TESTS OF LINEAR HYPOTHESES AND $l_1$ ESTIMATION

BY ROGER KOENKER AND GILBERT BASSETT

1. INTRODUCTION

ALTHOUGH $l_1$ ESTIMATION METHODS based on minimizing sums of absolute residuals have a long history in econometrics, a serious limitation to their wider application has been the lack of any $l_1$ hypothesis testing apparatus comparable to classical least squares procedures.\footnote{The seminal contributions of Edgeworth to $l_1$ estimation in econometrics are mentioned in Stigler [19]. Taylor's [20] survey discusses several econometric applications of $l_1$ methods. Blattberg and Sargent [7] conclude on the basis of their Monte-Carlo investigations that the $l_1$ estimator "performs sufficiently well that it deserves further study and elaboration. In particular, the sampling theory of the estimator needs to be developed." We have previously studied the asymptotic behavior of the $l_1$ estimator in linear models; see Bassett and Koenker [3]. Amemiya [2] has recently extended $l_1$ methods to simultaneous equation models.} This note investigates the asymptotic distribution of three alternative $l_1$ test statistics of a linear hypothesis in the standard linear model. These test statistics, which correspond to Wald, likelihood ratio, and Lagrange multiplier tests, are shown to have the same limiting chi-square behavior under mild regularity conditions on design and the distribution of errors. The asymptotic theory of the tests is derived for a large class of error distributions; thus in Huber's [10] terminology we investigate the behavior of the likelihood ratio test under "non-standard" conditions. The asymptotic efficiency of the $l_1$ tests involves a modest sacrifice of power compared to classical tests in cases of strictly Gaussian errors but may yield large efficiency gains in non-Gaussian situations. The Lagrange multiplier test seems particularly attractive from a computational standpoint.

We derive the asymptotic distribution of the three alternative $l_1$ test statistics for a simple linear exclusion hypothesis. Extension of these results to hypotheses of the form $R\beta = r$ is a straightforward exercise. When the density of the error distribution is strictly positive at the median, all three test statistics have the same limiting central $\chi^2$ behavior at the null and noncentral $\chi^2$ behavior for local alternatives to the null. When the variance of the error distribution is bounded, analogous results are well known for classical forms of the Wald, likelihood ratio, and Lagrange multiplier tests based on least-squares methods. See, for example, Silvey [18] and the discussion in Section 4 below.

2. NOTATION AND ASSUMPTIONS

We consider the familiar linear model

\begin{equation}
(2.1) \quad y = X\beta + \epsilon,
\end{equation}

which we will partition as

\begin{equation}
(2.2) \quad y = X_1 \beta_1 + X_2 \beta_2 + \epsilon,
\end{equation}

where $y$ is an $n$-vector of independent observations, $X$ is an $n$ by $p$ matrix of known constants, and $\beta$ is a $p$-vector of unknown parameters with components $\beta_1$ and $\beta_2$ of dimension $p - k$ and $k$ respectively. The errors $\epsilon_1, \epsilon_2, \ldots$ are assumed to be independent and identically distributed with common distribution function $F$. Interest will focus on tests of the null hypothesis

\begin{equation}
(2.3) \quad H_0 : \beta_2 = 0.
\end{equation}
We will require the following additional assumptions:

**Assumption A1** (Design): There exists a positive definite matrix $D$ such that $\lim_{n \to \infty} X'X/n = D$.

**Assumption A2** (Density): $F$ has continuous and strictly positive density $f(\cdot)$ at the median $F^{-1}(\frac{1}{2}) = 0$.

When $X$ includes an intercept the hypothesis that the error distribution has median zero involves no loss in generality.

The objective function for the $l_1$ estimation problem is defined as

\begin{equation}
V(b) = \sum_{i=1}^{n} |y_i - x_i b|.
\end{equation}

The unrestricted $l_1$ estimator, $\hat{\beta}_n$, is any minimizer of $V(b)$ over $b \in \mathbb{R}^p$. The restricted $l_1$ estimator $\hat{\beta}_n = (\hat{\beta}_1, 0)'$ is any minimizer of $V((b_1, 0))$ over $b_1 \in \mathbb{R}^{p-k}$. We assume there exists some rule to choose a unique estimate when minimum sets of (2.4) are not unique. Our asymptotic results are invariant to the choice of this rule. The sample-size subscript will be omitted below.

Under the foregoing assumptions, all of the $l_1$ tests we will consider are consistent, that is for any fixed alternative $\beta_2 \neq 0$, they will reject the null hypothesis with probability one as $n$ tends to infinity. Classical $F$ tests are also consistent in this sense if we impose the additional condition that the error distribution exhibit finite variance. In order to provide a meaningful comparison of the asymptotic power of the tests studied we adopt the classical convention of considering a sequence of alternatives which tends to the null hypothesis at rate $1/\sqrt{n}$. The statistical foundations of this approach are discussed by Hajek and Sidak [9]. The dependence of $\beta_2$ on $n$ will be notationally suppressed, but is made explicit in the following assumption.

**Assumption A3** (Sequence of Alternatives): There exists a fixed $\gamma \in \mathbb{R}^k$ such that $\beta_2 = \gamma/\sqrt{n}$ for samples of size $n$.

We also adopt the following standard notation for partitioning the matrix $D$. $D_{ij} : i, j = 1, 2$ denotes the $i, j$th block of $D$, while $D^{ij}$ denotes the $i, j$th block of $D^{-1}$. To illustrate, recall the useful identity

\begin{equation}
D^{22} = (D_{22} - D_{21}D_{11}^{-1}D_{12})^{-1}
\end{equation}

which plays an important role in subsequent developments. Finally, the symbol $\to$ will denote convergence in distribution, $G_p(\mu, \Sigma)$ denotes the $p$-variate Gaussian distribution with mean vector $\mu$ and covariance matrix $\Sigma$ and $\chi^2_\nu(\eta)$ denotes the noncentral chi-squared distribution with $\nu$ degrees of freedom and noncentrality parameter $\eta$.

### 3. Tests of $H_0$

The Wald test is based on the estimated coefficients of the unrestricted model. The test statistic is given by

\begin{equation}
\xi = n\omega^{-2}\hat{\beta}_2'(D^{22})^{-1}\hat{\beta}_2.
\end{equation}
where $\omega$ denotes the scale parameter $1/(2f(0))$. The likelihood ratio test is based on the difference between the sum of absolute residuals in the restricted and unrestricted models. Let $\hat{V}$ and $\hat{V}$ denote these respective sums, and set

$$
(3.2) \quad \xi_{LR} = 2\omega^{-1}(\hat{V} - \hat{V}).
$$

Finally, the Lagrange multiplier test is based on the gradient of the unrestricted $l_1$ objective function evaluated at the restricted estimate. Let

$$
W(\delta) = \sum |e_i - x_i\delta/\sqrt{n}|
$$

so $W(\sqrt{n}(b - \beta)) = V(b)$, set $\psi(x) = \text{sgn}(x)$ and define the normalized gradient of $W$,

$$
(3.3) \quad g_n(\delta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_i \psi(\epsilon_i - x_i\delta/\sqrt{n}).
$$

Actually, $g(\delta)$ is one element of a potentially set valued subgradient, but since $F$ is continuous this gradient is well-defined with probability one. We partition $g_n(\delta) = (g_{1n}(\delta), g_{2n}(\delta))$ conforming to the subvectors $(x_1, x_2)$. Let $\delta = \sqrt{n}(\beta_1 - \beta_{1n} - \beta_2) = (\sqrt{n}(\beta_1 - \beta_{1n}), -\gamma)$, and set $\tilde{g}_i = g_i(\delta)$, $i = 1, 2$. If $\tilde{g}_2$ is large then the null hypothesis is implausible. We define the test statistic

$$
(3.4) \quad \xi_{LM} = \tilde{g}_2 D^{22} \tilde{g}_2.
$$

The asymptotic behavior of the three tests under the null and the contiguous alternatives to the null specified in Assumption A3 is summarized in the following theorem.

**Theorem:** The test statistics $\xi_W$, $\xi_{LR}$, and $\xi_{LM}$ are asymptotically equivalent and converge in distribution to $\chi^2(\eta)$ where $\eta = \omega^{-2}\lambda(D^{22})^{-1}\lambda$.

**Remark:** In practice of course we would replace $n(D^{-1})^{-1}$ with its sample analogue $X'_2(I - X_1'X_1)^{-1}X_2$ in (3.1), (3.2), and (3.4). Consistent estimators for $\omega$ are proposed in Koenker and Bassett [14], based on techniques introduced in Bassett and Koenker [5]. The Lagrange multiplier test statistic (3.4) offers the significant advantage that it does not require an estimate of the nuisance parameter $\omega$.

**Proof:** Each test is treated in turn.

$W$: Koenker and Bassett [3] prove $\sqrt{n}(\beta - \beta) \rightarrow G_p(0, \omega^2 D^{-1})$. Thus using Assumption A3, $\sqrt{n} \beta_2 \rightarrow G_2(\gamma, \omega^2 D^{22})$ and the result follows.

$LR$: This is essentially a special case of a general result due to Schrader and Hettmansperger [17] on LR tests for robust $M$-estimators. Unfortunately their regularity conditions (due to Huber [11]) do not cover the $l_1$ case. We sketch an alternative method of proof using results of Jaeckel [12] and Ruppert and Carroll [16].

Define the quadratic objective function,

$$
(3.5) \quad Q(\delta) = \frac{1}{\omega} \delta^'D\delta - \delta^'g(0) + V(0).
$$

Using Ruppert and Carroll's [16, Lemma A.3] asymptotic linearity result on the gradient,

$$
(3.6) \quad \sup_{\|\delta\| < L} \|g(\delta) - g(0) + 2f(0)D\delta\| = o_p(1) \quad \text{for all} \quad L > 0.
$$
The argument leading to Lemma 1 of Jaeckel [12] yields,

\[ \max_{\|\delta\| < L} |Q(\delta) - V(\delta)| = o_p(1), \]

and furthermore any sequence of minimizers of \( V(\delta) \), say \( \{\delta_k\} \) is asymptotically equivalent to the unique minimizer of \( Q(\delta) \), i.e. to the pseudo-estimator,

\[ \delta_0 = \omega D^{-1}g(0). \]

(Since \( g(0) \to G_{02}(0, D) \) this line of argument yields a simple proof that \( \delta = \sqrt{n} (\beta - \beta) \to G(0, \omega^2 D^{-1}). \) Thus \( Q(\delta) \to Q(\delta_0) \) and similarly \( Q(\delta) \to Q(\delta_0) \) where \( \delta_0 \) is the restricted pseudo-estimator

\[ \delta_0 = \begin{bmatrix} D_{11}^{-1}D_{12}\gamma + \omega D_{11}^{-1}g_1(0) \\ -\gamma \end{bmatrix} \]

obtained by maximizing \( Q(\delta) \) subject to \( \delta_2 = -\sqrt{n} \beta_2 = -\gamma. \) Now, writing as in Schrader and Hettmansperger,

\[ \delta - \hat{\delta} = V(\delta) - Q(\delta) + Q(\delta) - Q(\delta_0) + Q(\delta_0) - Q(\delta) + Q(\delta) - V(\delta) = Q(\delta_0) - Q(\delta_0) + o_p(1). \]

Substituting for \( \delta_0 \) and \( \hat{\delta} \) and simplifying, we have

\[ \frac{2}{\omega} [Q(\delta_0) - Q(\delta_0)] = h'D^{22}h, \]

where

\[ h = \omega^{-1}(D^{22})^{-1}\gamma + g_{22}(0) - D_{21}D_{11}^{-1}g_1(0), \]

but

\[ h \to G(\omega^{-1}(D^{22})^{-1}\gamma, (D^{22})^{-1}) \]

so

\[ \xi_L \to h'D^{22}h \to \chi^2_1(\eta). \]

\( LM: \) Under somewhat more stringent regularity conditions on design, Adichie [1] obtains results on a general class of rank tests of \( H_0 \) which includes \( \xi_{LM}. \) The limiting distribution of \( \xi_{LM} \) under the null hypothesis \( H_0 \) is also given in Cogger [8]. As with \( \xi_L \) we rely heavily upon results of Ruppert and Carroll [16]. The definition of \( \delta \) implies

\[ \tilde{\xi}_1 = o_p(1), \]

but for any \( \epsilon > 0 \), there exist \( M > 0, \xi > 0, \) and integer \( n_0 \) such that

\[ \Pr \left\{ \inf_{\|\delta\| > M} \|G_1(\delta)\| < \xi \right\} < \epsilon, \quad n > n_0. \]
Thus $\|\hat{\beta}\| = O_p(1)$ and (3.6) implies

$$ (3.15) \quad \| g(\hat{\beta}) - g(0) + \omega^{-1}D\hat{\beta} \| = o_p(1). $$

Using (3.13) we have

$$ (3.16) \quad \omega^{-1}\left[ \sqrt{n} D_{11}(\hat{\beta}_1 - \beta_1) - D_{12} \gamma \right] = \frac{1}{\sqrt{n}} X_i^\prime \psi + o_p(1), $$

where $\psi = (\psi(\epsilon_i - x_i\tilde{\beta}/\sqrt{n}))$, hence,

$$ (3.17) \quad \sqrt{n} (\hat{\beta}_1 - \beta_1) = D_{11}\gamma + \frac{1}{\sqrt{n}} \omega D_{11}^{-1}X_i^\prime \psi + o_p(1). $$

Substituting in (3.15) and simplifying we have

$$ (3.18) \quad \tilde{g}_2 = \omega^{-1}(D_{22})^{-1} \gamma + \frac{1}{\sqrt{n}} \left[ X_i^2 - D_{21}D_{11}^{-1}X_i \right] \psi + o_p(1). $$

Assumption A1 implies the standard multivariate Lindeberg condition and we conclude

$$ (3.19) \quad \tilde{g}_2 \rightarrow G(\omega^{-1}(D_{22})^{-1} \gamma, (D_{22})^{-1}), $$

and therefore that $\xi_{LM} \rightarrow \chi^2(\eta)$. \(Q.E.D.\)

Under fairly mild regularity conditions these results may be extended to other maximum likelihood (M-) estimators for linear models with iid errors. Let $\hat{\beta} = (\hat{\beta}_1, 0)^\prime$ and $\beta$ denote restricted and unrestricted solutions to (i) the problem:

$$ (3.20) \quad \min \sum \rho(y_i - x_i, b), $$

or equivalently, letting $\psi = \rho'$, to (ii) the “normal” equations,

$$ (3.21) \quad \sum x_i \psi(y_i - x_i, b) = 0. $$

Under conditions discussed in Huber [11]

$$ (3.22) \quad \xi_M = n\sigma^{-2}(\psi, F) \hat{\beta}_2(D_{22})^{-1} \beta_2 \rightarrow \chi^2_\eta(\psi, F) $$

where

$$ (3.23) \quad \eta(\psi, F) = \sigma^{-2}(\psi, F)\gamma'(D_{22})^{-1} \gamma $$

and

$$ (3.24) \quad \sigma^2(\psi, F) = E_F \psi^2(\epsilon)/[E_F \psi'(\epsilon)]^2. $$

Schrader and Hettmansperger [17] show under Huber’s conditions that the likelihood ratio statistic

$$ (3.25) \quad \xi_{LR} = \lambda^{-1}(\psi, F)[\hat{V} - \hat{V}] \rightarrow \chi^2_\eta(\psi, F) $$

where $\hat{V}$ and $\hat{V}$ denote the values taken by the objective function (6.1) at the restricted
and unrestricted estimates respectively, and

\begin{equation}
\lambda = \frac{1}{2} E_F \psi^2 / E_F \psi'.
\end{equation}

Finally, similar methods may be used to show that the Lagrange multiplier statistic,

\begin{equation}
\xi_{LM} \equiv \psi^{-1} \tilde{g}_2(D^{22}) \tilde{g}_2 \rightarrow \chi^2(n - \nu, \nu)
\end{equation}

where \( \tilde{g} = X' \tilde{\psi}/\sqrt{n} \), \( \tilde{\psi} = (\psi(y_i - x_i \tilde{\beta})) \), and \( \nu = E_F \psi^2 \).

In general, the nuisance parameters \( \sigma^2(\psi, F) \), \( \lambda(\psi, F) \), and \( \nu(\psi, F) \) are nontrivial functions of both the assumed form of the likelihood function, and the true, but probably unknown, error distribution. The \( l_1 \) and \( l_2 \) examples are rather striking special cases. In the \( l_2 \) case \( \sigma^2(\psi, F) = \lambda(\psi, F) = \nu(\psi, F) = \sigma^2(F) \), the variance of the errors. In the \( l_1 \) case, using Assumption A2, \( E_F \psi^2 = 1 \), while \( E_F \psi' = 2f(0) \). In the former case \( E_F \psi' \) is independent of \( F \); in the latter case \( E_F \psi^2 \) is independent of \( F \).

In the rather unrealistic case in which \( F \) is known and \( \psi = f'/f \), an integration by parts yields \( E_F \psi = E_F \psi^2 = I(F) \), the Fisher information number of \( F \), and \( \sigma^2(\psi, F) = I^{-1}(F) \). Judicious choice of \( \psi \) can keep \( \sigma^2(\psi, F) \) near the lower bound \( I^{-1}(F) \) for a wide class of \( F \)’s. The least squares choice, \( \psi(x) = x \) is unbounded thus not robust and therefore not judicious in this sense.

4. ASYMPTOTIC RELATIVE EFFICIENCY

We may now compare the efficiency of the \( l_1 \) tests discussed above with classical tests based on least-squares residuals. Let \( \hat{S} \) and \( \hat{S}' \) denote the error sum of squares from the least-squares fit of the restricted and unrestricted forms of model (2.2). If the variance of the error distribution, say \( \sigma^2(F) \), is bounded, then

\begin{equation}
\xi = (\hat{S} - \hat{S}') / \sigma^2(F) \rightarrow \chi^2(\sigma^{-2}(F) \gamma(D^{22})^{-1} \gamma).
\end{equation}

Substituting \( \sigma^2 = \hat{S}/(n - p) \) for \( \sigma^2(F) \) and dividing by \( k \) yields the usual \( F \) statistic for the hypothesis \( H_0 \). Wald and Lagrange multiplier tests based on least-squares methods have the same limiting \( \chi^2 \) distribution under the null and local alternatives to the null. Obviously the distinction between \( \chi^2 \) and \( F \) and the degrees of freedom correction will be unimportant asymptotically, however in practice they may loom large. In applications we suggest using \( (\xi_{LM}/k)(n - p + k)/n \) which is a degrees-of-freedom corrected Lagrange multiplier statistic and may be compared with critical values of \( F \) with \( k \) and \( n - p + k \) degrees of freedom. A very small scale Monte-Carlo experiment designed to evaluate the asymptotic approximations to the size and power of this test statistic is reported in Bassett and Koenker [4]. The degrees-of-freedom correction seems to be quite successful in achieving the correct nominal size for the test in small samples.

The asymptotic relative efficiency (ARE) of the \( l_1 \) tests to the least squares test \( \xi \) is simply the ratio of the noncentrality parameters of the limiting \( \chi^2 \) distributions. This in turn is simply the ratio of the asymptotic variances of the least-squares and \( l_1 \) estimators, \( \sigma^2(F)/\omega^2(F) \). The ratio ARE may be interpreted as the ratio of sample sizes required to achieve a specified power for both tests for a specified level and alternative. See, for example, Pitman [15].

At the Gaussian distribution \( \text{ARE} = 2/\pi \approx 0.64 \) so the \( l_1 \) tests need (asymptotically) 36% more observations than the \( l_1 \) test. However \( \lambda \) is unbounded as \( F \) becomes long-tailed. As in the estimation problem, the choice of test procedures involves the question, “Can we be sufficiently confident of the Gaussian model to ignore an opportunity to insure against the disastrous behavior of least-squares methods under non-Gaussian conditions?” Admittedly the \( l_1 \) insurance premium at the Gaussian distribution is quite high; this is to be
expected since it is maximally robust in a certain sense. $l_i$ estimators and tests are most effective in very inclement statistical weather. For mild departures from Gaussian conditions, less radical methods are probably more appropriate. Huber's $M$-estimator and the tests suggested by Schrader and Hettmansperger, or tests based on Bickel's [6] pseudo-observations offer attractive options. The trimmed least-squares estimators proposed in Koenker and Bassett [13] and studied by Ruppert and Carroll [16] also appear attractive. The $l_i$ estimator and its tests may be viewed as limiting forms of these “trimmed least-squares” alternatives. They have several advantages over their robust competitors: they are somewhat easier to compute, no preliminary estimate of scale is required, and at least in the case of the $l_i$ Lagrange multiplier test the difficulties of estimating nuisance parameters for robust tests are nicely finessed.

Bell Laboratories

and

University of Illinois at Chicago Circle

Manuscript received May, 1981; revision received December, 1981.

REFERENCES
