GLOBAL SENSITIVITY RESULTS FOR GENERALIZED LEAST SQUARES ESTIMATES

By

Edward E. Leamer
University of California, Los Angeles

Discussion Paper Number 296
June 1983

GLOBAL SENSITIVITY RESULTS FOR GENERALIZED LEAST SQUARES ESTIMATES

by Edward E. Leamer

Department of Economics

University of California, Los Angeles

Los Angeles, California 90024

ABSTRACT

The covariance matrix for the residuals of a regression process is written as the identity matrix plus a matrix V. The matrix V is bounded from above, and the corresponding set of generalized least-squares estimates is identified. The extreme estimates in this set are functions of the usual t-statistics; in particular the number $((T-k)/8)^{1/2}/|t|$ measures the influence of reweighting extreme observations, where T-k is the degrees-of-freedom, t is the t-statistic, and where the weights on observations are allowed to vary by a factor of at most two.

Acknowledgement: Support from NSF grant SES82-07532 is acknowledged.

July 1983

GLOBAL SENSITIVITY RESULTS FOR GENERALIZED LEAST-SQUARES ESTIMATES

by Edward E. Leamer

UCLA

Outlying observations are often discarded or otherwise reweighted when a regression equation is estimated. Both formal and informal techniques for dealing with outliers can give fragile results in the sense that minor changes in the assumptions can imply major changes in the inferences. In order to study the sensitivity of estimates to choice of procedure, I consider here the class of generalized least squares estimators $\hat{\beta}(\Sigma) = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}Y$ which are maximum likelihood estimators of β if the vector Y is normally distributed with mean $X\beta$ and known covariance matrix Σ. The covariance matrix is written as $\Sigma = I + V$ where I is the covariance matrix of the "usual" errors and V is the covariance matrix of the "unusual" errors. If V is proportional to the identity matrix, then $\hat{\beta}(\Sigma)$ is just the least-squares estimator, but otherwise, by choice of V, we can reweight some observations or adjust for correlation among the errors.

This paper reports results on the global sensitivity of $\hat{\beta}(I+V)$ to choice of V. A global sensitivity analysis characterizes the set S of values for $\hat{\beta}(I+V)$ corresponding to a set F of values for V:

$$S = \{\hat{\beta} \mid \hat{\beta} = (X'(I+V)^{-1}X)^{-1}X'(I+V)^{-1}Y,\ V \in F\}$$

where F is a class of covariance matrices of analytical interest. Here, we consider two such classes:

$$F_1 = \{V | V = \text{diag}\{v_1, v_2, \ldots, v_T\};\ v_i < v_i^*,\ i = 1, \ldots, T\}$$

$$F_2 = \{V | \text{symmetric},\ 0 < V < V^*,\ V^* \text{ sym. pos def.}\}$$

where $V < V^*$ means $V^* - V$ is positive semi-definite.

The set of estimates S_1 corresponding to the family of diagonal covariance matrices F_1 is a set of weighted regressions with weights on individual observations bounded between $(1+v_i^*)^{-1}$ and 1. If $v_i^* = 1$, for example, weights on individual observations can vary by at most a factor of two. Although this may be considered a reasonably wide set of weights, many "robust" estimators proposed by Huber (1964) and others allow the weights on extreme observations to go to zero, and can produce estimates outside of S_1. For studies of global sensitivity with other classes of weight matrices see Gilstein and Leamer (1982).

An important problem with the set S_1 is that it is rather difficult to characterize numerically. However, the set S_2 is much more tractible since it is necessarily elliptical. Moreover, if V^* is proportional to the identity matrix, then the boundary of the set can be described in terms of the traditional t-values. Thus a t-statistic is a measure of the resistance to reweighting. In addition to yielding this simple result, the class F_2 of covariance matrices often makes more sense than the class F_1. The statement $0 < V < V^*$ means that the variance of a linear combination of residuals, say $d'u$, where d is a vector of constants and u is the vector of residuals, is bounded from above and below $d'd < \text{Var}(d'u) < (d'd+d'V^*d)$. If $V^* = I$, for example, all weighted regressions with weights between 1 and 2 are included in the set S_2, but also included in it are various generalized least squares estimators with Σ close to a diagonal matrix $I < \Sigma < 2I$.

This seems actually to make the set S_2 more interesting than the set S_1 since in most settings the assumption that the covariance matrix Σ is diagonal is not compelling but the assumption that the covariance matrix is nearly diagonal is likely to be acceptable. In any case, the set S_2 contains the set S_1, and since it implies computationally simpler bounds for estimates, it may be of practical interest even when S_1 is preferred.

1. Results

The first result reveals that $\hat{\beta}(I+V)$ can be written as a least-squares regression with the data Y corrected for outliers.

Theorem 1

The generalized least-squares estimator with $\Sigma = I + V$ can be written as

$$\hat{\beta}(\Sigma) = (X'(I+V)^{-1}X)^{-1}X'(I+V)^{-1}Y$$

$$= (X'X)^{-1}X'(Y-\hat{\gamma}) = b - (X'X)^{-1}X'\hat{\gamma} \tag{1}$$

where
$$\hat{\gamma} = (M+V^{-1})^{-1}MY \tag{2}$$

with
$$b = (X'X)^{-1}X'Y,$$

and
$$M = I + X(X'X)^{-1}X' \tag{3}$$

Proof: Conceptually, the easiest way to get to this result is to write the regression process as $Y = X\beta + I\gamma + u$ where γ is a $T \times 1$ vector of constants with a normal prior distribution with mean vector zero and covariance $\sigma^2 V$. The posterior mean of (β,γ) is in the usual matrix-weighted average form (e.g., Leamer (1978, p. 78))

$$E\left(\left|\begin{matrix}\beta\\ \gamma\end{matrix}\right| \,\middle|\, Y,\sigma^2\right) = \left(\sigma^{-2}\left|\begin{matrix}X'X & X\\ X & I\end{matrix}\right| + \sigma^{-2}\left|\begin{matrix}0 & 0\\ 0 & V^{-1}\end{matrix}\right|\right)^{-1}\sigma^{-2}\left|\begin{matrix}X'Y\\ Y\end{matrix}\right|$$

$$= \left|\begin{matrix}X'X & X'\\ X & I+V^{-1}\end{matrix}\right|^{-1}\left|\begin{matrix}X'Y\\ Y\end{matrix}\right| \qquad (4)$$

Equation (1) is just the conditional mean of β given Y and γ and Equation (2) is the mean of γ given Y. Equation (2) follows from the partitional inverse rule applied to (4)

$$\left|\begin{matrix}X'X & X'\\ X & I+V^{-1}\end{matrix}\right| = \left|\begin{matrix}\cdot & \cdot\\ -C^{-1}X(X'X)^{-1} & C^{-1}\end{matrix}\right|$$

where $C = (I+V^{-1}-X(X'X)^{-1}X') = (M+V^{-1})$. Thus $E(\gamma|Y,\sigma^2)$ $= C^{-1}(-X(X'X)^{-1}X'Y+Y) = (M+V^{-1})^{-1}MY$.

This proof may leave non-Bayesians uncomfortable and an alternative can be built on the matrix result

$$(A+BCB')^{-1} = A^{-1} - A^{-1}B(B'A^{-1}B+C^{-1})^{-1}B'A^{-1}B'A^{-1}. \qquad (5)$$

From this it follows that

$$(I+V)^{-1} = I - (I+V^{-1})^{-1} \quad ,$$

$$X'(I+V)^{-1}X = X'X - X'(I+V^{-1})^{-1}X$$

$$(X'(I+V)^{-1}X)^{-1} = (X'X)^{-1} + (X'X)^{-1}X'(M+V^{-1})^{-1}X(X'X)^{-1}.$$

Then

$$\hat{\beta} = (X'(I+V)^{-1}X)^{-1}X'(I+V)^{-1}Y$$

$$= [(X'X)^{-1} + (X'X)^{-1}X'(M+V^{-1})^{-1}X(X'X)^{-1}]$$

$$[X'Y - X'(I+V^{-1})^{-1}Y]$$

$$= (X'X)^{-1}X'Y + (X'X)^{-1}X'[(M+V^{-1})^{-1}X(X'X)^{-1}X'$$

$$- (I+V^{-1})^{-1} - (M+V^{-1})^{-1}X(X'X)^{-1}X'(I+V^{-1})^{-1}]Y$$

The term in brackets can be written as

$$(M+V^{-1})^{-1}[X(X'X)^{-1}X'(I+V^{-1}) - (M+V^{-1}) - X(X'X)^{-1}X'](I+V^{-1})^{-1}$$

$$= (M+V^{-1})^{-1}[X(X'X)^{-1}X'(I+V^{-1}) - (I+V^{-1})](I+V^{-1})^{-1}$$

$$= -(M+V^{-1})^{-1}M.$$

Inserting this into the formula above produces

$$\hat{\beta} = (X'X)^{-1}X'Y - (X'X)^{-1}X'(M+V^{-1})^{-1}MY$$

which implies Equations (1) and (2).

Equation (1) is the least-squares regression using data Y corrected for outliers. Equation (2) selects the outlier correction depending on X, V and MY, the vector of least-squares residuals. The matrix-weighted average form (2) has been studied by Leamer and Chamberlain (1976), Leamer (1978) and Leamer (1982). If V is a diagonal matrix, Leamer and Chamberlain (1976) or Leamer (1978, p. 153) can be used to produce:

<u>Theorem 2</u>

The generalized least-squares estimator $\hat{\beta}(\Sigma)$ with $V = \text{diag}\{v_1, v_2, \ldots, v_T\}$ can be written as

$$\hat{\beta}(\Sigma) = \sum_I w_I \, b_I$$

where I indexes the 2^T subsets of the first T integers, and selects the observations which are included, b_I is least-squares with observations $i \in I$ included but observations $i \notin I$ excluded, and

$$w_I = \left(\prod_{i \in I} v_i^{-1}\right) \left| I - X_{\bar{I}}(X'X)^{-1}X'_{\bar{I}} \right| / \left| M + V^{-1} \right|$$

$$\sum_I w_I = 1,$$

where $X_{\bar{I}}$ is the matrix of observations formed from excluded elements of X. (Note w_I as defined is zero if I has fewer than k elements.)

Theorem 2 describes the weighted regression estimators as a weighted average of least-squares estimates based on subsets of the data. This theorem suggests that the set of estimates of β corresponding to the set of matrices $V = \text{diag}\{v_1, v_2, \ldots, v_T\}$ with $0 < v_i < v_i^*$ will be difficult to characterize numerically, since estimates based on any subset of the data would have to be considered. If attention is restricted to one-at-a-time deletions with $v_i = 0$ for $i \neq j$, then $\hat{\beta}(\Sigma)$ is a weighted average of least-squares, b, and least-squares with observation j omitted, b_{I_j}:

$$\hat{\beta}(\Sigma) = w_o b + w_1 b_{I_j}$$

with weights

$$w_0 \; \alpha \quad v_j^{-1}$$

$$w_1 \; \alpha \quad 1 - x_j'(X'X)^{-1}x_j.$$

where x_j' is row j of X. As v_j varies from 0 to v_j^*, $\hat{\beta}(\Sigma)$ then sweeps out a line segment from b to $b + v_j^*(1-x_j'(X'X)^{-1}x_j)b_{I_j} /(1+v_j^{*-1}(1-x_j'(X'X)^{-1}x_j))$.

The other set S_2 is much more easily described. By application of theorem (3) in Leamer (1982) we obtain

<u>Theorem 3</u>

Given $\hat{\gamma} = (M+V^{-1})^{-1}MY$ with $V < V^*$ with V^* symmetric positive definite, then $\hat{\gamma}$ lies in the ellipsoid

$$(\hat{\gamma}-f)'H(\hat{\gamma}-f) < c \tag{6}$$

where

$$H = V^{*-1} + M$$

$$f = (V^{*-1}+M)^{-1}MY/2$$

$$c = Y'M(V^{*-1}+M)^{-1}MY/4.$$

Conversely, for any value of $\hat{\gamma}$ in ellipsoid (6), there exists a V such that $V \leq V^*$ and $\hat{\gamma} = (M+V^{-1})^{-1}\ MY$.

A corollary of theorem 3 which can be used to bound estimates of linear combinations of parameters $\psi'\beta$ is:

Theorem 4

The extreme estimates $\psi'\hat{\beta} = \psi'b - \psi'(X'X)^{-1}X'\hat{\gamma}$ with $\hat{\gamma}$ constrained to ellipsoid (6) are

$$\psi'b - \psi'(X'X)^{-1}X'(V^{*-1}+M)^{-1}MY/2$$

$$\pm(\psi'(X'X)^{-1}X'(V^{*-1}+M)^{-1}X(X'X)^{-1}\psi c)^{1/2}$$

A special case of theorem 4 produces a surprising result. Let $V^* = qI$. Then

$$V^{*-1} + M = (1+q^{-1})I - X(X'X)^{-1}X'$$

$$(V^{*-1}+M)^{-1} = (1+q^{-1})^{-1}(I+qX(X'X)^{-1}X').$$

Then, since $X'M = 0$,

$$f = (1+q^{-1})^{-1}MY/2$$

$$c = (Y'MY)/4(1+q^{-1});$$

and the extreme estimates of $\psi'\beta$ become

$$\psi'b \pm [(1+q)\ \psi'(X'X)^{-1}\psi Y'MY]^{1/2}/2(1+q^{-1})$$

$$= \psi'b(1 \pm t^{-1}q(T-k)^{1/2}/2(1+q)^{1/2})$$

where t is the t value for testing $\psi'\beta = 0$

$$t = \psi'b/(\psi'(X'X)^{-1}\psi\ Y'MY/(T-k))^{1/2}.$$

What is surprising about this result is that the usual t value tells you all you need to know about the sensitivity of an estimate to the choice of weights. A coefficient with a large t-value is relatively insensitive to reweighting of observations to deal with outliers. One choice for q would be one, in which case some observations are allowed to have twice the weight of others. Then the maximum percentage change in the estimate that could be induced by reweighting is $((T-k)/8)^{1/2}/|t|$. Note that constant resistance to reweighting as sample size increases requires an increasing t-statistic at a rate equal to the square root of the degrees of freedom. For example, the sign of the estimate is insensitive to reweighting if the t-statistic exceeds the critical value in the following table.

Degrees of Freedom = T-k	Critical t = $((T-k)/8)^{1/2}$
50	2.5
100	3.5
1000	11.2

The traditional reason for requiring the critical t-statistic to increase with sample size is to have a sensible tradeoff between Type I and Type II error. A Bayesian treatment (e.g., Leameer (1978, p. 114) has the critical t increasing like $(T-k)^{1/2}(T^{1/T} - 1)^{1/2}$, which grows less rapidly than $(T-k)^{1/2}$.

It may also be noted that by choice of q, the interval of estimates can be made to include any value whatever. This is a special case of the general result that the set of generalized least-squares estimates $\{\hat{\beta}(\Sigma) \mid \Sigma \text{ p.d.}\}$ is the whole space, Leamer (1981).

In some cases interest may focus on uncovering the outlying observations, more than on correcting the estimates for the effects of outliers. The outlier correction γ indicates adjustments to the data Y that are required before applying the usual least-squares formula. Estimates of γ are given by Equation (2) and bounds by Theorem 3. The extreme values of $\psi'\hat{\gamma}$ on the ellipsoid (6) are

$$\psi'\hat{\gamma} = \psi'f \pm (\psi'H^{-1}\psi\ c)^{1/2}.$$

If ψ selects the ith residual, and if $V^* = qI$, then a bound for $\hat{\gamma}_i$ is

$$\hat{\gamma}_i = (1+q^{-1})^{-1}e_i/2 \pm \sqrt{(1+q^{-1})^{-1}(1+qh_i)ESS/4(1+q^{-1})}$$

$$= (1+q^{-1})^{-1}2^{-1}(e_i \pm \sqrt{(1+qh_i)ESS}) \qquad (7)$$

where h_i is the ith diagonal element of $X(X'X)^{-1}X$, e_i is the ith residual and ESS is the error sum-of-squares, $ESS = Y'MY = e'e$. This compares with the t-value for a dummy variable which selects the ith observation, referred to by Belsley, Kuh and Welsch (1980, p. 20) as the Studentized residual

$$t_i = e_i/s_i(1-h_i)^{1/2} \tag{8}$$

where s_i^2 is the estimate of the residual variance if the ith observation is omitted, $(T-k-1)s_i^2 = ESS - e_i^2/(1-h_i) = ESS_i$. If $q=1$, Equation (7) can be written as

$$\hat{\gamma}_i = (e_i/4)(1 \pm t_i^{-1}(T-k-1)^{1/2}\, g_i^{1/2}$$

where $g_i = ESS(1+h_i)/ESS_i(1-h_i)$. Thus the extreme estimate of the outlier adjustment factor is the least-squares residual divided by four, times a factor that depends on the Studentized residual in basically the same way as the estimates of coefficients depend on their t values, the biggest difference being that the interval (7) is a function of $1 + h_i$ whereas the Studentized residual uses $1 - h_i$.

REFERENCES

Belsley, D., E., Kuh and R. Welsch (1980), Regression Diagnostics, New York: John Wiley.

Gilstein, E. and E. Leamer (1982), "Robust Sets of Regression Estimates," Econometrica, 51 (March), 321-333.

Huber, Peter J. (1964), "Robust Estimation of a Location Parameter," Annals of Math. Stat., 35, 73-101.

Leamer, E. E. (1982), "Sets of Posterior Means with Bounded Variance Priors," Econometrica, 50, No. 3 (May), 725-736.

Leamer, E. E. (1978), Specification Searches: Ad Hoc Inference With Nonexperimental Data, New York: John Wiley.

Leamer, E. E. (1981), "Techniques for Estimation with Incomplete Assumptions," Proceedings of IEEE Meeting on Decision and Control, San Diego, December.

Leamer, E. E. and G. Chamberlain (1976), "A Bayesian Interpretation of Pretesting," J. Royal Statistical Society, 13, 38, 85-94.