Influencial Observations in Ridge Regression

by

Robert L. Obenchain

We augment traditional trace displays with dynamic graphics that can reveal dramatic changes in influence/outliers/leverages of observations resulting from shrinkage of coefficients along generalized ridge paths. First, a Visual Re-Regression (VRR) plot provides new twists on interpretation of the familiar display of predicted responses versus their observed values. And a second, linked Leverage/Outlier plot displays contours of constant Cook influence as hyperbolas. An implementation of this approach in XLisp-Stat is also illustrated.

Key Words: generalized ridge shrinkage, influential observations, outlying responses, high leverage regressor combinations, Cook's distance.

[0]Robert L. Obenchain is a Research Scientist in the Statistical and Mathematical Sciences department of Lilly Research Laboratories, Eli Lilly Corporate Center 2233, Indianapolis, IN 46285-0001, ochain@lilly.com.

1 Introduction

We extend well known least squares measures of the size of residuals [Ellenberg(1973), Beckman and Trussell(1974)], leverage ratios, and overall influence [Cook(1977)] by tailoring them specifically to generalized ridge estimation, Hoerl and Kennard(1970). We also discuss highly informative ways to view these ridge influence statistics using dynamic graphics.

Our graphical approach does not attempt to automatically "adjust" ridge fits for influential observations, as do the methods of Holland(1973) or Askin and Montgomery(1980). Rather, our graphics simply show how each observation influences each ridge fit and how this influence can change with the form and extent of shrinkage. On the other hand, these displays certainly may motivate regression practitioners to reweight and/or "setting aside" observations before performing subsequent analyses.

2 Visual Re-Regression

It is relatively easy to spot response outliers and predictor leverage points in "simple" regression because only one response variable and only one predictor variable is involved in the model. In this case, one can "see" everything by simply drawing a scatter plot with, say, the response (y) values along the vertical axis and the predictor (x) values along the horizontal axis. Regression methods fit a line or curve to this scatter that represents the locus of conditional expected values for y over the given

range of values for x. An outlying response then corresponds to a relatively large vertical deviation in an observed y value from the fitted line or curve. And the points relatively near the left-hand and right-hand ends of this x range are the ones with relatively high leverage in determining this best fit.

The goal of Visual Re-Regression (VRR) is to use this same sort of display to evaluate the influence of individual observations on multiple regression fits. Specifically, VRR retains the response variable to define vertical coordinates for observations but replaces the horizontal coordinates by predicted values from the ridge fit that is to be evaluated. Because regression predictions are linear combinations of the given regressor X coordinates, VRR rightly interprets the horizontal axis on this plot as defining a single, **composite** predictor (x) variable.

Next, VRR superimposes a least-squares (or robust) fit onto this predicted versus observed response plot. And, just as in the $p = 1$ dimensional case, outliers and (to a lesser extent) leverage points then tend to become visually "obvious."

On the other hand, loss-of-information occurs when $p > 1$ predictor variables are represented by any single ($p = 1$) composite predictor. In fact, to counteract this loss, we recommend that our VRR plot be augmented with a second (linked) plot of full-fledged ($p > 1$ dimensional) measures of the extent to which each observation is an outlier, a leverage point or both.

3 Composite Predictor Coordinates

Consider the standard model for multiple linear regression in which the conditional expectation of a $n \times 1$ vector of values for a nonconstant response variable, y, given the values of a $n \times p$ matrix of nonconstant regressor variables, X, is written as $E(y|X) = 1\mu + X\beta$, where 1 is a $n \times 1$ vector of all ones, μ is the unknown intercept, and β is the $p \times 1$ vector of unknown regression coefficients. To simplify notation, we assume that all variables will be "centered" by subtracting off column means, so that $1'y = 0$ and $1'X = 0'$. In particular, the intercept term is then implicitly $\mu = \bar{y} - \bar{x}'\beta$, and the regression can be rewritten as $E(y|X) = X\beta$.

We also assume that the coordinates for each variable will be rescaled by being divided by their sample standard deviation. This rescaling implies that $y'y = (n - 1)$ and $Diag(X'X) = (n - 1)I$; the sample sum-of-squares for each centered and rescaled variable is 1 fewer than the number of observations, n. Finally, we assume that $rank(X) = p$; after centering and rescaling, the p columns of X are linearly independent, where $1 < p \leq n - 1$.

Any generalized ridge estimator, b^*, of the regression coefficient β vector [equation (8) of the Appendix] defines a corresponding $n \times 1$ vector of fitted/predicted response values, $y^* = Xb^*$. Note that this y^* vector can almost always be rewritten as

$$y^* = b^{=}x^{=}, \tag{1}$$

where $b^=$ is a scalar and $x^=$ is a $n \times 1$ vector with mean 0 and sum-of-squares $x^{=\prime}x^= = (n-1)$. The only exception occurs when $x^= = 0$ due to shrinkage all of the way to $b^* = 0$; the $x^=$ vector cannot be rescaled to sum-of-squares $(n-1)$ in this extreme shrinkage case. Our use of $=$ signs for superscripts in the above notation will be motivated in the next section, below.

In drawing a VRR plot, we interpret the $x^=$ vector as yielding a univariate $(p = 1)$ set of composite regressor coordinates derived from the original $(p > 1)$ regressors, X. After all, $y^* = Xb^*$ is a linear combination of the columns of X and, thus, contains derived coordinates along a certain direction within the $p-$dimensional column space of X.

Note from equation (1) that $b^=$ is the **sample standard deviation** of predicted response values

$$b^= = \sqrt{y^{*\prime}y^*/(n-1)}. \tag{2}$$

But this $b^=$ is also the **slope of the ridge fit** on the VRR plot. When b^* is a shrinkage estimate of β, $b^=$ is usually less than the slope of the OLS line for the regression of y onto $x^=$ defined by

$$b^{VRR} = x^{=\prime}y/x^{=\prime}x^= = y^{*\prime}y/\sqrt{(n-1)(y'^*y^*)}. \tag{3}$$

Note that b^{VRR} is the **correlation** between y and $x^=$ (and between y and y^*.) It turns out that $b^= \leq b^{VRR}$;see equation (14) of the Appendix. But $b^=$ will be much less than b^{VRR} only when the extent of shrinkage in generalized ridge regression is

excessive. Intuitive explanations of why the ridge slope, $b^{=}$, cannot exceed the VRR slope, b^{VRR}, are the topic of the next section.

4 Orientation and Length of the Ridge b* Vector

A key characteristic of the composite $x^{=}$ coordinates is that they depend only upon the **orientation** of the b^{*} vector in $p-$dimensional space (i.e. only upon the **relative magnitudes** of the elements of b^{*}). Specifically, $x^{=}$ does not depend upon the **length** of the b^{*} vector. After all, it is clear from $x^{=} = Xb^{*}/b^{=}$ and equation (2), that $x^{=}$ and b^{VRR} would be unchanged if b^{*} were to be replaced by $f \times b^{*}$ for any positive scalar factor, f. (But y^{*} and $b^{=}$ would change to $f \times y^{*}$ and $f \times b^{=}$, respectively, under this rescaling.) In other words, the length of the b^{*} vector determines only the ridge slope, $b^{=}$. Thus the superscripts of $=$ remind us that $x^{=}$ is determined by the direction in $p-$space **parallel** to b^{*}, while $b^{=}$ is determined by the length of b^{*} in this parallel direction.

The above observations also help us understand why the vector of VRR predictions, $y^{VRR} = b^{VRR}x^{=}$, tends to be longer than the vector of ridge predictions, $y^{*} = b^{=}x^{=}$. After all, ridge methods deliberately shrink b^{*}, which reduces $b^{=}$. But b^{VRR} corresponds to a least squares estimator on the **Feasible Ellipsoid** of Leamer(1978). Specifically, the maximum likelihood estimate of β constrained to be strictly parallel to b^{*} is $(b^{VRR}/b^{=}) \times b^{*}$. [Technically, Leamer's Theorem 1, page 582,

speaks only of least squares estimates constrained to being orthogonal to a given matrix. But being parallel to a nonzero b^* is, of course, equivalent to being orthogonal to $I - b^*b^{*\prime}/(b^{*\prime}b^*)$.]

The facet of ridge regression that I, personally, have always found most interesting is that the relative magnitudes of the fitted b^* coefficients usually tend to change during the "shrinkage" process. In other words, the shrinkage path usually starts out at the overall least squares solution, b^o of (7), and usually terminates where the coefficient for every nonconstant regressor has been reduced to zero. But ridge shrinkage paths rarely follow a straight line between these two extremes. In fact, a straight shrinkage path generally occurs only in the special case where the path "shape" parameter of Goldstein and Smith(1974) is $q = +1$ [see equation (9) of the Appendix.] Obenchain(1975,1996a) describes how to estimate the MSE risk optimal path $q-$shape using normal-theory maximum likelihood.

In "ordinary" ridge regression, Hoerl and Kennard(1970), the shrinkage path $q-$shape is zero by definition. This shrinkage path is straight only in the extremely rare case where all of the eigenvalues of the predictor $X'X$ matrix happen to be equal; all $q-$shape paths are then equivalent [again, see equation (9) of the Appendix.] The conceptual distinction between $b^=$ and b^{VRR} is, perhaps, most straight forward when the path $q-$shape$= 0$. Every ridge b^* is then more likely to be β than is any other vector of the **same or shorter length**. However, some strictly **longer vector par-**

allel to b^* will be more likely to be the true β whenever the given b^* is shorter than the overall least squares solution. [This property is indeed obvious when the path is straight and, thus, shrinkage is "uniform."]

5 Questions about Alternative VRR Fits and Plots

Q: Rather than computing $b^=$ or b^{VRR}, why not simply superimpose the line with slope $b = 1$ in VRR? After all, the ideal situation would be $y = y^* = x^=$, and we are trying to "visualize" departures from this ideal case, right?

A: The problem with this potentially over-simplified approach is that $b^= \leq b^{VRR} \leq R$, where R is the multiple correlation between y and the columns of X [again, see equation (14) of the Appendix.] Thus the line with slope $b = 1$ can be quite inappropriate in problems where the familiar R^2 statistic is not large. On the other hand, when R^2 is large (greater than .9, say), the $b = 1$ line probably would be adequate for visualization. Still, when software that computes both $b^=$ and b^{VRR} is readily available, why not at least see how different these two slopes are? In fact, if you have the necessary software, why not also ask what a truly **robust fit** [Dallal(1991), Stata(1995), etc.] of y to $x^=$ would look like?

Q: Wouldn't it be better to base VRR on the plot of ridge residuals versus ridge predictions?

A: If your graphical display doesn't have very good resolution and/or the number,

n, of observations is very large, you probably would be able to see many more details by plotting residuals (rather than response observations) along the vertical axis on your VRR plot. But a horizontal reference line, $b = 0$, on this plot could be "too flat" for the same reason that the $b = 1$ line can be "too steep" when plotting observed responses versus ridge predictions! Besides, my VRR proposal allows you to visualize the **relative** sizes of three "additive" components: (1) the linear ridge fit (quantified by $b^{=}$), (2) the linear lack-of-ridge-fit (quantified by $b^{VRR} - b^{=}$), and (3) the non-linear lack-of-fit (quantified by deviations from the line of slope b^{VRR}.) Basing VRR on a plot using ridge residuals would essentially discard all information about that first component.

6 More on Ridge Outliers, Leverages and Influence

To compensate for loss of information implied by trying to use only the coordinates for a single ($p = 1$ dimensional) composite regressor to represent the joint effects of $p > 1$ predictor variables, we propose linking the VRR plot discussed above to a second plot that displays ridge leverages and residuals, respectively.

6.1 Ridge Leverages

The leverage ratio (predictive variance divided by residual variance) of the $i-$th observation in the ridge fit of y onto X can be written as

$$\Lambda_i^2(\Delta) = \frac{1/n + h_i'\Delta^2 h_i}{(n-1)/n - h_i'(2\Delta - \Delta^2)h_i}, \tag{4}$$

where h_i' is the i-th row of the semi-orthogonal matrix, H, of standardized principal-axis predictor coordinates [see equation (7) of the Appendix.] These ridge leverage ratios, $\Lambda_i^2(\Delta)$, are plotted along the horizontal axis on our second plot.

Note in equation (4) that ridge shrinkage (starting at $\Delta = I$ and ending at $\Delta = 0$) **systematically reduces the leverages of all observations**. However, ridge shrinkage does not necessarily reduce relatively large leverage ratios any faster than the already smaller ones.

6.2 Ridge Residuals

Generalizing the results of Ellenberg(1973) or Beckman and Trussell(1974), the $i-$th studentized ridge residual, $t_i(\Delta)$, and the $i-$th standardized ridge residual, $r_i(\Delta)$, can be written as

$$t_i(\Delta) = r_i(\Delta)\sqrt{\frac{(n-p-2)}{[n-p-1-r_i^2(\Delta)]}} \tag{5}$$

and

$$r_i(\Delta) = \frac{y_i - y_i^*(\Delta)}{s\sqrt{(n-1)/n - h_i'(2\Delta - \Delta^2)h_i}}\ , \tag{6}$$

where h_i' is again the i-th row of the semi-orthogonal matrix, H, of standardized principal-axis predictor coordinates (see Appendix) and $s^2 = (y'y)[1-R^2]/(n-p-1)$ is the OLS residual mean square for error. Since $t_i(\Delta)$ and $r_i(\Delta)$ are monotonically related, an outlying response value is indicated when either its studentized or standardized residual is large (positive or negative) relative to those of other observations.

Since the numerical signs of outlying residuals will be obvious from the VRR plot, there is no real need to retain signs on our second plot. In fact, ignoring signs here helps us make visual "size" comparisons between residuals that differ in sign. Anyway, an advantage of using squared, standardized residuals, $r_i^2(\Delta)$, along our second vertical axis (rather than either absolute values or studentized residuals) is that it will then be easy to visualize Cook's overall influence, as explained next.

6.3 Ridge Influences on the Leverage/Outlier Plot

The most widely used measure of overall influence for the $i-$th observation in least squares fitting is probably Cook's distance, $D_i = r_i^2 \Lambda_i^2/(p+1)$, Cook(1977). When our second plot displays ridge leverages, $\Lambda_i^2(\Delta)$, versus ridge squared, standardized residuals, $r_i^2(\Delta)$, contours of constant Cook's distance will be **hyperbolas**. After all, each ridge $D_i(\Delta)$ is then proportional to the product of its horizontal and vertical coordinates.

7 Longley Data in XLisp-Stat

Figure 1 is a **coefficient trace** for the infamous Longley(1967) dataset, a well known benchmark for ill-conditioning in multiple regression. This model regresses average yearly U.S. employment onto gross national product, gnp price deflation index, unemployment percentage, size of the armed forces, total U.S. population, and year; thus there are $p = 6$ predictor variables and $n = 16$ observations for the years 1947 through1962. With all seven variables centered and rescaled, the 2−parameter generalized ridge estimator most likely to yield overall, minimum MSE risk, Obenchain(1975,1996a), is on the $q-$shape$= -1.5$ path at the rather extreme shrinkage extent of $m = 4.00$, where $m = 6$ is total shrinkage to zero. A much more conservative approach here would be to utilize the "2/p-ths rule-of-thumb" of Obenchain(1978) to limit shrinkage to the $m \leq (2/6) \times 4 = 1.33$ range that is much more likely to yield a ridge estimator that dominates OLS in every (matrix valued) MSE sense. Although not reproduced here, trace displays of estimated **scaled MSE risk**, **excess eigenvalues** (OLS minus ridge) and of the **inferior direction**, Obenchain(1978), confirm that the $q-$shape$= -1.5$ ridge estimates at $m = 1.00$ and $m = 1.25$ do indeed appear to have more desirable MSE risk characteristics than those near $m = 4.00$.

(Insert Figure 1 here.)

The above trace was created in XLispStat, Tierney(1990), by an implementation of shrinkage/ridge methodology, Walter(1994) and Obenchain(1996b). On the linked

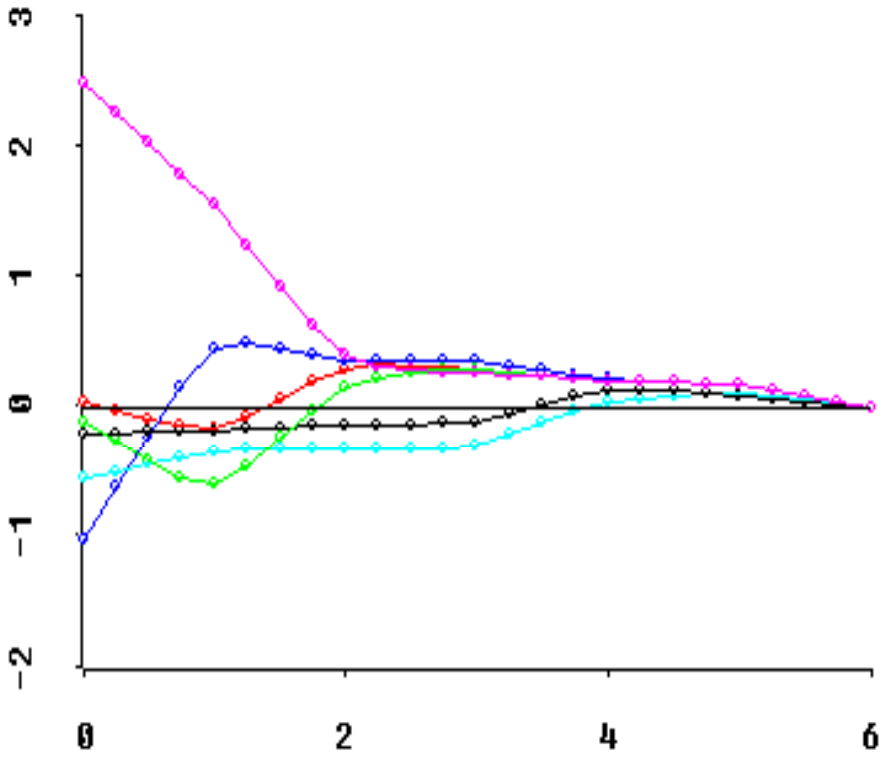

Figure 1: Coefficient trace for the Longley(1967) data and the $q-$shape$= -1.5$ path that is most likely to lead to overall minimum MSE risk. In fact, the extent of shrinkage along this path that is most likely to be MSE optimal is $m = 4$, Obenchain(1996a).

VRR and Leverage/Outlier plots, a slider controls the extent of shrinkage, m, allowing these plots to change dynamically as m varies over the range from $m = 0$ to $m = rank(X)$. The default, initial position of this slider is at $m = 0$, and the corresponding VRR and Leverage/Outlier plots for least squares estimates are displayed in Figures 2 and 3. A second slider controls the level of Cook influence displayed using the hyperbolic contour, as in Figure 3.

Note that year 5=1951 provides a good illustration of information loss about leverage in VRR plots. Figure 3 shows that only year 16=1962 has higher leverage (in $p = 6$ dimensional predictor space) on the least squares solution than does 5=1951. However, four other years (1947, 1949, 1948 and 1950) are represented as being further than 1951 towards the left-hand extreme of the 1-dimensional least squares composite predictor space displayed in Figure 2.

The extent of ridge shrinkage along the $q-$shape$=-1.5$ path can be increased, using the $m-$slider, until it becomes excessive, as in Figures 4 and 5, below, for $m = 5$.

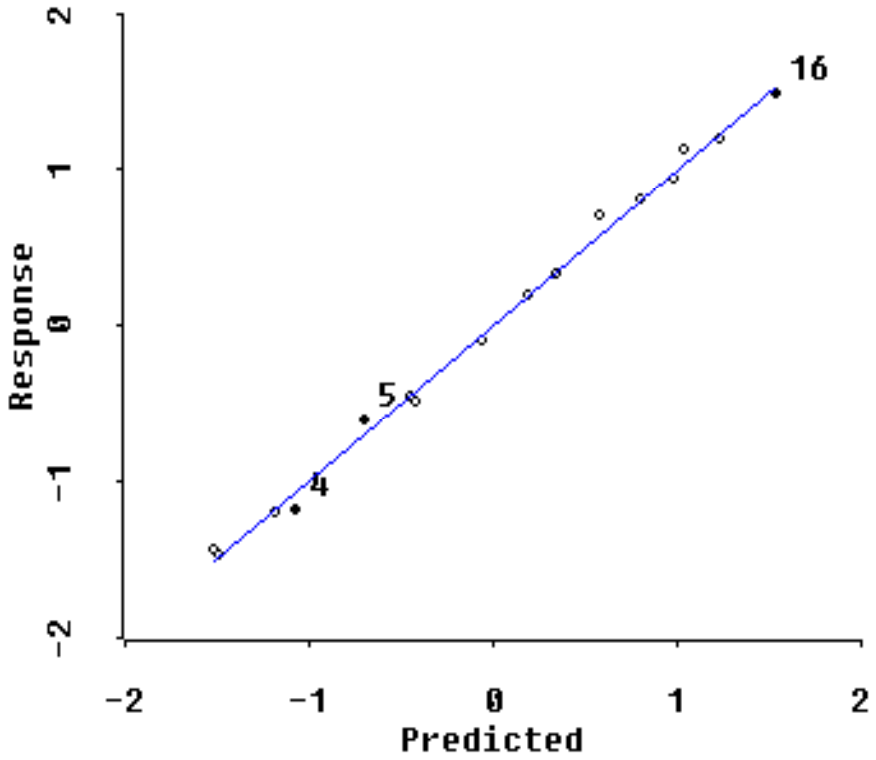

Figure 2: VRR plot for least squares ($m = 0$) on the Longley(1967) data; $b^{=} = b^{VRR} = R = 0.9977$. The highlighted points (5=1951, 16=1962 and 4=1950) are the three years with largest Cook(1977) influence.

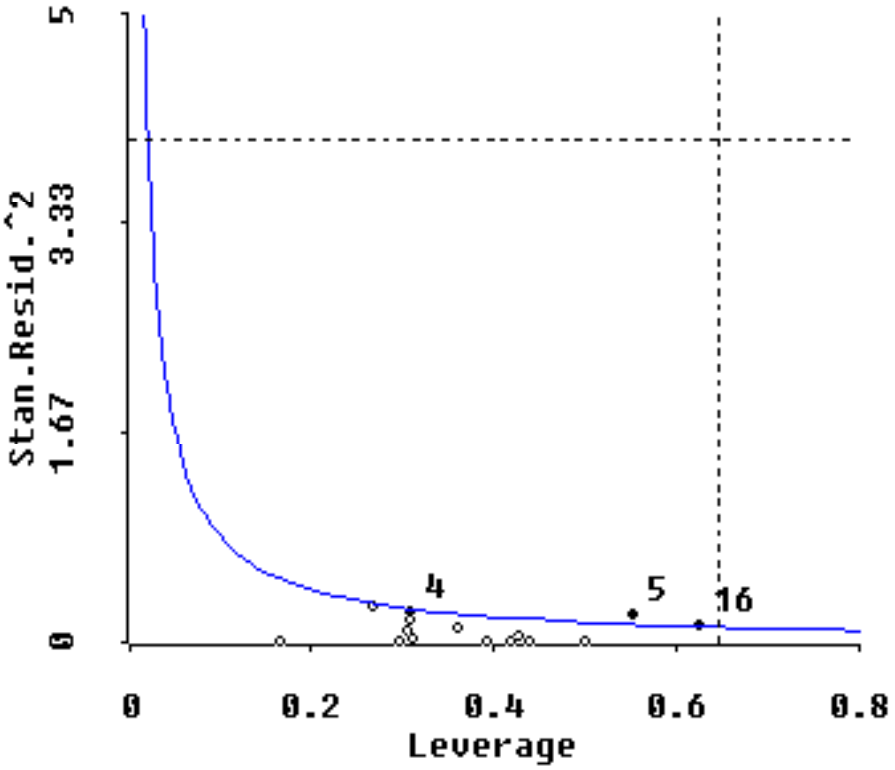

Figure 3: Leverage/Outlier plot for least squares ($m = 0$) on the Longley(1967) data, with a hyperbolic contour of constant Cook(1977) influence. Note how the majority of influence at $m = 0$ comes from large leverages rather than from large residuals for this dataset.

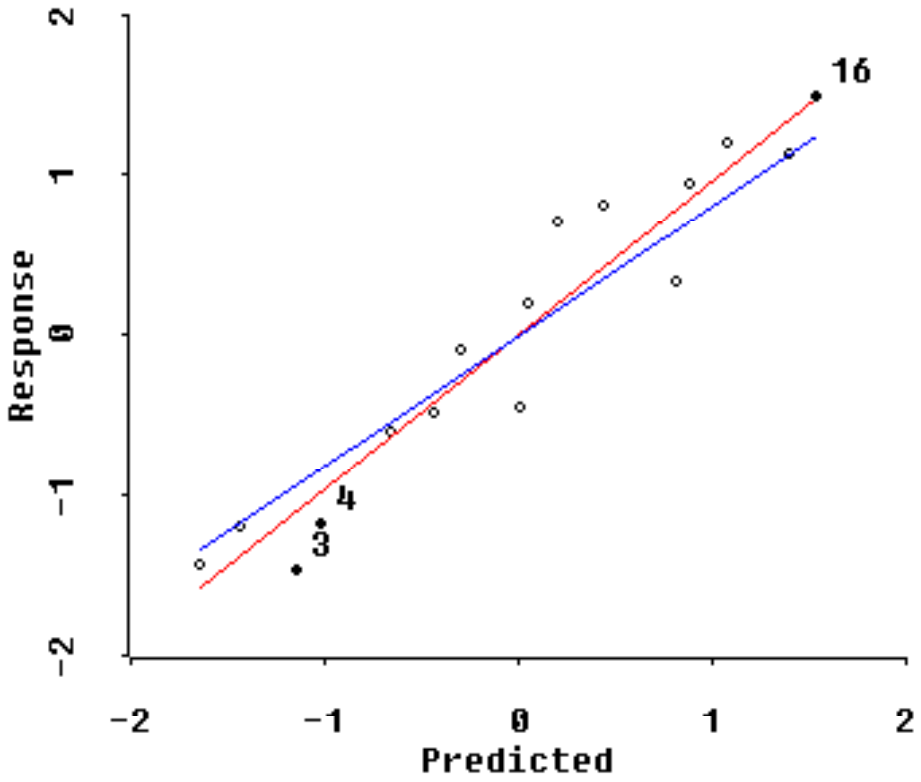

Figure 4: VRR plot for shrinkage extent $m = 5.00$ along the $q-$shape$=-1.5$ path on the Longley(1967) data. Shrinkage has become excessive here, and $b^{=} = 0.8084$ is distinctly smaller than $b^{VRR} = 0.9588$. The highlighted points (3=1949, 16=1962 and 4=1950) are the three years with largest Cook influence at $m = 5.00$.

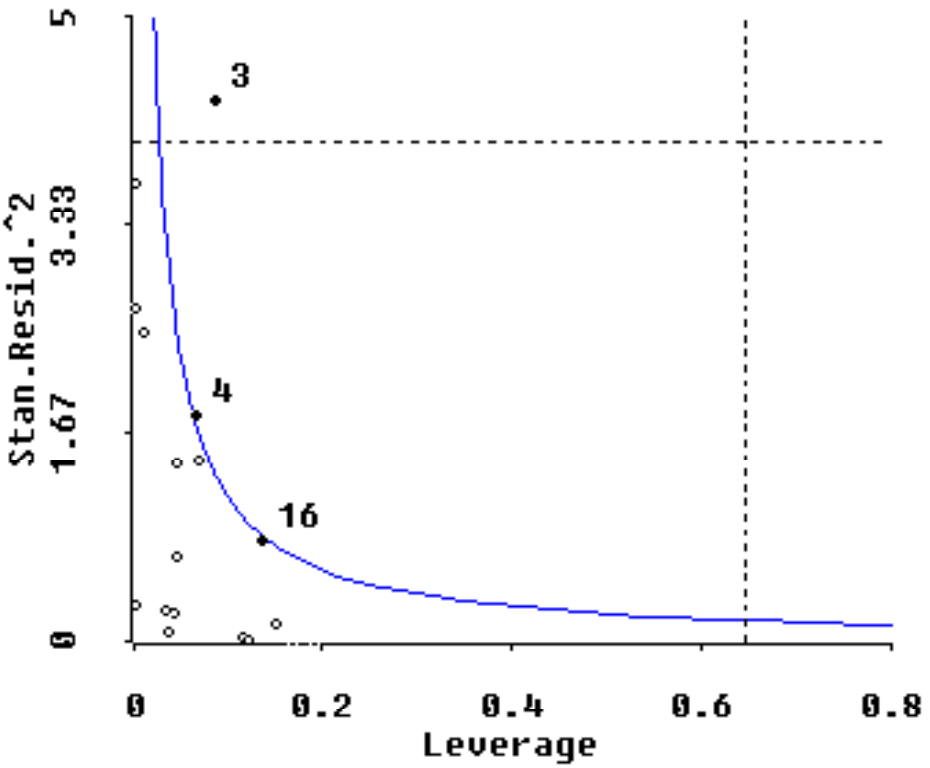

Figure 5: Leverage/Outlier plot for shrinkage extent $m = 5.00$ along the $q-$shape$=-1.5$ path, with a superimposed hyperbola of constant Cook influence. For this excessive extent of shrinkage, note how leverages have decreased dramatically, at least relative to the least squares solution depicted in Figure 3. In fact, overall influence here at $m = 5.00$ comes much more from residuals than from leverages.

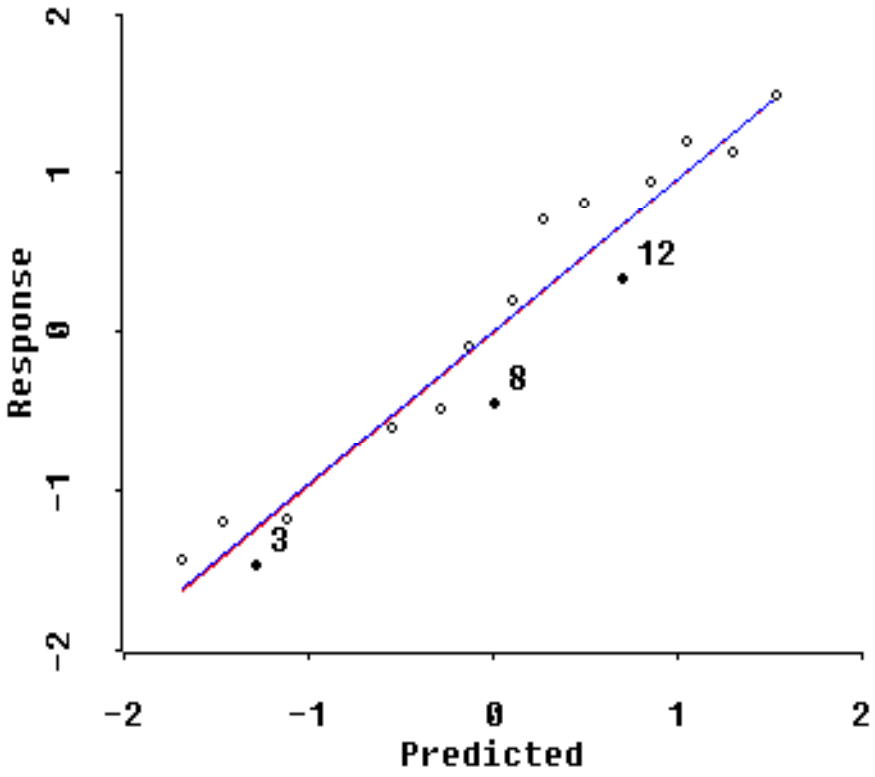

Figure 6: VRR plot for shrinkage extent $m = 4.00$ along the $q-$shape$= -1.5$ path on the Longley(1967) data. The ridge and VRR fits are almost indistinguishable ($b^{=} = 0.9593$ and $b^{VRR} = 0.9695$), so shrinkage is not excessive here. Two of the three points with largest Cook influence here at $m = 4.00$ (namely, 12=1958 and 8=1954) are distinct from the most influential points at either $m = 0.00$ or $m = 5.00$ (namely 5=1951, 3=1949, 4=1950 and 16=1962.)

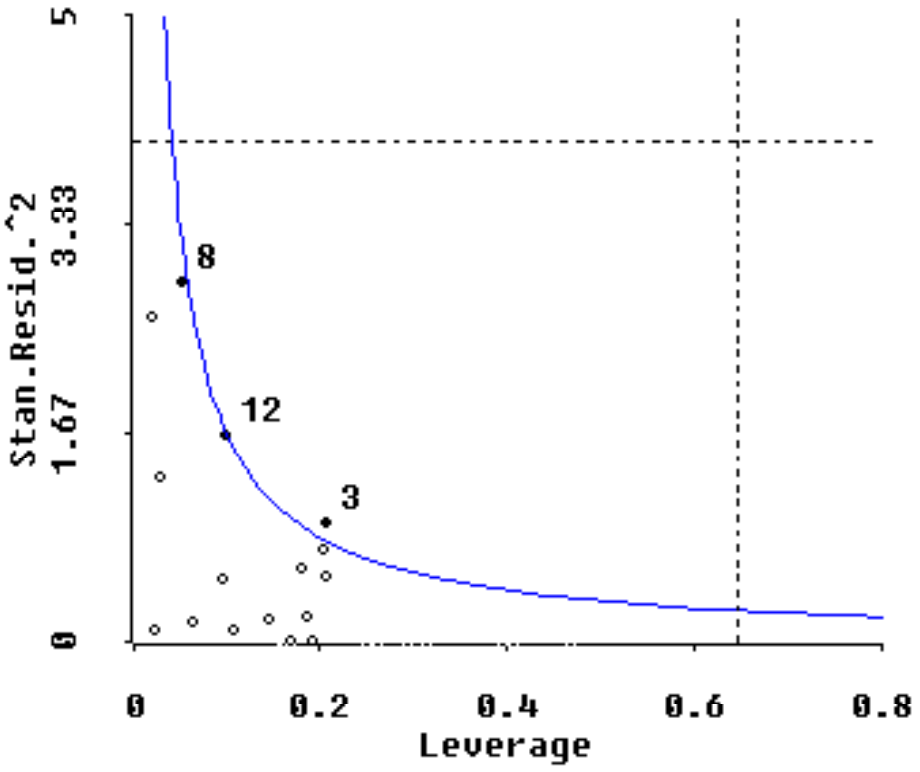

Figure 7: Leverage/Outlier plot at $m = 4.00$ along the $q-$shape$= -1.5$ path on the Longley(1967) data, with superimposed hyperbola of constant Cook influence. Note that the majority of influence here is coming from residuals (as at $m = 5.00$) rather than from leverages (as at $m = 0.00$.)

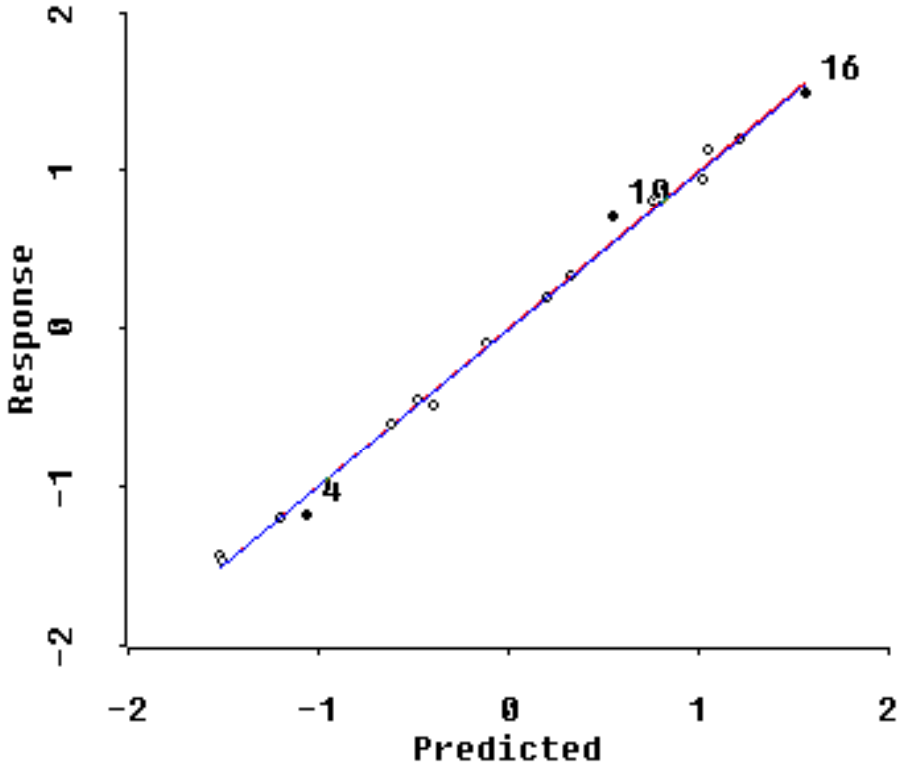

Figure 8: VRR plot for shrinkage extent $m = 1.00$ along the $q-$shape$= -1.5$ path for the Longley(1967) data. The ridge $b^{=} = 0.9964$ and $b^{VRR} = 0.9971$ slopes are very close numerically. The highlighted points (16=1962,10=1956and 4=1950) are the three years with largest Cook influence here at $m = 1.00$.

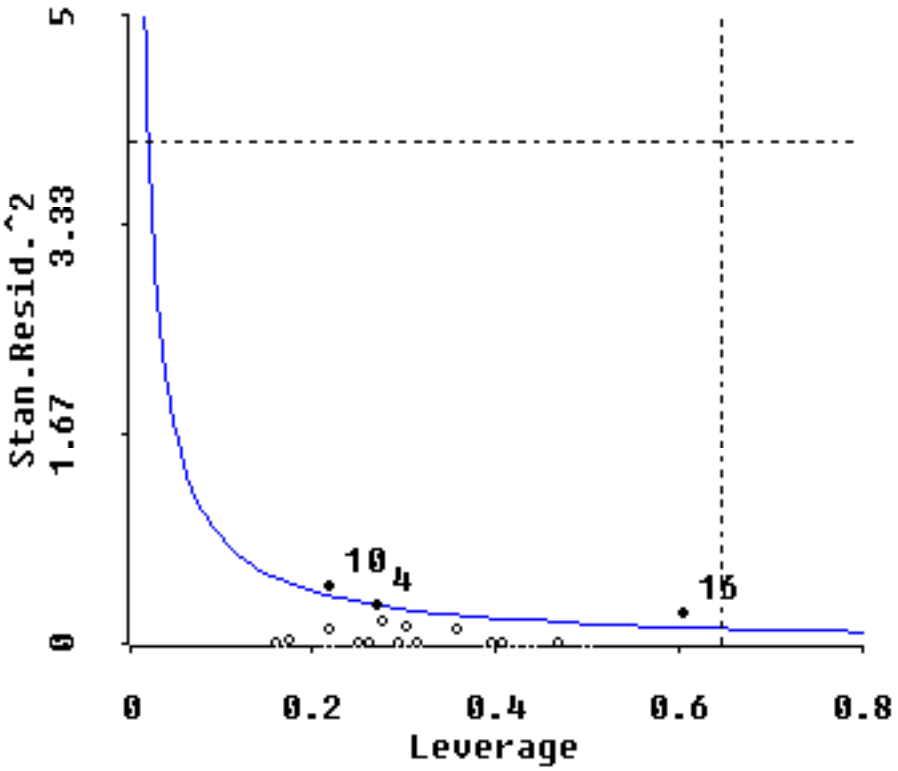

Figure 9: Leverage/Outlier plot for shrinkage extent $m = 1.00$ along the $q-$shape$=-1.5$ path for the Longley(1967) data. As at $m = 0.00$, note that the majority of influence here at $m = 1.00$ comes from large leverages rather than from large residuals, as at $m = 4.00$ and $m = 5.00$. And here point 10=1956 joins 4=1950 and 16=1962 as the three most influential years (replacing either 5=1951 of $m = 0.00$ or 3=1949 of $m = 5.00$.)

Table 1. Summary of Shrinkage Effects on Influence Statistics

Extent of Shrinkage	Largest Std. Residuals	Largest Leverages	Largest Influences
Least Squares: m = 0			
first	10=1956	16=1962	5=1951
second	4=1950	5=1951	16=1962
third	5=1951	2=1948	4=1950
2/p-ths Rule: m = 1			
first	10=1956	16=1962	16=1962
second	4=1950	2=1948	10=1956
third	16=1962	12=1958	4=1950
Maximum Likelihood: m = 4			
first	10=1956	1=1947	3=1949
second	8=1954	3=1949	12=1958
third	12=1958	6=1952	8=1954
Excessive Shrinkage: m = 5			
first	10=1956	1=1947	3=1949
second	3=1949	16=1962	16=1962
third	8=1954	15=1961	4=1950

We now see that ridge shrinkage can make dramatic differences in influence statistics. Observation 5=1951 is most influential on the least squares solution ($m = 0$), while observation 3=1949 is most influential when ridge shrinkage becomes excessive (at about $m = 5$.) And shrinkage has also produced a distinct shift in emphasis from influence dominance by leverages at $m = 0$ in Figure 3 to influence dominance by residuals at $m = 5$ in Figure 5.

Next, let us examine what is happening when shrinkage is limited to the $m = 4$ extent that is most likely to minimize overall MSE risk along the $q = -1.5$ path; see Figures 6 and 7. Finally, we show the situation where shrinkage is limited by the 2/p-ths rule-of-thumb to about $m = 1$ along the $q = -1.5$ path; see Figures 8 and 9.

A summary of what we have learned about the effects of shrinkage on the influence of individual observations within the Longley(1967) dataset is presented in Table 1. The Longley(1967) data do not contain any dramatically obvious outliers or distinctly high leverage points. As expected, we have seen that shrinkage tends, generally, to decrease leverages and increase residuals for all observations. And shrinkage can cause major shifts in the relative influence/outliers/leverages of observations.

8 Curvature and Heteroscedasticity in VRR

Our discussion has focussed on use of VRR to see the influence of individual observations on a multiple regression model. But we should, perhaps, remind the reader

that the $x^{=}$(or y^{*}) versus y plot is also well known to be useful in revealing lack-of-fit and/or heteroscedasticity.

When plotted against their predicted values from multiple regression, no "curvature" in the general pattern of response values should remain. In fact, a key property of VRR is that the regression of y onto $x^{=}$ can always, without loss of generality, be restricted to be **linear** in $x^{=}$. Any hint of curvature when y is plotted against $x^{=}$ is almost surely an indication that the multiple regression model that yielded the $x^{=}$ predictions is an inadequate (misspecified) model. Some sort of transformation of the response and/or predictor variables is needed.

The $x^{=}$ versus y plot can also reveal that the variability of the responses changes with $x^{=}$, which is an indication of heteroscedasticity. This calls for, say, use of observation weights inversely proportional to variances.

9 Conclusions

Linked Visual Re-Regression and Leverage/Outlier plots help practitioners **see** the influence of individual points on multiple regression fits. VRR heuristics are somewhat more reliable for showing the effects of outlying responses than those of high leverage regressor combinations.

10 Appendix: Technical Details.

10.1 Principal Axis Rotation

The singular value decomposition of X can be written as $X = H\Lambda^{1/2}G'$, where H is the $n \times p$ semi-orthogonal matrix of *principal coordinates* of X and G is the $p \times p$ orthogonal matrix of *principal axis direction cosines*. The principal axes are ordered such that the eigenvalues, $\lambda_1 \geq \cdots \geq \lambda_p > 0$, of $X'X$ are non-increasing. Note that $1'X = 0'$ implies that the columns of the principal coordinates matrix also sum to zero, $1'H = 0'$. The least squares estimate of the regression coefficient β vector can then be written as

$$b^o = (X'X)^{-1}X'y = G\Lambda^{-1/2}H'y = \sqrt{y'y}G\Lambda^{-1/2}r = Gc. \tag{7}$$

where $r = H'y/\sqrt{y'y}$ is the vector of *principal correlations* between the $y$vector and the columns of the Hmatrix of principal coordinates of X and $c = \sqrt{y'y}\Lambda^{-1/2}r$ is the vector of *uncorrelated components* of b^o. Note also that the familiar least squares R-squared statistic is the sum-of-squares of the principal correlations: $R^2 = r'r$.

10.2 Shrinkage/Ridge Estimates

Generalized ridge estimators shrink the uncorrelated components of the least-squares solution along the principal axes of $X'X = G\Lambda G'$ using multiplicative shrinkage

factors, $\delta_1, \cdots, \delta_p$, where each shrinkage factor lies in the closed interval from zero to one, $0 \leq \delta_i \leq 1$. With $\Delta = Diag(\delta_1, \cdots, \delta_p)$ denoting the $p \times p$ diagonal matrix of shrinkage factors, generalized ridge estimates are seen to be of the general form:

$$b^* = G\Delta c. \tag{8}$$

Attention is commonly restricted to a $1-$or $2-$parameter family, Goldstein and Smith(1973), in which the shrinkage (δ) factor applied to the $i-$th uncorrelated component of the least-squares solution is of the form

$$\delta_i = \lambda_i/(\lambda_i + k\lambda_i^q) = 1/(1 + k\lambda_i^{q-1}), \tag{9}$$

where $k$is non-negative and q is a finite power that determines the shape (or curvature) of the ridge path through $p-$dimensional space.

The **extent** of ridge shrinkage is measured by the *multicollinearity allowance* parameter

$$m = p - \delta_1 - \ldots - \delta_p = Rank(X) - Trace(\Delta). \tag{10}$$

Every $q-$shape ridge family starts with the least squares solution at $m = 0[k = 0, \delta_1 = \ldots = \delta_p = 1]$ and ends with all ridge coefficients zero at $m = p[k = +\infty, \delta_1 = \ldots = \delta_p = 0]$.

10.3 Ridge Predictions

The $n \times 1$ vector of ridge predicted values is written as

$$y^* = y^*(\Delta) = Xb^* = H\Delta H'y = \sqrt{y'y}H\Delta r = b^{=}x^{=}. \tag{11}$$

where $b^{=}$ is again a scalar and $x^{=}$ is a $n \times 1$ vector with mean 0 and, except when $\Delta = 0$, sum-of-squares $x^{=\prime}x^{=} = (n-1)$. The $b^{=}$ scalar factor can thus be expressed as

$$(b^{=})^2 = \frac{1}{(n-1)}\sum_{i=1}^{n} y_i^{*2} = r'\Delta^2 r = \sum_{j=1}^{p} \delta_j^2 r_j^2 \ . \tag{12}$$

Since $b^{=}$ is the slope of the ridge fit, $y^* = b^{=}x^{=}$, it is of interest to note that $b^{=}$ is usually distinct from the slope of the OLS regression of y onto $x^{=}$:

$$b^{VRR} = x^{=\prime}y/(n-1) = r'\Delta r/\sqrt{r'\Delta^2 r}. \tag{13}$$

Note that $b^{VRR} \leq R$ because b^{VRR}/R is the cosine of the angle between r and Δr. Furthermore, $y^{VRR} = b^{VRR}x^{=} = k^{VRR}y^*$ where $k^{VRR} = b^{VRR}/b^{=} = (r'\Delta r)/(r'\Delta^2 r)$ is well defined and ≥ 1 whenever at least one shrinkage factor is strictly positive and its corresponding principal correlation is non-zero. In fact, $k^{VRR}b^*$ is the estimate of β that yields the y^{VRR} predictions vector. Finally, (12) and (13) together imply that $b^{VRR} - b^{=} = \sqrt{r'\Delta^2 r} \times (k^{VRR} - 1) \geq 0$. Thus, we end up with the sequence of bounds:

$$0 \leq b^{=} \leq b^{VRR} \leq R \leq 1 \leq k^{VRR}. \tag{14}$$

For example, in the special case of *uniform shrinkage*, $\Delta = \delta \times I$, we have that $b^{=} = \delta \times R$, $b^{VRR} \equiv R$ and $k^{VRR} = 1/\delta$ for all $0 < \delta \leq 1$.

It is easily shown that both the ridge and the least squares VRR fits pass through the origin, $(0, 0)$, of $(x_i^{=}, y_i)$ coordinates when all n points are weighted equally. After all, $(0, 0)$ is then also the centroid of the scatter of $(x_i^{=}, y_i)$ points for $1 \leq i \leq n$. Since b^{VRR} increases the length of the b^* estimate vector, VRR revises the intercept estimate to be $\hat{\mu} = \bar{y} - k^{VRR} \times \bar{x}' b^*$. If a VRR line were to be fit to the $(x_i^{=}, y_i)$ scatter using robust regression, the fit might not pass through $(0, 0)$; the revised intercept estimate would then of the more general form $\hat{\mu} = \bar{y} - k \times \bar{x}' b^* - \bar{\varepsilon}$, where the average error term could be non-zero, $\bar{\varepsilon} \neq 0$.

References

Askin, R. G. and Montgomery, D. C. (1980). "Augmented robust estimators." **Technometrics** 22, 333-341.

Beckman, R. J. and Trussell, H. J. (1974), "The distribution of an arbitrary studentized residual and the effects of updating in multiple regression," **Journal of the American Statistical Association** 69, 199-201.

Cook, R. D. (1977), "Detection of influential observations in linear regression," **Technometrics** 19, 15-18.

Dallal, G. E. (1991), "LMS: Least Median of Squares Regression," **The American Statistician**, 45, 74.

Ellenberg J. H. (1973), "The joint distribution of the studentized least squares residuals from a general linear regression," **Journal of the American Statistical Association** 68, 941-943.

Goldstein, M. and Smith, A. F. M. (1974). "Ridge-type estimators for regression analysis." **Journal of the Royal Statistical Society** B 36, 284-291.

Hoerl, A .E. and Kennard, R. W. (1970), "Ridge regression: biased estimation for nonorthogonal problems," **Technometrics** 12, 55-67.

Holland, P. (1973). "Weighted ridge regression: combining ridge and robust regression methods." National Bureau of Economic Research, Working Paper #11, Cambridge, MA.

Leamer, E. E. (1978). "Regression selection strategies and revealed priors." **Journal of the American Statistical Association** 73, 580-587.

Longley, J. W. (1967). "An appraisal of least squares programs for the electronic computer from the point of view of the user." **Journal of the American Statistical Association** 62, 819-841.

Obenchain, R. L. (1975). "Ridge analysis following a preliminary test of the shrunken hypothesis." **Technometrics** 17, 431-441.

Obenchain, R. L. (1978). "Good and optimal ridge estimators." **Annals of Statistics** 6, 1111-1121.

Obenchain, R. L. (1996a). "Maximum likelihood shrinkage in regression." Submitted to **The American Statistician.**

Obenchain, R. L. (1996b). XLisp-Stat code for shrinkage/ridge regression and visual re-regression, Eli Lilly and Company, Statistical and Mathematical Sciences.

Stata Corporation. (1995), **Stata Reference Manual**, Release 4.0, [rreg - Robust regression, Volume 3, 132-137; rreg.ado version 3.0.2], College Station, Texas: Stata Press.

Tierney, L. (1990), **LISP-STAT: An Object Oriented Environment for Statistical Computing and Dynamic Graphics**, New York: John Wiley and Sons.

Walter, B. (1994). XLisp-Stat code for shrinkage/ridge regression, Technische Universitaet Muenchen, Weihenstephan, Germany.