf{L}_n$ *is bounded away from zero in probability with* $\hat{\sigma}_n^2 = (n-k_n)^{-1} \sum_{t=1}^{n} \hat{\varepsilon}_{n,t}^2$,

$$\begin{pmatrix} \hat{\psi}_{n,1} \\ \vdots \\ \hat{\psi}_{n,n} \end{pmatrix} \equiv (\mathbf{\Gamma}_n + \mathbf{L}_n)^{-1} \begin{pmatrix} \hat{\varepsilon}_{n,1}^4 - \hat{\sigma}_n^4 \left(6P_{n,11} - 18P_{n,11}^2 + 12P_{n,11}^3 - 3\sum_{s=1}^{n} P_{n,s1}^4 \right) \\ \vdots \\ \hat{\varepsilon}_{n,n}^4 - \hat{\sigma}_n^4 \left(6P_{n,nn} - 18P_{n,nn}^2 + 12P_{n,nn}^3 - 3\sum_{s=1}^{n} P_{n,sn}^4 \right) \end{pmatrix},$$

$\mathbf{\Gamma}_n$ *the diagonal matrix with elements* $1 - 4P_{n,tt} + 6P_{n,tt}^2 - 4P_{n,tt}^3$ *and* $\mathbf{L}_n$ *the matrix with* (s,t) *element* $P_{n,st}^4$. *If, additionally,* $\kappa_{n,t} = \kappa_n$ *for* $t = 1, \ldots, n$, *then*

$$2(1+c_n) + \hat{\kappa}_n q_n^{-1} \sum_{t=1}^{n} D_{n,t} = \eta_n^2 + o_p(1) \tag{1.2.6}$$

with

$$\hat{\kappa}_n = n^{-1} \sum_{t=1}^{n} \frac{\hat{\varepsilon}_n^4 / \hat{\sigma}_n^4 - 6k_n/n + n^{-1} \sum_{s=1}^{n} \left(18P_{n,ss}^2 - 12P_{n,ss}^3 + 3\sum_{u=1}^{n} P_{us}^4 \right)}{n - 4\,k_n + \sum_{s=1}^{n} \left(6P_{n,ss}^2 - 4P_{n,ss}^3 + \sum_{u=1}^{n} P_{n,us}^4 \right)}. \tag{1.2.7}$$

■

All three estimators, $\hat{\psi}_{n,t}$, $\hat{\sigma}_n^2$, and $\hat{\kappa}_n$, can be estimated with $P_n - P_n^*$ replacing P_n in the formulae. This replacement amounts to using the residuals from the restricted model instead of the unrestricted model, which is appropriate when the null hypothesis is true. Denote these alternative estimators as $\tilde{\psi}_{n,t}$, $\tilde{\sigma}_n^2$, and $\tilde{\kappa}_n$. The Cauchy-Schwarz inequality implies that the excess kurtosis must be greater

than -2, so we recommend using the estimators $\max\{\hat{\kappa}, -2\}$ and $\max\{\tilde{\kappa}, -2\}$ instead of $\hat{\kappa}$ or $\tilde{\kappa}$ on their own.

The asymptotic distribution of the feasible test statistic is an immediate corollary.

Corollary 1.2.4. *Suppose that the conditions of Theorem 1.2.2 are satisfied, let $\hat{v}_n$ be a consistent estimator of v_n, and define*

$$\hat{G}_n = \hat{v}_n \hat{F}_n + (1 - \hat{v}_n). \tag{1.2.8}$$

If (1.2.3) *holds,* $\mathbf{P}[\hat{G}_n > z_{n,\alpha}] \to \alpha$. ■

The preceding discussion has focused on the validity of the F-test and on proposing a valid alternative test, but we also care about the power of these tests. The asymptotic theory for G and $\hat{G}$ is based fundamentally on asymptotically normal random variables, so it is relatively easy to derive the distribution under local alternatives of the form

$$R_n \beta_n = r_n + \delta_n. \tag{1.2.9}$$

Corollary 1.2.6 shows that the test based on G has nontrivial power if $\delta_n' \delta_n = O(q_n^{1/2}/n)$ and is consistent if δ_n converges to zero more slowly. An important special case is if $R_n = I_k$, $r_n = \mathbf{0}$, and $\delta_n = (1, 0, \ldots, 0)'$ — i.e. there is a single, nonzero regressor and test is for the joint significance of the regression. In that case, the test has unit power asymptotically.

Lemma 1.2.5. *Suppose that the conditions of Lemma 1.2.1 hold but that the alternative hypothesis (1.2.9) holds with $\delta_n' \delta_n = O(q_n^{1/2}/n)$. Then*

$$\frac{\sqrt{q_n}}{\eta_n} \left(\hat{F}_n - 1 \right) - \theta_n \to N(0, 1) \tag{1.2.10}$$

in distribution, with

$$\theta_n \equiv \sigma^{-2} \eta_n^{-1} (n/\sqrt{q_n}) \delta_n' (R(n^{-1} \mathbf{X}_n \mathbf{X}_n)^{-1} R')^{-1} \delta_n = O(1). \tag{1.2.11}$$

■

The behavior of $\hat{F}$ under local alternatives is sufficient to describe the first-order local power of $\hat{G}$. Exploring the higher-order behavior of G or $\hat{G}$ is beyond the scope of this paper. Consistency of the test is an immediate corollary.

Corollary 1.2.6. *Suppose that the conditions of Corollary 1.2.4 hold, but that the alternative hypothesis* (1.2.9) *holds with* $\delta_n'\delta_n \sim 1$. *Then* $\mathbf{P}[\hat{G}_n > z_{n,\alpha}] \to 1$.

1.2.3 Behavior of the F-statistic

Lemma 1.2.1 and Theorem 1.2.2 show that the F-test is invalid and propose a corrected replacement test statistic, but if the correction, v, is near one, the F-test may do well in practice. This section looks at the correction term in more detail. If the effect of v on the F-test is small, researchers might prefer to use the uncorrected F-test out of convenience. However, this section shows that the size of that test can be compromised, suggesting that the new test is preferable. In this section, we assume that the fourth moments of the errors are all identical. In this case, the correction simplifies considerably:

$$v = 2(1+c) + \kappa\bar{D}, \qquad \bar{D} = \textstyle\sum_{t=1}^{n} D_t/q. \tag{1.2.12}$$

Section 1.2.3 runs a series of simulations to study the distribution of $\bar{D}$, and Section 1.2.3 looks at the effect of $\kappa\bar{D}$ on the size of the uncorrected F-test.

Distribution of $\bar{D}$

Unless the kurtosis is near zero, the value of $\bar{D}$ determines the extent to which a correction is necessary for valid testing. In practice, researchers can calculate this statistic to check whether it is large. This subsection uses simulations to study the distribution of $\bar{D}$ and looks for systematic patterns based on the marginal distribution of the regressors and on n, k, and q.

We draw 600 realizations of the statistic $\bar{D}$ when the predictors are drawn from each of three different simple distributions — the Normal(0,1), Cauchy, and Exponential distributions — and include an intercept. Comparing the Normal and Cauchy distributions allows us to see how $\bar{D}$ is affected by heavy-tailed distributions; the Normal distribution is thin-tailed, the Cauchy is fat-tailed, and the

Table 1.1: Values of n, k, and q used for simulations for distribution of $\bar{D}$.

n:	50	100	200	500
k:	$n/20$	$n/10$	$n/4$	$n/2$
q:	1	$k/2$	$k-1$	

Exponential distribution falls in-between. We then draw matrices of predictors with different values of n, k, and q for each distribution. Table 1.1 contains the precise values; when fractions resulted in non-integer values, we rounded up to the next largest integer.

The results of the simulations are presented as boxplots. Figure 1.3 gives the results for Normal predictors, Figure 1.4 for Cauchy, and Figure 1.5 for Exponential. Each boxplot contains the results for a particular combination of n, k, and q, and they are grouped by n, then k/n, and then q/k, with larger values above lower values. The boxplots are constructed as usual, except we do not label any observations as "outliers." The boxes themselves mark the interquartile range of the simulated distribution, and the dark line in the middle of each box marks the median. Each whisker extends to the largest or smallest observed value.[4] Other optional enhancements, such as notches or varying the width of the plots to represent the number of observations, were not used in constructing these plots.

Some patterns emerge from the simulations. The value of $\bar{D}$ increases dramatically as q/k increases, for fixed values of n and k. This behavior is most visually apparent for $k/n = 0.5$, but can be seen generally by comparing the medians for any values of n and k (see, especially, Figure 1.5). Moreover, the dispersion of the distribution of $\bar{D}$ decreases with all of the variables, n, k/n, and q/k. The relationship between q and the dispersion is most clearly seen in Figures 1.4 and 1.5.

We also compare the distributions of $\bar{D}$ for different distributions of the predictors. The distribution of $\bar{D}$ has much larger dispersion and is on average much larger for the Cauchy distribution, and is somewhat larger and more dispersed for

[4]For most boxplots, the whiskers do not extend beyond 1.5 times the interquartile range, and observations beyond the edge of the whisker are marked individually as outliers. In these simulations, there are many such "outliers," and including them separately is visually distracting. Moreover, they are not really outliers; they are known to be drawn from the same distribution as the other observations, so there is no point in studying them individually.

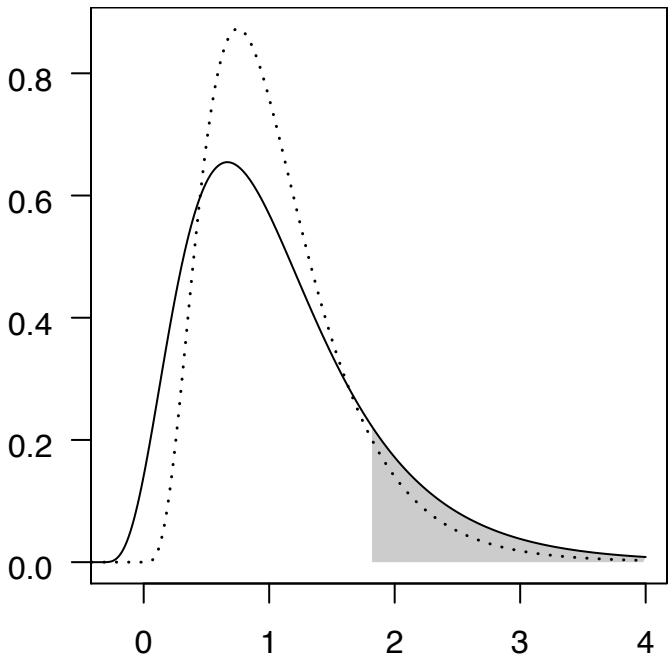

Figure 1.1: $F(15, 70)$ density (dotted line) and the density $f_F(\cdot)$ given in Equation (1.2.13) for $n = 100$, $k = 30$, $q = 15$, and $v = 0.5$ (solid line). The shaded region is the area to the right of the 0.90 quantile of the $F(15, 70)$ distribution.

the Exponential distribution compared to the Normal, suggesting that the tails of the distribution of the regressors plays a large role in the degree of correction necessary. We still see some large values of $\bar{D}$ for the Gaussian distribution, but they require tests of the joint significance of many predictors at once.

Taken together, these simulations indicate that the corrected statistic, $\hat{G}$, is most necessary when testing a hypothesis with a large number of restrictions, especially if the regressors are heavy-tailed. The number of additional regressors beyond those in the null hypothesis does not seem to affect degree of correction necessary for the test.

Effect of the Infeasible Correction on the Size of the Test

This section looks at the relationship between the magnitude of the infeasible correction, v, and the size of the uncorrected F-test. This paper's asymptotic theory implies that G has approximately an $F(q, n-k)$ distribution; in this sub-

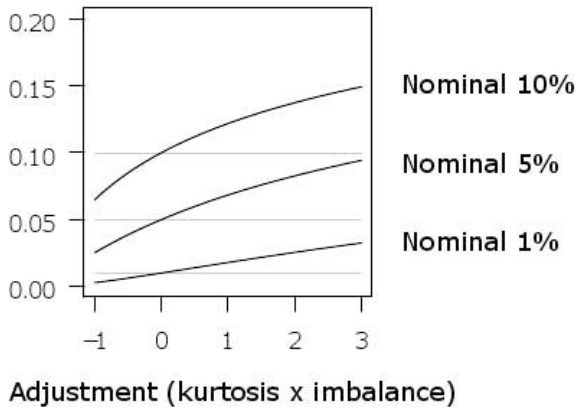

Figure 1.2: Plot of area to the right of the 0.90, 0.95, and 0.99 quantiles of the $F(10, 30)$ distribution, for different values of $\kappa\bar{D}$ (horizontal axis).

section, we assume that G has this distribution exactly, and derive the implied density of the uncorrected F-statistic. We then calculate the mass of that density above the 0.90, 0.95, and 0.99 quantiles of the F-distribution for different values of n, k, q, and $\kappa\bar{D}$, with κ the excess kurtosis of the errors. This mass gives an indication of the extent to which the naïve F-test over-rejects; if 5% of the distribution of the F-statistic lies to the right of the 0.99 quantile, then the true size of a nominal 1% test is 5%. These calculations are approximate since G does not have this distribution in finite samples, but they do allow the impact of $\kappa\bar{D}$ to be isolated.

If the distribution of G is known and v is fixed, it is straightforward to derive the distribution of $\hat{F}$. Suppose that $f_{q,n-k}(\cdot)$ is the density of G, an $F(q, n-k)$ random variable. Then we have a formula for $\hat{F}$,

$$\hat{F} = (G + v - 1)/v,$$

and the density of $\hat{F}$, denoted $f_F(\cdot)$, is trivially calculated to be

$$f_F(x) = v \cdot f_{q,n-k}(vx + 1 - v). \tag{1.2.13}$$

Figure 1.1 plots a representative graph of the densities $f_{q,n-k}$ and f_F for $n = 100$,

$k = 30$, $q = 15$, and $v = 0.5$. We see that both densities are centered at one, and that the density of $\hat{F}$ is more dispersed than that of G. The shaded part of graph is the area under the density of $\hat{F}$ that lies to the right of the 90% critical value of the F-distribution. This area is equal to 0.15, so the F-test would over-reject.

To better understand these size distortions, we calculate that area for different values of n, k, q, and $\kappa\bar{D}$. Table 1.1 contains the values of n, k, and q, they are the same values used to construct Figures 1.3 through 1.5. We consider all values of $\kappa\bar{D}$ between -1 and 3 and consider tests of size 1%, 5%, and 10%. These values are plotted in Figures 1.6, 1.7, 1.8, and 1.9; Figure 1.2 gives a key to the diagrams. Each figure presents the true area for a different value of n. The plots are arranged in a grid, and each plot in the grid contains three curves, each depicting the area for a different quantile. The true area is the vertical axis and the value of $\kappa\bar{D}$ is the horizontal axis. When $\kappa\bar{D}$ is zero, no correction is necessary and each area is equal to the corresponding area of the F density function. In each plot, the top curve presents the values for a test of size 10%, the middle curve 5%, and the bottom curve 1%.

The broad patterns are the same for each test size do not depend heavily on n, although the size distortions are larger for larger values of n. These distortions are, obviously, smaller when $\kappa\bar{D}$ is near zero, and increase as $\kappa\bar{D}$ increases. For $\kappa\bar{D}$ near 3, the true size of each test is roughly 5 percentage points higher than the test's nominal size. For any values of n and k, the distortions increase with q, and for any values of n and q, the distortions do not seem to vary with k.

1.2.4 Summary

When many restrictions are tested simultaneously, the naïve F statistic over-rejects. This tendency is not so pronounced that the F-test is useless, but it can be large enough to effect empirical conclusions. The rescaled statistic, $\hat{G}$, avoids this size distortion and should be used instead of the F statistic for testing. The next section studies the finite sample properties of this statistic through monte carlo simulations.

1.3 Monte Carlo comparison

Although the asymptotic properties of the new statistic, $\hat{G}$, are superior to the conventional F-test, researchers are far more concerned about these tests' finite sample properties. One possible concern is the dependence of this statistic on an estimate of the kurtosis of the regression errors; if this estimate is poor, the entire test may do poorly as well and may do worse than the uncorrected F-test. This section presents two monte carlo simulations. The first studies the accuracy of the kurtosis estimator, and the second studies the size of the test statistic based on $\hat{G}$.

1.3.1 Kurtosis Estimates

This subsection studies the performance of the kurtosis estimator proposed in Lemma 1.2.3. This estimator is constructed for different distributions of the predictors, different numbers of observations and variables, and different magnitudes of the population kurtosis. The predictors are drawn from either the Cauchy, Standard Normal, or Exponential distributions, and the errors are drawn from Student's t distribution with either 5, 10, or 30 degrees of freedom. Varying the distribution of the regressors allows us to determine the effect of the imbalance of the regressors on the kurtosis estimator, and varying the degrees of freedom of the error distribution allows us to determine the effect of the kurtosis of these errors on the estimator.

The values of the number of predictors and observations that were used in these simulations are listed in Table 1.1. We ran 1000 simulations; for each simulation, the linear regression

$$y_t = \beta_0 + x_t'\beta + \varepsilon_t$$

was fit, and the estimator $\hat{\kappa}$, defined in Lemma 1.2.3, was calculated. The Median Absolute Deviation (MAD) and the Root Mean-Squared Error (RMSE) of the kurtosis estimator were calculated for each experiment, and the results are displayed

in Tables 1.2, 1.3 and 1.4. For comparison, the naïve estimator

$$(n-k)^{-1}\sum_{t=1}^{n}\hat{\varepsilon}_n^4/\hat{\sigma}^4 - 3$$

is also calculated for each experiment, and the MAD and RMSE of this estimator is tabulated in Tables 1.5, 1.6, and 1.7.

We can see that the estimators are terrible for very large values of k/n, here $1/2$. For more moderate values of k/n, even $1/4$, $\hat{\kappa}$ improves considerably, and continues to improve as k/n decreases. The performance of the estimator (in terms of RMSE and MAD) improves as the kurtosis decreases, and improves as the number of observations increases. For five degrees of freedom, the kurtosis estimator is poor; its RMSE is comparable to the value of the excess kurtosis itself. For smaller values of excess kurtosis, the estimator improves. The distribution of the predictors, on the other hand, does not seem to affect the kurtosis estimators very much. Although the kurtosis estimator is more accurate for the Cauchy predictors, this effect is small compared to that of the other factors. In comparison, the naïve kurtosis estimator performs poorly for all values of q, k, and n and should be avoided.

These results are not as discouraging as they first appear. In practice, researchers can choose to estimate the kurtosis using the residuals from either the unrestricted model or from the restricted model that imposes the null hypothesis. Using the restricted model decreases the effective number of predictors, so the case $k/n = 1/2$ should not arise in practice.

1.3.2 Test Size

We perform a similar monte carlo experiment to study the size of the test statistic based on $\hat{G}$. As before, design matrices are generated with q, k, and n set according to Table 1.1, but do not consider $n = 500$. The regressors are drawn independently from the Cauchy, Standard Normal, or Exponential distribution with an intercept, and the dependent variables are drawn independently of the regressors from Student's t distribution with 5, 10, and 30 degrees of freedom. For

each simulation, we estimate the regression

$$y_t = x_t'\beta + \varepsilon_t$$

and calculate the following random variables: the F-test; the infeasible test based on G; the feasible test based on $\hat{G}$, using the restricted residuals to estimate the kurtosis of the errors ($\tilde{\kappa}$); the feasible test based on $\hat{G}$, using the unrestricted residuals to estimate the kurtosis; and a test statistic based on the Normal approximation for $\hat{F}$ in Lemma 1.2.1. For all but the last statistic, the test rejects if its corresponding random variable exceeds the 0.90 quantile of the $F(q, n-k)$ distribution. The last test, based on the Normal distribution, rejects if $\sqrt{q}(\hat{F}-1)/\hat{\eta}$ exceeds the 0.90 quantile of the standard normal distribution. The variance for the Normal approximation, $\hat{\eta}^2$, is estimated using the restricted residuals. We only study the size of nominal 10% tests; Section 1.2.3 indicates that we would get similar results for tests of different size.

Tables 1.8, 1.9, and 1.10 contain the results of these simulations. Each entry lists the percentage of the 1500 simulations ran that rejects the null hypothesis. The naïve F-test and the feasible corrected test, $\hat{G}$, both perform well for almost all of the simulations we consider; their simulated size is very close to the nominal size of 10%. The exception is for Cauchy predictors with 5 degrees of freedom. In this case, unless very few restrictions are tested, the F-test over-rejects by roughly 5 percentage points, and the degree of over-rejection increases with n, k/n, and q. The simulated size of $\hat{G}$, on the other hand, is much closer to its nominal size and over-rejects by only one or two percentage points.

The other statistics perform worse. Using the unrestricted residuals to estimate the kurtosis gives a statistic that under- or over-rejects by up to seven percentage points, suggesting that both the size and power of the test are worse than the naïve F-test. The test that uses a normal approximation performs the worst, with higher than nominal size for any distribution of regressors. Surprisingly, it performs worse for high values of q than for small values. It also seems to perform better for larger values of n.

Generally, these simulations demonstrate that the proposed statistic, $\hat{G}$, preserves size well in finite samples, even for models with many regressors and hy-

potheses with many restrictions. This performance may appear puzzling, given the demonstrated poor performance of the kurtosis estimator in the previous section, but it is not. When q is large, k_o — the number of regressors used by the restricted model — is small, since $k_o = k - q$. In that case, the kurtosis can be estimated precisely, and $\hat{G}$ should be expected to perform well. On the other hand, when q is small, k_o is large but $\bar{D}$ is small. In this case, the kurtosis may be estimated poorly, but its estimate has only a small effect on the final statistic. These compensating forces are not present in $\hat{G}_a$, since it uses the unrestricted residuals to estimate the kurtosis, explaining its poor performance in these simulations.

1.4 Empirical Exercise

This section presents two empirical studies that illustrate the new test statistic based on $\hat{G}$. The first is a macroeconomic application based on Olivei and Tenreyro's (2007) study of monetary policy shocks and the second is a cross-sectional application based on Levine and Renelt's (1992), Sala-i-Martin's (1997), and Sala-i-Martin, Doppelhofer, and Miller's (2004) studies of economic growth. Although this paper's theory has not yet been extended to time series applications with lagged dependent variables, which is the econometric model used by Olivei and Tenreyro, their study is a natural application of this paper's statistic and it is unlikely that the form of the test statistic based on $\hat{G}$ will need to change to be appropriate in these applications.

1.4.1 Monetary Policy Shocks

Macroeconomic models often impose rigidities in price and wage contracts so that monetary policy has an effect on real economic variables, output and unemployment in particular. Taylor (1980) and Calvo (1983) pioneered models with many agents who set wages or prices rationally but infrequently; the length for which the price is set is exogenous, and the agents set these prices at different, staggered, times. In this framework, aggregate prices and wages do not change instantaneously, and these frictions can cause agents to change their consumption

and labor supply in response to changes by the Federal Reserve to the money supply or interest rate.

Olivei and Tenreyro (2007) argue that these models could be missing important seasonal effects in the price rigidity. These seasonalities could cause monetary policy to have a stronger effect at some times of the year than others. Sticky price models are often motivated by citing union wage negotiations and other similar contracts. Olivei and Tenreyro cite survey evidence that most firms renegotiate these contracts in the fourth quarter of the calendar year, and these changes are enacted in the first quarter of the next year. Consequently, actions by the Federal Reserve could have less impact in the first and fourth quarters, when wages and prices are the most flexible.

Olivei and Tenreyro formalize this argument in two ways. They develop a variation of a Calvo sticky-wage Dynamic Stochastic General Equilibrium (DSGE) model that allows the probability of wage renegotiation to vary over the year. They also estimate a structural vector autoregression (SVAR) using GDP, the GDP deflator, an index of commodity prices, and the Federal Funds rate, and allow the coefficients of this model to be different in different quarters.[5] They find that the impulse response functions of the DSGE model match those estimated from the SVAR, supporting their intuition. Moreover, the impulse response functions of the VAR show a more pronounced effect from monetary policy shocks in the second and third quarters.

Although Olivei and Tenreyro focus on developing and studying this extended Calvo model, this section will focus on a relatively small aspect of their work: whether there is evidence that the VAR coefficients are truly different across quarters. Olivei and Tenreyro use quarterly data from 1959 to 2005 to estimate the vector autoregression

$$y_t = B_{0,Q(t)} + B_1 \cdot t + \sum_{k=1}^{4} A_{k,Q(t)} y_{t-k} + \varepsilon_t \tag{1.4.1}$$

with y_t containing log GDP for quarter t, the log of the GDP deflator, the log of the

[5]Olivei and Tenreyro report that the BEA is the source of the GDP and the GDP deflator series, and that the Commodity Research Bureau is the source of the commodity price index. The full dataset used by Olivei and Tenreyro is available through the website of the AER.

commodity price index, and the Federal Funds rate. The calendar quarter of period t is given by $Q(t)$, so each equation has 69 different unknown regression coefficients and is estimated with 144 total observations. To test that the coefficients $B_{0,j}$ and $A_{k,j}$ are equal across j for any one of the equations in (1.4.1) requires imposing 51 constraints, giving $q/n \sim 0.35$. This ratio is large enough that the naïve F-test could over-reject.

The null hypothesis of no seasonal effects in equation i can be written formally as

$$\begin{aligned} H_o: \qquad & B_{0,1}^{(i)} = B_{0,m} && m = 1, \ldots, 4 \\ & A_{k,1}^{(ij)} = A_{k,m}^{(ij)} && m = 1, \ldots, 4, \quad j = 1, \ldots, 4 \end{aligned}$$

with $B_{0,1}^{(i)}$ the ith element of the vector $B_{0,1}$, and $A_{k,1}^{(ij)}$ the (i,j) element of the matrix $A_{k,1}$. To test this hypothesis, we calculated both $\hat{F}$ and $\hat{G}$, using the restricted VAR to estimate the excess kurtosis of the errors. These statistics, and the supplementary statistics used to construct $\hat{G}$, are presented in Table 1.11.

These tests somewhat support Olivei and Tenreyro's results. The test statistic rejects at the 10% level for the equations with GDP Deflator, the commodity price index, and the Federal Funds Rate as the dependent variables. However, one-period-ahead GDP does not seem to be subject to these seasonal effects — the new statistic, $\hat{G}$, has a p-value of 0.105 and so fails to reject. Notice that the naïve F-test has a p-value of 0.098 and so it would reject.

In macroeconomic applications, like Olivei and Tenreyro's, the desire to flexibly model the dynamics of the economy leads researchers to use vector autoregressions that have many unknown coefficients. The paucity of data makes it especially difficult to accurately estimate these models, but it is not clear that a smaller model would be able to capture the dynamics of interest. This paper's theory suggests that one can reliably test the significance of these coefficients even if they are estimated imprecisely, and this section's analysis indicates that failing to account for the complexity of these models can give misleading results. It is important to note that, even though the one-step-ahead forecasts that we study and the impulse response functions that Olivei and Tenreyro study are tightly related, there can be seasonal effects in the response of output to structural shocks, even

if the relationship between output and past values of the observed series is not seasonal. Moreover, until we verify that the test based on $\hat{G}$ is appropriate in time series regressions with lagged dependent variables, this section's results should be viewed as promising but unconfirmed.

1.4.2 Cross-Country Growth Regressions

This second application looks at the literature on the determinants of economic growth. Over long periods of time, the welfare benefits from a high growth rate dominate other determinants of a region's welfare, so understanding the factors that cause economic growth is important. This interest has led researchers to estimate equations of the from

$$\text{growth}_j = \beta_0 + \beta_1 x_j + \varepsilon_j \tag{1.4.2}$$

with growth_j the average rate of per capita GDP growth in country j between two specified years and x_j a vector of country-level explanatory variables. A concern in this literature is that there are many potential variables that cause economic growth, so the dimension of x_j can be large. In practice, researchers often select a small subset of those predictors and test the smaller model; this approach makes it hard to compare studies and hard to know the importance of any one variable while controlling for the effects of all of the others.

Levine and Renelt (1992) propose one solution to these problems; they use a variation of Leamer's Extreme Bounds Analysis using the average growth rate from 1960 to 1989. To use this approach, Levine and Renelt estimate (1.4.2) using different subsets of the regressors and label the relationship between, for example, the ith variable and economic growth "fragile" if any two of the estimated coefficients for this variable have different signs. To make this approach computationally feasible, they restrict the subsets of the regressors they consider. Each regression includes four variables — a measure of initial per-capita income in 1960, primary school enrollment in 1960, the investment share of GDP in 1960, and the average annual population growth rate; arguing that these variables have broad support and are included in most prior empirical studies. Permutations of the other re-

gressors are then examined, subject to the constraint that at most three additional variables enter the equation. Levine and Renelt find, perhaps unsurprisingly, that the relationship between most variables and economic growth is "fragile," that there are different subsets of additional regressors for which the sign of almost any estimated coefficient switches.

Sala-i-Martin (1997) and Sala-i-Martin, Doppelhofer, and Miller (2004) argue that this Extreme Bounds Analysis is too strict of a assessment, so Levine and Renelt's finding does not really reflect the relationship between economic growth and these regressors. Sala-i-Martin (1997) proposes a model-averaging approach instead, that Sala-i-Martin, Doppelhofer, and Miller (2004) build on, and finds that many of these variables are strongly correlated with economic growth. Similarly to Levine and Renelt (1992), Sala-i-Martin splits the regressors into a set of three that are included in every regression (replacing population growth and the investment share of GDP with life expectancy in 1960) and estimates each possible regression that includes the other variables, again imposing a limit to mitigate the computational complexity. Instead of comparing the two most extreme point estimates, though, Sala-i-Martin uses the empirical distribution of the estimated coefficients to determine the relationship between that regressor and economic growth. Sala-i-Martin, Doppelhofer, and Miller (2004) use a related Bayesian procedure and emphasize the posterior distributions of the regression coefficients. Both studies find that many of these regressors are correlated with growth, contradicting Levine and Renelt (1992).

In this section, we take a much different perspective: this is not a model selection problem at all, but is a conventional estimation and testing problem. The relationship of interest is

$$\text{growth}_j = \beta_0 + \beta_w' w_j + \beta_z' z_j + \varepsilon_j \tag{1.4.3}$$

with w_j the three undisputed determinants of growth and z_j the additional explanatory variables of interest. If $\beta_z = 0$, these additional predictors are not correlated with growth; if $\beta_z \neq 0$ they are; so it is natural to test it. Although this simple analysis does not tell us any details about how the elements of z_j are related to growth, as the original studies aim to, it does support one set of conclusions

over the other: if the variables z_j are jointly significant, we should not treat their relationship with growth as "fragile," and if they are insignificant there may be very little relationship to explain. We also test the hypothesis $\beta_w = 0$ to determine whether the favored variables are correlated with growth after controlling for the others.

We use Sala-i-Martin, Doppelhofer, and Miller's (2004) dataset for this analysis. It includes data on economic growth from 1960 to 1996 for 88 countries, and includes 67 other country level variables, giving $k/n = 0.77$. The vector w_j includes an intercept, the enrollment rate in primary education in 1960, the level of GDP per capita in 1960, and the life expectancy in 1960. These are the three variables that Sala-i-Martin (1997) included in all of his regressions. Please see Sala-i-Martin, Doppelhofer, and Miller (2004) for a full description of the countries and variables contained in this dataset.

The statistics are presented in Table 1.12. The second test, for the significance of the "consensus" variables that Sala-i-Martin (1997) includes in every regression, fails to reject at 10%. While these variables could have an important structural relationship, they do not seem to have strong partial correlations with growth, and the data do not seem to justify favoring these regressors over the others. For this hypothesis, the test statistics $\hat{F}$ and $\hat{G}$ have similar values: the null hypothesis imposes only three restrictions, so the degree of correction, $\hat{v}$, is small.

The test of the main hypothesis, that the additional regressors do not help explain economic growth, reject at the 10% level, supporting Sala-i-Martin's (1997) and Sala-i-Martin, Doppelhofer, and Miller's (2004) conclusion that there is a meaningful relationship between these variables and growth. It is somewhat surprising that the correction estimate, $\hat{v}$, is not further from one — it is estimated to be 0.98 — given the number of restrictions tested. The F-test rejects the null hypothesis as well, agreeing with the corrected statistic. One would not normally be confident in the F-test here, but its close agreement with $\hat{G}$ should give us some confidence in this result.

1.5 Conclusion

Often researchers are concerned that using too large a model will bias their results — that they will find spurious and nonexistent patterns in a dataset simply because the model has many unknown parameters. This paper shows that this concern has been well founded. The naïve F-test has a tendency to over-reject for models with many parameters. However, this tendency can be understood and modeled, and this paper derives a new statistic that controls for model size and yields a valid test for regression models with many coefficients. Our theory suggests that this correction is especially important when the number of restrictions being tested is large, when the regressors are fat-tailed, and when the regression errors have high excess kurtosis — when those conditions are not met, both the original F-test and our corrected version are reliable. This paper's monte carlo evidence suggests that the F-test can over-reject in finite samples, and the empirical exercises demonstrate that the F-test and our new statistic can give different answers in practice when the original F statistic is near the test's critical values.

The asymptotic theory underlying this new statistic builds on and extends similar results for the F-test in the ANOVA literature. The statistic that we present has several advantages over the ANOVA test statistics, the most important of which is its proximity to the F-test in situations where the F-test performs well. In that light, we also suggest that the statistic $\hat{G}$ also be used for homoskedastic ANOVA when the number of groups is large, and the number of observations per group is small.

1.A Proofs and Additional Results

Lemma 1.A.1. *Suppose the conditions of Lemma 1.2.1 hold. Then*

$$\text{var}(\boldsymbol{\varepsilon}_n' P_n^* \boldsymbol{\varepsilon}_n \mid \mathbf{X}_n)^{-1/2} (\boldsymbol{\varepsilon}_n' P_n^* \boldsymbol{\varepsilon}_n - q_n \sigma^2) \xrightarrow{d} N(0,1) \tag{1.A.1}$$

and

$$\text{var}(\boldsymbol{\varepsilon}_n'(I_n - P_n)\boldsymbol{\varepsilon}_n \mid \mathbf{X}_n)^{-1/2} \Big[\boldsymbol{\varepsilon}_n'(I_n - P_n)\boldsymbol{\varepsilon}_n - (n - k_n)\sigma^2\Big] \xrightarrow{d} N(0,1). \tag{1.A.2}$$

Proof of Lemma 1.A.1. The proofs of (1.A.1) and (1.A.2) are identical, so we only present the proof of (1.A.1). Since the errors are strictly exogenous and the limiting distribution does not depend on $\mathbf{X}_n$, we can treat the regressors as deterministic for this proof, which simplifies the notation. Observe that

$$q_n^{-1/2}\Big(\boldsymbol{\varepsilon}_n' P_n^* \boldsymbol{\varepsilon}_n - q_n\sigma^2\Big) = q_n^{-1/2}\sum_{t=1}^{n}(\varepsilon_{n,t}^2 - \sigma^2)P_{n,tt}^* + q_n^{-1/2}\sum_{t=1}^{n}\sum_{s\neq t}\varepsilon_{n,t}\varepsilon_{n,s}P_{n,st}^*.$$

Since the errors are independent with mean zero, these two sums are independent. It then suffices to prove that each term is individually asymptotically normal.

The proof that $q_n^{-1/2}\sum(\varepsilon_{n,t}^2-\sigma^2)P_{n,tt}^{*2}$ is asymptotically normal is immediate. Each summand is independent, and $P_{n,tt}^*$ is bounded between zero and one. Since $\varepsilon_{n,t}$ has bounded rth moments, the summation satisfies the Lindeberg-Feller central limit theorem.

The proof for

$$q_n^{-1/2}\sum_{t=1}^{n}\sum_{s\neq t}\varepsilon_{n,s}\varepsilon_{n,t}P_{n,st}^*$$

is only slightly more difficult and follows from a central limit theorem for quadratic forms developed by Hall (1984) and de Jong (1987). Define

$$\varsigma_n^2 = \operatorname{var}\left(q_n^{-1/2}\sum_{t=1}^{n}\sum_{s\neq t}\varepsilon_{n,s}\varepsilon_{n,t}P_{n,st}^*\right).$$

Straightforward calculations give

$$\begin{aligned}\varsigma_n^2 &= (2\sigma^4/q_n)\sum_{t=1}^{n}\sum_{s\neq t}(P_{n,st}^*)^2\\ &= (2\sigma^4/q_n)\sum_{t=1}^{n}(P_{n,tt}^* - (P_{n,tt}^*))\\ &= 2\sigma^4 - (2\sigma^4/q_n)\sum_{t=1}^{n}P_{n,tt}^{*2}.\end{aligned}$$

We can assume without loss of generality that ς_n^2 remains uniformly positive; if not, this term vanishes and the proof is complete. To apply de Jong's central limit theorem (Theorem 5.2 of de Jong 1987), we must prove that there exists a sequence of numbers M_n such that $M_n \to \infty$ and the following three conditions hold.

1. $\varsigma_n^{-2}M_n^4 \max_{s=1,\ldots,n}\sum_{t\neq s}\left(q_n^{-1/2}P_{n,st}^*\right)^2 \to 0$ in probability.

2. $\max_{s=1,\ldots n} \mathrm{E}\left(\varepsilon_{n,t}^2 1\left\{|\varepsilon_{n,t}| > M_n\right\}\right) \to 0$

3. $\varsigma_n^{-2} q_n^{-1} \lambda_{\max}(P_n^* - \Lambda_n)^2 \to 0$ in probability, with Λ_n the diagonal matrix with elements $(P_{n,tt}^*)$.

Since P_n^* is idempotent, $\sum_{t \neq s} P_{n,st}^{*2} = P_{n,tt}^* - P_{n,tt}^{*2}$ almost surely, which is in turn less than one. The first condition, then, is satisfied for any $M_n = o(q_n^{1/4})$. The second conditions is satisfied automatically because $\varepsilon_{n,t}$ has bounded rth moments. Finally,

$$\lambda_{\max}(P_n^* - \Lambda_n) \leq \lambda_{\max}(P_n^*) = 1$$

by construction, ensuring that the third condition is met. ■

Proof of Lemma 1.2.1. Under the null hypothesis, we have

$$\begin{aligned}\sqrt{q_n}(\hat{F} - 1) &= \frac{q_n^{-1/2} \boldsymbol{\varepsilon}_n' P_n^* \boldsymbol{\varepsilon}_n}{(n-k_n)^{-1} \boldsymbol{\varepsilon}_n' (I_n - P_n) \boldsymbol{\varepsilon}_n} - q_n^{1/2} \\ &= \frac{q_n^{-1/2} \left[\boldsymbol{\varepsilon}_n' P_n^* \boldsymbol{\varepsilon}_n - c_n \boldsymbol{\varepsilon}_n' (I_n - P_n) \boldsymbol{\varepsilon}_n\right]}{(n-k_n)^{-1} \boldsymbol{\varepsilon}_n' (I_n - P_n) \boldsymbol{\varepsilon}_n}.\end{aligned}$$

Straightforward calculations give

$$\mathrm{E}\left(\boldsymbol{\varepsilon}_n' P_n^* \boldsymbol{\varepsilon}_n - c_n \boldsymbol{\varepsilon}_n' (I_n - P_n) \boldsymbol{\varepsilon}_n\right) = 0$$

and

$$\mathrm{var}\left(\boldsymbol{\varepsilon}_n' P_n^* \boldsymbol{\varepsilon}_n - c_n \boldsymbol{\varepsilon}_n' (I_n - P_n) \boldsymbol{\varepsilon}_n \mid \mathbf{X}_n\right) = \eta_n^2$$

(For formulae for the mean and variance of quadratic forms, see Seber and Lee 2003). Lemma 1.A.1 implies that the numerator is asymptotically normal and the denominator converges in probability to σ^2, completing the proof. ■

Proof of Theorem 1.2.2. Define a sequence of random variables $\{F_n^*\}$ such that $F_n^* \sim F(q_n, n - k_n)$. Lemma 1.2.1 implies that if $q_n \to \infty$,

$$\sqrt{\frac{q_n}{2\left(1 + c_n\right)}}\left(F_n^* - 1\right) \xrightarrow{d} N(0,1)$$

and

$$\frac{\sqrt{q_n}}{\eta_n}(\hat{F}_n - 1) \xrightarrow{d} N(0,1).$$

As a result,

$$\frac{\sqrt{2\,(1+c_n)\,q_n}}{\eta_n}\left(\hat{F}_n - 1\right) \stackrel{d}{=} \sqrt{q_n}\,(F_n^* - 1) + o_p(1)$$

and convergence in distribution follows. Now, suppose that q_n is bounded. Then we can apply, for example, White's (2000) Theorem 5.3 to prove that $\hat{\beta}_n$ is asymptotically normal, so $q_n\hat{F}_n \stackrel{d}{=} \chi^2_{q_n} + o_p(1)$ and, since $(q\hat{F}_n)^2$ is uniformly integrable, $\mathrm{var}(q_n\hat{F}) - 2q_n \to 0$ as $n \to \infty$. Since $c_n \to 0$ as well, $v_n \to 1$. ■

Proof of Lemma 1.2.3. The proofs for the restricted and unrestricted versions of these estimators are identical and we present the proof for the unrestricted estimator. As with previous proofs, we assume that the predictors are deterministic to streamline notation, but the identical proof also holds for stochastic regressors. From Lemma 1.A.1, $\hat{\sigma}_n^2 \to \sigma^2$ in probability, so it suffices to prove that

$$q_n^{-1}\sum_{t=1}^{n}\left[\psi_{n,t} - \mu_{n,t}^{(4)}\right]D_{n,t} \xrightarrow{p} 0,$$

with $\mu_{n,t}^{(4)} = \mathrm{E}(\varepsilon_{n,t}^4 \mid \mathbf{X}_n)$ and

$$\begin{pmatrix}\psi_{n,1}\\ \vdots \\ \psi_{n,n}\end{pmatrix} \equiv (\mathbf{\Gamma}_n + \mathbf{L}_n)^{-1}\begin{pmatrix}\hat{\varepsilon}_{n,1}^4 - \sigma_n^4\left(6P_{n,11} - 18P_{n,11}^2 + 12P_{n,11}^3 - 3\sum_{s=1}^n P_{n,s1}^4\right)\\ \vdots \\ \hat{\varepsilon}_{n,n}^4 - \sigma_n^4\left(6P_{n,nn} - 18P_{n,nn}^2 + 12P_{n,nn}^3 - 3\sum_{s=1}^n P_{n,sn}^4\right)\end{pmatrix}.$$

To show this convergence, we will prove first that

$$\mathrm{E}\, q_n^{-1}\sum_{t=1}^{n}(\psi_{n,t} - \mu_{n,t}^{(4)})D_{n,t} \to 0, \tag{1.A.3}$$

then prove that the variance of $q_n^{-1}\sum_{t=1}^n(\psi_{n,t} - \kappa_{n,t})D_{n,t}$ converges to zero under the auxiliary assumption the eighth moment for $\varepsilon_{n,t}$ is bounded. A truncation argument allows us to extend that result to the case with unbounded eighth moments and completes the proof. The proof for the special case of the lemma when the errors all have identical kurtosis is similar and not presented. Without loss of generality, assume that $\beta_n = 0$ for all n.

To prove (1.A.3), we expand $\hat{\varepsilon}_{n,t}^4$ and take the expectation of each term

separately. Observe that

$$\begin{aligned}\hat{\varepsilon}_{n,t}^4 &= \left(\varepsilon_{n,t} - x'_{n,t}\hat{\beta}_n\right)^4 \\ &= \varepsilon_{n,t}^4 - 4\varepsilon_{n,t}^3 x'_{n,t}\hat{\beta}_n + 6\varepsilon_{n,t}^2 (x'_{n,t}\hat{\beta}_n)^2 \\ &\quad - 4\varepsilon_{n,t}(x'_{n,t}\hat{\beta}_n)^3 + (x'_{n,t}\hat{\beta}_n)^4.\end{aligned}$$

Then

$$\mathrm{E}\left(\varepsilon_{n,t}^3 x'_{n,t}\hat{\beta}_n\right) = \mu_{n,t}^{(4)} P_{n,tt}.$$

Similarly,

$$\begin{aligned}\mathrm{E}\left(\varepsilon_{n,t}^2 (x'_{n,t}\hat{\beta}_n)^2\right) &= \mu_{n,t}^{(4)} P_{n,tt}^2 + \sigma_n^4 (P_{n,tt} - P_{n,tt}^2) \\ \mathrm{E}\left(\varepsilon_{n,t}(x'_{n,t}\hat{\beta}_n)^3\right) &= \mu_{n,t}^{(4)} P_{n,tt}^3 + 3\sigma_n^4 (P_{n,tt}^2 - P_{n,tt}^3) \\ \mathrm{E}\left((x'_{n,t}\hat{\beta}_n)^4\right) &= \sum_{s=1}^n \mu_{n,s}^{(4)} P_{n,st}^4 + 3\sigma_n^4 \left(P_{n,tt}^2 - \textstyle\sum_{s=1}^n P_{n,ts}^4\right),\end{aligned}$$

where many of these terms are simplified because the matrix P_n is idempotent. As a result,

$$\begin{aligned}\mathrm{E}(\hat{\varepsilon}_{n,t}^4) = \mu_{n,t}^{(4)} \left(1 - 4P_{n,tt} + 6P_{n,tt}^2 - 4P_{n,tt}^3\right) + \textstyle\sum_{s=1}^n \mu_{n,s}^{(4)} P_{n,st}^4 \\ + \sigma_n^4 \left(6P_{n,tt} - 18P_{n,tt}^2 + 12P_{n,tt}^3 - 3\textstyle\sum_{s=1}^n P_{n,st}^4\right)\end{aligned}$$

so each $\psi_{n,t}$ has mean $\mu_{n,t}^{(4)}$ as required for the first step of the proof.

If the eighth moment of $\varepsilon_{n,t}$ is bounded, it can be shown similarly through tedious algebra that

$$\mathrm{E}\left(q_n^{-1} \textstyle\sum_{t=1}^n \left(\psi_{n,t} - \mu_{n,t}^{(4)}\right) D_{n,t}\right)^2 \to 0 \tag{1.A.4}$$

as $n \to \infty$, and the proof is omitted. Finally, to prove (1.A.4) for bounded rth moments, define, for any fixed constant C,

$$\begin{aligned}\tilde{\varepsilon}_{n,t}^c &= \varepsilon_{n,t} 1\{|\varepsilon_{n,t}| < C\} - \mathrm{E}(\varepsilon_{n,t} 1\{|\varepsilon_{n,t}| < C\} \mid \mathbf{X}_n) \\ \tilde{\beta}_n &= \left(\mathbf{X}_n'\mathbf{X}_n\right)^{-1} \sum_{t=1}^n x_{n,t}\tilde{\varepsilon}_{n,t}^c\end{aligned}$$

and

$$\begin{pmatrix} \tilde{\psi}^c_{n,1} \\ \vdots \\ \tilde{\psi}^c_{n,n} \end{pmatrix} = (\mathbf{\Gamma}_n + \mathbf{L}_n)^{-1} \times \begin{pmatrix} (\tilde{\varepsilon}^c_{n,1} - x'_{n,1}\tilde{\beta}_n)^4 - \sigma_n^4 \left(6P_{n,11} - 18P^2_{n,11} + 12P^3_{n,11} - 3\sum_{s=1}^n P^4_{n,s1}\right) \\ \vdots \\ (\tilde{\varepsilon}^c_{n,n} - x'_{n,n}\tilde{\beta}_n)^4 - \sigma_n^4 \left(6P_{n,nn} - 18P^2_{n,nn} + 12P^3_{n,nn} - 3\sum_{s=1}^n P^4_{n,sn}\right) \end{pmatrix}.$$

It follows that

$$q_n^{-1} \sum_{t=1}^n \left(\tilde{\psi}^c_{n,t} - \mathrm{E}(\tilde{\varepsilon}^c_{n,t})^4\right) D_{n,tt} \xrightarrow{p} 0$$

for any finite C. For any $\delta > 0$, there is a value of C such that

$$\mathbf{P}[|\tilde{\psi}^c_{n,t} - \hat{\psi}_{n,t}| < \delta] > 1 - \delta$$

and

$$\mathbf{P}[|(\tilde{\varepsilon}^c_{n,t})^4 - (\varepsilon_{n,t})^4| < \delta] > 1 - \delta$$

for all n and t; choose such a C to complete the proof. ■

Proof of Lemma 1.2.5. Under (1.2.9), the numerator of $\hat{F}_n$ becomes

$$q_n^{-1}\mathbf{Y}'_n P^*_n \mathbf{Y}_n = q_n^{-1}\Big[\boldsymbol{\varepsilon}'_n P^*_n \boldsymbol{\varepsilon}_n + 2\delta'_n (R_n(\mathbf{X}'_n\mathbf{X}_n)^{-1}R'_n)^{-1} R_n(\mathbf{X}'_n\mathbf{X}_n)^{-1}\mathbf{X}'_n\boldsymbol{\varepsilon}_n + \delta'_n (R_n(\mathbf{X}'_n\mathbf{X}_n)^{-1}R'_n)^{-1}\delta_n\Big],$$

and so

$$\frac{\sqrt{q_n}}{\eta_n}(\hat{F} - 1) = \eta_n^{-1}\sigma^{-2}q_n^{-1/2}\boldsymbol{\varepsilon}'_n P^*_n \boldsymbol{\varepsilon}_n + 2q_n^{-1/2}\sigma^{-2}\eta_n^{-1}\delta'_n (R_n(\mathbf{X}'_n\mathbf{X}_n)^{-1}R'_n)^{-1} R_n(\mathbf{X}'_n\mathbf{X}_n)^{-1}\mathbf{X}'_n\boldsymbol{\varepsilon}_n + q_n^{-1/2}\sigma^{-2}\eta_n^{-1}\delta'_n (R_n(\mathbf{X}'_n\mathbf{X}_n)^{-1}R_n)^{-1}\delta_n + o_p(1).$$

Lemma 1.2.1 ensures that the first term converges to a standard normal. The second term has mean zero and variance (conditional on $\mathbf{X}_n$) equal to

$$2q_n^{-1}\eta_n^{-2}\delta'_n (R_n(\mathbf{X}'_n\mathbf{X}_n)^{-1}R'_n)^{-1}\delta_n.$$

This variance is in turn of order less than

$$(n/q_n)\delta_n'\delta_n\,\lambda_{\max}((R_n(n^{-1}\mathbf{X}_n'\mathbf{X}_n)^{-1}R_n')^{-1}) \xrightarrow{p} 0$$

by assumption. Similarly,

$$\eta_n^{-1}q_n^{-1/2}\delta_n'(R_n(\mathbf{X}_n'\mathbf{X}_n)^{-1}R_n')^{-1}\delta_n = O_p(1).$$

■

Proof of Corollary 1.2.6. Suppose that q_n is bounded. Then $\hat{G}$ behaves like $\hat{F}$, and standard results give the result. If $q_n \to \infty$, the result holds as a consequence of Lemma 1.2.5 ■

Table 1.2: Root Mean Squared Error (RMSE) and Median Absolute Deviation (MAD) for the kurtosis estimator, $\hat{\kappa}$, for different dimensions of the matrix of regressors and different values of excess kurtosis of the errors; n is the number of observations; k is the number of predictors, including an intercept; and "error df" is the degrees of freedom of the Student's t distribution used to generate the errors. The predictors are drawn from a standard normal distribution, and each entry in the table is based on 1000 simulations.

		RMSE			MAD		
	error df:	5	10	30	5	10	30
	excess kurtosis:	6.00	1.00	0.23	6.00	1.00	0.23
k/n = 0.05	n = 50	5.2	1.4	0.9	5.2	1.0	0.6
	n = 100	5.1	1.3	0.7	4.8	0.8	0.4
	n = 200	5.8	1.8	0.5	4.3	0.6	0.3
	n = 500	5.8	0.8	0.3	3.7	0.5	0.2
k/n = 0.1	n = 50	5.4	1.3	0.9	5.2	1.1	0.6
	n = 100	5.2	1.3	0.7	4.7	0.8	0.5
	n = 200	5.4	1.1	0.6	4.2	0.6	0.3
	n = 500	6.9	0.9	0.3	3.7	0.5	0.2
k/n = 0.25	n = 50	5.7	1.6	1.3	5.5	1.2	0.9
	n = 100	5.4	1.5	1.0	4.9	1.0	0.7
	n = 200	5.6	2.0	0.7	4.3	0.8	0.5
	n = 500	8.6	1.2	0.5	3.7	0.5	0.3
k/n = 0.5	n = 50	1083.3	5638.5	2350.9	8.0	3.0	2.2
	n = 100	746.3	5137.8	244.4	60.6	3.0	2.2
	n = 200	1179.5	412.6	218.6	256.5	32.7	2.2
	n = 500	3247.9	829.6	406.6	1116.0	329.0	28.1

Table 1.3: Root Mean Squared Error (RMSE) and Median Absolute Deviation (MAD) for the kurtosis estimator, $\hat{\kappa}$, for different dimensions of the matrix of regressors and different values of excess kurtosis of the errors; n is the number of observations; k is the number of predictors, including an intercept; and "error df" is the degrees of freedom of the Student's t distribution used to generate the errors. The predictors are drawn from a Cauchy distribution, and each entry in the table is based on 1000 simulations.

		RMSE			MAD		
	error df:	5	10	30	5	10	30
	excess kurtosis:	6.00	1.00	0.23	6.00	1.00	0.23
k/n = 0.05	n = 50	5.3	1.3	0.8	5.2	1.0	0.6
	n = 100	5.9	1.4	0.6	4.8	0.8	0.4
	n = 200	6.5	1.4	0.5	4.3	0.6	0.3
	n = 500	5.6	0.7	0.3	3.8	0.4	0.2
k/n = 0.1	n = 50	5.2	1.3	0.9	5.3	1.1	0.6
	n = 100	5.1	1.2	0.7	4.8	0.8	0.5
	n = 200	6.5	1.3	0.5	4.4	0.7	0.3
	n = 500	7.6	1.0	0.4	3.7	0.5	0.2
k/n = 0.25	n = 50	5.6	26.6	1.2	5.4	1.3	0.8
	n = 100	5.8	2.3	1.0	4.8	1.0	0.7
	n = 200	5.8	1.6	0.9	4.3	0.7	0.5
	n = 500	6.7	18.4	2.5	3.6	0.5	0.3
k/n = 0.5	n = 50	92.6	27.3	20.7	8.0	3.0	4.5
	n = 100	6.9	5.2	14.1	8.0	3.0	2.2
	n = 200	7.3	3.3	3.6	8.0	3.0	2.2
	n = 500	7.9	2.6	2.2	8.0	3.0	2.1

Table 1.4: Root Mean Squared Error (RMSE) and Median Absolute Deviation (MAD) for the kurtosis estimator, $\hat{\kappa}$, for different dimensions of the matrix of regressors and different values of excess kurtosis of the errors; n is the number of observations; k is the number of predictors, including an intercept; and "error df" is the degrees of freedom of the Student's t distribution used to generate the errors. The predictors are drawn from an Exponential distribution, and each entry in the table is based on 1000 simulations.

		RMSE			MAD		
	error df:	5	10	30	5	10	30
	excess kurtosis:	6.00	1.00	0.23	6.00	1.00	0.23
k/n = 0.05	n = 50	5.3	1.6	0.9	5.2	1.0	0.6
	n = 100	5.2	1.3	0.7	4.7	0.8	0.4
	n = 200	6.4	1.1	0.5	4.3	0.6	0.3
	n = 500	10.6	0.8	0.3	3.8	0.4	0.2
k/n = 0.1	n = 50	5.2	1.6	0.9	5.2	1.1	0.6
	n = 100	5.2	1.6	0.7	4.8	0.8	0.4
	n = 200	5.6	1.5	0.5	4.2	0.7	0.3
	n = 500	6.1	0.9	0.3	3.7	0.5	0.2
k/n = 0.25	n = 50	5.5	1.8	1.3	5.5	1.3	0.8
	n = 100	5.9	1.7	1.1	4.9	1.0	0.6
	n = 200	5.4	1.4	0.7	4.4	0.8	0.5
	n = 500	6.6	1.0	0.5	3.7	0.5	0.3
k/n = 0.5	n = 50	18216.2	2803.9	514.7	8.0	3.0	7.8
	n = 100	10839.8	4709.1	1093.5	8.0	3.1	2.2
	n = 200	10198.1	2668.7	3405.4	212.3	23.3	2.2
	n = 500	10154.4	5537.1	2383.4	1569.8	375.4	8.4

Table 1.5: Root Mean Squared Error (RMSE) and Median Absolute Deviation (MAD) for the naive kurtosis estimator for different dimensions of the matrix of regressors and different values of excess kurtosis of the errors; n is the number of observations; k is the number of predictors, including an intercept; and "error df" is the degrees of freedom of the Student's t distribution used to generate the errors. The predictors are drawn from a standard normal distribution, and each entry in the table is based on 1000 simulations.

		RMSE			MAD		
	error df:	5	10	30	5	10	30
	excess kurtosis:	6.00	1.00	0.23	6.00	1.00	0.23
k/n = 0.05	n = 50	14.2	2.7	1.2	4.1	1.1	0.7
	n = 100	9.4	2.2	0.9	3.5	0.8	0.5
	n = 200	14.9	3.3	0.7	2.7	0.7	0.4
	n = 500	12.4	1.4	0.4	2.2	0.6	0.2
k/n = 0.1	n = 50	37.0	2.2	1.1	4.3	1.1	0.7
	n = 100	10.3	2.1	0.8	3.6	0.8	0.5
	n = 200	12.1	1.6	0.7	2.8	0.6	0.4
	n = 500	13.9	1.4	0.4	2.2	0.6	0.3
k/n = 0.25	n = 50	10.2	1.7	1.2	4.4	1.0	0.7
	n = 100	9.1	1.5	0.8	3.6	0.8	0.5
	n = 200	8.2	2.6	0.6	3.0	0.6	0.4
	n = 500	14.9	1.2	0.4	2.4	0.4	0.2
k/n = 0.5	n = 50	5.6	1.8	1.2	4.6	1.1	0.8
	n = 100	9.5	1.2	0.9	4.0	0.8	0.6
	n = 200	4.4	1.0	0.6	3.7	0.6	0.4
	n = 500	4.3	0.7	0.4	3.2	0.4	0.2

Table 1.6: Root Mean Squared Error (RMSE) and Median Absolute Deviation (MAD) for the naive kurtosis estimator for different dimensions of the matrix of regressors and different values of excess kurtosis of the errors; n is the number of observations; k is the number of predictors, including an intercept; and "error df" is the degrees of freedom of the Student's t distribution used to generate the errors. The predictors are drawn from a Cauchy distribution, and each entry in the table is based on 1000 simulations.

		RMSE			MAD		
	error df:	5	10	30	5	10	30
	excess kurtosis:	6.00	1.00	0.23	6.00	1.00	0.23
k/n = 0.05	n = 50	20.0	2.5	1.1	4.0	1.0	0.7
	n = 100	30.7	2.5	0.8	3.4	0.8	0.5
	n = 200	19.5	2.5	0.7	2.7	0.8	0.4
	n = 500	11.8	1.5	0.5	2.0	0.7	0.3
k/n = 0.1	n = 50	10.5	2.3	1.2	4.0	1.1	0.7
	n = 100	9.7	2.1	1.0	3.5	0.8	0.5
	n = 200	17.8	2.4	0.7	2.8	0.8	0.4
	n = 500	18.2	1.9	0.6	2.2	0.9	0.3
k/n = 0.25	n = 50	13.1	2.6	1.3	4.0	1.0	0.7
	n = 100	12.5	3.2	1.2	3.2	0.9	0.6
	n = 200	9.9	2.3	1.1	2.5	0.9	0.5
	n = 500	10.3	2.1	0.9	2.0	1.1	0.6
k/n = 0.5	n = 50	10.2	3.1	2.0	4.1	1.2	1.0
	n = 100	34.0	3.1	1.8	3.1	1.1	0.9
	n = 200	15.2	3.5	1.8	2.5	1.5	1.2
	n = 500	43.9	2.8	1.5	1.9	1.7	1.2

Table 1.7: Root Mean Squared Error (RMSE) and Median Absolute Deviation (MAD) for the naïve kurtosis estimator for different dimensions of the matrix of regressors and different values of excess kurtosis of the errors; n is the number of observations; k is the number of predictors, including an intercept; and "error df" is the degrees of freedom of the Student's t distribution used to generate the errors. The predictors are drawn from an Exponential distribution, and each entry in the table is based on 1000 simulations.

		RMSE			MAD		
	error df:	5	10	30	5	10	30
	excess kurtosis:	6.00	1.00	0.23	6.00	1.00	0.23
k/n = 0.05	n = 50	12.3	3.0	1.2	4.1	1.0	0.7
	n = 100	13.7	2.3	0.9	3.5	0.8	0.5
	n = 200	19.5	1.9	0.6	2.8	0.7	0.4
	n = 500	38.5	1.3	0.5	2.1	0.6	0.3
k/n = 0.1	n = 50	17.7	2.6	1.1	4.1	1.0	0.7
	n = 100	11.2	2.4	0.9	3.4	0.8	0.5
	n = 200	12.6	2.2	0.6	2.9	0.7	0.4
	n = 500	11.7	1.3	0.4	2.1	0.6	0.3
k/n = 0.25	n = 50	17.2	2.1	1.2	4.2	1.0	0.7
	n = 100	12.4	2.1	0.9	3.7	0.8	0.5
	n = 200	7.3	1.5	0.6	3.1	0.6	0.4
	n = 500	8.7	1.0	0.4	2.4	0.5	0.3
k/n = 0.5	n = 50	6.4	2.3	1.3	4.6	1.1	0.7
	n = 100	5.7	1.4	1.0	3.8	0.8	0.6
	n = 200	5.6	1.0	0.7	3.7	0.6	0.4
	n = 500	5.6	0.7	0.4	3.2	0.4	0.3

Table 1.8: Simulated size for a nominal 10% test, based on 1500 simulations. The regressors are a $k \times n$ matrix drawn from the Normal distribution and include an intercept; the null hypothesis of each test imposes q restrictions; "df" denotes the degrees of freedom of the t distribution used to generate the errors. Each column contains the size for a given test statistic and error df.

			df = 5					df = 10					df = 30				
q	k/n	n	$\hat{F}$	G	$\hat{G}$	$\hat{G}_a$	N	$\hat{F}$	G	$\hat{G}$	$\hat{G}_a$	N	$\hat{F}$	G	$\hat{G}$	$\hat{G}_a$	N
1	0.1	50	9	9	9	9	9	11	10	11	11	10	9	9	9	9	9
1	0.1	100	13	12	12	12	12	11	11	11	11	11	10	10	10	10	10
1	0.1	200	9	9	9	9	9	12	12	12	12	11	10	10	10	10	9
1	0.25	50	10	9	10	10	10	11	11	11	11	11	9	9	9	9	9
1	0.25	100	9	9	9	9	9	11	11	11	11	11	11	11	11	11	11
1	0.25	200	10	9	10	10	9	12	12	12	12	12	10	10	10	10	10
1	0.5	50	9	8	9	7	9	10	10	10	8	10	10	10	10	9	10
1	0.5	100	11	11	10	7	10	10	10	10	9	10	10	10	10	10	10
1	0.5	200	10	10	10	6	9	10	10	10	9	10	11	11	11	9	11
$k/2$	0.1	50	8	7	8	8	10	11	10	10	11	12	13	13	13	12	14
$k/2$	0.1	100	10	10	10	10	12	11	11	11	11	14	11	11	11	11	13
$k/2$	0.1	200	12	11	11	11	13	9	9	9	9	10	10	10	10	10	11
$k/2$	0.25	50	11	10	11	11	13	9	9	9	9	12	12	12	12	12	15
$k/2$	0.25	100	11	10	11	11	12	10	10	10	10	11	11	11	11	11	12
$k/2$	0.25	200	11	11	11	10	12	9	9	9	9	11	9	9	9	9	11
$k/2$	0.5	50	10	9	10	8	14	11	11	11	9	16	10	10	10	8	13
$k/2$	0.5	100	11	11	11	7	13	11	11	11	9	13	9	9	9	7	12
$k/2$	0.5	200	9	9	9	4	12	12	12	12	9	14	9	9	9	7	11
$k-1$	0.1	50	10	10	10	10	12	10	10	10	10	11	9	9	9	9	11
$k-1$	0.1	100	11	11	11	11	13	9	9	9	9	11	11	11	11	11	13
$k-1$	0.1	200	10	10	10	10	12	9	9	9	9	10	11	11	11	11	12
$k-1$	0.25	50	8	7	7	7	11	10	10	10	10	14	10	10	10	10	13
$k-1$	0.25	100	11	11	11	11	13	10	10	10	10	13	10	10	10	10	13
$k-1$	0.25	200	10	10	10	10	12	9	9	9	9	11	8	8	8	8	9
$k-1$	0.5	50	10	8	9	8	14	11	10	11	9	17	10	10	10	8	15
$k-1$	0.5	100	11	10	11	7	15	10	10	10	9	14	12	12	12	10	16
$k-1$	0.5	200	9	9	9	5	12	8	8	8	6	11	9	9	9	8	12

Table 1.9: Simulated size for a nominal 10% test, based on 1500 simulations. The regressors are a $k \times n$ matrix drawn from the Cauchy distribution and include an intercept; the null hypothesis of each test imposes q restrictions; "df" denotes the degrees of freedom of the t distribution used to generate the errors. Each column contains the size for a given test statistic and error df.

			df = 5					df = 10					df = 30				
q	k/n	n	$\hat{F}$	G	$\hat{G}$	$\hat{G}_a$	N	$\hat{F}$	G	$\hat{G}$	$\hat{G}_a$	N	$\hat{F}$	G	$\hat{G}$	$\hat{G}_a$	N
1	0.1	50	11	7	10	10	10	10	9	10	10	10	9	9	10	9	10
1	0.1	100	7	5	7	7	7	10	9	9	9	9	11	11	11	11	11
1	0.1	200	10	7	9	9	9	8	7	8	8	7	9	9	9	9	8
1	0.25	50	9	7	9	9	9	11	10	11	10	11	13	12	13	13	13
1	0.25	100	9	6	8	8	8	11	11	11	11	11	10	10	10	11	10
1	0.25	200	10	8	9	9	9	9	9	9	9	9	10	10	10	10	9
1	0.5	50	9	7	7	8	7	11	10	8	9	9	11	11	10	8	10
1	0.5	100	11	8	10	11	10	11	10	9	10	9	11	11	9	9	8
1	0.5	200	10	7	11	11	11	11	10	11	11	11	9	9	9	9	8
$k/2$	0.1	50	10	6	10	10	11	8	7	8	7	9	11	11	11	11	12
$k/2$	0.1	100	12	7	10	10	11	12	10	11	11	13	11	10	10	10	12
$k/2$	0.1	200	11	6	9	9	10	12	11	11	11	12	12	11	12	11	13
$k/2$	0.25	50	11	7	10	10	12	12	10	11	11	14	9	8	9	10	13
$k/2$	0.25	100	13	9	12	12	14	12	10	11	11	14	9	9	9	9	11
$k/2$	0.25	200	13	6	10	10	11	10	9	9	10	10	13	12	12	12	14
$k/2$	0.5	50	12	8	11	10	15	10	9	10	7	14	11	10	11	8	16
$k/2$	0.5	100	12	6	11	13	14	11	9	10	9	13	11	11	11	10	15
$k/2$	0.5	200	15	8	12	17	15	12	10	11	12	13	12	11	11	11	13
$k-1$	0.1	50	11	8	11	11	12	10	9	10	10	12	11	10	11	11	13
$k-1$	0.1	100	13	7	11	12	12	12	11	11	11	12	10	9	9	10	11
$k-1$	0.1	200	13	7	10	11	12	13	10	11	11	12	8	8	8	9	9
$k-1$	0.25	50	12	6	10	11	13	13	11	12	13	16	11	10	11	11	14
$k-1$	0.25	100	16	9	12	14	15	11	10	10	10	12	11	10	10	10	13
$k-1$	0.25	200	14	7	11	11	12	12	9	10	10	11	11	10	11	11	13
$k-1$	0.5	50	14	8	12	11	18	11	10	10	9	15	9	9	9	6	15
$k-1$	0.5	100	15	8	12	14	15	12	10	11	10	15	11	10	10	9	14
$k-1$	0.5	200	17	8	12	17	15	12	9	9	12	13	10	9	9	9	12

Table 1.10: Simulated size for a nominal 10% test, based on 1500 simulations. The regressors are a $k \times n$ matrix drawn from the Exponential distribution and include an intercept; the null hypothesis of each test imposes q restrictions; "df" denotes the degrees of freedom of the t distribution used to generate the errors. Each column contains the size for a given test statistic and error df.

			df = 5					df = 10					df = 30				
q	k/n	n	$\hat{F}$	G	$\hat{G}$	$\hat{G}_a$	N	$\hat{F}$	G	$\hat{G}$	$\hat{G}_a$	N	$\hat{F}$	G	$\hat{G}$	$\hat{G}_a$	N
1	0.1	50	11	10	11	11	11	12	11	11	11	11	10	10	11	11	10
1	0.1	100	11	10	11	11	11	9	9	9	9	9	10	10	10	10	10
1	0.1	200	10	10	10	10	10	9	9	9	9	8	10	10	10	10	10
1	0.25	50	10	8	10	10	10	9	9	9	9	9	10	10	10	10	10
1	0.25	100	9	8	9	9	8	10	9	10	10	9	9	9	9	9	9
1	0.25	200	9	8	8	8	8	10	10	10	10	9	10	10	10	10	10
1	0.5	50	8	8	8	7	9	10	10	10	8	10	10	10	10	8	10
1	0.5	100	10	9	8	6	8	10	10	10	8	10	11	11	11	8	11
1	0.5	200	10	10	9	5	8	10	10	10	8	10	10	10	10	7	10
$k/2$	0.1	50	11	8	10	10	12	8	8	8	8	10	11	11	11	11	12
$k/2$	0.1	100	10	8	10	10	10	11	10	10	10	11	10	10	10	10	11
$k/2$	0.1	200	9	8	9	9	11	11	11	11	11	12	10	10	10	10	12
$k/2$	0.25	50	12	10	11	12	15	10	9	9	9	13	10	10	10	10	13
$k/2$	0.25	100	9	8	9	9	10	10	10	10	10	12	10	10	10	10	12
$k/2$	0.25	200	10	10	9	10	11	8	8	8	8	10	10	10	10	10	13
$k/2$	0.5	50	10	9	10	7	14	10	10	10	6	15	8	8	8	5	13
$k/2$	0.5	100	9	9	9	6	12	9	9	9	5	12	9	9	9	6	13
$k/2$	0.5	200	11	10	10	5	13	9	9	9	6	12	12	12	12	8	15
$k-1$	0.1	50	11	9	11	11	12	9	8	9	9	9	9	9	9	9	11
$k-1$	0.1	100	10	9	10	10	11	11	10	11	11	12	11	11	11	11	12
$k-1$	0.1	200	10	9	10	10	11	11	11	11	11	12	10	10	10	10	11
$k-1$	0.25	50	9	8	9	9	12	12	12	12	12	15	11	11	11	11	14
$k-1$	0.25	100	10	9	10	10	13	13	12	13	13	15	10	10	10	10	12
$k-1$	0.25	200	11	10	10	10	13	9	8	8	9	10	11	11	11	11	13
$k-1$	0.5	50	10	9	10	7	16	10	10	10	7	15	12	12	12	7	17
$k-1$	0.5	100	11	10	11	7	14	10	10	10	6	13	10	10	10	6	14
$k-1$	0.5	200	12	11	11	7	14	9	9	9	5	13	11	11	11	6	15

Table 1.11: Statistics for equation-by-equation hypothesis tests of coefficient equality for Olivei and Tenreyro's (2007) monetary policy VAR. $\hat{F}$ is the F-statistic, and $\hat{G}$ is this paper's proposed corrected statistic. $p.$ is each statistic's corresponding p-value.

	$\hat{\kappa}$	$\bar{D}$	$\hat{v}$	$\hat{F}$	$\hat{G}$	$p_{\hat{F}}$	$p_{\hat{G}}$
GDP	1.8	0.15	0.96	1.39	1.37	0.098	0.105
GDP Deflator	0.9	0.15	0.98	1.55	1.54	0.042	0.044
Commodity Index	0.8	0.15	0.98	1.81	1.79	0.010	0.010
Fed. Funds	11.5	0.15	0.82	1.79	1.65	0.010	0.024

Table 1.12: Statistics for equation-by-equation hypothesis tests of coefficient equality for cross-country growth regressions using Sala-i-Martin *et al.*'s (2004) dataset. $\hat{F}$ is the F-statistic, and $\hat{G}$ is this paper's proposed corrected statistic. $p.$ is each statistic's corresponding p-value.

	$\hat{\kappa}$	$\bar{D}$	$\hat{v}$	$\hat{F}$	$\hat{G}$	$p_{\hat{F}}$	$p_{\hat{G}}$
Main Hypothesis	0.8	0.34	0.98	1.74	1.73	0.084	0.086
Comparison	0.1	0.03	1.00	1.22	1.22	0.328	0.328

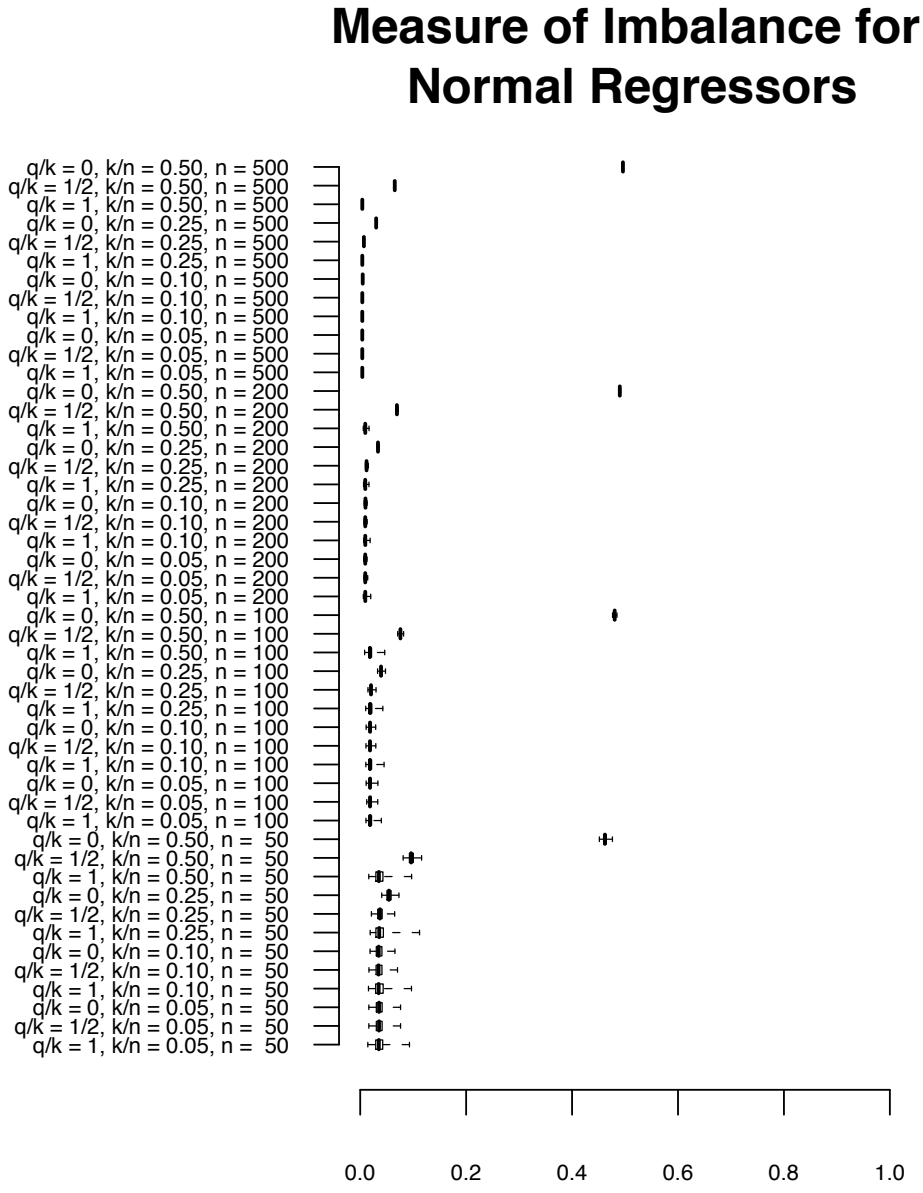

Figure 1.3: Boxplots for the statistic $\bar{D}$, a measure of the imbalance of the design matrix of the regressors, for i.i.d. Normal regressors (with a constant as the first column). The boxplots are based on 600 simulations. When q/k is labeled "1", the test corresponding to $\bar{D}$ is a test of joint significance of all of the regressors except for the intercept. When q/k is labeled "0", the associated test is of the significance of a single regressor, and when q/k is labeled "1/2", the test is for the significance of half of the regressors.

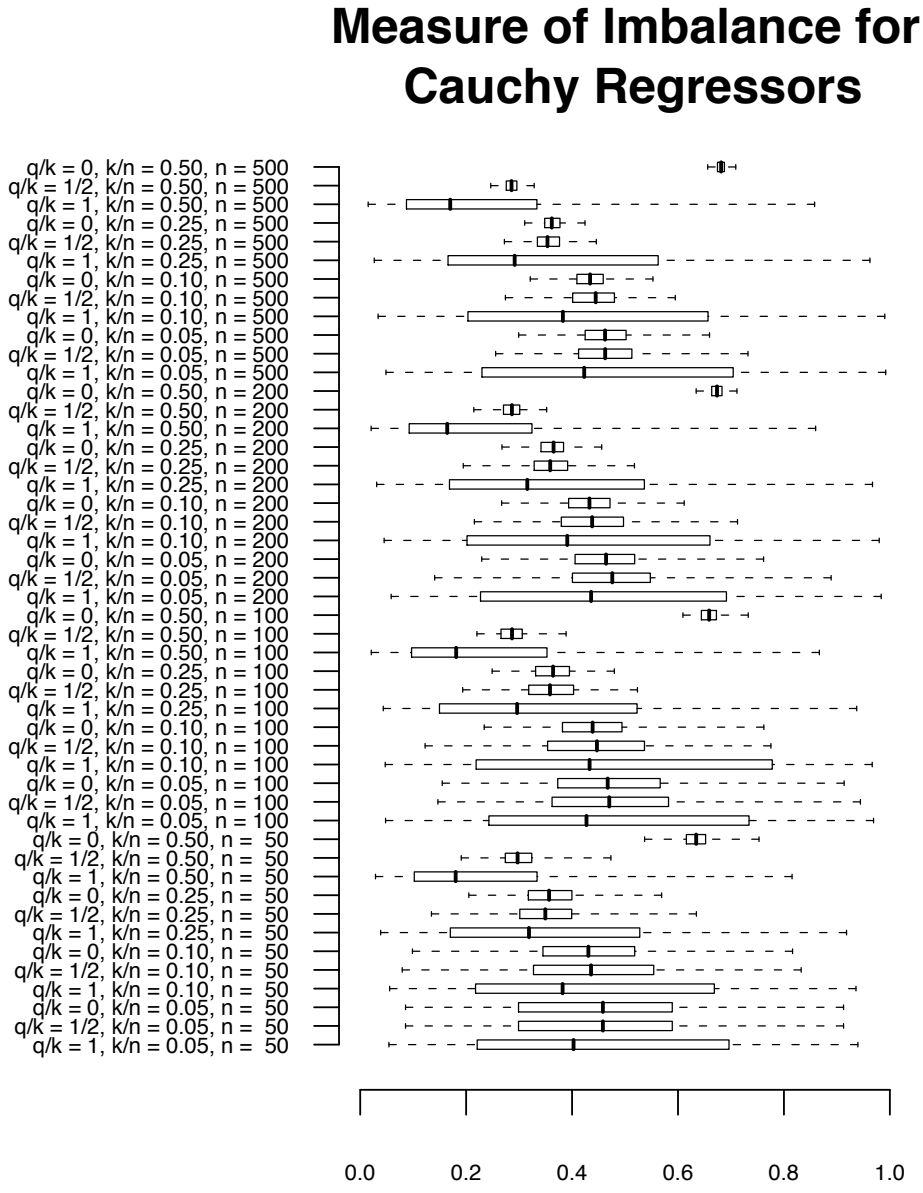

Figure 1.4: Boxplots for the statistic $\bar{D}$, a measure of the imbalance of the design matrix of the regressors, for i.i.d. Cauchy regressors (with a constant as the first column). The boxplots are based on 600 simulations. When q/k is labeled "1", the test corresponding to $\bar{D}$ is a test of joint significance of all of the regressors except for the intercept. When q/k is labeled "0", the associated test is of the significance of a single regressor, and when q/k is labeled "1/2", the test is for the significance of half of the regressors.

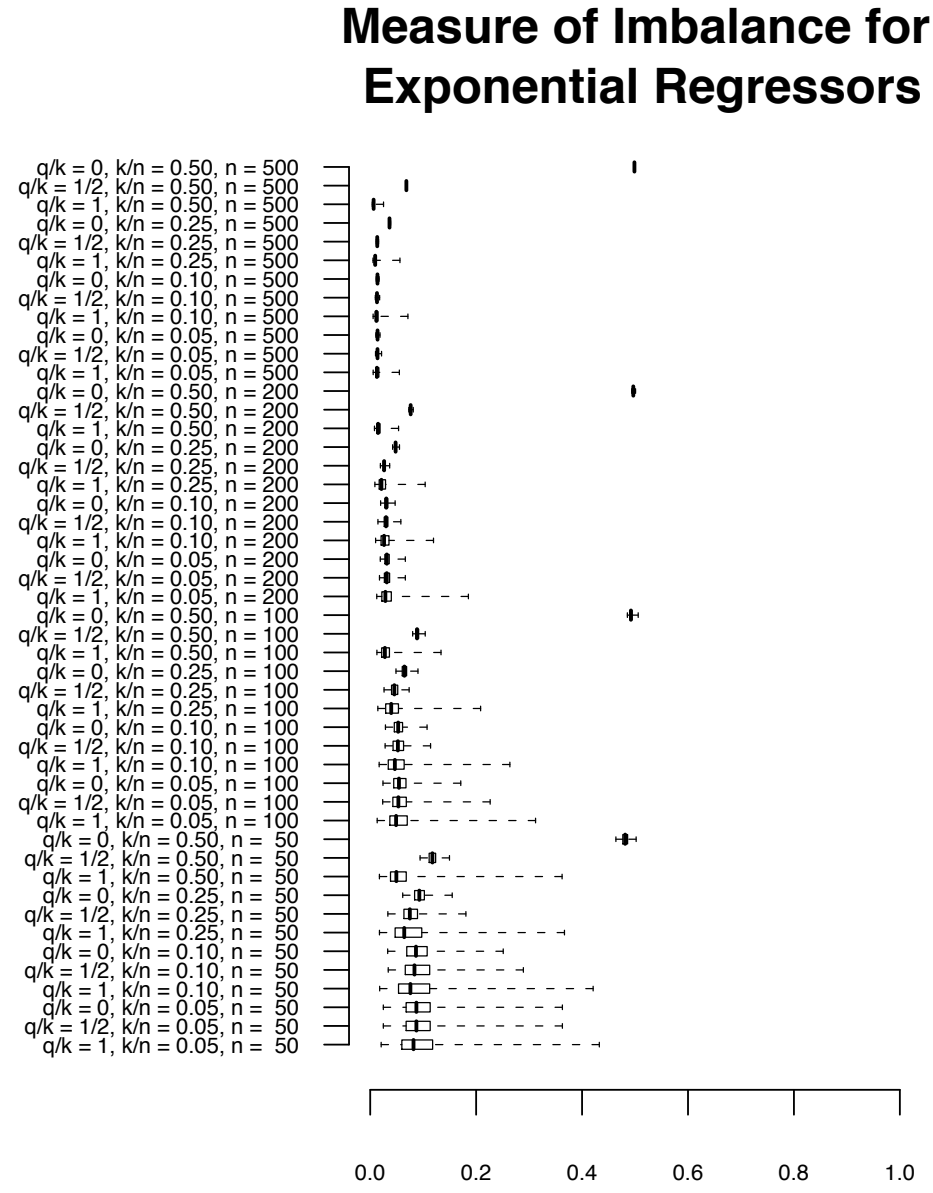

Figure 1.5: Boxplots for the statistic $\bar{D}$, a measure of the imbalance of the design matrix of the regressors, for i.i.d. Exponential regressors (with a constant as the first column). The boxplots are based on 600 simulations. When q/k is labeled "1", the test corresponding to $\bar{D}$ is a test of joint significance of all of the regressors except for the intercept. When q/k is labeled "0", the associated test is of the significance of a single regressor, and when q/k is labeled "1/2", the test is for the significance of half of the regressors.

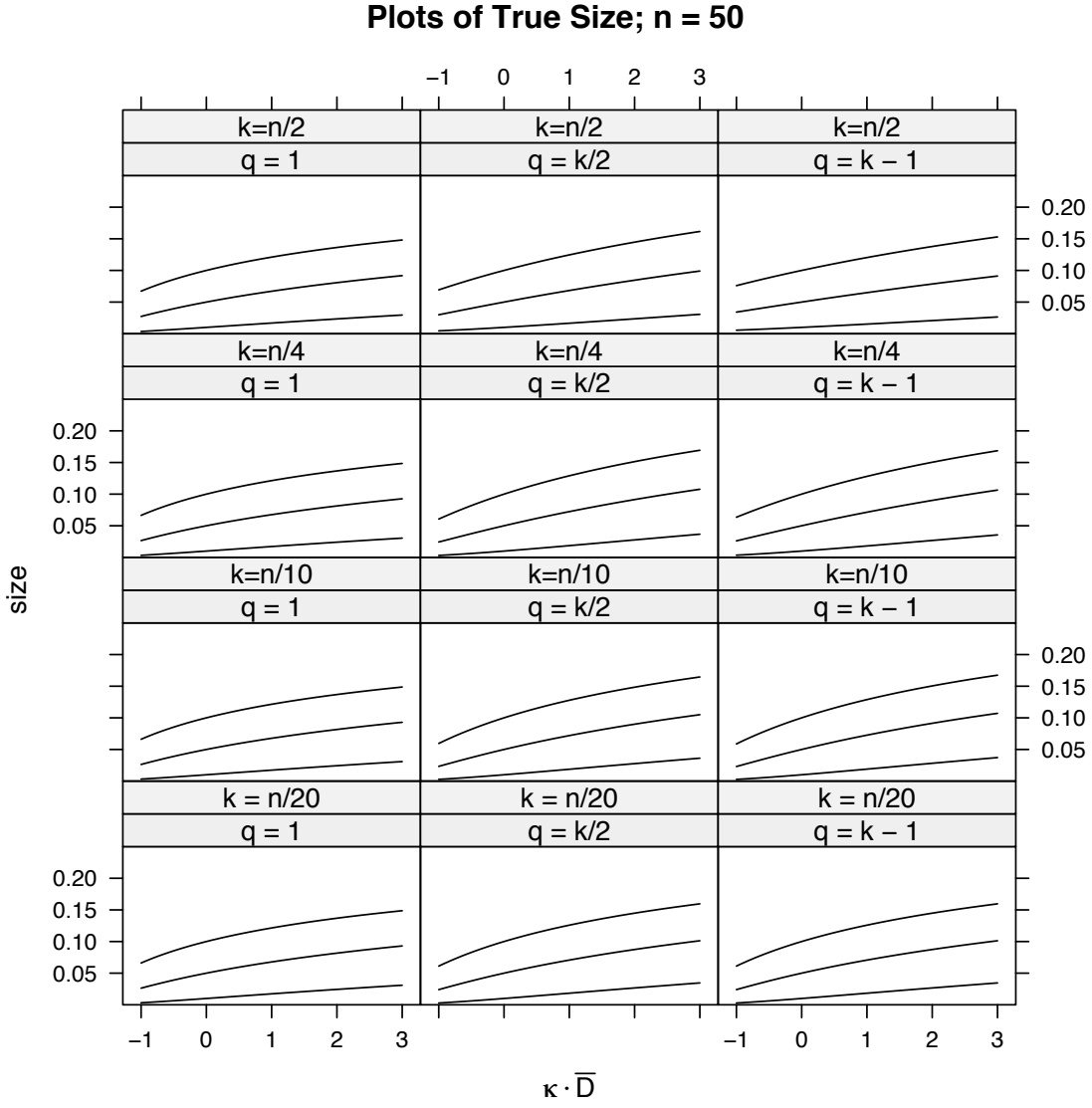

Figure 1.6: Approximate size of the F-test for different values of k and q. See Figure 1.2 for legend.

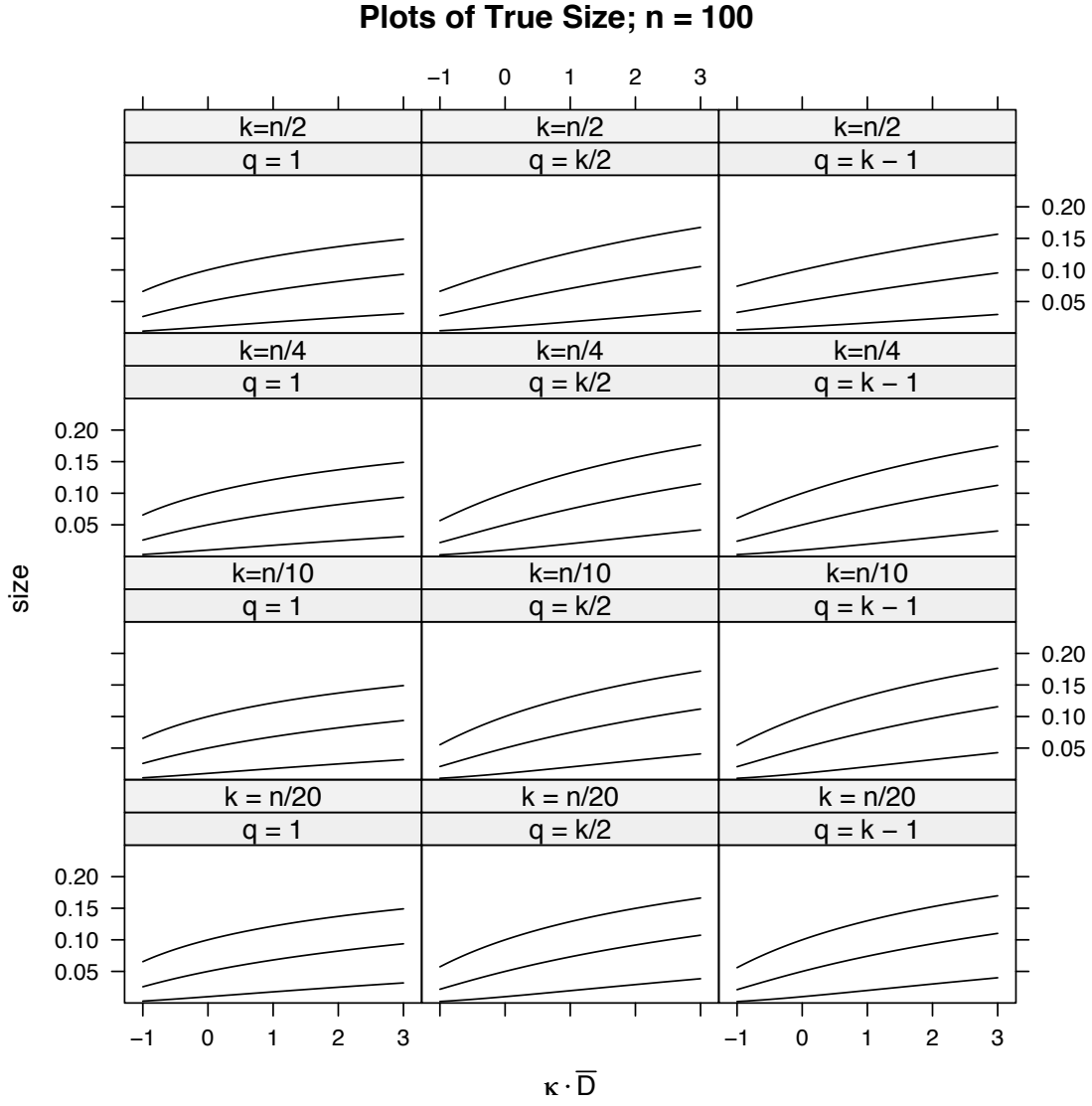

Figure 1.7: Approximate size of the F-test for different values of k and q. See Figure 1.2 for legend.

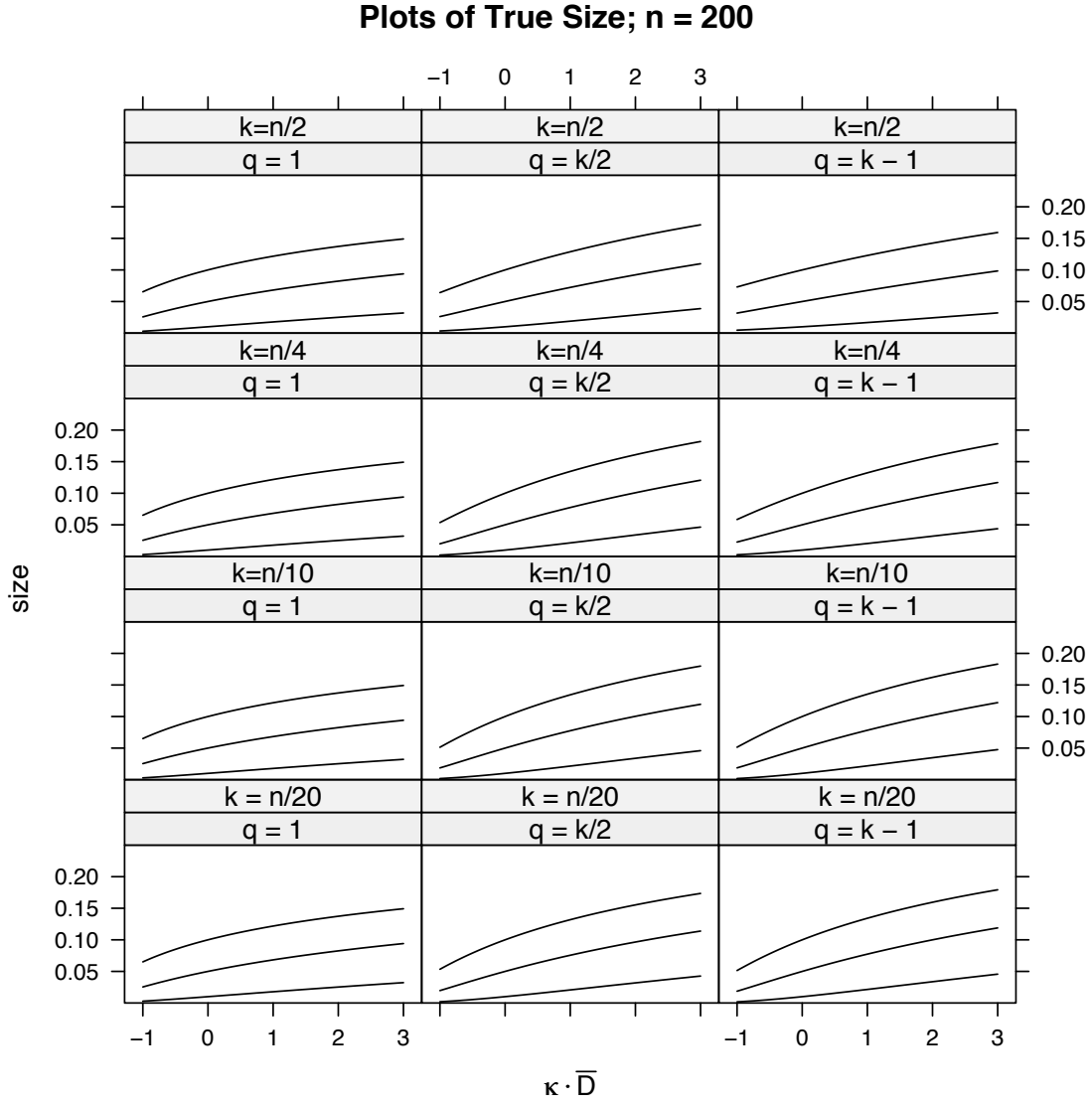

Figure 1.8: Approximate size of the F-test for different values of k and q. See Figure 1.2 for legend.

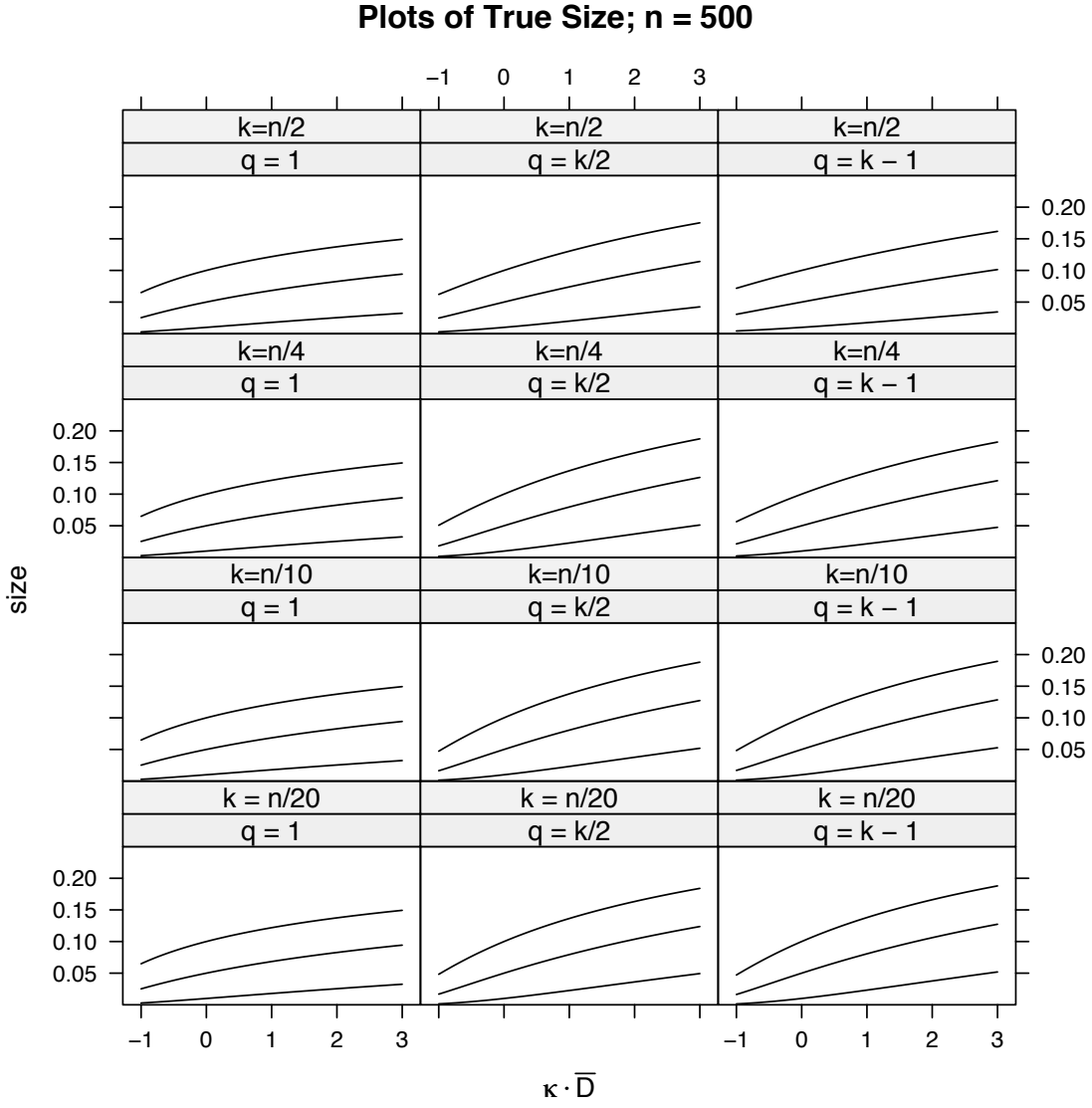

Figure 1.9: Approximate size of the F-test for different values of k and q. See Figure 1.2 for legend.

Chapter 2

Limit theory for comparing overfit models out-of-sample

2.1 Introduction

Consider two sequences of prediction errors, each of length P, the result of forecasting the same variable with two different estimated models. The R observations used to estimate the models are called, collectively, the estimation window, and the P observations used to produce the errors are called the test sample. There are T observations in all, and $R+P=T$. This paper introduces a new asymptotic theory for the sample moments of these prediction errors that assumes at least one model is overfit, that statistics calculated from that model's sample residuals are unreliable because its estimated coefficients match the data too well.

If one of these models nests the other, a researcher can determine whether the additional regressors help predict the dependent variable (in population) by using one of two representative methods. He or she can run a regression over the full dataset and use a robust F-test to directly test whether the coefficients on the extra predictors equal zero. Or he or she could instead produce a series of forecasts with each model, and then test whether the models' predictive mean squared errors (PMSE) are equal. The first test is simple and easy, and the second test is not; Clark and McCracken (2001) and McCracken (2007) show that the

test statistic is not asymptotically normal and that the critical values must be calculated by monte-carlo or bootstrap. But the out-of-sample test is much more popular among forecasters because it is believed to be more reliable: see Meese and Rogoff (1983), Stock and Watson (2003) and Lettau and Ludvigson (2001) for prominent examples.

The asymptotic theory used to derive these out-of-sample statistics, however, does not imply that they are any more reliable than their in-sample counterparts. There are two main approaches to approximate these statistics' distributions. The first, begun by Diebold and Mariano (1995) and West (1996) and extended to nested models by Clark and McCracken (2001), Chao, Corradi, and Swanson (2001), Corradi and Swanson (2002), and McCracken (2007), finds the limit distribution as though all of values of the models' coefficients were known, and then adjusts that distribution to account for estimation error.[1] This adjustment requires the estimated coefficients to be consistent and asymptotically normal. But if they were, a forecaster could instead just use an F-test directly on the coefficients.

The second method, proposed by Giacomini and White (2006), is becoming a popular alternative to West's (1996) approach because it always gives an asymptotically normal statistic. Their method requires that the forecaster use a fixed-length rolling estimation window — the model for each forecast is estimated over the previous R observations — and they derive a statistic to test whether the difference between the models' forecasts is predictable.[2] This rolling window can be appropriate if the series exhibit substantial instability, so Giacomini and White's (2006) out-of-sample statistic should be reliable in settings where an in-sample statistic is not. But, by its nature, their statistic measures the forecasting performance of the estimated models, and does not reflect the predictive content

[1] Diebold and Mariano (1995) assume that the coefficients are known. West (1996) introduces the adjustment.

[2] Although Giacomini and White (2006) claim that their results also apply to fixed-length fixed window schemes, their proofs do not apply to that situation. For a rolling window of length R, the period-s prediction error only depends on the observations of period $s - R - 1$ through period s, and so, for any finite R, this sequence inherits mixing from the underlying observations. But for a fixed window of length R, the period-s prediction error depends on all of the past observations, periods 1 through s, and so does not inherit mixing from the underlying observations. Giacomini and White's proofs require the prediction errors to be mixing.

of their variables in population. So Giacomini and White's method suggests a new criterion to use to choose between forecasting methods, but does not indicate why a researcher who wants to study the true relationship between several series should use an out-of-sample statistic.

Although the foundation of West's approach — asymptotic normality of the coefficient estimates — implies that in-sample and out-of-sample statistics should behave similarly, in practice they do not. Stock and Watson (2003), for example, show that in-sample tests of Granger Causality are much more likely to find predictive relationships when predicting output and inflation than out-of-sample tests. As Inoue and Kilian (2005) argue, this fact can support either approach: either the in-sample tests do not preserve size, or the out-of-sample tests have low power. The monte carlo evidence is mixed.[3]

This paper introduces a new approximation for out-of-sample sample averages that does not require the coefficient estimates to be consistent or asymptotically normal. This approximation allows us to study the behavior of out-of-sample statistics when one of the models is overfit, and we find that these statistics can be reliable even when in-sample statistics are not. We impose overfit by studying the limit distribution of a sequence of linear regression models in which the ratio K/T remains positive; K is the number of predictors. Huber (1973) shows that the OLS estimators are not asymptotically normal under this limit theory and that they have positive variance in the limit, so robust F-tests should be invalid. This new approximation gives evidence that the out-of-sample average loss is a reliable statistic for comparing complex, overfit models to a simple benchmark and that ad hoc critical values, such as Clark and McCracken's (2001), are not necessary; using the standard normal critical values gives valid but conservative tests. The approximation also leads to a new criterion to use to evaluate a model's expected forecasting performance.

This increasing-dimension asymptotic theory is discussed in detail in Sections 2.2 and 2.3. Section 2.2 presents the notation, assumptions, and models that the rest of the paper will use, and Section 2.3 studies the behavior of West's

[3]This issue has been studied both analytically and in simulations by Inoue and Kilian (2005, 2006), McCracken (1998), Clark (2004), Clark and McCracken (2005), and Chen (2005).

(1996) approximation under this new limit theory. West's asymptotic distribution no longer holds because the test statistics are not centered on the expected loss of the pseudo-true models. Section 2.3 also presents the results of a brief simulation demonstrating that conventional fixed-K asymptotic theory can be inaccurate in finite samples when K/T is as small 0.03.

Section 2.4 presents our new approximation. The fixed-window average loss converges to a measure of the expected performance of the estimated models and is asymptotically normal in most applications. Section 2.5 shows that this average can be used to compare nested models when the larger model is overfit, and that using standard normal critical values leads to an asymptotically valid (if conservative) test statistic.

Section 2.6 discusses the relationship between the models' average loss and their future forecasting performance and shows that the conventional wisdom articulated by Hastie, Tibshirani, and Friedman (2003) among others, that one should split the dataset in half or in thirds is wrong: to estimate a model's future performance (under this paper's limit theory) P/T must converge to zero and to construct a confidence interval, P^2/T must converge to zero.

In short, this paper studies out-of-sample averages to see if they are more robust to overfit than in-sample statistics and finds that they are. It shows that using a one-sided t-test to compare nested models is conservative but preserves size as long as the benchmark is simple, consistent with simulations presented in Clark and West (2007) but in contrast to existing theory. It also shows that the test sample must be much smaller than is commonly used if one hopes to accurately compare the performance of the full-sample models that will be used to produce real forecasts.

2.2 Notation and Assumptions

This paper does not define a specific data generating process but instead assumes that there are two competing forecasting models and both are misspecified. The underlying data are represented as a stationary and absolutely regular

stochastic array:

$$\{(y_{T,t}, x_{1T,t}, x_{2T,t});\ t = 1, \ldots, T+1;\ T \text{ an integer}\}, \tag{2.2.1}$$

and $\mathcal{F}_{T,t}$ denotes the information available in period t:

$$\mathcal{F}_{T,t} \equiv \sigma(y_{T,1}, x_{1T,1}, x_{2T,1}, \ldots, y_{T,t}, x_{1T,t}, x_{2T,t}).$$

The first model uses $x_{1T,t}$ as predictors for $y_{T,t}$ and the second uses $x_{2T,t}$. Each vector $x_{jT,s}$ has K_{jT} elements; K_{jT} can change with T, but is always assumed to be less than R_T (the sequence of estimation window sizes, $\{R_T\}$, will be discussed in detail later in the paper). Moreover, T and R_T are implicitly assumed to be large enough that all of the operations in this paper are well defined. To keep the presentation relatively clean, the T subscript will be removed whenever possible. Although we only present results for one-period forecasts, these results can be generalized easily to multi-period forecasts.

Assumption 4 states the moment and dependence conditions that the array (2.2.1) must satisfy.

Assumption 4. The random array (2.2.1) is stationary and absolutely regular with coefficients β_τ of size $-\rho/(\rho-2)$; ρ is greater than two and discussed further in Assumption 6. The variance of $y_{T,t}$ is uniformly positive and finite, and all of the eigenvalues of the covariance matrices of $x_{jT,T}$ are uniformly positive and finite as well. ■

Each forecast is produced by a linear model; model j's forecast for period $T+1$ is $x'_{jT+1}\hat{\theta}_{jT}$, with $\hat{\theta}_{jS}$ the OLS estimate using observations one through S. The models' pseudo-true coefficients are denoted θ^*_j and defined by

$$\theta^*_{jT} \equiv argmin_\theta \, \mathrm{E}(y_{T,T+1} - x'_{jT,T+1}\theta)^2.$$

Assumption 5 rules out the uninteresting cases where the forecast error vanishes.

Assumption 5. The Euclidean norm of the pseudo-true coefficients satisfies $|\theta^*_{jT}|_2 = O(1)$, and the population residuals, $\varepsilon_{jT,t} \equiv y_{T,t} - x'_{jT,t}\theta^*_{jT}$, have uniformly positive and finite variance. ■

Since this paper will apply limit theory to the average loss over the test sample, the observed test sample loss satisfies moment restrictions. The observed loss for model j in period s is L_{js} which is defined by

$$L_{js} = L(e_{jt}) \equiv L(y_s - x'_{js}\hat{\theta}_{jR});$$

L is the loss function of interest. The period-s loss for the forecast produced by model j's pseudo-true coefficients is $L^*_{js} \equiv L(\varepsilon_{js})$, and the period-$s$ loss for the forecast produced by the full-sample estimates is $L^T_{js} \equiv L(y_s - x'_{js}\hat{\theta}_{jT})$.

The vector $\bar{L}$ denotes the average loss of the estimated models over the test sample:

$$\bar{L}_j \equiv P^{-1} \sum_{s=R+1}^{T} L_{js}.$$

Assumption 6 restricts the moments of L_s.

Assumption 6. The loss function L is convex and there is a constant B_L such that $\|L_{jT,s}\|_\rho \leq B_L$ for all j, s, and T. Moreover, the function $\mathrm{E}\, L(y_{T,T+1} - x'_{jT,T+1}\theta)$ is continuously differentiable in θ. ■

Finally, $|\cdot|_v$ is the l_v-norm for vectors in $\mathbb{R}^p$ (p arbitrary) and $\|\cdot\|_v$ the L_v-norm for L_v-integrable random variables. The functions $\lambda_i(\cdot)$ take a square matrix argument and return its ith eigenvalue. All limits are stated for $T \to \infty$ unless explicitly labeled otherwise.

2.3 Background

A large sample does not guarantee the accuracy of an asymptotic result; other factors come into play, among them the complexity of the model estimated. Clark and West (2006, 2007) show that parameter estimation error affects the analysis of nested models in practice, making West's (1996) limit theory unreliable. This section shows that the same problems occur without nesting and motivates the use of increasing-K asymptotics as a method of studying that estimation error formally.[4]

[4]For this section, assume that there is only one model under consideration.

It is natural to use sequences of models with K/T positive to study the failure of in-sample asymptotics because applied researchers choosing between models routinely make adjustments that are proportional to K/T. Statisticians have proposed model selection criteria, such as Mallow's C_p (Mallows 1973), the AIC, (Akaike 1973) and cross validation, because a model's *apparent error*, its average loss over the dataset used to estimate its unknown coefficients, is a biased estimator of the expected loss one encounters after using that model to predict new observations. The bias is proportional to K/T, and these criteria are estimated by correcting the apparent error by a term proportional to K/T (see, for example, Efron 1986, 2004).

Moreover, since the variance of each element of $\hat{\theta}_T$ is of order T^{-1}, the variance of this vector does not vanish if K/T remains positive. The behavior of M-estimators in this setting has been studied by Huber (1973), Yohai and Marona (1979), and Portnoy (1985, 1986). If the coefficients are estimated by OLS, they are consistent and asymptotically normal only if K/T converges to zero; if they are general M-estimates they require a faster rate of convergence. By keeping K/T positive, one can keep the variance of the coefficient estimates positive. Since the coefficients are not asymptotically normal, the F-test is not necessarily asymptotically chi-square, so in-sample tests are unavailable.

Expansions of $\bar{L}$ around the pseudo-true coefficients, like West's (1996) and Clark and McCracken's (2001), also do not hold when K/T remains positive. These approximations require that the coefficient estimates be root-R consistent. For a short illustration, suppose that the underlying series are stationary; that the same loss function is used to estimate the coefficients and evaluate the forecasts; that P and R both equal $T/2$; and that a fixed-window is used for evaluation.[5] In this special case, West's theory gives the approximation

$$P^{1/2}(\bar{L} - \mathrm{E}\, L^*_{T+1}) = P^{-1/2} \sum_{s=R+1}^{T} (L^*_s - \mathrm{E}\, L^*_s) + o_p(R^{1/2}|\hat{\theta}_R - \theta^*|_2).$$

Under conventional (fixed-K) asymptotic theory, $R^{1/2}|\hat{\theta}_R - \theta^*|_2$ is tight, and the last term of the approximation converges to zero in probability. Then

[5]These assumptions are not necessary, but simplify equation (**??**).

Table 2.1: Simulated Size. Nominal size is 10%

	P =	40	80	120	160
R = 80		15.4	18.5	19.0	21.0
160		12.0	14.5	15.9	17.2
		.	.	.	.

According to West's (1996) theory, the random variable (2.3.1) is asymptotically standard normal. Each cell lists the percentage of simulations for which that r.v. exceeds 1.282, the 90%-quantile of the standard normal distribution.

$P^{1/2}(\bar{L} - \mathrm{E}\, L^*_{T+1})$ is asymptotically normal as long as $P^{-1/2}\sum_s (L^*_s - \mathrm{E}\, L^*_s)$ is. But when K grows proportionally to T, $R^{1/2}|\hat{\theta}_R - \theta^*|_2$ is not tight. And since $\hat{\theta}_R$ is not asymptotically normal, the distribution of the remainder term is not known. In general, the variance of the fixed-window average remains positive asymptotically, so $\bar{L}$ does not converge to any non-random value. Sections 2.4 and 2.5 will explore this convergence in more detail.

In practice, K/T will always be positive, so this discussion indicates that West's approximation is unreliable unless this ratio is close to zero, but it is not necessarily clear how small the ratio must be to make West's results accurate. Some brief simulations demonstrate that the size of his tests can be too high even when K/T is as small as 0.03, and empirically relevant values of K/T seem to be from 0.01 to about 0.2. Economists are sometimes interested in studying tightly parameterized models, but they are often interested in larger structural models. Meese and Rogoff's (1983) seminal study of exchange rate models, for example, includes some models for which the K/T is almost 0.3. More recently, Stock and Watson (2003) consider a range of output and inflation forecasts with this ratio between 0.01 and 0.08, and Negro, Schorfheide, Smets, and Wouters (2007) study Bayesian DSGE models for which it is roughly 0.15.

To see whether West's theory is accurate for these values of K/T, we ran a brief monte-carlo experiment to study forecasting with a bivariate VAR(4); T ranges from 120 to 320, so K/T is between 0.025 and 0.075. The DGP is taken from Clark and West (2007) and is designed to represent macroeconomic forecasting. Each entry in Table 1 is the simulated size of a test at the 10%-level of the null

hypothesis that the expected PMSE of the bivariate VAR(4) is less than or equal to its true value. We ran 3000 simulations: for each simulation we drew 320 observations from the stationary process with independent innovations:

$$\begin{aligned} y_t &= 2.237 + 0.261 y_{t-1} + \varepsilon_{1t} \\ z_t &= (0.804, -0.221, 0.226, -0.205) \cdot (z_{t-1}, \ldots, z_{t-4}) + \varepsilon_{2t} \\ \varepsilon &\sim N\left(\mathbf{0}, \begin{pmatrix} 10.505 & 1.036 \\ \cdot & 0.366 \end{pmatrix}\right). \end{aligned}$$

For each R and P, we constructed sequences of forecasts, $\hat{y}_t$, $t = R+1, \ldots, R+P$ from the bivariate VAR(4):

$$\hat{y}_t = \hat{\alpha}_0 + \sum_{j=1}^{4} \hat{\alpha}_j y_{t-j} + \sum_{j=1}^{4} \hat{\beta}_j z_{t-j}$$

with the coefficients estimated recursively by OLS.

Each cell in Table 1 is the percentage of simulations for which the random variable

$$P^{1/2}(\bar{L} - 10.505)/\hat{\sigma} \tag{2.3.1}$$

is greater than 1.282, with

$$\bar{L} \equiv P^{-1} \sum_{s=R+1}^{R+P} (y_s - \hat{y}_s)^2$$

and

$$\hat{\sigma}^2 \equiv (P-1)^{-1} \sum_{s=R+1}^{R+P} [(y_s - \hat{y}_s)^2 - \bar{L}]^2.$$

Under West's (1996) limit theory, this random variable is approximately standard normal, so each entry should be close to 10%; but the true size is higher, ranging from 12% to 21%. Since the parameter values of this simulation were chosen by Clark and West (2007) to represent quarterly macroeconomic data, the values $R = 80$ and $P = 160$ roughly correspond to the common practice of estimating forecasting models with pre-1970 data and assessing its performance recursively from 1970 on:[6] the true size is roughly twice the nominal size when testing this one-sided hypothesis with such a procedure. In general, increasing P while keeping

[6] As in, for example, Stock and Watson (2003)

R constant distorts the size more, suggesting that the variance of the out-of-sample average decreases, but that the average is not centered at $\mathrm{E}\, L^*_{T+1}$.

2.4 The New Approximation

Instead of using an expansion, we re-center each random sequence $\{L_s\}$ by subtracting a random vector $M(\hat{\theta}_R)$ from each term. The new sequence is a mixingale, and $M(\hat{\theta}_R)$ can be interpreted as a conditional expectation after a change-of-measure.

Lemma 2.4.1. *Suppose that Assumptions 4 to 6 hold and define*

$$M_{iT}(\theta) \equiv \mathrm{E}\, L(y_{T,T+1} - x'_{iT,T+1}\theta).$$

Then, for any T, any positive j, and any τ between zero and j,

$$\| \mathrm{E}(L_{iT,R_T+j} \mid \mathcal{F}_{T,R_T+j-\tau}) - M_{iT}(\hat{\theta}_{iT,R_T})\|_2 \leq 2^{1+1/\rho} B_L\, \zeta_\tau \tag{2.4.1}$$

with $\zeta_\tau = O(\tau^{-1/2-\delta})$ for some positive δ. So the array

$$\{L_{T,R_T+j} - M_T(\hat{\theta}_{T,R_T}), \mathcal{F}_{T,R_T+j}\}$$

is an L_2-mixingale array of size $-1/2$. ■

The proof relies on a coupling argument due to Merlevède and Peligrad (2002) that builds on Berbee's Lemma (Berbee 1979). Merlevède and Peligrad's (2002) statement of the Lemma is repeated here verbatim for reference.

Berbee's Lemma (Merlevède and Peligrad 2002). *Let X and Y be random variables defined on $(\Omega, \mathcal{T}, \mathbb{P})$ with values in a Polish space S. Let $\sigma(X)$ be a σ-field generated by X and U be a random variable uniformly distributed on $[0,1]$ independent of (X,Y). Then there exists a random variable Y^* measurable with respect to $\sigma(X) \vee \sigma(Y) \vee \sigma(U)$, independent of X and distributed as Y, and such that*

$$\mathbb{P}(Y \neq Y^*) = \beta(X,Y).$$

■

The coefficient β is the coefficient of absolute regularity.

Merlevède and Peligrad (2002) use this result to bound the L_p-norm of the distance between Y and Y^*. We prove Lemma 2.4.1 by defining (y^*, x^*) to satisfy:

1. $(y^*, x^*) \stackrel{d}{=} (y_{T,R_T+j}, x_{iT,R_T+j})$
2. (y^*, x^*) is independent of $\mathcal{F}_{T,R_T+j-\tau}$.
3. $\mathbf{P}[(y^*, x^*) \neq (y_{T,R_T+j}, x_{iT,R_T+j})] = \beta_\tau$.

Since $M_{iT}(\hat{\theta}_{iT,R_T}) = \mathrm{E}(L(y^* - x^* \cdot \hat{\theta}_{iR_T}) \mid \mathcal{F}_{T,R_T+j-\tau})$ almost surely, the left side of (2.4.1) is bounded by

$$\|L_{iR_T+j} - L(y^* - x^* \cdot \hat{\theta}_{iR_T})\|_2.$$

We can directly use Merlevede and Peligrad's method of proof to bound this last distance.

An immediate consequence of Lemma 2.4.1 is that $\bar{L}_{iT} - M_{iT}(\hat{\theta}_{iT,R_T})$ converges to zero almost surely as P_T increases. We can also show that $\sqrt{P_T}[\bar{L}_T - M_T(\hat{\theta}_{T,R_T})]$ is asymptotically normal under a slightly stronger condition that its asymptotic variance is positive definite.

Lemma 2.4.2. *Suppose that the conditions of Lemma 2.4.1 hold and that*

$$\lambda_{min}(\Sigma_T(\hat{\theta}_{T,R_T}))^{-1} = O_p(1),$$

with each element of $\Sigma_T(\cdot)$ *defined by*

$$[\Sigma_T((\theta_1, \theta_2))]_{ij} \equiv P^{-1} \sum_{s,t=R+1}^{T} \left[\mathrm{E}\, L(y_{T,s} - x'_{iT,s}\theta_i)L(y_{T,t} - x'_{jT,t}\theta_j) - M_{iT}(\theta_i)M_{jT}(\theta_j)\right].$$

If $P_T \to \infty$ *as* $T \to \infty$, *then*

$$P_T^{1/2}\Sigma_T(\hat{\theta}_{T,R_T})^{-1/2}\left[\bar{L}_T - M_T(\hat{\theta}_{T,R_T})\right] \xrightarrow{d} N(0, I) \quad \text{as } T \to \infty. \tag{2.4.2}$$

■

The proof is based on de Jong's (1997) Theorem 1, a mixingale Central Limit Theorem.

This mixingale approximation allows us to work with sequences of estimators that do not converge. As in Giacomini and White (2006), such sequences can ensure that the asymptotic variance matrix remains positive definite, even for nested models.

It is useful to compare this approximation to McCracken's (2000). McCracken assumes that the estimators $\hat{\theta}_{T,R_T}$ are asymptotically normal and that the function $M_T(\cdot)$ is smooth enough that

$$\sqrt{R_T}[M_{iT}(\hat{\theta}_{iT,R_T}) - M_{iT}(\theta^*_{iT})]$$

is asymptotically normal as a consequence of the delta-method. As a result, he, like West (1996), can apply a central limit theorem to $P_T^{1/2}[\bar{L}_T - M_T(\theta^*_T)]$ and then adjust its covariance matrix to account for the difference between $M_{iT}(\hat{\theta}_{iT,R_T})$ and $M_{iT}(\theta^*_{iT})$.

Lemmas 2.4.1 and 2.4.2 show that we can apply a Central Limit Theorem directly to $P_T^{1/2}[\bar{L}_T - M_T(\hat{\theta}_{T,R_T})]$. This extra generality allows us to study how out-of-sample averages perform when their models are estimated imprecisely. However, we need to impose more restrictions before we can relate $M_T(\hat{\theta}_{T,R_T})$ to objects that a researcher should be interested in, such as $M_T(\theta^*_T)$.

2.5 Comparing Nested Models

In most applications, the benchmark model is very simple (a random walk, for example) so we can treat the smaller model as having fixed K. Only the alternative model is complex. In this case, we can construct a conservative one-sided test for the null hypothesis that smaller model is more accurate (in terms of MSE) in population.

Theorem 2.5.1. *Suppose that the conditions of Lemmas 2.4.1 and 2.4.2 hold and that*

(i) $K_{1T} = K_1$ *and does not change with* T*;* $K_{2T}/T \to k_2 > 0$*; and* $P_T/R_T \to \pi < \infty$*.*

(ii) $\hat{\sigma}_T$ *is a consistent estimator of* $(1,-1)\Sigma_T(1,-1)'$*.*

(iii) Each of the smaller model's predictors is also used by the larger model (model two).

(iv) $L(e) = e^2$*.*

Then, under the null hypothesis

$$H_0: \qquad \mathrm{E}\, L^*_{1T,T+1} = \mathrm{E}\, L^*_{2T,T+1} \qquad \textit{for all } T,$$

the one-sided t-test satisfies

$$\lim_{T\to\infty} \mathbf{P}\left[P_T^{1/2}(\bar{L}_{1T} - \bar{L}_{2T})/\hat{\sigma}_T \geq z_\alpha\right] \leq \alpha,$$

with z_α *the* $(1-\alpha)$*-quantile of the standard normal distribution.* ■

In short, one-sided tests for nested models that (erroneously) act as though the out-of-sample average were normal with mean $\mathrm{E}\, L^*_{T+1}$ are asymptotically valid.

This theorem relies on the inequality

$$(\bar{L}_1 - \bar{L}_2) - (\mathrm{E}\, L^*_{1T+1} - \mathrm{E}\, L^*_{2T+1}) \leq \left[\bar{L}_1 - M_1(\hat{\theta}_{1R})\right] - \left[\bar{L}_2 - M_2(\hat{\theta}_{2R})\right] + o_p(R^{-1/2}), \quad (2.5.1)$$

which holds because $\mathrm{E}\, L^*_{1T+1} = M_1(\hat{\theta}_{1R}) + o_p(R^{-1/2})$ and $\mathrm{E}\, L^*_{2T+1} \leq M_2(\hat{\theta}_{2R})$.

We should discuss the variance estimator, $\hat{\sigma}_T$, further. Since the asymptotic variance, $\Sigma_T(\hat{\theta}_{T,R_T})$, is a random element that depends on $\hat{\theta}_{T,R_T}$, the usual proofs that HAC estimators are consistent do not apply. Moreover, those proofs require NED sequences, not mixingales, so they would not apply anyway. Because of the special structure of our mixingale process, though, it is straightforward to prove the consistency of HAC estimators using a coupling argument similar to the one used to prove Lemmas 2.4.1 and 2.4.2. In fact, one can simply mimic the available NED proofs. In this paper, we will simply assume the existence of a consistent

estimator. Lemma 2.A.2 (in the Appendix) contains the basic argument for how to modify existing proofs, Davidson and de Jong (1998, 2000) in particular.

In this case, there are more basic assumptions that guarantee that the covariance matrix $\Sigma_T(\hat{\theta}_{T,R_T})$ is uniformly positive definite. Remember that the null hypothesis imposes that the coefficients on the additional predictors used by the larger model are zero. Under Theorem 2.5.1's assumptions, the models' prediction errors satisfy the relationship

$$\begin{aligned} e_{1s} &= \varepsilon_{1s} + o_p(1) \\ e_{2s} &= \varepsilon_{1s} + z_s' \hat{\alpha}_{R_T} + o_p(1) \end{aligned}$$

with z_s the additional predictors and $\hat{\alpha}_{R_T}$ their coefficient estimates. As long as z_s and $\hat{\alpha}_{R_T}$ are almost surely not zero, the two forecasts are almost surely different.

This example is generalized slightly by the next lemma.

Lemma 2.5.2. *Suppose that the conditions of Lemma 2.4.1 and the additional conditions of Theorem 2.5.1 hold. In addition, suppose that*

(i) The maximum eigenvalues of

$$R_T^{-1} \boldsymbol{X}'_{iT,R_T} \boldsymbol{X}_{iT,R_T}$$

and

$$R_T^{-1} \boldsymbol{X}'_{iT,R_T} \boldsymbol{\varepsilon}_{iR_T} \boldsymbol{\varepsilon}'_{iR_T} \boldsymbol{X}_{iT,R_T}$$

are $O_p(1)$ and their minimum eigenvalues are bounded away from zero in probability.

(ii) The first model's innovations, $\varepsilon_{1T,t}$, are sequentially exogenous and independent of $x_{1T,s}$ and $x_{2T,s}$ for $s = 1, \ldots, t$.

(iii) The elements of $x_{2T,t}$ and $\hat{\theta}_{2T,R_T}$ are continuous random variables.

Then $[(1,-1)\Sigma_T(\hat{\theta}_{T,R_T})(1,-1)']^{-1} = O_p(1)$. ■

These eigenvalue conditions are analogous to the usual restrictions made on asymptotic variance matrices. The strong assumption of sequential exogeneity simplifies the proof but is not crucial. The fact that the random variables

are continuous rules out the possibility that, for example, the estimation error, $x'_{2T,T+1}(\hat{\theta}_{2T,R_T} - \theta^*_{2T})$, is zero.

To summarize this section: researchers can use fixed-window out-of-sample averages to compare the pseudo-true performance of an overfit model to a simple benchmark. If the benchmark is also overfit (i.e., K_{1T}/T remains positive), this approach does not work because inequality (2.5.1) does not hold. Also, if different loss functions are used to estimate and evaluate the model, this approach again does not work because inequality (2.5.1) does not hold. But Theorem 2.5.1 and Lemma 2.5.2 justify using the naive out-of-sample one-sided t-test for most empirical research.

2.6 Comparing Finite-Sample Performance

This section gives conditions under which $M_T(\hat{\theta}_{T,R_T})$ converges to $M_T(\hat{\theta}_{T,T})$ and suggests that $M_T(\hat{\theta}_{T,T})$ be used as a criterion for choosing between forecasting models.

Ideally, a forecaster choosing between two models to use for period $T+1$ would chose the one that minimizes the expected loss given the information available in period T, $\mathrm{E}(L^T_{jT+1} \mid \mathcal{F}_{n,T})$. Even if the underlying random variables are independent, $\mathcal{F}_T$ has valuable information about the models' performance — the values of the coefficient estimates $\hat{\theta}_{1T}$ and $\hat{\theta}_{2T}$. But usually a model will use lagged variables as predictors, and their values are also included in $\mathcal{F}_T$. Consequently, a forecaster choosing the model that minimizes the conditional expectation will make better forecasts than one who minimizes the unconditional expected loss, $\mathrm{E}\, L^T_{jT+1}$.

For i.i.d. series, $M_T(\hat{\theta}_{T,T}) = \mathrm{E}(L^T_{T+1} \mid \mathcal{F}_{n,T})$ almost surely. With dependence, $M_T(\hat{\theta}_{T,T})$ ignores the past information beyond the value of the coefficient estimate, and so is a biased proxy for the true conditional expectation. However, the true conditional expectation can be difficult to estimate, and the fixed-window out-of-sample average can be used to estimate $M_T(\hat{\theta}_{T,T})$.

Lemma 2.6.1. *Suppose that Assumptions 4 to 6 hold and that*

(i) For each sequence $\{s_T\}$ *with* s_T *between* R_T *and* T*, the maximum eigenvalues of* $s_T^{-1}\boldsymbol{X}'_{iT,s_T}\boldsymbol{X}_{iT,s_T}$ *and* $s_T^{-1}\boldsymbol{X}'_{iT,s_T}\boldsymbol{\varepsilon}_{is_T}\boldsymbol{\varepsilon}'_{is_T}\boldsymbol{X}_{iT,s_T}$ *are* $O_p(1)$ *and their minimum eigenvalues are bounded away from zero in probability.*

(ii) $P_T/R_T \to 0$*,* $K_{1T}/T \to k_1$*, and* $K_{2T}/T \to k_2$*. Both* k_1 *and* k_2 *are less than one.*

(iii) L has finite left- and right-derivatives at every point, denoted $D_L(\cdot)$ *and* $D_R(\cdot)$ *respectively.*

Then

$$M_{iT}(\hat{\theta}_{iT,R_T}) - M_{iT}(\hat{\theta}_{iT,T}) = O_{L_1}(\sqrt{P_T/R_T}),$$

and so $\bar{L}_T - M_T(\hat{\theta}_{T,T}) \to 0$ *in probability.* ■

Theorem 2.6.2 is an immediate corollary.

Theorem 2.6.2. *Suppose the conditions of Lemmas 2.4.2 and 2.6.1 hold. If* $P_T^2/R_T \to 0$*, then*

$$P_T^{1/2}\Sigma_T(\hat{\theta}_{T,R_T})^{-1/2}\left[\bar{L}_T - M_T(\hat{\theta}_{T,T})\right] \xrightarrow{d} N(0,I) \quad \text{as } T \to \infty.$$

■

These results suggest two new ideas. When the models are overfit, out-of-sample averages implicitly condition on the coefficient estimates.[7] In economic forecasting applications, such conditioning is desirable — we will make predictions for the same series using those coefficients, so averaging the forecasts' performance over other hypothetical values for those coefficients is inappropriate. However, unless P/R is very small, the forecasts comparative performance can change when the models are re-estimated over the entire dataset. Since, in practice, P must also be large enough to justify a Law of Large Numbers or Central Limit Theorem, these out-of-sample statistics may have limited use in macroeconomics.

[7] Efron (1986) makes a similar point about cross-validation in finite samples. Our result, though, is the first that we are aware of to study such conditioning asymptotically.

2.7 Conclusion

By studying the behavior of the fixed-window out-of-sample average under a new limit theory that increases the number of predictors with the number of observations, this paper shows that these out-of-sample tests can prevent overfit and are properly sized when in-sample tests are not. Many of the previously known results on these statistics do not carry over to this setting, though: the performance of the model's pseudo-true coefficients can not be estimated, but researchers can still construct some one-sided confidence intervals; nested comparisons are asymptotically normal; and the test sample must be extremely small if this out-of-sample exercise will estimate how well the models perform when they are re-estimated over the full dataset.

Future research should study whether it is possible to improve the power of out-of-sample tests while preserving size under this asymptotic theory in the manner of Clark and West (2006, 2007); whether resembling techniques can improve the restrictions on P/R; and how these results can be extended to M-estimators and nonlinear models.

2.A Additional Technical Results and Proofs

Lemma 2.A.1. *Suppose Assumptions 4 to 6 hold. Then, for any T, s, t, and u, with $t \geq s > u \geq R_T$ there exists an array $\{\tilde{L}_{jv}; v = s, \ldots, t;\ j = 1, 2\}$ such that*

$$\mathrm{E}\left(\phi(\tilde{L}_s, \ldots, \tilde{L}_t) \mid \mathcal{F}_{T,u}\right) = \int \phi(L_{T,s}, \ldots, L_{T,t})\, \mathbf{P}(dx_{T,s}, dy_{T,s}, \ldots, dx_{T,t}, dy_{T,t}) \tag{2.A.1}$$

almost surely for all measurable functions ϕ such that the expectations are finite. Moreover,

$$\mathbf{P}[\tilde{L}_{jv} \neq L_{jT,v} \text{ for at least one } v \text{ and } j] = \beta_{s-u} \tag{2.A.2}$$

and

$$\|\tilde{L}_v - L_{T,v}\|_2 \leq 2^{1+1/\rho} B_L \beta_{s-u}^{(\rho-2)/2\rho} \quad \text{for each } v \text{ and } j. \tag{2.A.3}$$

Proof. Fix T, t, s, and u. The array $\mathcal{A} \equiv \{(y_{T,\tau}, x_{T,\tau}, \ldots, y_{T,\tau+t-s}, x_{T,\tau+t-s});\ \tau\}$ is also absolutely regular of size $\rho/(\rho-2)$, so Berbee's Lemma allows us to construct

a new array $\mathcal{A}^* \equiv \{(y^*_\tau, x^*_\tau, \ldots, y^*_{\tau+t-s}, x^*_{\tau+t-s})\}$ that is independent of $\mathcal{F}_{T,u}$, equal to $\mathcal{A}$ in distribution, and satisfies $\mathbf{P}[\mathcal{A}^* \neq \mathcal{A}] = \beta_{s-u}$.

Now it is easy to construct $\{\tilde{L}_{jv}\}$:

$$\tilde{L}_{jv} \equiv L(y^*_v - x^*_{jv} \cdot \hat{\theta}_{jT,R_T}), \quad j = 1, 2, \quad v = s, \ldots, t. \tag{2.A.4}$$

Equations (2.A.1) and (2.A.2) are satisfied by construction, so it remains to prove (2.A.3). But (2.A.3) follows immediately from Merlevède and Peligrad's (2002) Proposition 2.3 — this proposition only uses (2.A.2) and moment restrictions, not the equality of distributions. As noted by Dedecker and Prieur (2005), Merlevède and Peligrad's constant, 2^{p+2} can be reduced when $p = 2$. ■

Lemma 2.A.2. *Suppose $\{b_T\}$ is a sequence of positive integers such that $b_T \to \infty$ and $b_T/P_T \to 0$, and define*

$$Z_{Ti} \equiv \sum_{s=R_T+(i-1)b_T+l_T+1}^{R_T+ib_T} [L_{T,s} - M_T(\hat{\theta}_{T,R_T})].$$

If Assumptions 4 to 6 hold then

$$\sum_i \left[Z^2_{Ti} - \mathrm{E}^*_{\mathrm{R}}(Z^2_{Ti} \mid \mathcal{F}_{T,R_T}) \right] \to 0 \quad \textit{in } L_1,$$

where $\mathrm{E}^*_{\mathrm{R}}$ is the integral with respect to a new probability measure $\mathrm{P}^*_{\mathrm{R}}$ that imposes independence between $\mathcal{F}_{T,R_T}$ and the sigma-field generated by the random variables

$$\{y_{T,R_T+j}, x_{1T,R_T+j}, x_{2T,R_T+j};\ j > 0\}$$

but otherwise preserves the original probability measure.

Proof. Much of the proof mimics that of de Jong's (1997) Lemma 5. For clarity, suppose that Z_{Ti} is a scalar. Define the function

$$h_c(x) = \begin{cases} \mathrm{sign}(x) c\sqrt{b_T/P_T} & \text{if } |x| > c\sqrt{b_T/P_T} \\ x & \text{otherwise} \end{cases}$$

for an arbitrary constant c. McLeish (1975a) shows that (in this paper's notation) $\{P_T Z^2_{Ti}/b_T\}$ is uniformly integrable,[8] so it is sufficient to prove that

$$\sum_i \left[h_c(Z_{Ti})^2 - \mathrm{E}^*_{\mathrm{R}}(h_c(Z_{Ti})^2 \mid \mathcal{F}_{T,R_T}) \right] \to 0$$

[8] Also see the remarks after Davidson's (1992) Lemma 3.2.

in L_1 for any choice of c.

We prove this by showing that the array

$$\{h_c(Z_{Ti})^2 - \mathrm{E}^*_{\mathrm{R}}(h_c(Z_{Ti})^2 \mid \mathcal{F}_{T,R_T}), \mathcal{F}_{T,R_T+i\,b_T}\} \tag{2.A.5}$$

is another L_2-mixingale of size $-1/2$ with constants d_{Ti} that satisfy $\sum_i d_{Ti}^2 \to 0$. Fix T, i, and κ, and define the array $\{\tilde{L}_{jv}; v = R_T+(i-1)b_T+1, \ldots, R_T+ib_T;\ j = 1,2\}$ independent of $\mathcal{F}_{T,R_T+(i-\kappa)b_T}$ using Lemma 2.A.1. Now let

$$W_{Ti} = \sum_{s=R_T+(i-1)b_T+l_T+1}^{R_T+ib_T} [\tilde{L} - M_T(\hat{\theta}_{T,R_T})],$$

so

$$\begin{aligned}
&\Big\| \mathrm{E}\left[h_c(Z_{Ti})^2 - \mathrm{E}^*_{\mathrm{R}}(h_c(Z_{Ti})^2 \mid \mathcal{F}_{T,R_T}) \mid \mathcal{F}_{T,R_T+(i-\kappa)b_T}\right]\Big\|_2 \\
&\qquad = \Big\| \mathrm{E}\left[h_c(Z_{Ti})^2 - h_c(W_{Ti})^2 \mid \mathcal{F}_{T,R_T+(i-\kappa)b_T}\right]\Big\|_2 \leq \Big\|h_c(Z_{Ti})^2 - h_c(W_{Ti})^2\Big\|_2
\end{aligned}$$

and it suffices to bound the last quantity.

As in de Jong, we have the inequalities:

$$\begin{aligned}
\|h_c(Z_{Ti})^2 - h_c(W_{Ti})^2\|_2 &\leq 2c\sqrt{b_T/P_T}\|h_c(Z_{Ti}) - h_c(W_{Ti})\|_2 \\
&\leq 2c\sqrt{b_T/P_T}\,\Big\| \sum_{s=R_T+(i-1)b_T+1}^{R_T+ib_T} (L_{T,s} - \tilde{L}_s)\Big\|_2 \\
&\leq \left[2cb_T^{3/2}P_T^{-1}\beta_{b_T}^{(\rho-2)/2\rho}\right]\beta_\kappa^{(\rho-2)/2\rho} \\
&\equiv d_{Ti}\ \beta_\kappa^{(\rho-2)/2\rho}.
\end{aligned}$$

Since $\beta_\kappa^{(\rho-2)/2\rho}$, the array (2.A.5) is an L_2-mixingale and has size $-1/2$, $\sum_i d_{Ti}^2 \to 0$. Then McLeish's (1975b) Theorem 1.6 gives

$$\Big\| \sum_i \left[h_c(Z_{Ti})^2 - \mathrm{E}^*_{\mathrm{R}}(h_c(Z_{Ti})^2 \mid \mathcal{F}_{T,R_T})\right]\Big\|_1 = O\Big(\sum_i d_{Ti}^2\Big),$$

to complete the proof. ■

Proof of Lemma 2.4.1. We'll prove that

$$\|\,\mathrm{E}(L_{iT,R_T+j} - M_{jT}(\hat{\theta}_{jT,R_T}) \mid \mathcal{F}_{T,R_T+j-\tau})\|_2 \leq 2^{1+1/\rho}\ B_L\ \beta_\tau^{(\rho-2)/2\rho}$$

Notice that $\beta_\tau^{(\rho-2)/2\rho} = O(\tau^{-1/2-\delta}$. Define $\tilde{L}_s$ as in Lemma 2.A.1 to be independent of $\mathcal{F}_{T,s-\tau}$. Then

$$\begin{aligned}\|\,\mathrm{E}(L_{iT,s} - M_{iT}(\hat{\theta}_{iT,R_T}) \mid \mathcal{F}_{T,s-\tau})\|_2 &= \|\,\mathrm{E}(L_{iT,s} - \tilde{L}_{is} \mid \mathcal{F}_{T,s-\tau})\|_2 \\ &\leq \|L_{T,s} - \tilde{L}_{is}\|_2 \\ &\leq 2^{1+1/\rho} B_L \beta_\tau^{(\rho-2)/2\rho}\end{aligned}$$

by Lemma 2.A.1. ■

Proof of Lemma 2.4.2. Without loss of generality, assume that $L_{T,s}$ is a scalar. We will modify de Jong's (1997) Theorem 1 to establish normality. The only part of de Jong's proof that needs to be changed is the handling of the covariance matrix; here it is a random element and in de Jong's theorem it is a constant.

Let $\{b_T\}$, $\{l_T\}$, and $\{m_T\}$ be sequences of positive integers that satisfy $P_T \geq b_T \geq l_T + 1$, $b_T \to \infty$, $l_T \to \infty$, $\lfloor P_T/b_T \rfloor \to \infty$, and $l_T/b_T \to 0$. Then, de Jong proves that (in our notation)

$$P_T^{-1/2}\Big[\bar{L}_T - M_T(\hat{\theta}_{T,R_T})\Big] = \sum_i Z_{Ti} + o_p(1)$$

with

$$Z_{Ti} \equiv \sum_{s=R_T+(i-1)b_T+l_T+1}^{R_T+ib_T} \Big[L_{T,s} - \mathrm{E}(L_{T,s} \mid \mathcal{F}_{T,R_T+(i-1)b_T})\Big].$$

The array $\{Z_{Ti}, \mathcal{F}_{T,R_T+ib_T}, i = 1, \ldots, m_T\}$ is a martingale difference array by construction and it suffices to apply a Central Limit Theorem to $\sum_i Z_{Ti}$.

We apply Hall and Heyde's (1980) Theorem 3.3 to complete the proof.[9] De Jong's condition (9) ensures that Hall and Heyde's (3.18) and (3.20) are satisfied, so it remains to prove that

$$\sum_i Z_{Ti}^2 = \Sigma_T + o_p(1).$$

This last step is an immediate consequence of Lemma 2.A.2 and De Jong's Lemmas 3 and 4. ■

[9]The covariance matrix, Σ_T, is measurable in all of the sub-sigma-fields $\mathcal{F}_{T,s}$, so Hall and Heyde's nesting condition is unnecessary. See the remarks surrounding their Theorem for more details. This measurability also allows us to use a sequence of covariance matrices that does not necessarily converge.

Proof of Theorem 2.5.1. Under the null hypothesis,

$$\begin{aligned}\bar{L}_{1T} - \bar{L}_{2T} &= [\bar{L}_{1T} - M_{1T}(\hat{\theta}_{1T,R_T})] - [\bar{L}_{2T} - M_{2T}(\hat{\theta}_{2T,R_T})] \\ &\quad + [M_{1T}(\hat{\theta}_{1T,R_T}) - M_{1T}(\theta^*_{1T})] + [M_{2T}(\theta^*_{2T}) - M_{2T}(\hat{\theta}_{2T,R_T})] \\ &\leq [\bar{L}_{1T} - M_{1T}(\hat{\theta}_{1T,R_T})] - [\bar{L}_{2T} - M_{2T}(\hat{\theta}_{2T,R_T})] + o_p(R_T^{-1/2})\end{aligned}$$

since the derivative of $M_{1T}(\cdot)$ at θ^*_{1T} is zero and $[M_{2T}(\theta^*_{2T}) - M_{2T}(\hat{\theta}_{2T,R_T})]$ is positive. Lemma 2.4.2 ensures that this last quantity is asymptotically normal with asymptotic variance $\hat{\sigma}_T$. ■

Proof of Lemma 2.5.2. Let $\{(v_T, z_{1T}, z_{2T})\}$ be a sequence of random vectors, independent of $\hat{\theta}_{T,R_T}$ and equal in distribution to $\{(\varepsilon_{1T,t}, x_{1T,t}, x_{2T,t})\}$. The prediction errors satisfy

$$\begin{aligned}e_{1T,t} &= \varepsilon_{1T,t} + x'_{1T,t}(\hat{\theta}_{1T,R_T} - \theta^*_{1T}) \\ &= \varepsilon_{1T,t} + O_p(R_T^{-1/2}) \\ e_{2T,t} &= \varepsilon_{1T,t} + x'_{2T,t}(\hat{\theta}_{2T,R_T} - \theta^*_{2T}).\end{aligned}$$

Since $z'_T(\hat{\theta}_{2T,R_T} - \theta^*_{2T})$ is a continuous random variable, the probability of it taking a value that guarantees constant loss is zero. To show that

$$[(1,-1)\Sigma_T(\hat{\theta}_{T,R_T})(1,-1)']^{-1} = O_p(1),$$

it suffices to prove that the conditional variance (given $\hat{\theta}_{T,R_T}$) of the vector $(v_T^2,\ [v_T + z'_T(\hat{\theta}_{2T,R_T} - \theta^*_{2T})]^2)'$ satisfies the same relationship. Since z_T has uniformly positive variance, we only need to prove that $|\hat{\theta}_{2T,R_T} - \theta^*_{2T}|_2$ is uniformly a.s. positive. This follows from the inequality

$$|\hat{\theta}_{2T,R_T} - \theta^*_{2T}|_2^2 \geq \lambda_{max}(\boldsymbol{X}'_{2T,R_T}\boldsymbol{X}_{2T,R_T})^{-1}\lambda_{min}(\boldsymbol{X}'_{2T,R_T}\boldsymbol{\varepsilon}_{2R_T}\boldsymbol{\varepsilon}'_{2R_T}\boldsymbol{X}_{2T,R_T}).$$

■

Proof of Lemma 2.6.1. Observe that

$$M_{jT}(\hat{\theta}_{jT,R_T}) - M_{jT}(\hat{\theta}_{jT,T}) = \\ \mathrm{E}\left[L(\psi_T - z'_T\hat{\theta}_{jT,R_T}) - L(\psi_T - z'_T\hat{\theta}_{jT,T}) \mid \hat{\theta}_{jT,R_T}, \hat{\theta}_{jT,T}\right]$$

almost surely, with $(\psi_T, z_T) \stackrel{d}{=} (y_{T,T+1}, x_{jT,T+1})$ and independent of $(\hat{\theta}_{jT,R_T}, \hat{\theta}_{jT,T})$. As a result, it suffices to show that

$$\|L(\psi_T - z_T'\hat{\theta}_{jT,R_T}) - L(\psi_T - z_T'\hat{\theta}_{jT,T})\|_1 = O(\sqrt{P/R}).$$

Since L has finite left- and right-derivatives and is convex,

$$L(\psi_T - z_T'\hat{\theta}_{jT,R_T}) - L(\psi_T - z_T'\hat{\theta}_{jT,T}) = O_p(1)z_T'(\hat{\theta}_{jT,R_T} - \hat{\theta}_{jT,T}),$$

and, because this difference is uniformly integrable, we only need to prove that

$$|\hat{\theta}_{jT,R_T} - \hat{\theta}_{jT,T}|_2 = O_p(\sqrt{P_T/R_T}).$$

Now, we can express this last difference as

$$\hat{\theta}_{jT,R_T} - \hat{\theta}_{jT,T} = \left[(\boldsymbol{X}'_{jT,T}\boldsymbol{X}_{jT,T})^{-1} - (\boldsymbol{X}'_{jT,R_T}\boldsymbol{X}_{jT,R_T})^{-1}\right]\boldsymbol{X}'_{jT,T}\boldsymbol{\varepsilon}_{jT} + (\boldsymbol{X}'_{jT,R_T}\boldsymbol{X}_{jT,R_T})^{-1}\sum_{s=R_T+1}^{T} x_{jT,s}\varepsilon_{jT,s}.$$

The square of each of these terms is $O_p(\sqrt{P_T/R_T})$. First, observe that

$$\left|\left[(\boldsymbol{X}'_{jT,T}\boldsymbol{X}_{jT,T})^{-1} - (\boldsymbol{X}'_{jT,R_T}\boldsymbol{X}_{jT,R_T})^{-1}\right]\boldsymbol{X}'_{jT,T}\boldsymbol{\varepsilon}_{bjT}\right|_2^2 = O_p(T)\sum_{i=1}^{K_{jT}}\lambda_i\left[(\boldsymbol{X}'_{jT,T}\boldsymbol{X}_{jT,T})^{-1} - (\boldsymbol{X}'_{jT,R_T}\boldsymbol{X}_{jT,R_T})^{-1}\right]^2,$$

which is $O_p(P_T/R_T)$ since $(\boldsymbol{X}'_{jT,T}\boldsymbol{X}_{jT,T})^{-1} - (\boldsymbol{X}'_{jT,R_T}\boldsymbol{X}_{jT,R_T})^{-1}$ has rank P_T and its largest eigenvalue is $O_p(1/T)$. A similar argument proves that the second term is $O_p(\sqrt{P_T/R_T})$ as well, completing the proof. ■

Chapter 3

The empirical behavior of out-of-sample forecast comparisons

3.1 Introduction

Empirical macroeconomics and finance have been heavily influenced by the conclusions drawn from pseudo out-of-sample forecast comparisons. These statistics are so influential that when an in-sample and out-of-sample comparison disagree, the results of the in-sample comparison are usually discarded in favor of those of the out-of-sample comparison; this is exemplified by Meese and Rogoff's (1983) comparison of exchange rate models. But, despite their influence, there have been no studies of the empirical properties of out-of-sample comparisons themselves and it is unclear whether the statistics perform well in practice. This concern is present to some degree with all statistical techniques; but it matters more in areas like macroeconomics and finance, where replication of empirical studies is difficult, if not impossible. In fields where it is possible to perform many similar studies independently, flaws in statistical methodology can often be detected. When such studies are not conducted, it can take a long time before methodological flaws are discovered.

In this paper, we examine whether the empirical behavior of these statistics matches their asymptotic properties. We use the asymptotic distributions derived by Clark and West (2006, 2007), Giacomini and White (2006), McCracken (2007), and Calhoun (2009) to construct confidence intervals for the recent performance of the Phillips curve relative to a random walk for nineteen different OECD countries, and we calculate the frequency with which these intervals contain the observed difference in MSE. If this observed frequency is much smaller than the intervals' nominal confidence level, the intervals are too small and these out-of-sample comparisons are unreliable; if instead the observed frequency is higher than the confidence level, then out-of-sample tests have lower power in practice than their asymptotic theory indicates.

This paper's analysis mimics the problem facing a forecaster who has to choose between two models to produce a sequence of forecasts for a known number of periods. The observations up to the end of the first out-of-sample period can be viewed as the data available to that forecaster when choosing between the models, and the second out-of-sample period can be viewed as the values of the series that will determine the real-time performance of the models. In general, the forecaster will use the available data to construct a lower bound for difference in the average loss of the two models over the second period. If this interval does not contain zero, the forecaster can be confident (at a predetermined confidence level) that forecasts produced by the alternative model will give a smaller average loss than forecasts produced by the benchmark model.

Although this type of practical forecasting application is a natural setting for a pseudo out-of-sample comparison, the theoretical research into the behavior of these comparisons has focused on testing whether the population versions of the models forecast equally well.[1] Asymptotic approximations for these test statistics were first derived by Diebold and Mariano (1995) and West (1996) — Diebold and Mariano (1995) derive the asymptotic distributions of several out-of-sample test statistics under the assumption that none of the models' coefficients are estimated, and West (1996) extended those results to allow for estimated coefficients.

[1]By "population version," we mean the infeasible models that use the pseudo-true values of the coefficients.

West's (1996) asymptotic results have the practical limitation that they require each model's estimated coefficients to converge to different limits; this condition is violated whenever the true DGP can be expressed as a particular parameterization of the models being compared. One example is when one of the models is a generalization of the other, and the null hypothesis is that the smaller model is more accurate. Subsequent research has focused on extending West's (1996) asymptotic theory to apply to those nested models. These extensions include Chao, Corradi, and Swanson (2001), Corradi and Swanson (2002, 2004) Clark and McCracken (2001) and McCracken (2007). These papers show that the asymptotic distribution of the out-of-sample statistics is nonstandard and derive critical values for the different test statistics.

A second approach for dealing with nested models has been to propose a different asymptotic approximation under which the coefficient estimates do not converge. Under those asymptotics, the limiting distribution is Gaussian because the models still produce different forecasts in the limit. Giacomini and White (2006) and Clark and West (2006, 2007) propose using a finite-length rolling window to achieve this effect, and Calhoun (2009) proposes using the limiting distribution where both the number of regressors and the number of observations increase at the same time.

In this paper, we study several of these asymptotic approximations: Giacomini and White's (2006), Clark and West's (2006, 2007), McCracken's (2007), and Calhoun's (2009). These approximations lead us to consider three different statistics for each of the three basic window schemes: the rolling, recursive, and fixed windows; we use only three different statistics because Giacomini and White (2006) and Calhoun (2009) both recommend using the same naïve Gaussian approximation for the out-of-sample average. The two models that we are comparing are nested, so statistics based on Diebold and Mariano's (1995) and West's (1996) original approximation are inappropriate.

Since Giacomini and White (2006), Clark and West (2006, 2007), and McCracken (2007) do not derive the joint distribution of a pair of adjacent out-of-sample averages, we extend their results to apply in this setting. These extensions

are mathematically simple, but are of independent interest beyond this paper. In academic research, the population quantities that the original papers study are of primary interest, but in applied forecasting the actual forecasting performance of the models is usually more important.

By doing this study, we hope to discover whether any of these approximations are systematically more or less reliable than the others. This question is obviously important to applied forecasters and also sheds light on whether the original out-of-sample approximations are useful for academic research focusing on the population models. Although we use slightly different statistics than are proposed in the original papers by Giacomini and White (2006), Clark and West (2006, 2007), and McCracken (2007), the asymptotic theory that motivates our statistics is identical to the asymptotic theory used in the original papers; if the intervals do poorly in our empirical exercise, it is likely that they do poorly when applied as originally designed, and vice versa

In addition, there are several factors that can potentially have a large effect on the reliability of an out-of-sample analysis but are incompletely understood. Some of these can be chosen by the forecaster, such as the division of the available data into estimation and test windows, and others are out of the forecaster's control to some degree. This second group includes the complexity of the underlying models, the time-periods available to estimate the models and to forecast chosen for the analysis, and the particular window scheme to be used.[2] We also hope to understand the impact of these variables on the quality of the out-of-sample comparison through this paper's analysis as well.

In this paper, we study the accuracy of the theoretical approximations for a particular pseudo out-of-sample comparison. However, implications of this study are more broadly applicable to other such comparisons because our analysis uses the same maintained assumptions as these other statistics and our confidence intervals are both conceptually and numerically similar to the statistics proposed

[2]Obviously, the forecaster controls all of these different factors at some stage of the analysis. However, a forecaster will choose them to try to produce the best forecasts possible and to improve the quality of the pseudo out-of-sample comparison. So it is better to think of those variables as outside the forecasters control when analyzing procedures to choose between a pair of forecasts.

for other applications. Obviously, there are some differences between the statistics and these differences may turn out to be significant in unpredictable ways, so one should not take evidence that the statistics perform well uncritically.

A limitation of our analysis is that McCracken's (2007) and Clark and West's (2006, 2007) approximations are only valid under the null hypothesis that the benchmark model is correctly specified – that the errors from the benchmark model form a martingale difference sequence. This is a more restrictive than imposing that the smaller model be more accurate, so intervals based on their approximations could break down because of the failure of that maintained assumption. We include these intervals despite this possibility for two reasons. Although McCracken's (2007) and Clark and West's (2006, 2007) theoretical results require the additional martingale difference sequence assumption, there is no evidence (empirical or Monte Carlo) that indicates whether the results would hold under weaker assumptions. The statistics may still perform well. And, even though the theory does not support using these statistics to choose a model for forecasting, they are often used this way in practice, so it is important to document how well they perform.

The application we choose, the forecasting performance of a Phillips curve relative to a random walk, of widespread interest. Inflation is the primary series targeted by the central banks and government policy-makers are interested in predicting the effect of different policy choices on inflation. This goal necessitates a model that is both accurate and theoretically grounded, so the accuracy of the Phillips curve is significant. Moreover, since inflation and expected inflation both have strong effects on the real economy and on the financial markets, businesses in the private sector have strong financial incentives to forecast inflation accurately and academic economists have interest in the accuracy of models that relate inflation to potential output, unemployment, and other series.

The question of whether out-of-sample forecast comparisons are reliable in this setting is especially pertinent because of Atkeson and Ohanian's (2001) and Stock and Watson's (2007, 2008) demonstrations that the Phillips curve has forecast worse than a random walk since the early 1980s. Atkeson and Ohanian

show that a random walk outperformed several Phillips curve models from 1985 through 1999, and Stock and Watson verify and refine that finding with a detailed comparison of the Phillips curve to several univariate models over several different time periods. Of course, these findings rest on the reliability of out-of-sample comparisons themselves, which hasn't yet been demonstrated.

Our choice of inflation forecasting brings up a further goal of our study. The out-of-sample statistics that we use assume that the underlying series are stationary to some degree, that the difference in the models' performance is constant across time periods. One of the proposed explanations for the deterioration of the Phillips curve is that inflation exhibits some form of instability, which could violate such an assumption. It is often claimed, however, that out-of-sample statistics provide a guard against this sort of instability, even though the theory behind these approximations does not yet incorporate that generality. In that vein, by studying the behavior of the statistics in a potentially unstable environment, we can test this claim to some degree. Moreover, because these statistics are believed to be reliable when the underlying series are unstable, they are often used to study models of inflation (as in the papers cited above, and in many more: see Stock and Watson 2003 and 2008 for recent surveys). Therefore, this analysis also indicates whether that literature is based on solid statistical foundations.

The rest of the paper proceeds as follows. The next section describes our empirical exercise and setup in more detail. The third section presents our results, and the fourth section concludes. Our theoretical results and their proofs are presented in the appendix.

3.2 Setup

3.2.1 Introduction and Some Notation

We will denote period-t inflation as π_t and a forecast for that period's inflation as $\hat{\pi}_t$. We are interested in the difference in squared error between the

random walk and Phillips curve forecasts, which we'll call D_t:

$$D_t = (\pi_t - \hat{\pi}_t^{RW})^2 - (\pi_t - \hat{\pi}_t^{PC})^2. \tag{3.2.1}$$

We will discuss the details behind these forecasts later in this section. For now, assume that there are T total observations in the dataset and that those observations are split into an *estimation window* (the first R observations), a *test sample* (the next P_1 observations), and a *forecast sample* (the remaining P_2 observations). We will refer to the first $R + P_1$ observations as the *interval sample* for reasons that will become clear and define $T_1 = R + P_1$.

Each of the approximations we study gives a method for constructing intervals of the form

$$I = [\bar{D}_1 - c_\alpha \times \hat{\sigma}, \infty) \tag{3.2.2}$$

such that, in the limit, $Pr[\bar{D}_2 \in I] = \alpha$ with α the predetermined coverage probability of the interval, where $\bar{D}_1$ is the average of D_t over the test sample and $\bar{D}_2$ is the average over the forecast sample:

$$\bar{D}_1 \equiv {P_1}^{-1} \sum_{t=R+1}^{R+P_1} D_t, \qquad \text{and} \qquad \bar{D}_2 \equiv {P_2}^{-1} \sum_{t=R+P_1+1}^{T} D_t. \tag{3.2.3}$$

The random variable $\hat{\sigma}$ is an estimator of the standard deviation of $\bar{D}_1$.

To determine whether these approximations are accurate, we estimate the interval I and the average $\bar{D}_2$ for each country j, each window w, and every reasonable division of the interval sample into an estimation window and test sample (which is determined by R/T_1).[3] Our estimate of $Pr[\bar{D}_2 \in I; R, w]$ is the frequency with which the intervals contain $\bar{D}_2$,

$$\widehat{Pr}[\bar{D}_2 \in I; R, w] = J^{-1} \sum_{j=1}^{J} 1\{\bar{D}_2^{j,w,R} \in I^{j,w,R}\}, \tag{3.2.4}$$

with J the number of countries. The rest of this section fills in the necessary details.

[3]The size of the forecast sample, P_2, will be determined in practice by the particular application, so we consider two different choices of P_2 in our study, but do not conduct a rigorous analysis of the effect of P_2 on the quality of the approximation.

3.2.2 Description of the Dataset

We use the OECD's data on inflation and unemployment to estimate these forecasts. In particular, we use the first difference of the natural log of the Consumer Price Index (CPI) as our measure of inflation,

$$\pi_t = \ln \mathrm{CPI}_t - \ln \mathrm{CPI}_{t-1} \tag{3.2.5}$$

and use the seasonally adjusted survey-based unemployment rate. We use the quarterly series of both to be consistent with empirical practice. To ensure that we have data on enough countries that our analysis is reasonably accurate, we use data from the first quarter of 1975 through the fourth quarter of 2008. This starting value allows us to include fourteen countries in our analysis. We also consider a shorter sample, starting in the first quarter of 1992, that allows us to use nineteen countries. Table 3.3 on page 98 lists the countries that comprise each sample.

3.2.3 Construction of the forecasts

In this paper, we focus on forecasting at the one-quarter horizon and assume that inflation has a unit root. We impose the unit root to be consistent with empirical practice and so that our benchmark random walk and alternative Phillips curve models agree on the order of integration. Restricting our forecast horizon to the one quarter horizon simplifies the test statistics and some of the forecasting decisions, so we focus on that horizon even though longer horizon forecasts of inflation have been studied more extensively in the literature.[4] The two alternative models are Autoregressive Distributed Lag (ADL) models that use the log of the unemployment rate as an additional predictor.

The random walk forecasts are given by the equation

$$\hat{\pi}_{t+1} = \pi_t. \tag{3.2.6}$$

[4]If the forecast horizon were more than a single period ahead, the forecast errors would have a moving average dependency structure even if the smaller model were the true DGP, so our pseudo out-of-sample intervals would have to account for that in estimating the variance of the average loss.

Atkeson and Ohanian (2001) and Stock and Watson (2007) demonstrate the strong performance of the random walk for forecasting annual inflation, making it the natural benchmark for this study. Our random walk model is slightly different than the model studied by those papers because of the forecast horizon. Since this model has no parameters to be estimated, the forecasts are the same for each choice of window scheme and estimation window length.

The Phillips curve models use lags of the change in inflation and the log of unemployment as predictors. We impose that the same number of lags are used for each series, giving a forecasting relationship of the form

$$\widehat{\Delta\pi}_{t+1} = \beta_0 + \sum_{j=1}^{p} \beta_j \Delta\pi_{t+1-j} + \sum_{j=1}^{p} \beta_{p+j} u_{t+1-j} + \varepsilon_{t+1}. \tag{3.2.7}$$

The variable u_t denotes the natural log of the unemployment rate in period t and ε_{t+1} is the population forecast error; the unknown coefficients are estimated by OLS. For simplicity, and to ensure that the models we are studying agree as closely as possible with those described by the asymptotic theory in the original papers, we do not try to choose the number of lags optimally, but instead conduct a separate analysis for two different lag choices: one lag and six lags. These choices let us study the behavior of these intervals for a tightly parameterized model and a potentially overfit model.

3.2.4 Construction of the Intervals

The difference in the squared error of the random walk and Phillips curve forecasts is determined by the estimates of the unknown coefficients, β. Out-of-sample statistics mimic the actual forecasting process by estimating β each period using only the information available in that period, so we can refine (3.2.1) by substituting the forecasting models into the right hand side of the equation, giving

$$D_t = (\pi_t - \pi_{t-1})^2 - \left(\pi_t - \pi_{t-1} - \hat{\beta}_{t-1,0} - \sum_{j=1}^{p} (\hat{\beta}_{t-1,j} \Delta\pi_{t-j} + \hat{\beta}_{t-1,p+j} u_{t-j})\right)^2. \tag{3.2.8}$$

We consider three different window schemes for estimating β: the rolling window, recursive window, and fixed window. For the fixed window, β is estimated

$t = R+1, \ldots, R+P_1$		
recursive window	$\hat{\beta}_t =$	$\left(\sum_{s=1}^{t} x_s x_s'\right)^{-1} \sum_{s=1}^{t} x_s y_s$
fixed window		$\left(\sum_{s=1}^{R} x_s x_s'\right)^{-1} \sum_{s=1}^{R} x_s y_s$
rolling window		$\left(\sum_{s=t-R}^{t} x_s x_s'\right)^{-1} \sum_{s=t-R}^{t} x_s y_s$
$t = R+P_1+1, \ldots, T$		
recursive window	$\hat{\beta}_t =$	$\left(\sum_{s=1}^{t} x_s x_s'\right)^{-1} \sum_{s=1}^{t} x_s y_s$
fixed window		$\left(\sum_{s=1}^{T} x_s x_s'\right)^{-1} \sum_{s=1}^{T} x_s y_s$
rolling window		$\left(\sum_{s=t-R}^{t} x_s x_s'\right)^{-1} \sum_{s=t-R}^{t} x_s y_s$

Table 3.1: This table displays the construction of $\hat{\beta}_t$ for each window scheme and period. The variables x_t denote the vector of all of the regressors (in this application, lags of $\Delta\pi_t$ and of u_t) and y_t denotes the values of the target, $\Delta\pi_t$.

once using the data from periods one through R and all of the forecasts are constructed from those estimates. For the recursive scheme, β is estimated repeatedly for each period t, using the information from periods one through $t-1$. For the rolling scheme, β is also estimated repeatedly for each period, but is estimated using the most recent R observations only.

Table 3.1 displays how $\hat{\beta}_t$ is determined by the window scheme and by the particular value of t. For the recursive and fixed window schemes, the size of the estimation window, R, only affects the coefficient estimates in the test sample and not in the second out-of-sample period. That second period is meant to reflect observations that the forecaster can not observe when choosing a model, and in practice the forecaster would want to reestimate the model with all of the available data before making those truly out-of-sample forecasts. For the rolling window, the choice of R affects the coefficient estimates in both samples, because the window is an intrinsic part of the forecasting method, and the forecaster would choose to use a rolling window for the truly out-of-sample forecasts if he or she were concerned about instability.

In practice, the fixed window is used infrequently – since the coefficients are only calculated once, it can be useful if the model is computationally difficult to evaluate, but forecasters usually want to use the most recent data to estimate

their models. The rolling window is used when forecasters are concerned about unmodeled heterogeneity in the underlying series; it is believed that using only the most recent data can improve the forecasts in those settings. When forecasters are less concerned about instability than about small sample sizes, and when computing the model is not unduly difficult, the recursive window is preferred because it uses as many observations as possible to construct each forecast, so the forecasts should be more accurate.

In analyzing these empirical results, we'll take the particular window scheme as given. Although much of the asymptotic theory for out-of-sample inference assumes that the windows are interchangeable – West's (1996), Clark and West's (2006, 2007), and McCracken's (2007) approximations imply that all three windows test the same hypothesis – it's clear that each window scheme has a different practical application in this study. The fixed-window interval is appropriate when the model will be estimated once and put in place for several periods – as can happen when the model is developed by an outside party such as a consulting firm. The recursive window is appropriate when the forecasting model is repeatedly re-estimated over the entire dataset, and the rolling scheme is appropriate when a rolling-window model will be used to construct the actual forecasts.

We consider a few different options for the division of the data into interval and forecast samples. These options are listed in 3.4 on page 98. For one analysis, we use the OECD data starting in the first quarter of 1975, and for a second analysis, we use the data starting in 1992. For each of these start periods, we construct confidence intervals for two different forecast samples: from the first quarter of 2000 through the fourth quarter of 2008, and from the first quarter of 2006 through 2008. The intervals are constructed using the difference in the two models' squared error, D_t, over the preceding dataset. Varying the dates of these samples allows us to informally determine whether our results are affected by unmodeled heterogeneity – if the quality of the underlying approximation depends on the sample period, these out-of-sample statistics probably do not adequately control for instability in the underlying series.

After deciding on the interval and forecast samples, we need to split the T_1

observations in the interval sample into an estimation and test sample of sizes R and P_1. We look at every division such that R is at least twenty four and P_1 is at least six. The restriction on R rules out extremely inaccurate forecasts, and the restriction on P_1 is necessary to calculate HAC variance estimators of the variance of $\bar{D}_1$.

3.2.5 Details of the interval construction

Finally, we present the formulae for the pseudo out-of-sample intervals that we study. All of the intervals in this paper are constructed using an 80% confidence level. The intervals are of the form

$$I \equiv [\bar{D}_1 - c_\alpha \hat{\sigma}, \infty), \tag{3.2.9}$$

and both c_α and $\hat{\sigma}$ depend on the window scheme, the particular asymptotic theory behind the approximation, and the sample sizes P_1 and P_2. The critical value, c_α of these intervals follows directly from the asymptotic distribution of $\bar{D}_1 - \bar{D}_2$. We will first discuss a naïve Gaussian interval that can be motivated by Giacomini and White's (2006) and Calhoun's asymptotic theories, as well as (informally) Diebold and Mariano's (1995) and West's (1996). We then look at a similar Gaussian interval motivated by Clark and West (2006, 2007), and finally an interval based on McCracken's (2007) nonstandard limiting distribution.

Naïve Gaussian Intervals

Under Giacomini and White's (2006) and Calhoun's (2009) asymptotic approximations, and under Diebold and Mariano's (1995) and West's (1996) if the models are not nested, both $\bar{D}_1$ and $\bar{D}_2$ are asymptotically normal. Moreover, as we show in the appendix, the two averages are independent in the limit.

The motivation behind this independence is different for each of the four approximations. Diebold and Mariano's (1995) approximation assumes that the coefficients are known and do not need to be estimated, so independence follows from the weak dependence of the underlying series. In West's (1996) approximation the coefficients are estimated consistently and the interval behaves in the limit like

Diebold and Mariano's (1995) statistic.[5] Under Giacomini and White's (2006) approximation, the forecasts are constructed using a fixed-length rolling window; in the limit, completely different samples are used to construct each of the two out-of-sample averages, so they are clearly independent.[6] Finally, under Calhoun (2009) the coefficient estimators are assumed to fail to converge to a non-stochastic limit and the two out-of-sample averages are highly interdependent; however the dependence is removed after conditioning on the coefficient estimates and so it does not affect the validity of the intervals.

This independence is enough to determine the asymptotic variance $\hat{\sigma}$ and the value c_α. The asymptotic variance of $\bar{D}_1 - \bar{D}_2$ is straightforward to calculate:

$$\operatorname{avar}(\bar{D}_1 - \bar{D}_2) = \operatorname{avar}(\bar{D}_1) + \operatorname{avar}(\bar{D}_2) = (1 + P_1/P_2)\operatorname{avar}(\bar{D}_1), \tag{3.2.10}$$

making $c_\alpha = z_\alpha \cdot \sqrt{1 + P_1/P_2}$ with z_α the α-quantile from the standard normal distribution. We estimate the asymptotic variance using the Newey-West HAC variance estimator (Newey and West 1987) of $\bar{D}_1$, setting the number of lags used by the kernel to be the smallest integer greater than or equal to ${P_1}^{1/4}$.

Although Giacomini and White's (2006) and Calhoun's (2009) approximations are only derived for a restricted class of window schemes – Giacomini and White's for a finite-length rolling window and Calhoun's for a fixed window with a small test sample – we will present results for the naïve Gaussian approximation for all of the window schemes. There are two reasons. Although these approximations have not been formally extended to other window schemes, doing so may be possible and so their accuracy is worth examining. The second reason is that Giacomini and White's (2006) approximation is sometimes cited informally for recursive window comparisons, and whether or not this use is appropriate is a valid empirical question.

[5] This argument only holds when the same loss function is used to estimate and evaluate the forecasting models. When a different loss function is used for the two purposes, the variance of $\hat{\beta}$ needs to be accounted for explicitly and would introduce another source of dependence between $\bar{D}_1$ and $\bar{D}_2$. In this paper, squared error is used for both estimation and evaluation, so the asymptotic independence holds.

[6] A formal proof of this argument is presented as Lemma 3.A.2 in the appendix.

Clark and West

Clark and West (2006, 2007) use a slightly different approximation that's based on Giacomini and White's (2006) fixed-R approach. While Giacomini and White drop the assumption that the benchmark model is the true DGP and impose a secondary assumption that the difference between the models' loss is unpredictable (that it is a martingale difference sequence), Clark and West (2006, 2007) maintain the assumption that the benchmark model is true. Since R is fixed, they show that the larger model can be expected to perform worse than the smaller model because of the noise introduced by estimating its coefficients. This leads Clark and West to introduce a correction term that converges in probability to the expected performance difference due to that estimation error.

In particular, suggest that one introduce the correction term

$$\bar{f}_1 = {P_1}^{-1} \sum_{t=R+1}^{T} (\hat{\pi}_t^{RW} - \hat{\pi}_t^{PC})^2 \tag{3.2.11}$$

and show that the corrected out-of-sample average,

$$\frac{\sqrt{P_1}(\bar{D}_1 - E\bar{f}_1)}{\sqrt{{P_1}^{-1} \sum_{t=R+1}^{R+P_1} (D_t - E\bar{f}_1)^2}} \tag{3.2.12}$$

is asymptotically standard normal. Note that under Clark and West's (2006, 2007) asymptotic theory, a HAC estimate of the standard deviation is inappropriate.

Clark and West's (2006, 2007) results are derived for a single out-of-sample period. When we extend their results to two samples, we observe that the only difference between these intervals and the naïve Gaussian intervals is the estimator of the variance. We define $\bar{f}_2$ to be the equivalent correction term from Equation (3.2.11) calculated over the second test sample. Under Clark and West's asymptotic theory, both $\bar{D}_1 - E\bar{f}_1$ and $\bar{D}_2 - E\bar{f}_2$ are asymptotically normal and, under stationarity, $E\bar{f}_1 = E\bar{f}_2$. As a result, we can add and subtract the correction terms before applying Clark and West's limit theory:

$$\bar{D}_2 - \bar{D}_1 = (\bar{D}_2 - E\bar{f}_2) - (\bar{D}_1 - E\bar{f}_1). \tag{3.2.13}$$

which is asymptotically normal.

This asymptotic normality implies that c_α is also calculated using the α-quantile of the standard normal distribution. Clark and West suggest estimating the asymptotic variance of $\bar{D}_1$ with

$$\hat{\sigma}^2 = {P_1}^{-1} \sum_{t=R+1}^{R+P_1} (D_t - \bar{D}_1 - \bar{f}_1)^2. \tag{3.2.14}$$

This estimator corrects the difference in the MSE to reflect estimation error in the alternative model, and maintains the assumption that the smaller model is correctly specified so there is no serial correlation in the sequence of differences in squared error.

McCracken

McCracken's (2007) is the only approximation we consider that leads to a non-Gaussian interval. McCracken uses a standard asymptotic setup like West's (1996) but assumes the coefficient estimates of the larger model converge to those of the smaller model because the smaller model is the true DGP. As a result, West's (1996) Gaussian approximation is inaccurate because variance of difference between the two models vanishes.

McCracken (2007) derives the distribution a custom distribution for the behavior of the out-of-sample t-statistic we consider in this paper under this different assumption. An important feature of this distribution is that it is not centered at zero. When both models are equally accurate in population, McCracken's results imply that the larger model will perform substantially worse on average in finite samples. This is the same behavior described by Clark and West (2006, 2007) and by Calhoun (2009).

Under McCracken's (2007) asymptotic theory, the numerator and the denominator of out-of-sample t-statistics converge to zero at the same rate, so the asymptotic distribution is nonstandard. In particular, McCracken shows that

$$\sum_{t=R+1}^{T} D_t \xrightarrow{d} (\Gamma_1 - 0.5\Gamma_2) \quad \text{and} \quad \sum_{t=R+1}^{T} (D_t - \bar{D}_1)^2 \xrightarrow{d} \Gamma_2 \tag{3.2.15}$$

where Γ_1 and Γ_2 are random variables that depend on the window scheme and the window lengths. Table 3.2 on page 90 gives the definitions of Γ_1 and Γ_2.

Window	R.V.	Formula
Recursive	Γ_1	$\int_{\lambda_1}^{1} s^{-1}W(s)'dW(s)$
	Γ_2	$\int_{\lambda_1}^{1} s^{-2}W(s)'W(s)ds$
	Λ_1	$\int_{1}^{1/\lambda_2} s^{-1}W(s)'dW(s)$
	Λ_2	$\int_{1}^{1/\lambda_2} s^{-2}W(s)'W(s)ds$
Fixed	Γ_1	$\lambda_1^{-1}[W(1) - W(\lambda_1)]'W(\lambda_1)$
	Γ_2	$\pi_1\lambda_1^{-1}W(\lambda_1)'W(\lambda_1)$
	Λ_1	$[W(\lambda_2^{-1}) - W(1)]'W(1)$
	Λ_2	$\pi_2 W(1)'W(1)$
Rolling	Γ_1	$\lambda_1^{-1}\int_{\lambda_1}^{1}[W(s) - W(s-\lambda_1)]'dW(s)$
	Γ_2	$\int_{\lambda_1}^{1} s^{-2}[W(s) - W(s-\lambda_1)]'[W(s) - W(s-\lambda_1)]ds$
	Λ_1	$\lambda_1^{-1}\int_{1}^{1/\lambda_2}[W(s) - W(s-\lambda_1)]'dW(s)$
	Λ_2	$\int_{1}^{1/\lambda_2} s^{-2}[W(s) - W(s-\lambda_1)]'[W(s) - W(s-\lambda_1)]ds$

Table 3.2: This table displays the components of the limiting distribution of the out-of-sample t-test McCracken's (2007) limit theory. $W(\cdot)$ is a $K_2 - K_1$ dimensional Brownian Motion, where K_1 is the number of regressors used by the benchmark model and K_2 is the of regressors used by the larger model. The two variables λ_1 and λ_2 denote R/T_1 and T_1/T respectively.

As with the other approximations, McCracken's (2007) asymptotic theory immediately gives a limiting distribution for $\bar{D}_2$. Moreover, since the coefficient estimates are assumed to converge to their pseudo-true values, the averages $\bar{D}_1$ and $\bar{D}_2$ are asymptotically independent.[7] Since McCracken recommends using the sample variance of D_t over the test sample to estimate σ^2, we can find c_α directly from the asymptotic distribution of the random variable M, where

$$M \equiv \frac{P_1\left(\bar{D}_2 - \bar{D}_1\right)}{\sqrt{\sum_{t=R+1}^{R+P_1}(D_t - \bar{D}_1)^2}} \tag{3.2.16}$$

$$\xrightarrow{d} \frac{\phi(\Lambda_1 - 0.5\Lambda_2) - (\Gamma_1 - 0.5\Gamma_2)}{\sqrt{\Gamma_2}} \tag{3.2.17}$$

and $\phi = \lim\sqrt{P_1/P_2}$. The limiting distributions of each Γ_i and Λ_i are listed in Table 3.2.

[7] A derivation is presented in the appendix as Lemma 3.A.1.

3.3 Results

The empirical coverage for the recursive, fixed, and rolling windows are depicted in Figures 3.1 to 3.3. Each figure depicts twelve different panels: each panel graphs the empirical coverage for intervals constructed for the difference between the MSE of the random walk benchmark model and the single lag and the six-lag Phillips curve models as a function of R/T_1. Each interval has nominal coverage of 0.8. The twelve panels present the results for all of the different combinations of asymptotic approximation and choice of interval and forecast samples. Each row contains the results for a different interval sample and each column the results for a different statistic.

We start by looking at Figure 3.1, the coverage for the recursive window. Overall the empirical coverage varies strongly with the choice of sample period and the lag structure. On average, the coverage for the single lag Phillips curve is 22 percentage points smaller than the coverage for the six lag Phillips curve, but the difference is much smaller for the 1975 to 2000 interval sample than for the others. In general, the observed coverage is close to the nominal coverage of 0.8 for each approximation for the six-lag alternative. For the single-lag alternative, the intervals are generally too small, leading to empirical coverage lower than the nominal coverage.

The coverage does not depend on R/T_1 very much, and the effect of R/T_1 is much smaller than the effect of the lag structure or the sample period. We can see a slight tendency for the coverage to decrease as R/T_1 increases. This tendency is most visible in the 1992–2005 interval sample, but is not uniform and is extremely mild in the other samples. The coverage plots for the 1992–1999 interval sample are too short to make meaningful statements about the relationship between the coverage and R/T_1 for that sample.

The most striking pattern is that the choice of approximation does not seem to matter much. The difference between the graphs for different approximations is hard to distinguish visually—not only is the coverage similar, but the dependence on the ratio R/T_1, on the lag structure, and the choice of the sample periods are all virtually identical. The single exception is that the empirical coverage for

Clark and West's (2006, 2007) approximation is slightly lower than the coverage of the naïve Gaussian approximation, especially for the single lag alternative. This difference occurs by construction, since the statistics are the same except that Clark and West's approximation uses a variance estimator that gives smaller estimates. The result is that these intervals give a less accurate approximation than the naïve Gaussian intervals.

The coverage plots for the fixed window, Figure 3.2, behave similarly to those for the recursive window. The coverage varies with the lag structure and the choice of sample periods; as with the recursive window, the coverage for the six-lag comparison is roughly twenty two percentage points higher than the coverage for the single lag comparison. Moreover, the overall patterns are similar for both window schemes across the different interval samples.

The coverage of the fixed window intervals decreases slightly as R/T_1 increases. The pattern is more uniform for the fixed window and is roughly the same magnitude as for the recursive window plots. The coverage of the intervals based on McCracken's (2007) approximation decreases for the 1975–1999 interval sample and did not for the recursive window.

As with the recursive window, the choice of approximation seems to have almost no impact on the coverage. By construction intervals constructed using Clark and West's (2006, 2007) limit theory are again smaller than the naïve Gaussian intervals, so the coverage of their statistic is smaller. The practical impact of this tendency is negligible for the six-lag alternative, but makes the coverage much too small for the single lag alternative for the 1975–2005 and 1992–2005 interval samples in particular.

Figure 3.3 depicts the coverage for the rolling window. Again, the empirical coverage varies with the choice of sample period and the lag structure. The variation is comparable to that of the recursive window and the fixed window. The behavior of the coverage plot is very similar to the behavior in the other graphs for intervals based on Clark and West's (2006, 2007) approximation and the naïve Gaussian approximation, but is different for those based on McCracken's (2007) approximation; for these intervals, the coverage for the single lag and six lag alter-

native models are very close and the coverage is above the nominal coverage for small and moderate values of R/T_1.

The behavior of the coverage as R/T_1 changes is similar to the behavior with other window schemes. The decrease in coverage as R/T_1 increases is more pronounced for McCracken's approximation using the rolling window than other window schemes. With Clark and West's and the naïve Gaussian approximation, the coverage is flat as R/T_1 varies, but there is a slight tendency for the coverage to decrease with R/T_1 for the 1992 to 2006 interval sample.

The choice of the approximation matters for the rolling window, unlike the fixed and recursive windows. As we would expect, intervals derived from Clark and West's (2006, 2007) approximation behave similarly to those using the naïve Gaussian approximation and have slightly lower coverage. But we see a dramatic difference between those intervals and the intervals based on McCracken's approximation. The coverage for McCracken's intervals is much higher, and exceeds the nominal coverage for most of the range of R/T_1, while the naïve Gaussian approximation has empirical coverage very close to the nominal coverage for the six lag comparison over the 2006 to 2008 forecast samples and for both alternative models for the 1975 to 1999 interval sample. For the single lag model, the Gaussian intervals have very small coverage in the other samples.

As we discussed earlier, we consider the choice of the window to be dictated by the particular application and forecasting models, so it is not something that a forecaster could choose in practice. In that light, we will summarize how these empirical results can inform the decision on the approximation to use – Clark and West's (2006, 2007), the naïve Gaussian, or McCracken's (2007) — and, for the recursive and fixed schemes, the split between estimation and test samples. For the rolling window scheme, the choice of window length is also dictated by the application since it also determines the forecasting model that will be used in practice.

As we can see in Figures 3.1, 3.2, and 3.3, the choice of approximation matters very little. The naïve Gaussian approximation advocated by Giacomini and White (2006) and Calhoun (2009) achieves empirical coverage slightly closer to

the intervals' nominal coverage than the other approximations, but McCracken's (2007) performs very similarly and is better for some of the samples and comparison models. In particular, the naïve Gaussian approximation gives intervals that are too small for the rolling window using a forecasting sample of 2006 to 2008, and McCracken's (2007) approximation does not. Clark and West's (2005, 2006) approximation has systematically lower coverage than the naïve Gaussian intervals by construction, so Clark and West's approximation should be avoided when choosing between models for applied forecasting. It is important to remember that Clark and West's approximation is derived under the hypothesis that the random walk model is the true DGP, so the approximation's poor performance in our setting is neither surprising nor indicative of the quality of that approximation for testing their null hypothesis. It does, however, indicate that forecasters should not use their statistic as a general measure of the models' forecasting performance.

This analysis does not give clear recommendations for the choice of R/T_1; that ratio does not matter for most of the periods, models, and windows. When it does matter, there are no values that are clearly superior across the different sample periods and comparison models. In a later section, we will look at the length of the individual confidence intervals for each country as another method of comparing different choices of R/T_1 recommending one choice of window selection.

The two factors that most affect the quality of the intervals' coverage have not been the focus of much theoretical research. Those factors are the exact model used as the alternative model (in this case, either a single lag or a six lag ADL), and the particular sample chosen for the analysis. The impact of these factors may be indicative of a single underlying cause. Instability in the underlying series could cause both of these factors to influence the quality of the intervals. The effect of the choice of the interval and forecast sample, at least, is an indication that these statistics do not generally account for instability in the underlying series and they should not be applied uncritically as if they do.

3.4 Conclusion

In this paper, we constructed confidence intervals for the difference in Predictive Mean Squared Error for two forecasting models of quarterly inflation over two recent periods, 2000 through 2008 and 2006 through 2008, for nineteen OECD countries. We estimated those intervals by conducting a pseudo out-of-sample comparison using the data available just before those periods. We constructed different intervals that are asymptotically valid under the asymptotic approximations proposed by several recent papers: Giacomini and White (2006), Clark and West (2006, 2007), McCracken (2007), and Calhoun (2009) and extended the results of those papers to apply to a pair of out-of-sample averages. We then calculated the average frequency with which each interval contains the actual difference in MSE. Since this difference is observed, that frequency estimates the actual coverage probability of those intervals.

The intervals' actual coverage was heavily influenced by factors that have been largely ignored in the theoretical literature. Two factors that have been considered important, the choice of asymptotic approximation to use for the limiting distribution of the statistics and the division of the available data into an estimation and a test window, had little effect on the coverage. The factors that did affect the coverage were the particular choice of alternative model and the choice of the particular sample periods.

These factors are likely to influence any pseudo out-of-sample comparison. Although the particular statistics we considered in this paper are new, they are numerically similar to statistics that are in current use and are based on the same limit theories. Despite what is often claimed, these out-of-sample comparisons do not seem to automatically control for unmodeled instability. Future theoretical research should explicitly include such instability to better understand its impact on out-of-sample forecast comparisons.

3.5 Mathematical Appendix

Lemma 3.5.1. *Suppose that McCracken's (2007) Assumptions 1–3 hold and that* $P_1/R \to \pi_1$ *and* $P_2/(R+P_1) \to \pi_2$ *with* $\pi_1, \pi_2 \in (0, \infty)$. *Then*

$$\frac{1}{\sqrt{P_1^{-1}\sum_{t=R+1}^{R+P_1}(D_t - \bar{D}_1)^2}}\left(P_1^{-1/2}\sum_{t=R+1}^{T} D_t, P_2^{-1/2}\sum_{t=R+P_1+1}^{R+P_1+P_2} D_t\right) \xrightarrow{d} \left(\frac{\Gamma_1 - \frac{1}{2}\Gamma_2}{\sqrt{\Gamma_2}}, \phi\frac{\Lambda_1 - \frac{1}{2}\Lambda_2}{\sqrt{\Lambda_2}}\right) \quad (3.5.1)$$

where Γ_1, Γ_2, Λ_1, *and* Λ_2 *are defined in Table 3.2 and* $\phi = \lim\sqrt{P1/P2}$.

Proof. We can apply the approach used in McCracken's (2007) Theorem 3.1 to show that (using McCracken's notation)

$$\sum_{t=R+P_1+1}^{R+P_1+P_2} D_t = \sigma^2 \sum_{t=R+P_1+1}^{R+P_1+P_2} \tilde{H}_{2,t}'\tilde{h}_{2,t+1} - \frac{\sigma^2}{2}\sum_{t=R+P_1+1}^{R+P_1+P_2} \tilde{H}_{2,t}'\tilde{H}_{2,t} + o_p(1), \quad (3.5.2)$$

with $\tilde{h}_{2,t+1} = \sigma^{-1}\tilde{A}x_t y_t$ and

$$\tilde{H}_{2,t} = \begin{cases} t^{-1}\sum_{s=1}^{t}\tilde{h}_{2,t+1} & \text{recursive window} \\ R^{-1}\sum_{s=1}^{R}\tilde{h}_{2,t+1} & \text{fixed window, } t = R_1, \ldots, T_1 \\ T_1^{-1}\sum_{s=1}^{T_1}\tilde{h}_{2,t+1} & \text{fixed window, } t = R_1, \ldots, T \\ R^{-1}\sum_{s=t-R+1}^{t}\tilde{h}_{2,t+1} & \text{rolling window} \end{cases} \quad (3.5.3)$$

The matrix $\tilde{A}$ any is any $K_1 \times K_2$ matrix that satisfies

$$\Sigma^{1/2}(\Sigma^{-1} - J\Sigma_{11}J')\Sigma^{1/2} \quad (3.5.4)$$

with K_1 the number of predictors used by the smaller model and K_2 the number used by the larger model, $J = (I_{K_1\times K_1}, 0_{K_1\times K_2})$, $\Sigma = Ex_t x_t'$ and Σ_{11} the square matrix of the upper left K_1 elements of that matrix. In the same theorem, McCracken establishes that

$$\sum_{t=R+1}^{R+P_1} D_t = \sigma^2 \sum_{t=R+1}^{R+P_1} \tilde{H}_{2,t}'\tilde{h}_{2,t+1} - \frac{\sigma^2}{2}\sum_{t=R+1}^{R+P_1} \tilde{H}_{2,t}'\tilde{H}_{2,t} + o_p(1). \quad (3.5.5)$$

and

$$\sum_{t=R+1}^{R+P_1}(D_t-\bar{D}_1)^2=\sum_{t=R+1}^{R+P_1}\tilde{H}'_{2,t}\tilde{H}_{2,t}+o_p(1). \tag{3.5.6}$$

The result then follows from McCracken's Lemmas A1, A2, and A3 (which prove that each sum converges individually) and the continuous mapping theorem. ■

Lemma 3.5.2. *Supose that* $\{y_t, x_t\}$ *is strong mixing of size* $-r/(r-1)$ *for* $r>1$, *that* D_t *is* $L_{r+\delta}$*-bounded for each* t *and for some* $\delta>0$, *and that the asymptotic variances of* $P_1^{-1/2}\sum_{t=R+1}^{R+P_1}D_t$ $P_2^{-1/2}\sum_{t=R+P_1+1}^{R+P_1+P_2}D_t$ *are equal to* $\sigma^2>0$. *If* $ED_t=0$ *for all* t, *then*

$$\frac{1}{\hat{\sigma}}\left(P_1^{-1/2}\sum_{t=R+1}^{R+P_1}D_t, P_2^{-1/2}\sum_{t=R+P_1+1}^{R+P_1+P_2}D_t\right)\xrightarrow{d}N(0,I) \tag{3.5.7}$$

as $T\to\infty$, *where* $\hat{\sigma}^2$ *is a consistent estimator of* σ^2.

Proof. Giacomini and White's (2006) Theorem 1 esnures that

$$P_1^{-1/2}\sum_{t=R+1}^{R+P_1}D_t\xrightarrow{d}N(0,\sigma^2)\quad\text{and}\quad P_2^{-1/2}\sum_{t=R+P_1+1}^{R+P_1+P_2}D_t\xrightarrow{d}N(0,\sigma^2), \tag{3.5.8}$$

so it suffices to prove that these two random variables are asymptotically independent. This independence follows from the fact that $\{D_t\}$ is a mixing sequence when R is fixed and a rolling window is used for the forecasts. ■

Table 3.3: Countries in dataset available at each start date.

From 1975	From 1992
Australia	Australia
Austria	Austria
Canada	Canada
Finland	Finland
France	France
Germany	Germany
	Hungary
	Ireland
Italy	Italy
Japan	Japan
	Korea
	New Zealand
Norway	Norway
	Portugal
Spain	Spain
Sweden	Sweden
Switzerland	Switzerland
United Kingdom	United Kingdom
United States	United States

Table 3.4: Sample sizes and number of available countries for each choice of interval sample and forecast sample.

First observation	Start of Forecast Sample	Number of countries	P_2
1975 q 1	2001 q 1	14	32
1975 q 1	2006 q 1	14	8
1992 q 1	2001 q 1	19	32
1992 q 1	2006 q 1	19	8

3.A Mathematical Appendix

Lemma 3.A.1. *Suppose that McCracken's (2007) Assumptions 1–3 hold and that* $P_1/R \to \pi_1$ *and* $P_2/(R+P_1) \to \pi_2$ *with* $\pi_1, \pi_2 \in (0, \infty)$. *Then*

$$\frac{1}{\sqrt{P_1^{-1}\sum_{t=R+1}^{R+P_1}(D_t - \bar{D}_1)^2}}\left(P_1^{-1/2}\sum_{t=R+1}^{T} D_t, P_2^{-1/2}\sum_{t=R+P_1+1}^{R+P_1+P_2} D_t\right) \xrightarrow{d} \left(\frac{\Gamma_1 - \frac{1}{2}\Gamma_2}{\sqrt{\Gamma_2}}, \phi\frac{\Lambda_1 - \frac{1}{2}\Lambda_2}{\sqrt{\Lambda_2}}\right) \quad (3.A.1)$$

where Γ_1, Γ_2, Λ_1, *and* Λ_2 *are defined in Table 3.2 and* $\phi = \lim\sqrt{P1/P2}$.

Proof. We can apply the approach used in McCracken's (2007) Theorem 3.1 to show that (using McCracken's notation)

$$\sum_{t=R+P_1+1}^{R+P_1+P_2} D_t = \sigma^2 \sum_{t=R+P_1+1}^{R+P_1+P_2} \tilde{H}_{2,t}'\tilde{h}_{2,t+1} - \frac{\sigma^2}{2}\sum_{t=R+P_1+1}^{R+P_1+P_2} \tilde{H}_{2,t}'\tilde{H}_{2,t} + o_p(1), \quad (3.A.2)$$

with $\tilde{h}_{2,t+1} = \sigma^{-1}\tilde{A}x_t y_t$ and

$$\tilde{H}_{2,t} = \begin{cases} t^{-1}\sum_{s=1}^{t}\tilde{h}_{2,t+1} & \text{recursive window} \\ R^{-1}\sum_{s=1}^{R}\tilde{h}_{2,t+1} & \text{fixed window, } t = R_1, \ldots, T_1 \\ T_1^{-1}\sum_{s=1}^{T_1}\tilde{h}_{2,t+1} & \text{fixed window, } t = R_1, \ldots, T \\ R^{-1}\sum_{s=t-R+1}^{t}\tilde{h}_{2,t+1} & \text{rolling window} \end{cases} \quad (3.A.3)$$

The matrix $\tilde{A}$ any is any $K_1 \times K_2$ matrix that satisfies

$$\Sigma^{1/2}(\Sigma^{-1} - J\Sigma_{11}J')\Sigma^{1/2} \quad (3.A.4)$$

with K_1 the number of predictors used by the smaller model and K_2 the number used by the larger model, $J = (I_{K_1\times K_1}, 0_{K_1\times K_2})$, $\Sigma = Ex_t x_t'$ and Σ_{11} the square matrix of the upper left K_1 elements of that matrix. In the same theorem, McCracken establishes that

$$\sum_{t=R+1}^{R+P_1} D_t = \sigma^2 \sum_{t=R+1}^{R+P_1} \tilde{H}_{2,t}'\tilde{h}_{2,t+1} - \frac{\sigma^2}{2}\sum_{t=R+1}^{R+P_1} \tilde{H}_{2,t}'\tilde{H}_{2,t} + o_p(1). \quad (3.A.5)$$

and

$$\sum_{t=R+1}^{R+P_1} (D_t - \bar{D}_1)^2 = \sum_{t=R+1}^{R+P_1} \tilde{H}'_{2,t}\tilde{H}_{2,t} + o_p(1). \tag{3.A.6}$$

The result then follows from McCracken's Lemmas A1, A2, and A3 (which prove that each sum converges individually) and the continuous mapping theorem. ■

Lemma 3.A.2. *Supose that* $\{y_t, x_t\}$ *is strong mixing of size* $-r/(r-1)$ *for* $r > 1$*, that* D_t *is* $L_{r+\delta}$*-bounded for each* t *and for some* $\delta > 0$*, and that the asymptotic variances of* $P_1^{-1/2} \sum_{t=R+1}^{R+P_1} D_t$ $P_2^{-1/2} \sum_{t=R+P_1+1}^{R+P_1+P_2} D_t$ *are equal to* $\sigma^2 > 0$*. If* $ED_t = 0$ *for all* t*, then*

$$\frac{1}{\hat{\sigma}} \left(P_1^{-1/2} \sum_{t=R+1}^{R+P_1} D_t, P_2^{-1/2} \sum_{t=R+P_1+1}^{R+P_1+P_2} D_t \right) \xrightarrow{d} N(0, I) \tag{3.A.7}$$

as $T \to \infty$*, where* $\hat{\sigma}^2$ *is a consistent estimator of* σ^2*.*

Proof. Giacomini and White's (2006) Theorem 1 esnures that

$$P_1^{-1/2} \sum_{t=R+1}^{R+P_1} D_t \xrightarrow{d} N(0, \sigma^2) \quad \text{and} \quad P_2^{-1/2} \sum_{t=R+P_1+1}^{R+P_1+P_2} D_t \xrightarrow{d} N(0, \sigma^2), \tag{3.A.8}$$

so it suffices to prove that these two random variables are asymptotically independent. This independence follows from the fact that $\{D_t\}$ is a mixing sequence when R is fixed and a rolling window is used for the forecasts. ■

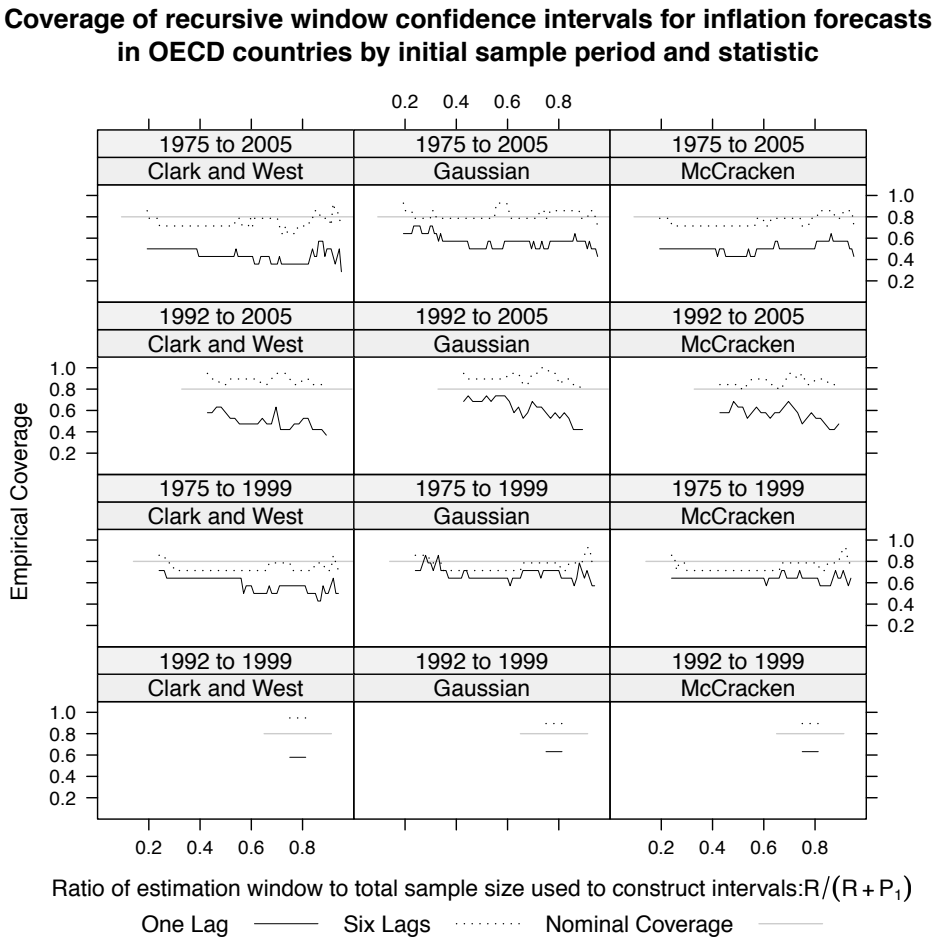

Figure 3.1: Each panel depicts the relative frequency with which the difference in MSE over the forecast sample is contained in its corresponding confidence interval. Intervals were constructed at the 80% level using the sample and asymptotic approximation listed in each panel's title; the difference in MSE is calculated from the end of the sample to the 4th quarter of 2008. Different intervals are constructed for each country (see Table 3.3 for a list) and each division of the sample into estimation and test windows of size R and P_1 respectively. The relative frequency is taken over the countries for each division.

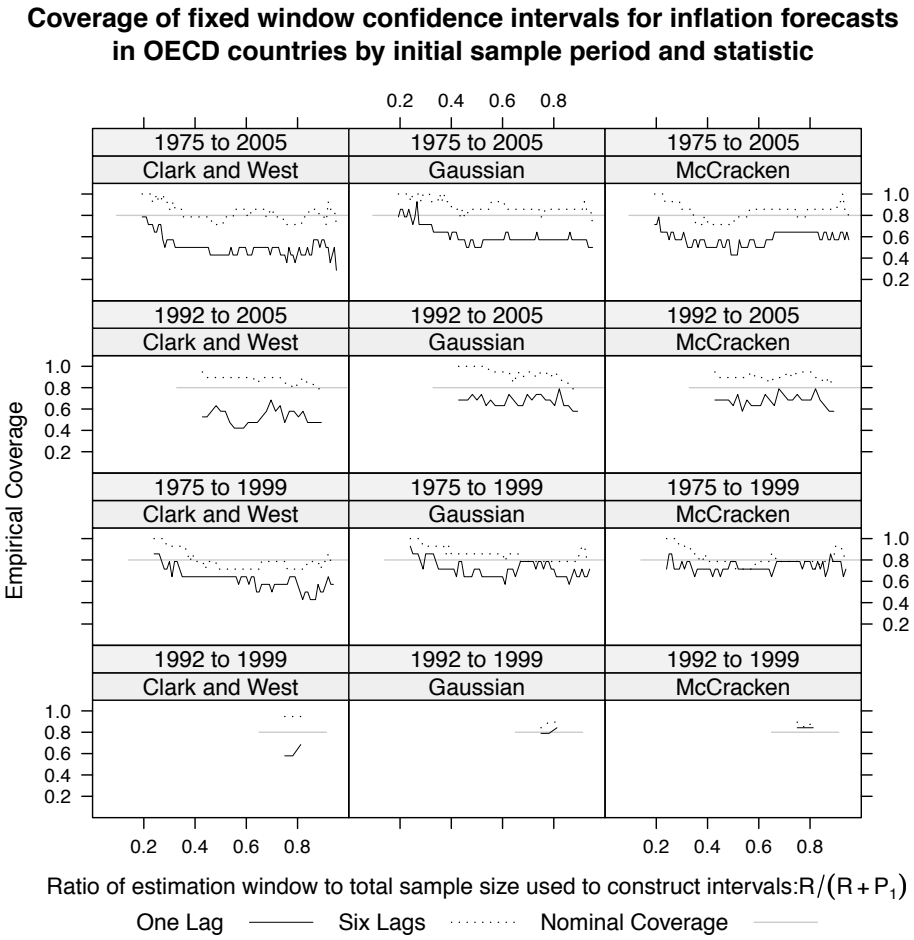

Figure 3.2: See caption for Figure 3.1

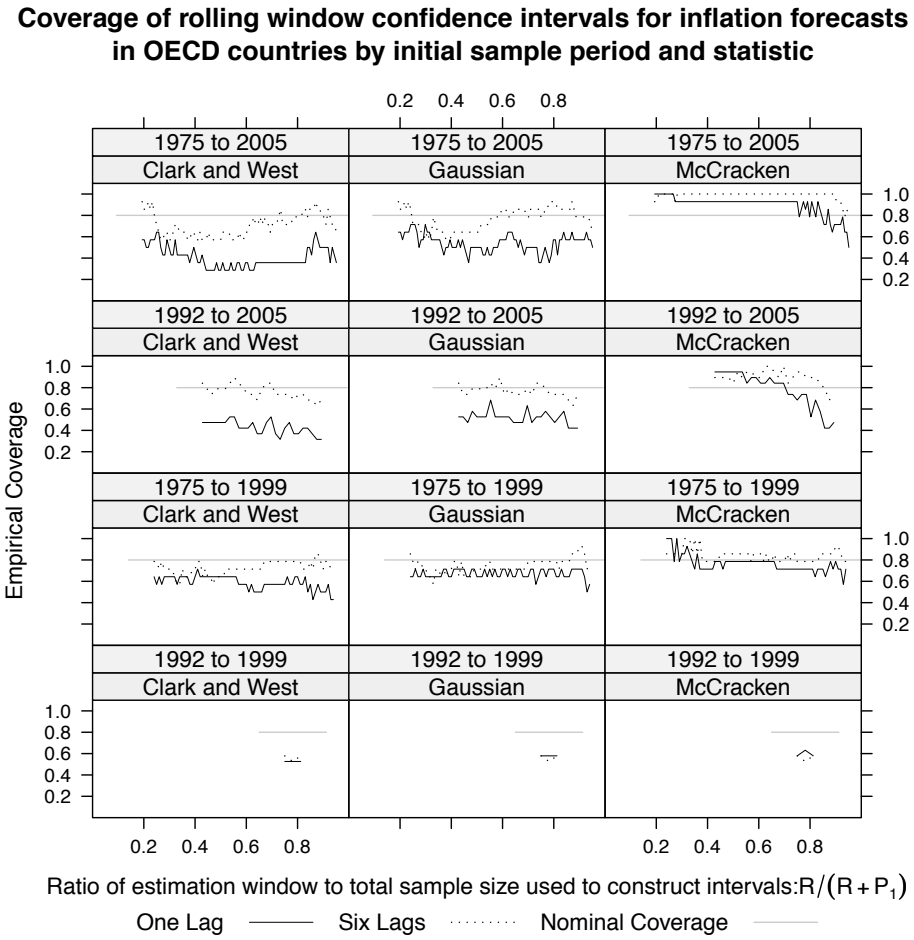

Figure 3.3: See caption for Figure 3.1

Bibliography

Akaike, H. (1973): "Information theory and an extension of the maximum likelihood principle," in *2nd International Symposium on Information Theory*, ed. by B. N. Petrov, and F. Csaki, pp. 267–281. Akademiai Kiado, Budapest.

Akritas, M., and S. Arnold (2000): "Asymptotics for analysis of variance when the number of levels is large," *Journal of the American Statistical Association*, 95, 212–226.

Akritas, M. G., and N. Papadatos (2004): "Heteroscedastic one-way ANOVA and lack-of-fit tests," *Journal of the American Statistical Association*, 99(466), 368–382.

Anatolyev, S. (2008): "Inference in regression models with many predictors," working paper.

Andrews, D. W., and J. H. Stock (2007): "Testing with many weak instruments," *Journal of Econometrics*, 138(1), 24–46.

Atkeson, A., and L. E. Ohanian (2001): "Are Phillips curves useful for forecasting inflation?," *Federal Reserve Bank of Minneapolis Quarterly Review*, 25(1), 2.

Bathke, A. (2004): "The ANOVA F test can still be used in some balanced designs with unequal variances and nonnormal data," *Journal of Statistical Planning and Inference*, 126, 413–422.

Bekker, P. A. (1994): "Alternative approximations to the distributions of instrumental variable estimators," *Econometrica*, 62(3), 657–681.

Berbee, H. C. P. (1979): *Random walks with stationary increments and renewal theory*, Mathematical Centre Tracts. Mathematisch Centrum, Amsterdam.

Boos, D. D., and C. Brownie (1995): "ANOVA and rank tests when the number of treatments is large," *Statistics & Probability Letters*, 23, 183–191.

Calhoun, G. (2009): "Asymptotic theory for overfit models," *working paper*.

Calvo, G. A. (1983): "Staggered prices in a utility-maximizing framework," *Journal of Monetary Economics*, 12(3), 383–398.

Chao, J., J. A. Hausman, W. K. Newey, N. R. Swanson, and T. Woutersen (2008): "Instrumental variable estimation with heteroskedasticity and many instruments," Working Paper.

Chao, J. C., V. Corradi, and N. R. Swanson (2001): "An out of sample test for granger causality," *Macroeconomic Dynamics*, 5(4), 598–620.

Chao, J. C., and N. R. Swanson (2005): "Consistent estimation with a large number of weak instruments," *Econometrica*, 73, 1673–1692.

Chen, G., and R. A. Lockhart (2001): "Weak convergence of the empirical process of residuals in linear models with many parameters," *The Annals of Statistics*, 29, 748–762.

Chen, S. S. (2005): "A note on in-sample and out-of-sample tests for granger causality," *Journal of Forecasting*, 24(6), 453–464.

Clark, T. E. (2004): "Can out-of-sample forecast comparisons help prevent overfitting?," *Journal of Forecasting*, 23(2), 115–139.

Clark, T. E., and M. W. McCracken (2001): "Tests of equal forecast accuracy and encompassing for nested models," *Journal of Econometrics*, 105(1), 85–110.

——— (2005): "The power of tests of predictive ability in the presence of structural breaks," *Journal of Econometrics*, 124(1), 1–31.

Clark, T. E., and K. D. West (2006): "Using out-of-sample mean squared prediction errors to test the martingale difference hypothesis," *Journal of Econometrics*, 135(1-2), 155–186.

——— (2007): "Approximately normal tests for equal predictive accuracy in nested models," *Journal of Econometrics*, 138(1), 291–311.

Corradi, V., and N. R. Swanson (2002): "A consistent test for nonlinear out of sample predictive accuracy," *Journal of Econometrics*, 110(2), 353–381.

——— (2004): "Some recent developments in predictive accuracy testing with nested models and (generic) nonlinear alternatives," *International Journal of Forecasting*, 20(2), 185–199.

Davidson, J. (1992): "A central limit theorem for globally nonstationary near-epoch dependent functions of mixing processes," *Econometric Theory*, 8(3), 313–329.

Davidson, J., and R. M. de Jong (1998): "Consistency of Newey-West type estimators with truncated kernels," working paper.

Davidson, J., and R. M. de Jong (2000): "Consistency of kernel estimators of heteroscedastic and autocorrelated covariance matrices," *Econometrica*, 68, 407–423.

de Jong, P. (1987): "A central limit theorem for generalized quadratic forms," *Probability Theory and Related Fields*, 75, 261–277.

de Jong, R. M. (1997): "Central limit theorems for dependent heterogeneous random variables," *Econometric Theory*, 13(3), 353–367.

Dedecker, J., and C. Prieur (2005): "New dependence coefficients. Examples and applications to statistics," *Probability Theory and Related Fields*, 132, 203–236.

Diebold, F. X., and R. S. Mariano (1995): "Comparing predictive accuracy," *Journal of Business and Economic Statistics*, 13(3), 253–63.

Efron, B. (1986): "How biased is the apparent error rate of a prediction rule?," *Journal of the American Statistical Association*, 81, 461–470.

——— (2004): "The estimation of prediction error: covariance penalties and cross-validation," *Journal of the American Statistical Association*, 99(467), 619–632.

Giacomini, R., and H. White (2006): "Tests of conditional predictive ability," *Econometrica*, 74(6), 1545–1578.

Hall, P. (1984): "Central limit theorem for integrated square error of multivariate nonparametric density estimators," *Journal of Multivariate Analysis*, 14(1), 1–16.

Hall, P., and C. C. Heyde (1980): *Martingale limit theory and its application*. Academic Press.

Han, C., and P. C. B. Phillips (2006): "GMM with many moment conditions," *Econometrica*, 74(1), 147–192.

Hansen, C., J. Hausman, and W. Newey (2008): "Estimation with many instrumental variables," *Journal of Business & Economic Statistics*, 26, 398–422.

Hastie, T., R. Tibshirani, and J. H. Friedman (2003): *The elements of statistical learning*. Springer, 1st ed. 2001. corr. 3rd printing edn.

He, X., and Q. Shao (2000): "On parameters of increasing dimensions," *Journal of Multivariate Analysis*, 73, 120–135.

Huber, P. J. (1973): "Robust regression: asymptotics, conjectures and monte carlo," *The Annals of Statistics*, 1(5), 799–821.

Inoue, A., and L. Kilian (2005): "In-sample or out-of-sample tests of predictability: which one should we use?," *Econometric Reviews*, 23(4), 371–402.

——— (2006): "On the selection of forecasting models," *Journal of Econometrics*, 130(2), 273–306.

Koenker, R., and J. A. F. Machado (1999): "GMM inference when the number of moment conditions is large," *Journal of Econometrics*, 93(2), 327–344.

Lettau, M., and S. Ludvigson (2001): "Consumption, aggregate wealth, and expected stock returns," *The Journal of Finance*, 56, 815–849.

Levine, R., and D. Renelt (1992): "A sensitivity analysis of cross-country growth regressions," *The American Economic Review*, 82(4), 942–963.

Mallows, C. L. (1973): "Some Comments on c_P," *Technometrics*, 15, 661–675.

Mammen, E. (1996): "Empirical process of residuals for high-dimensional linear models," *The Annals of Statistics*, 24(1), 307–335.

McCracken, M. W. (1998): "Data mining and out-of-sample inference," manuscript, Louisiana State University.

——— (2000): "Robust out-of-sample inference," *Journal of Econometrics*, 99(2), 195–223.

——— (2007): "Asymptotics for out of sample tests of granger causality," *Journal of Econometrics*, 140(2), 719–752.

McLeish, D. L. (1975a): "Invariance principles for dependent variables," *Probability Theory and Related Fields*, 32, 165–178.

——— (1975b): "A maximal inequality and dependent strong laws," *The Annals of Probability*, 3(5), 829–839.

Meese, R. A., and K. Rogoff (1983): "Empirical exchange rate models of the seventies: do they fit out of sample?," *Journal of International Economics*, 14(1-2), 3–24.

Merlevède, F., and M. Peligrad (2002): *On the coupling of dependent random variables and applications*pp. 171–193. Birkhauser Boston.

Negro, M. D., F. Schorfheide, F. Smets, and R. Wouters (2007): "On the fit of new keynesian models," *Journal of Business & Economic Statistics*, 25, 123–143.

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

Olivei, G., and S. Tenreyro (2007): "The timing of monetary policy shocks," *The American Economic Review*, 97, 636–663.

Orme, C. D., and T. Yamagata (2006): "The asymptotic distribution of the F-test statistic for individual effects," *Econometrics Journal*, 9, 404–422.

——— (2007): "A Simple heteroskedasticity and nonnormality robust F-test for individual effects," *working paper*.

Portnoy, S. (1984): "Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large. I. Consistency," *The Annals of Statistics*, 12(4), 1298–1309.

——— (1985): "Asymptotic behavior of M-Estimators of p regression parameters when p^2/n is large; II. normal approximation," *The Annals of Statistics*, 13(4), 1403–1417.

——— (1986): "Asymptotic behavior of the empiric distribution of M-Estimated residuals from a regression model with many parameters," *The Annals of Statistics*, 14(3), 1152–1170.

R Development Core Team (2009): *R: A Language and Environment for Statistical Computing*R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

Sala-i-Martin, X. X. (1997): "I just ran two million regressions," *The American Economic Review*, 87(2), 178–183.

Sala-i-Martin, X. X., G. Doppelhofer, and R. I. Miller (2004): "Determinants of long-term growth: a Bayesian Averaging of Classical Estimates (BACE) approach," *The American Economic Review*, 94, 813–835.

Scheffé, H. (1959): *The analysis of variance*. John Wiley & Sons.

Seber, G., and A. Lee (2003): *Linear regression analysis*. Wiley-Interscience, second edition edn.

Stock, J. H., and M. W. Watson (2003): "Forecasting output and inflation: the role of asset prices," *Journal of Economic Literature*, 41(3), 788–829.

Stock, J. H., and M. W. Watson (2007): "Why has U.S. inflation become harder to forecast?," *Journal of Money, Credit and Banking*, 39(s1), 3–33.

——— (2008): "Phillips cuvre inflation forecasts," working paper.

Stock, J. H., and M. Yogo (2005): "Asymptotic distributions of instrumental variable statistics with many weak instruments," *Identification and Inference for Econometric Models: Essays in Honor of Thomas J. Rothenberg*.

Taylor, J. B. (1980): "Aggregate dynamics and staggered contracts," *The Journal of Political Economy*, 88(1), 1âĂŞ23.

Wang, L., and M. G. Akritas (2006): "Two-way heteroscedastic ANOVA when the number of levels is large," *Statistica Sinica*, 16, 1387.

Welsh, A. H. (1989): "On M-processes and M-estimation," *The Annals of Statistics*, 17, 337–361.

West, K. D. (1996): "Asymptotic inference about predictive ability," *Econometrica*, 64(5), 1067–1084.

White, H. (2000): *Asymptotic theory for econometricians*. Academic Press.

Yin, Y. Q. (1986): "LSD' for a class of random matrices," *Journal of Multivariate Analysis*, 20, 50–68.

Yohai, V. J., and R. A. Maronna (1979): "Asymptotic behavior of M-estimators for the linear model," *The Annals of Statistics*, 7, 258–268.