Hypothesis testing in linear regression when $k/n$ is large

Gray Calhoun*

Iowa State University, Department of Economics 467 Heady Hall, Ames, IA 50011, USA

**A R T I C L E  S T A T I S T I C S**

Article history:
Received 18 January 2010
Received in revised form 13 May 2011
Accepted 15 July 2011
Available online 24 July 2011

JEL classification:
C12
C20

Keywords:
Dimension asymptotics
F-test
Ordinary least squares

**A B S T R A C T**

This paper derives the asymptotic distribution of the F-test for the significance of linear regression coefficients as both the number of regressors, $k$, and the number of observations, $n$, increase together so that their ratio remains positive in the limit. The conventional critical values for this test statistic are too small, and the standard version of the F-test is invalid under this asymptotic theory. This paper provides a correction to the F statistic that gives correctly-sized tests both under this paper’s limit theory and also under conventional asymptotic theory that keeps $k$ finite. This paper also presents simulations that indicate the new statistic can perform better in small samples than the conventional test. The statistic is then used to reexamine Olivei and Tenreyro’s results from [Olivei, G., Tenreyro, S., 2007. The timing of monetary policy shocks. The American Economic Review 97, 636–663] and Sala-i-Martin’s results from [Sala-i-Martin, X.X., 1997. I just ran two million regressions. The American Economic Review 87 (2), 178–183].

© 2011 Elsevier B.V. All rights reserved.

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 as $k$ increases, the behavior of Analysis of Variance (ANOVA) as the number of groups increases, and the behavior of instrumental variable 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 standard consistency and asymptotic normality results; $k = o(n)$ is necessary for these results to hold but is often insufficient. The second approach looks for alternative asymptotic distributions of the coefficient estimators keeping $k/n$ positive.

These increasing-$k$ asymptotics were first introduced in the context of $M$-estimation; Huber (1973) argues that assuming $k$ is fixed is unrealistic in practice. After proving that $k = o(n)$ is necessary for the OLS estimator to be consistent and asymptotically normal, Huber argues that this condition is likely to be needed by any tractable asymptotic theory and proves normality of the $M$-estimator of the coefficients of the linear regression model under the stronger condition that $k^{1/3}/n \to 0$. This rate was improved by Yohai and Maronna (1979) and Portnoy (1984, 1985) to $k^{1/5}/n \to 0$ for consistency and $(k\log k)^{1/3}/n \to 0$ for asymptotic normality. 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 (Chen and Lockhart, 2001; Mammen, 1996; Portnoy, 1986).

1. **Introduction**

Consider the linear regression model

$$y_t = x_t'\beta + \epsilon_t \quad t = 1, \ldots, n$$

with $x_t$ and $\epsilon_t$ 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, it is unreliable unless $k/n$ is close to zero; $k$ is the number of regressors in the model. 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 arbitrary linear hypotheses about these coefficients 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 rate was improved by Yohai and Maronna (1979) and Portnoy (1984, 1985) to $k^{1/5}/n \to 0$ for consistency and $(k\log k)^{1/3}/n \to 0$ for asymptotic normality. 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 (Chen and Lockhart, 2001; Mammen, 1996; Portnoy, 1986).

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), building on earlier results by Anderson (1976), Kunitomo (1980), and Morimune (1983),
studies the asymptotic behavior of Two-Stage Least Squares (2SLS) and variations of Limited Information Maximum Likelihood (LIML) in models with normal errors as \( l/n \) converges to a positive constant. These authors find that LIML is both consistent and asymptotically normal but that 2SLS is not. These results are extended to non-Gaussian errors by Hansen et al. (2008), Chao et al. (2008), and others. Koenker and Machado (1999) prove the consistency and asymptotic normality of GMM estimators with \( P/n \to 0 \). Stock and Yogo (2005), Chao and Swanson (2005), and Andrews and Stock (2007), among others, combine the many-instruments and the weak instrument literature and argue that the relationship between the concentration parameter and \( l \) is more important than that between the number of observations and \( l \).

These authors find that LIML is both consistent and asymptotically normal but that 2SLS is not. These results are extended to non-Gaussian errors by Hansen et al. (2008), Chao et al. (2008), and others. Koenker and Machado (1999) prove the consistency and asymptotic normality of GMM estimators with \( P/n \to 0 \). Stock and Yogo (2005), Chao and Swanson (2005), and Andrews and Stock (2007), among others, combine the many-instruments and the weak instrument literature and argue that the relationship between the concentration parameter and \( l \) is more important than that between the number of observations and \( l \).}

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 2 discusses the new test statistic and studies its asymptotic distributions under the null and alternative hypotheses. Section 3 presents Monte Carlo evidence in favor of the statistic. Section 4 presents the empirical exercises. Section 5 concludes. The proofs are presented in the Appendix.

### 2. Asymptotic theory and main results

This section derives the asymptotic distribution of the \( F \)-test of the null hypothesis \( R^2 \) for the linear equation

\[
y_t = x_t'\beta + \epsilon_t
\]

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 present 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 2.1 discusses the paper’s notation and assumptions. Section 2.2 presents asymptotic theory and the new test statistic, and Section 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. Unless otherwise indicated, all limits are taken as \( n \to \infty \).
دریافت فوری
متن کامل مقاله

امکان دانلود نسخه تمام متن مقالات انگلیسی
امکان دانلود نسخه ترجمه شده مقالات
پذیرش سفارش ترجمه تخصصی
امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
امکان دانلود رایگان ۲ صفحه اول هر مقاله
امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
دانلود فوری مقاله پس از پرداخت آنلاین
پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات