VALLEY REGRESSION: BIASED ESTIMATION FOR ORTHOGONAL PROBLEMS

by

Edward E. Leamer

Discussion Paper Number [redacted] 102
10/77

Valley Regression: Biased Estimation for Orthogonal Problems

by Edward E. Leamer*

The publication of Hoerl and Kennard's [1] paper on ridge regression has sparked a boomlet of activity among statisticians and their clients [1], [2], [5], [6], [7], [8]. For the standard linear regression model, $\underset{\sim}{Y} = \underset{\sim}{X}\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$ where $\underset{\sim}{X}(n\times p)$, $\underset{\sim}{Y}(n\times 1)$ are observed and $\underset{\sim}{\varepsilon}$ is normal with $E\underset{\sim}{\varepsilon} = 0$ and $E\underset{\sim}{\varepsilon}\underset{\sim}{\varepsilon}' = \sigma^2\underset{\sim}{I}$, the ridge estimator of $\underset{\sim}{\beta}$ is $\hat{\underset{\sim}{\beta}}^r = (\underset{\sim}{X}'\underset{\sim}{X} + k\underset{\sim}{I})^{-1}\underset{\sim}{X}'\underset{\sim}{Y}$. This procedure is most often sold as a cure for the "multi-collinearity problem" which plagues users of the traditional least-squares estimator. It is not recommended when $\underset{\sim}{X}'\underset{\sim}{X}$ is a diagonal matrix.

But whether the $\underset{\sim}{X}'\underset{\sim}{X}$ matrix is diagonal or not depends on the choice of parameterization. Whereas ridge regression might be recommended for estimating $\underset{\sim}{\beta}$, it would not be recommended for estimating $\underset{\sim}{\theta} = \underset{\sim}{F}\underset{\sim}{\beta}$ if $\underset{\sim}{F}'^{-1}\underset{\sim}{X}'\underset{\sim}{X}\underset{\sim}{F}^{-1}$ is diagonal. Thus the excitement of ridge regression is limited to those who are lucky enough to choose the right parameterization.

There <u>is</u> something that you can do if you are unlucky enough to be plagued by the lack of collinearity. The <u>off-diagonal</u> elements of the $\underset{\sim}{X}'\underset{\sim}{X}$ matrix may be augmented by a common factor. For reasons that may be made clear below, the off-diagonal elements of the $\underset{\sim}{X}'\underset{\sim}{X}$ matrix will be augmented positively only when the estimate vector $\underset{\sim}{b} = (\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}'\underset{\sim}{Y}$ is a positive vector. Otherwise the estimate is first rotated into the first orthant, then translated by the valley regression formula and last rotated back into its original orthant. In terms of a formula, this can be expressed as follows.

---

*Department of Economics, UCLA. Conversations with R.W. Clower and financial support from NSF grant SOC 76-08863 are gratefully acknowledged.

Let

$s_i = 1 \times \text{sign}(b_i)$; b_i = least squares estimate of β_i

$\underset{\sim}{S}$ = p x p diagonal matrix, diag $\{s_1, s_2, \ldots, s_p\}$

Then the <u>valley estimator</u> for a given k is

$$\hat{\underset{\sim}{\beta}}^{v}(k) = \underset{\sim}{S}[\underset{\sim}{S}'\underset{\sim}{X}'\underset{\sim}{X}\underset{\sim}{S} + k(\underset{\sim}{1}\underset{\sim}{1}'-\underset{\sim}{I})]^{-1}\underset{\sim}{S}'\underset{\sim}{X}'\underset{\sim}{Y}, \quad k \geq 0,$$

where $\underset{\sim}{1}$ is a vector of ones.

Given a data set $(\underset{\sim}{Y}, \underset{\sim}{X})$, as k varies, this formula will generate a curve of estimates known as the <u>stream bed</u>.

The valley estimator is recommended for precisely the same reason as the ridge estimator. Namely, it is possible to prove the powerful theorem that there exists a k such that the mean squared error of the valley estimator is less than least squares. This result is proved in section 1 of this paper. Concluding comments may be found in section 2.

1. The Existence Result

Given that Y is normal with mean $X\beta$ and variance $\sigma^2 I$, and given orthonormal data $X'X = I$, there exists a k such that the mean squared error of the valley estimator is less than the mean squared error of least squares.

The valley estimator is

$$\begin{aligned}\hat{\beta}^{v}(k) &= S(S'S + k(11'-I))^{-1}S'X'Y \\ &= S(I + k(11'-I))^{-1}S'b\end{aligned}$$

where $b = (X'X)^{-1}X'Y = X'Y$. The theorem states that there exists a k such that

$$E[(\hat{\beta}^{v}(k) - \beta)'(\hat{\beta}^{v}(k) - \beta)] < E[(\hat{\beta}^{v}(0) - \beta)'(\hat{\beta}^{v}(0) - \beta)],$$

Proof:

The valley estimator may be written as

$$\hat{\beta}^{v} = SC[C'C + kC'(11'-I)C]^{-1}C'Sb$$

for any invertible matrix C. This takes a convenient form when

$$C = \begin{bmatrix} 1 & 1 & 1 & 1 & \cdot & 1 \\ 1 & -1 & 1 & 1 & \cdot & 1 \\ 1 & 0 & -2 & 1 & \cdot & 1 \\ 1 & 0 & 0 & -3 & \cdot & 1 \\ \vdots & & & \vdots & & \vdots \\ 1 & 0 & 0 & 0 & \cdot & -(p-1) \end{bmatrix}$$

Then

$$C'C = \text{diag}\ \{p, 2, 6, 12, \ldots, p(p-1)\}$$

$$C'(11'-I)C = \text{diag}\ \{p^2-p, -2, -6, -12, \ldots, -p(p-1)\}$$

$$C'Sb = \{\sum_i |b_i|, |b_1| - |b_2|, |b_1| + |b_2| - 2|b_3|, \ldots, \sum_{i<p} |b_i| - (p-1)|b_p|\}$$

It is easiest to compute the estimator of the last coefficient but by symmetry the same formula applies for all coefficients.

$$\hat{\beta}_p^v = s_p\{\sum_i |b_i|(p + k[p^2 - p])^{-1} - (p - 1) \left(\sum_{i<p} |b_i| - (p - 1)|b_p|\right)$$

$$(1 - k)^{-1} p^{-1}(p - 1)^{-1}\}$$

$$= s_p\{(1 - k)\Sigma|b_i| - [1 + k(p - 1)][\Sigma|b_i| - p|b_p|]\}/(1 - k)p\ (1 + k(p - i))$$

$$= s_p\{(1 - k - 1 - kp + k)\Sigma|b_i| + p[1 + kp - k]|b_p|\}/(1 - k)p\ (1 + k(p - 1))$$

$$= s_p\ \{-k\Sigma|b_i| + [1 + kp - k]|b_p|\}/(1 - k)(1 + k(p - 1))$$

$$= s_p\ \{-k \sum_{i\neq p} |b_i| + (1 + kp - 2k)|b_p|\}/(1 - k)(1 + k(p - 1)).$$

The computation of the mean squared error of $\hat{\beta}_p(k)$ makesuse of a result on the moment of a truncated normal distribution. The least squares estimator b_i is normal with mean β_i and variance σ^2, and is independent of b_j, $j\neq i$. The density function of this normal distribution will be indicated by $n_i(\beta_i)$ and the cumulative function by $N_i(\beta_i)$. Then

$$E(|b_i|) = E(b_i|b_i \geq 0)\ P(b_i \geq 0) - E(b_i|b_i \leq 0)\ P(b_i \leq 0)$$

$$= [\beta_i + (n_i(0) - n_i(\infty))(1 - N_i(0))^{-1}\sigma^2][1 - N_i(0)]$$

$$- [\beta_i + (n_i(-\infty) - n_i(0)(N_i(0) - N_i(-\infty))^{-1}\sigma^2][N_i(0)]$$

$$= \beta_i(1 - 2N_i(0)) + 2\sigma^2 n_i(0)$$

The mean squared error of $\hat{\beta}_p^v$ is

$$E(\hat{\beta}_p^v - \beta_p)^2 = (1 - k)^{-2}(1 + k(p - 1))^{-2}E\{b_p(1 + kp - 2k)$$

$$-ks_p \sum_{i\neq p} |b_i| - \beta_p(1 - k)(1 + k(p - 1))\}^2$$

$$= (1 - k)^{-2}(1 + k(p - 1))^{-2}E\{(b_p - \beta_p)(1 + kp - 2k)$$

$$-ks_p \sum_{i\neq p} |b_i| - \beta_p k^2(1 - p)\}^2$$

$$= (1 - k)^{-2}(1 + k(p - 1))^{-2}\{\sigma^2(1 + kp - 2k)^2$$
$$-2k(1 + kp - 2k)E[s_p(b_p - \beta_p) \sum_{i \neq p} |b_i|]$$
$$+k^2 E[s_p \sum_{i \neq p} |b_i| - \beta_p k(1 - p)]^2$$

In order to show that there is a $k > 0$ such that $MSE[\hat{\beta}_p^v(k)] < MSE[\hat{\beta}_p^v(0)]$ it is necessary only to show that the derivative of this function evaluated at $k = 0$ is negative. Write the function as u/v with derivative $(vdu - udv)/v^2$ and note that

$$v(0) = 1, \qquad u(0) = \sigma^2$$
$$dv(0) = 2(p - 2)$$
$$du(0) = 2(p - 2)\sigma^2 - 2E[s_p(b_p - \beta_p) \sum_{i \neq p} |b_i|]$$

Thus $vdu - udv = -2E[s_p(b_p - \beta_p) \sum_{i \neq p} |b_i|]$.

But $E[s_p(b_p - \beta_p)] = E(|b_p|) - \beta_p E(s_p)$

$$= \beta_p(1 - 2N_p(0)) + 2\sigma^2 n_p(0) - \beta_p[1 - N_p(0) - N_p(0)]$$
$$= 2\sigma^2 n_p(0) > 0,$$

and the derivative at zero is therefore negative.

2. <u>Concluding Comments</u>

(a) There is a Bayesian justification for this!

(b) Epicurus to Menoeceus: "Pleasure [not truth] is the end and aim of life."

References

[1] Dempster, A.P., Schatzoff, M. and Wermuth, N., "A Simulation Study of Alternatives to Ordinary Least Squares," Journal of the American Statistical Association, 72 (March 1977), 77-97.

[2] Goldstein, M. and Smith, A.F.M., "Ridge-Type Estimations for Regression Analysis," Journal of the Royal Statistical Society, Sec B, 36, 2 (1974), 284-91.

[3] Hoerl, A.E. and Kennard, R.W., "Ridge Regression: Biased Estimation for Nonorthogonal Problems," Technometrics, 12 (February 1970), 55-67.

[4] Leamer, E.E., "Regression Selection Strategies and Revealed Priors," mimeo, 1975.

[5] McDonald, G.C. and Galarneau, D.I., "A Monte Carlo Evaluation of Some Ridge-Type Estimators," Journal of the American Statistical Association, 70 (June 1975), 407-16.

[6] Rolph, J.E., "Choosing Shrinkage Estimators for Regression Problems," Communications in Statistics - Theory and Methods, A5(9) (1976), 789-801.

[7] Swindel, B.F., "Instability of Regression Coefficients Illustrated," The American Statistician, 28 (May 1974), 63-5.

[8] Vinod, H.D., "Application of New Ridge Regression Methods to a Study of Bell System Scale Economies," Journal of the American Statistical Association, 71 (December 1976), 835-841.