Tests of Linear Hypotheses and l"1 Estimation

Roger Koenker; Gilbert Bassett

Econometrica, Volume 50, Issue 6 (Nov., 1982), 1577-1584.

Stable URL:
http://links.jstor.org/sici?sici=0012-9682%28198211%2950%3A6%3C1577%3ATOLHAL%3E2.0.CO%3B2-4

Econometrica, Vol. 50, No. 6 (November, 1982)

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.[1] 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 χ^2 behavior at the null and noncentral χ^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 multipler 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

$$y = X\beta + \epsilon, \tag{2.1}$$

which we will partition as

$$y = X_1\beta_1 + X_2\beta_2 + \epsilon, \tag{2.2}$$

where y is an n-vector of independent observations, X is an n by p matrix of known constants, and β is a p-vector of unknown parameters with components β_1 and β_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

$$H_0 : \beta_2 = 0. \tag{2.3}$$

[1] 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.

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

$$V(b) = \sum_{i=1}^{n} |y_i - x_i b|. \tag{2.4}$$

The unrestricted l_1 estimator, $\hat{\beta}_n$, is any minimizer of $V(b)$ over $b \in R^p$. The restricted l_1 estimator $\tilde{\beta}_n = (\tilde{\beta}_1, 0)'$ is any minimizer of $V((b_1, 0))$ over $b_1 \in R^{p-k}$. 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 β_2 on n will be notationally suppressed, but is made explicit in the following assumption.

ASSUMPTION A3 (Sequence of Alternatives): There exists a fixed $\gamma \in R^k$ 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,jth block of D, while D^{ij} denotes the i,jth block of D^{-1}. To illustrate, recall the useful identity

$$D^{22} = \left(D_{22} - D_{21} D_{11}^{-1} D_{12}\right)^{-1} \tag{2.5}$$

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 μ and covariance matrix Σ and $\chi^2_\nu(\eta)$ denotes the noncentral chi-squared distribution with ν degrees of freedom and noncentrality parameter η.

3. TESTS OF H_0

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

$$\xi_W = n\omega^{-2} \hat{\beta}_2' \left(D^{22}\right)^{-1} \hat{\beta}_2, \tag{3.1}$$

where ω 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 $\tilde{V}$ and $\hat{V}$ denote these respective sums, and set

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

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

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

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

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

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 $\delta = \sqrt{n}\,(\beta_1 - \beta_1, -\beta_2) = (\sqrt{n}\,(\tilde{\beta}_1 - \beta_1), -\gamma)$, and set $\tilde{g}_i = g_i(\tilde{\delta})$, $i = 1,2$. If $\tilde{g}_2$ is large then the null hypothesis is implausible. We define the test statistic

$$\xi_{LM} = \tilde{g}_2' D^{22}\tilde{g}_2\,. \tag{3.4}$$

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 ξ_W, ξ_{LR}, and ξ_{LM} are asymptotically equivalent and converge in distribution to $\chi_k^2(\eta)$ where $\eta = \omega^{-2}\gamma'(D^{22})^{-1}\gamma$.*

REMARK: In practice of course we would replace $n(D^{22})^{-1}$ with its sample analogue $X_2'(I - X_1(X_1'X_1)^{-1}X_1')X_2$ in (3.1), (3.2), and (3.4). Consistent estimators for ω 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 ω.

PROOF: Each test is treated in turn.

W: Bassett and Koenker [**3**] prove $\sqrt{n}\,(\hat{\beta} - \beta) \to G_p(0, \omega^2 D^{-1})$. Thus using Assumption A3, $\sqrt{n}\,\hat{\beta}_2 \to G_k(\gamma, \omega^2 D^{22})$ 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,

$$Q(\delta) = \tfrac{1}{2}\omega^{-1}\delta' D\delta - \delta' g(0) + V(0). \tag{3.5}$$

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

$$\sup_{\|\delta\| < L} \| g(\delta) - g(0) + 2f(0)D\delta \| = o_p(1) \qquad \text{for all} \quad L > 0. \tag{3.6}$$

The argument leading to Lemma 1 of Jaeckel [12] yields,

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

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

(3.8) $$\hat{\delta}_0 = \omega D^{-1} g(0).$$

(Since $g(0) \to G_k(0, D)$ this line of argument yields a simple proof that $\hat{\delta} = \sqrt{n}(\hat{\beta} - \beta) \to G(0, \omega^2 D^{-1})$.) Thus $Q(\hat{\delta}) \to Q(\hat{\delta}_0)$ and similarly $Q(\tilde{\delta}) \to Q(\tilde{\delta}_0)$ where $\tilde{\delta}_0$ is the restricted pseudo-estimator

(3.9) $$\tilde{\delta}_0 = \begin{bmatrix} D_{11}^{-1} D_{12}\gamma + \omega D_{11}^{-1} g_1(0) \\ -\gamma \end{bmatrix}$$

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

(3.10) $$\begin{aligned} \tilde{V} - \hat{V} &= V(\tilde{\delta}) - Q(\tilde{\delta}) + Q(\tilde{\delta}) - Q(\tilde{\delta}_0) \\ &\quad + Q(\tilde{\delta}_0) - Q(\hat{\delta}_0) \\ &\quad + Q(\hat{\delta}_0) - Q(\hat{\delta}) + Q(\hat{\delta}) - V(\hat{\delta}) \\ &= Q(\tilde{\delta}_0) - Q(\hat{\delta}_0) + o_p(1). \end{aligned}$$

Substituting for $\tilde{\delta}_0$ and $\hat{\delta}_0$ and simplifying, we have

(3.11) $$\frac{2}{\omega}[Q(\tilde{\delta}_0) - Q(\hat{\delta}_0)] = h' D^{22} h,$$

where

$$h = \omega^{-1}(D^{22})^{-1}\gamma + g_2(0) - D_{21} D_{11}^{-1} g_1(0),$$

but

$$h \to G\left(\omega^{-1}(D^{22})^{-1}\gamma, (D^{22})^{-1}\right)$$

so

(3.12) $$\xi_{LR} \to h' D^{22} h \to \chi_k^2(\eta).$$

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 ξ_{LM}. The limiting distribution of ξ_{LM} under the null hypothesis H_0 is also given in Cogger [8]. As with ξ_{LR} we rely heavily upon results of Ruppert and Carroll [16]. The definition of $\tilde{\delta}$ implies

(3.13) $$\tilde{g}_1 = o_p(1),$$

but for any $\epsilon > 0$, there exist $M > 0$, $\zeta > 0$, and integer n_0 such that

(3.14) $$\Pr\left\{ \inf_{\|\delta\| > M} \|g_1(\delta)\| < \zeta \right\} < \epsilon, \qquad n > n_0.$$

Thus $\|\tilde{\delta}\| = O_p(1)$ and (3.6) implies

$$\| g(\tilde{\delta}) - g(0) + \omega^{-1} D\tilde{\delta} \| = o_p(1). \tag{3.15}$$

Using (3.13) we have

$$\omega^{-1}\left[\sqrt{n}\, D_{11}(\tilde{\beta}_1 - \beta_1) - D_{12}\gamma\right] = \frac{1}{\sqrt{n}} X_1'\psi + o_p(1), \tag{3.16}$$

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

$$\sqrt{n}\,(\tilde{\beta}_1 - \beta_1) = D_{11}^{-1} D^{12}\gamma + \frac{1}{\sqrt{n}}\,\omega D_{11}^{-1} X_1'\psi + o_p(1). \tag{3.17}$$

Substituting in (3.15) and simplifying we have

$$\tilde{g}_2 = \omega^{-1}(D^{22})^{-1}\gamma + \frac{1}{\sqrt{n}}\left[X_2' - D_{21}D_{11}^{-1}X_1'\right]\psi + o_p(1). \tag{3.18}$$

Assumption A1 implies the standard multivariate Lindeberg condition and we conclude

$$\tilde{g}_2 \to G\left(\omega^{-1}(D^{22})^{-1}\gamma, (D^{22})^{-1}\right), \tag{3.19}$$

and therefore that $\xi_{LM} \to \chi_k^2(\eta)$. *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 $\tilde{\beta} = (\tilde{\beta}_1, 0)'$ and $\hat{\beta}$ denote restricted and unrestricted solutions to (i) the problem:

$$\min \sum \rho(y_i - x_i b), \tag{3.20}$$

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

$$\sum x_i \psi(y_i - x_i b) = 0. \tag{3.21}$$

Under conditions discussed in Huber [**11**]

$$\xi_W \equiv n\sigma^{-2}(\psi, F)\,\hat{\beta}_2'(D^{22})^{-1}\hat{\beta}_2 \to \chi_k^2(\eta(\psi, F)) \tag{3.22}$$

where

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

and

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

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

$$\xi_{LR} \equiv \lambda^{-1}(\psi, F)[\tilde{V} - \hat{V}] \to \chi_k^2(\eta(\psi, F)) \tag{3.25}$$

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

$$\lambda = \tfrac{1}{2} E_F \psi^2 / E_F \psi'. \tag{3.26}$$

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

$$\xi_{LM} \equiv \nu^{-1} \tilde{g}_2'(D^{22}) \tilde{g}_2 \rightarrow \chi_k^2(\eta(\psi, F)) \tag{3.27}$$

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 $\sigma^2(\psi, F) = \lambda(\psi, F) = \nu(\psi, F) \equiv \sigma^2(F)$, 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 $E\psi' = E\psi^2 = I(F)$, the Fisher information number of F, and $\sigma^2(\psi, F) = I^{-1}(F)$. Judicious choice of ψ 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 $\tilde{S}$ 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

$$\xi = (\tilde{S} - \hat{S})/\sigma^2(F) \rightarrow \chi_k^2\left(\sigma^{-2}(F)\gamma'(D^{22})^{-1}\gamma\right). \tag{4.1}$$

Substituting $\hat{\sigma}^2 = \hat{S}/(n-p)$ 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 χ^2 distribution under the null and local alternatives to the null. Obviously the distinction between χ^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 ξ is simply the ratio of the noncentrality parameters of the limiting χ^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 ARE $= 2/\pi \approx .64$ so the l_1 tests need (asymptotically) 36% more observations than the l_2 test. However λ 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_1 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_1 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_1 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

[1] ADICHIE, J. N.: "Rank Tests of Sub-hypotheses in the General Linear Regression," *Annals of Statistics*, 6(1978), 1012–26.
[2] AMEMIYA, T.: "Two Stage Least Absolute Deviations Estimators," *Econometrica*, 50(1982), 689–712.
[3] BASSETT, G. W., AND R. W. KOENKER: "The Asymptotic Distribution of the Least Absolute Error Estimator," *Journal of the American Statistical Association*, 73(1978), 618–622.
[4] ———: "Tests of Linear Hypotheses and l_1 Estimation," Bell Laboratories Economics Discussion Paper 191, 1981.
[5] ———: "An Empirical Quantile Function for Linear Models with iid Errors," *Journal of the American Statistical Association*, 77(1982), 407–415.
[6] BICKEL, P. J.: "Another Look at Robustness," *Scandanavian Journal of Statistics*, 3(1976), 145–168.
[7] BLATTBERG, R., AND T. SARGENT: "Regression with Non-Gaussian Stable Disturbances: Some Sampling Results," *Econometrica*, 39(1971), 501–11.
[8] COGGER, K. O.: "Statistical Inference in Minimum Absolute Deviation Regression," University of Kansas Discussion Paper, 1979.
[9] HAJEK, J., AND Z. SIDAK: *Theory of Rank Tests*. New York: Academic Press, 1967.
[10] HUBER, P. J.: "The Behavior of Maximum Likelihood Estimates under Nonstandard Conditions," *Proceedings of the 5th Berkeley Symposium on Probability and Statistics*, 1(1967), 221–233.
[11] ———: "Robust Regression: Asymptotics, Conjecture, and Monte-Carlo," *Annals of Statistics*, 1(1973), 799–821.
[12] JAECKEL, L. A.: "Estimating Regression Coefficients by Minimizing the Dispersion of the Residuals," *Annals of Mathematical Statistics*, 42(1972), 1328–38.
[13] KOENKER, R. W., AND G. W. BASSETT: "Regression Quantiles," *Econometrica*, 46(1978), 33–50.
[14] ———: "Robust Tests for Heteroscedasticity Based on Regression Quantiles," *Econometrica*, 50(1982), 43–62.
[15] PITMAN, E. J. G.: *Some Basic Theory for Statistical Inference*. London: Halsted, 1979.
[16] RUPPERT, D., AND R. J. CARROLL: "Trimmed Least Squares Estimation in the Linear Model," *Journal of the Amercan Statistical Association*, 75(1980), 828–38.
[17] SCHRADER, R. M., AND T. P. HETTMANSPERGER: "Robust Analysis of Variance Based on a Likelihood Ratio Criterion," *Biometrika*, 67(1980), 93–101.
[18] SILVEY, S. D.: "The Lagrangian Multiplier Test," *Annals of Mathematical Statistics*, 30(1959), 389–407.
[19] STIGLER, S. M.: "Francis Ysidro Edgeworth, Statistician," *Journal of the Royal Statistical Society A*, 141(1978), 287–322.
[20] TAYLOR, L. D.: "Estimation by Minimizing the Sum of Absolute Errors," in *Frontiers in Econometrics*, ed. by P. Zarembka. New York: Academic Press, Inc., 1974.