The Annals of Statistics
1985, Vol. 13, No. 4, 1625–1628

ON BOSCOVICH'S ESTIMATOR[1]

BY ROGER KOENKER AND GILBERT BASSETT

University of Illinois, Urbana and University of Illinois, Chicago

Boscovich's (1757) proposal to estimate the parameters of a linear model by minimizing the sum of absolute deviations subject to the constraint that the mean residual be zero is considered. The asymptotic theory of the estimator confirms a remark of Edgeworth who called it a "remarkable hybrid" between ℓ_1 and ℓ_2 methods.

1. Introduction. When Gauss discovered least squares in the twilight of the eighteenth century there were already several well-established proposals for estimating the bivariate linear models. See Plackett (1972) and Stigler (1981) for discussions of the least-squares priority debate between Gauss and Legendre. Perhaps the best known of these "precursors of least squares" is the proposal of Roger Boscovich in 1757 to minimize the sum of absolute residuals subject to the constraint that the mean residual is zero.

Boscovich's proposal attracted the attention of Thomas Simpson, a leading English eighteenth century analyst, who provided a partial solution to the problem of computing the Boscovich estimate. Stigler (1984) offers a fascinating glimpse of the Boscovich–Simpson interchange and describes an unpublished (1760) fragment in which Simpson develops his approach to the Boscovich problem. See Harter (1974) and Stigler (1973) for further background. Subsequently, in 1799 Laplace completely characterized the solution of the bivariate computational problem as a weighted median with weights $|x_i - \bar{x}|$ of the pairwise slopes $s_i = (y_i - \bar{y})/(x_i - \bar{x})$, $i = 1, 2, \ldots, n$. The term "weighted median" is apparently due to Edgeworth. Given an ordered sample $s_1, \ldots, s_n$, and associated weights, $w_1, \ldots, w_n$, the weighted median is simply s_m such that $m = \min\{j|\sum_{i=1}^{j}|w_i| \geq \sum_{i=1}^{n}|w_i|/2\}$.

After a long hiatus, Edgeworth (1887) revived the idea of the Boscovich estimator calling it a "remarkable hybrid between the *Method of Least Squares* and the *Method of Situation*," the latter being Laplace's rather vague term for ℓ_1 methods. In the next section, we develop an asymptotic theory of the Boscovich estimator for the general linear model and compare its asymptotic behavior with that of some of its better known but less venerable competitors.

2. Asymptotic theory of the Boscovich estimator. We will consider the classical linear model:

$$y_i = \sum_{j=1}^{p} x_{ij}\beta_j + u_i = x_i\beta + u_i, \tag{2.1}$$

Received March 1985; revised May 1985.
[1] Research supported by National Science Foundation Grant No. SES-8408567.
AMS 1980 *subject classifications.* Primary 62F12; secondary 62J07.
Key words and phrases. Least absolute error estimators, linear models, asymptotic theory.

where u_i: $i = 1, \ldots, n, \ldots$ are independent with common distribution function $F(\cdot)$, satisfying $F(1/2) = 0$, $Eu = \mu$, and having density f which is continuous and strictly positive at 0 and μ. We also need to assume that

$$\sigma^2 = E(u - \mu)^2 < \infty.$$

The design will be assumed to have an intercept (explicitly $x_{1j} = 1$ for all j) and to satisfy the usual condition:

$$\lim_{n \to \infty} \frac{1}{n} X'X \to D \tag{2.2}$$

for a positive definite matrix D. The objective function of the Boscovich estimator may be expressed in Lagrangian form as

$$\sum \left[|y_i - x_i b| + \lambda(y_i - x_i b) \right]. \tag{2.3}$$

Reparameterizing, set

$$\delta_0 = \sqrt{n}(\lambda - \lambda_0),$$
$$\delta_1 = \sqrt{n}(b - \beta - \mu e_1),$$

where $e_1' = (1, 0, \ldots, 0) \in R^p$ and $\lambda_0 = 2F(\mu) - 1$. Then (2.3) becomes

$$R(\delta) = \sum |u_i - x_i\delta_1/\sqrt{n} - \mu| + \left(\lambda_0 + \delta_0/\sqrt{n}\right)\left(u_i - x_i\delta_1/\sqrt{n} - \mu\right), \tag{2.4}$$

which we study employing the methods of Ruppert and Carroll (1980) and Jurečková (1977). The gradient of R is

$$g(\delta) = \nabla R(\delta) = \frac{1}{\sqrt{n}} \begin{pmatrix} \sum \left[u_i - x_i\delta_1/\sqrt{n} - \mu \right] \\ -\sum \left[\operatorname{sgn}\left(u_i - x_i\delta_1/\sqrt{n} - \mu\right) + \lambda_0 + \delta_0/\sqrt{n} \right] x_i \end{pmatrix}$$

and

$$Eg(\delta) = \frac{1}{\sqrt{n}} \begin{pmatrix} -\sum x_i\delta_1/\sqrt{n} \\ -\sum \left[1 - 2F\left(x_i\delta_1/\sqrt{n} + \mu\right) + \lambda_0 + \delta_0/\sqrt{n} \right] x_i \end{pmatrix}.$$

Using the methods of Ruppert and Carroll (1980) or Bickel (1975, Lemma 4.1), we have for fixed $M > 0$

$$\sup_{\|\delta\| < M} \|g(\delta) - g(0) - Eg(\delta) + Eg(0)\| = o_p(1).$$

It is then easily shown under our conditions on F that $Eg(\delta)$ has a unique root at $\hat{\delta} = 0$ which, following Jurečková (1977), implies that $\hat{\delta}$ solving (2.3) is $O_p(1)$ and hence $\hat{\beta} \to^p \beta - \mu e_1$ and $\hat{\lambda} \to^p \lambda_0$. Now expanding F around $\delta = 0$ and setting $\omega = 2f(\mu)$, yields

$$Eg(\delta) = \begin{pmatrix} 0 & -\bar{x} \\ -\bar{x}' & \omega D \end{pmatrix} \begin{pmatrix} \delta_0 \\ \delta_1 \end{pmatrix} + o(1)$$

and since $g(\hat{\delta}) = o_p(1)$ and $Eg(0) = 0$ we have that

$$\|Eg(\hat{\delta}_n) + g(0)\| = o_p(1).$$

Now

$$V(g(0)) = V\left[\frac{1}{\sqrt{n}}\begin{pmatrix}\sum(u_i-\mu)\\ -\sum[\operatorname{sgn}(u_i-\mu)+2F(\mu)-1]x_i\end{pmatrix}\right]$$
$$=\begin{bmatrix}\sigma^2 & G(\mu)\bar{x}\\ G(\mu)\bar{x}' & H(\mu)D\end{bmatrix},$$

where $G(\mu) = E|u-\mu|$ and $H(\mu) = 4(1-F(\mu))F(\mu)$. Condition (2.2) and the iid assumption on the errors imply that the summands of $g(0)$ satisfy the Lindeberg condition, and thus $\hat{\delta}$ converges in distribution to a $p+1$ variate normal distribution with mean vector 0 and covariance matrix

$$\begin{pmatrix}0 & -\bar{x}\\ -\bar{x}' & \omega D\end{pmatrix}^{-1}\begin{pmatrix}\sigma^2 & G\bar{x}\\ G\bar{x}' & HD\end{pmatrix}\begin{pmatrix}0 & -\bar{x}\\ -\bar{x}' & \omega D\end{pmatrix}^{-1}$$
$$=\begin{bmatrix}H+2\omega G+\omega^2\sigma^2 & (G+\omega\sigma^2)e_1'\\ (G+\omega\sigma^2)e_1 & \omega^{-2}H(D^{-1}-E_1)+\sigma^2E_1\end{bmatrix},$$

where E_1 denotes a $p\times p$ matrix with 1 in the (1,1) element and zeros elsewhere.

To interpret the result, consider first the symmetric case $\mu = 0$, so that

$$\omega=\omega_0=2f(0),$$
$$H(\mu)=4(1-F(0))F(0)=1,$$

and we have

$$\sqrt{n}(\hat{\beta}-\beta)\to N\left(0,\omega_0^{-2}(D^{-1}-E_1)+\sigma^2E_1\right).$$

Recall that the unconstrained ℓ_1 estimator under those conditions is asymptotically normal with covariance matrix $\omega_0^{-2}D^{-1}$. See Bassett and Koenker (1978) for details. Thus, the asymptotic theory of the Boscovich estimator, $\hat{\beta}$, in the symmetric case, is identical to that of the usual ℓ_1 estimator except that the asymptotic variance of the intercept is σ^2, the variance of F, instead of ω^{-2}, the asymptotic variance of the normalized sample median from F. This seems to vindicate Edgeworth's remark about the Boscovich estimator as a "remarkable hybrid" between ℓ_1 and ℓ_2 methods.

In asymmetric cases, $\hat{\beta}\to{}^p\beta-\mu e_1$ so that the regression surface is shifted to the conditional expectation of y rather than its conditional median as for the unconstrained ℓ_1 estimator. Secondly, the mean of the Lagrangian is nonzero in the asymmetric case; thus a diagnostic test for symmetry based on the Lagrange multiplier is possible. The covariance matrix of $\sqrt{n}(\hat{\beta}-\beta-\mu e_1)$ is fundamentally the same as in the simple ℓ_1 case except that the scale parameter on the covariance matrix of the slope parameters is $(2f(\mu))^{-2}4(1-F(\mu))F(\mu)$ instead of $(2f(0))^{-2}$.

A second, and perhaps more promising application of the Boscovich estimator, is to prediction problems for linear models. A possible objection to ℓ_1 methods for prediction is their failure to predict the *conditional expectation* of the response variable in asymmetric error situations. While a reasonable argument might be made for conditional median predictions, strict adherence to quadratic loss, for example, dictates prediction of conditional expectations. Nevertheless, to protect one's self against the consequences of heavy-tailed errors, one might prefer an estimation method which achieved median precision for the slope parameters, while sacrificing this precision for the intercept to remove the median bias effect. This is, in effect, what the Boscovich estimator achieves. It is easy to construct examples for which it is preferred to both its ℓ_1 *and* ℓ_2 competitors. Take $D = I_2$, $x' = (1,1)$ so that $x'D^{-1}x = 2$. We need $F(\mu)(1 - F(\mu))/f(\mu^2) < \sigma^2(F)$. This is satisfied for the Pareto distribution with parameter $\alpha = 3$, for which $F(\mu) = 1 - \mu^{-\alpha} = 19/27$, $f(\mu) = 3\mu^{-4} = 16/27$, $\mu = 3/2$, $\sigma^2 = 1$.

Finally, we might add that nothing we have done depends crucially on the form of the Boscovich estimator and could with appropriate modifications of regularity conditions be extended to problems of the general form

$$\min_{b \in \mathbb{R}^p} \sum \rho(y_i - x_i b) - \lambda\psi(y_i - x_i b)$$

for ρ and ψ corresponding to any plausible M estimators.

REFERENCES

BASSETT, G. and KOENKER, R. (1978). The asymptotic theory of the least absolute error estimator. *J. Amer. Statist. Assoc.* **73**, 618–622.

BICKEL, P. J. (1975). One-step Huber estimates in the linear model. *J. Amer. Statist. Assoc.* **70**, 428–434.

EDGEWORTH, F. Y. (1887). On observations relating to several quantities. *Hermathena* **6** 279–285.

HARTER, H. L. (1974). The method of least squares and some alternatives, Part I, *Internat. Statist. Rev.* **42** 147–74.

JUREČKOVÁ, J. (1977). Asymptotic relations of M-estimates and R-estimates in linear regression models. *Ann. Statist.* **5**, 464–472.

PLACKETT, R. L. (1972). The discovery of the method of least squares. *Biometrika* **59** 239–251.

RUPPERT, D. and CARROLL, R. (1980). Trimmed least squares estimation in the linear model. *J. Amer. Statist. Assoc.* **75** 828–838.

STIGLER, S. (1973). Laplace, Fisher, and the discovery of the concept of sufficiency. *Biometrika* **60** 439–445.

STIGLER, S. (1981). Gauss and the invention of least squares. *Ann. Statist.* **9** 465–474.

STIGLER, S. (1984). Boscovich, Simpson and a 1760 manuscript note on fitting a linear relation. *Biometrika* **71** 615–20.

DEPARTMENT OF ECONOMICS
UNIVERSITY OF ILLINOIS
CHAMPAIGN, ILLINOIS 61820

DEPARTMENT OF ECONOMICS
UNIVERSITY OF ILLINOIS
CHICAGO, ILLINOIS 60680