BEBR

FACULTY WORKING
PAPER NO. 1142

On Boscovich's Estimator

Roger Koenker
Gilbert Bassett

College of Commerce and Business Administration
Bureau of Economic and Business Research
University of Illinois, Urbana-Champaign

BEBR

FACULTY WORKING PAPER NO. 1142

College of Commerce and Business Administration

University of Illinois at Urbana-Champaign

April, 1985

On Boscovich's Estimator

Roger Koenker, Professor
Department of Economics

Gilbert Bassett
Univerity of Illinois at Chicago-Circle

On Boscovich's Estimator

By Roger Koenker and Gilbert Bassett

University of Illinois, Urbana Champaign and 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.

AMS 1980 subject classification. Primary 62F12 Secondary 62J07

Key words and phrases. Least absolute error estimators, linear models, asymptotic theory.

On Boscovich's Estimator

Roger Koenker
and
Gilbert Bassett

1. Introduction

When Gauss discovered least-squares in the twilight of the 18th century[1] there were already several well-established proposals for estimating bivariate linear models. 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 18th century analyst, who provided a partial solution to the problem of computing the Boscovich estimate.[2] Subsequently, in 1799 Laplace completed characterized the solution of the bivariate computational problem as a weighted median[3] with weights $|x_i - \bar{x}|$ of the pairwise slopes $s_i = (y_i - \bar{y})/(x_i - \bar{x})$, $i = 1,2,\ldots,n$.

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. The concluding section suggests some possible applications of the theory to diagnostic testing and prediction problems.

2. Asymptotic Theory of the Boscovich Estimator

We will consider the classical linear model

$$(2.1) \qquad y_i = \sum_{j=1}^{p} x_{ij}\beta_j + u_i = x_i\beta + u_i$$

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

$$(2.2) \qquad \lim_{n\to\infty} \frac{1}{n} X'X \to D$$

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

$$(2.3) \qquad \Sigma[\,|y_i - x_ib| + \lambda(y_i - x_ib)].$$

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,

$$(2.4) \qquad R(\delta) = \Sigma|u_i - x_i\delta_1/\sqrt{n} - \mu| + (\lambda_0 + \delta_0/\sqrt{n})(u_i - x_i\delta_1/\sqrt{n} - \mu)$$

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

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

and,

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

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

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

And using the methods of Ruppert and Carroll (1980) we have for fixed $M > 0$

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

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,

$$V(g(0)) = V\left[\frac{1}{\sqrt{n}}\begin{pmatrix}\Sigma(u_i - \mu) \\ \\ -\Sigma[\operatorname{sgn}(u_i - \mu) + 2F(\mu) - 1]x_i\end{pmatrix}\right]$$

$$= \begin{bmatrix}\sigma^2 & G(\mu)\overline{x} \\ \\ G(\mu)\overline{x}' & H(\mu)D\end{bmatrix}$$

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 implies that the summands of $g(\delta)$ 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

$$\begin{pmatrix}0 & -\overline{x} \\ -\overline{x} & \omega D\end{pmatrix}^{-1}\begin{pmatrix}\sigma_2 & G\overline{x} \\ G\overline{x}' & HD\end{pmatrix}\begin{pmatrix}0 & -\overline{x} \\ -\overline{x}' & \omega D\end{pmatrix}^{-1}$$

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

where E_1 denotes a pxp matrix with 1 in the (1,1)-element and zeros elsewhere.

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

$$\omega = \omega_0 = 2f(0)$$

$$H(\mu) = 4(1 - F(0))F(0) = 1$$

and we have

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

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 ususal 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 vindidate Edgeworth's remark about the Boscovich estimator as a "remarkable hybrid" between ℓ_1 and ℓ_2 methods.

In asymmetric cases $\hat{\beta} \rightarrow^{p} \beta - \mu e_1$ so 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 non-zero in the asymmetric case; thus a 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 $(2f(0))^{-2}$.

3. Applications

There are two applications of the foregoing theory which we would like to discuss briefly. The first is a test of symmetry of the error distribution in linear models which might be used as a diagnostic for ℓ_1 regression. The second is an application to prediction problems.

Consider the null hypothesis, H_0: $\mu = 0$, which given the intercept in the linear model is equivalent to H_0': $\mu = F^{-1}(1/2)$ and represents a salient necessary condition for symmetry of the errors. If we consider local alternatives of the form

$$H_A: \quad \mu = \mu_0/\sqrt{n}$$

then a test of H_0 is available using the test statistic

$$T = \sqrt{n}\hat{\lambda}/\sqrt{Q} \xrightarrow{D} N(\nu,1)$$

where $Q = H + 2\omega G + \omega^2\sigma^2$ and the parameter, ν, takes the form, $\mu_0/\sqrt{Q}$. Unfortunately, while finding a consistent estimator of Q is quite easy--one could use residuals from the ℓ_1-regression, or the empirical distribution function proposed in Bassett and Koenker (1982)--a reasonable estimate for small to modest size samples seems problematic.

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 <u>conditional expectation</u> 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.[4]

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

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

for ρ and ψ corresponding to any plausible m-estimators.

Footnotes

[1]We begin on a note of controversy. See Plackett (1972) and Stigler (1981) for discussions of the least-squares priority debate between Gauss and Legendre.

[2]Stigler (1984) offers a fascinating glimpse of the Boscovich-Simpon 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.

[3]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 $m = \min\{j \mid \sum_{i=1}^{j} |w_i| \geq \sum_{i=1}^{n} |w_i|/2\}$.

[4]Take $D = I_2$, $x' = (1,1)$ so $x'D^{-1}x = 2$. 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 $F(u) = 1 - u^{-\alpha} = 19/27$, $f(\mu) = 3u^{-4} = 16/27$, $\mu = 3/2$, $\sigma^2 = 1$.

References

Bassett, G. and Koenker, R. (1978) The Asymptotic Theory of the Least Absolute Error Estimator, Journal of the American Statistical Association, 73, 618-622.

Bassett, G. and Koenker, R. (1982) An Empirical Quantile Function for Linear Models with iid errors, Journal of the American Statistical Association, 77, 407-415.

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, International Statistical Review, 42, 147-174.

Jurecková, J. (1977) Asymptotic Relations of M-Estimates and R-Estimates in Linear Regression Models, Annals of Statistics, 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, Journal of the American Statistical Association, 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, Annals of Statistics, 465-474.

Stigler, S. (1984) Boscovich, Simpson and a 1760 Manuscript Note on Fitting a Linear Relation, Biometrika, 71, 615-20.