NBER WORKING PAPER SERIES

WEIGHTED RIDGE REGRESSION:

COMBINING RIDGE AND ROBUST REGRESSION METHODS

Paul W. Holland*

Working Paper No. 11

COMPUTER RESEARCH CENTER FOR ECONOMICS AND MANAGEMENT SCIENCE
National Bureau of Economic Research, Inc.
575 Technology Square
Cambridge, Massachusetts 02139

September 1973

*NBER Computer Research Center. Research supported in part by National Science Foundation Grant GJ-1154X2 to the National Bureau of Economic Research, Inc.

Abstract

Gives the formulas for and derivation of ridge regression methods when there are weights associated with each observation. A Bayesian motivation is used and various choices of k are discussed. A suggestion is made as to how to combine ridge regression with robust regression methods.

Contents

1. Introduction

We consider here the familiar regression problem specified by

$$y = X\beta + \varepsilon \tag{1-1}$$

where y is $N\times 1$, X is $N\times p$ and β is $p\times 1$. We use the notation

$$Z \sim \text{Gau}_N(\mu, \Sigma) \tag{1-2}$$

to mean that Z has an N-dimensional multivariate Gaussian distribution with mean vector $\mu = E(Z)$ and covariance matrix $\Sigma = \text{Cov}(Z)$. In this notation we assume that

$$\varepsilon \sim \text{Gau}_N(0\ , \sigma^2 \langle w \rangle^{-1}) \tag{1-3}$$

where $\langle w \rangle$ denotes a diagonal matrix with the vector w along its main diagonal. $\langle w \rangle$ is assumed to be a known matrix, σ^2 and β are unknown.

The weighted least squares estimate of β is given by

$$\hat{\beta}_{LS} = (X^T \langle w \rangle X)^{-1} X^T \langle w \rangle y \tag{1-4}$$

and the weighted least squares "fitted values,"

$$\hat{y}_{LS} = X\,\hat{\beta}_{LS}\ , \tag{1-5}$$

satisfy the following normal equations:

$$X^T \langle w \rangle y = X^T \langle w \rangle \hat{y}_{LS}\ . \tag{1-6}$$

The problem we wish to attack here is how to improve on $\hat{\beta}_{LS}$ as an estimator of β. Because, $\hat{\beta}_{LS}$ is the best linear unbiased estimator of β, to find an improvement we must consider estimators which are both non-linear

functions of y and biased. The fundamental work of Stein [1956] and later that of Baranchik [1970] and Sclove [1968] show that when the number of regression parameters is sufficiently large, then uniform imporvements over $\hat{\beta}_{LS}$ are possible using biased, non-linear estimators. The minimum p (not including the constant term) is p = 3. The results of Wermuth [1972] show that the degree of improvement possible increases substantially as the x's become more multicollinear.

The particular class of estimators we will discuss is a slight extension of the "ridge regression" estimators developed by Hoerl and Kennard [1970] and studied by Wermuth [1972], Sclove [1973], Marquardt [1970], Mayer and Willke [1973].

We use the "weighted least squares" framework because it allows us to use a suggestion of Tukey [1973] for doing robust regression and in effect to combine ridge and robust regression methods. This combination is discussed in Section 6.

2. Ridge Regression When There are Weights

We now give a prescription for doing ridge regression when there is a weight, w_i, associated with each observation. The weights are assumed to be non-negative and they need not sum to unity -- but they may if that is convenient. In addition we also assume that we have a "prior mean" for β. This will be amplified more fully in the next section. We denote the prior mean for β by δ. In the usual case of ridge regression, δ is taken as zero, and the weights, w_i, are all equal.

In this paper, we always assume there is a constant term in the regression equation, but that this is not reflected in the choice of X-matrix. Hence we assume that no column of X is constant. The constant term is

estimated separately from the other regression coefficients via the following formula

$$\hat{\beta}_0 = \bar{y} - \sum_{j=1}^{p} \hat{\beta}_j \bar{x}_j \tag{2-1}$$

where $\bar{y}$ and $\bar{x}_j$ denote the weighted means of y and the $j^{\underline{th}}$ column of X, respectively, i.e.,

$$\bar{y} = \frac{\sum_i w_i y_i}{\sum_i w_i} \quad ; \quad \bar{x}_j = \frac{\sum_i w_i x_{ij}}{\sum_i w_i} \quad . \tag{2-2}$$

In (2-1), $\hat{\beta}_j$ denotes the estimate of the $j^{\underline{th}}$ element of β which we will describe shortly.

In the calculation of the regression coefficients, as opposed to $\hat{\beta}_0$, we assume that all variables have had their weighted mean (2-2) subtracted out,

$$\tilde{y} = y - \bar{y} \; ; \; \tilde{x}_j = x_j - \bar{x}_j \quad . \tag{2-3}$$

The "weighted" length of $\tilde{x}_j$ is given by

$$s_j = \sqrt{\sum_i w_i \tilde{x}_{ij}^2} \tag{2-4}$$

thus we give all the x_j the same weighted length by setting

$$X^* = \tilde{X} \langle s \rangle^{-1} \tag{2-5}$$

This scaling of the x's implies a rescaling of the β's via

$$\beta^* = \langle s \rangle \beta \tag{2-6}$$

Hence the prior mean δ must be rescaled also,

$$\delta^* = \langle s \rangle \delta \quad . \tag{2-7}$$

Having properly centered and scaled all variables we may now give the ridge regression estimates of the rescaled parameter, β^*. This is given by

$$\hat{\beta}_R^* = \delta^* + (X^{*T}\langle w\rangle X^* + kI)^{-1}X^{*T}\langle w\rangle (y - X^*\delta^*) . \qquad (2\text{-}8)$$

The ridge regression estimator of β (rather than β^*) is given by

$$\hat{\beta}_R = \delta + (\tilde{X}^T\langle w\rangle \tilde{X} + k\langle s^2\rangle)^{-1}\tilde{X}^T\langle w\rangle (\tilde{y} - \tilde{X}\delta) \qquad (2\text{-}9)$$

We observe that $\hat{\beta}_R$ depends on the parameter, k. When k=0, then $\hat{\beta}_R = \hat{\beta}_{LS}$ no matter what δ is. When $k=\infty$, then $\hat{\beta}_R = \delta$. For intermediate values of k, $\hat{\beta}_R$ interpolates between these extreme values.

Hoerl and Kennard [1970] suggest using several values of k in a diagnostic mode to identify those least-square parameter estimates which might be improvable. Wermuth [1972] and Sclove [1973] suggest choosing k from the data and obtaining a single point estimator of β rather than a one-parameter family of estimators. In Section 4 we discuss various choices of k that are data dependent.

In summary, the method of estimation we propose here is as follows.

(a) Compute weighted means and subtract them from each variable, obtaining $\tilde{y}$ and $\tilde{X}$.

(b) Compute k via one of the methods discussed in Section 4.

(c) Estimate the regression coefficients via equation (2-9), obtaining $\hat{\beta}_R$.

(d) Estimate the constant term via equation (2-1) using $\hat{\beta}_R$ for the regression coefficients.

3. Motivations and Interpretations

In this section, we shall give the Bayesian motivation for ridge

regression as put forward in the previous section. In addition, we show how ridge regression may be interpreted as a "smooth" selection of variables method of estimating parameters.

Bayesian Background

We begin with the statement of a useful lemma that allows us to pass back and forth between conditioning U on V and then V on U when (U,V) has a multivariate Gaussian distribution.

The back-and-forth lemma: If $U|V \sim Gau_n[a + B(V-d), C]$ and $V \sim Gau_m(d, E)$ and if C and E are non-singular, then

$$\text{(a)} \quad V|U \sim Gau_m(d + (E^{-1}+B^TC^{-1}B)^{-1}B^TC(U-a), (E^{-1}+B^TC^{-1}B)^{-1})$$

and $$\text{(b)} \quad U \sim Gau_n(a, C+BEB^T).$$

The proof of the back-and-forth lemma is a straightforward exercise in properties of the multivariate Gaussian distribution and matrix algebra.

Now suppose we consider the Bayesian analysis of the general linear model given by

$$y|\beta \sim Gau_N(X\beta, \sigma^2 \langle w \rangle^{-1}) \tag{3-1}$$

$$\beta \sim Gau_p(\delta, \Delta) \tag{3-2}$$

Shortly, we shall specialize Δ to $\tau^2 I$, but for the moment we consider the more general setting given in (3-2). Note that we do not give a prior distribution to σ^2 in this development. This is to keep the analysis simple. In all cases we estimate σ^2 by the weighted residual mean square from the least squares fit. This is

$$\hat{\sigma}^2 = (N-p)^{-1} \sum_i w_i(\tilde{y}_i - (\hat{\tilde{y}}_{LS})_i)^2 \tag{3-3}$$

Furthermore, we will often regard σ^2 as known and equal to the estimated value, $\hat{\sigma}^2$. While this is not the way a full Bayesian analysis would proceed, it is adequate for our purpose which is to motivate the procedure given in the previous section.

From the model (3-1) and (3-2) and the back-and-forth lemma we may obtain the posterior distribution of β and the marginal distribution of y.

Theorem 1: If $y|\beta \sim \text{Gau}_N(X\beta, \sigma^2 \langle w \rangle^{-1})$ and $\beta \sim \text{Gau}_p(\delta, \Delta)$, then

(a) (Posterior distribution of β)

$$\beta|y \sim \text{Gau}_p(\delta + (X^T \langle w \rangle X + \sigma^2\Delta^{-1})^{-1} X^T \langle w \rangle (y - X\delta), \sigma^2(X^T \langle w \rangle X + \sigma^2\Delta^{-1})^{-1})$$

(b) (Marginal distribution of y)

$$y \sim \text{Gau}_N(X\delta, \sigma^2 \langle w \rangle^{-1} + X\Delta X^T).$$

From part (a) of Theorem 1 we see that if the prior covariance matrix of β is taken as $\Delta = \tau^2 I$, then except for the replacement of y by $\tilde{y}$ and X by X* the formula for $\hat{\beta}^*_R$ corresponds to the posterior mean of β with

$$k = \frac{\sigma^2}{\tau^2} \quad . \tag{3-4}$$

Suppose we assume that δ is given, what conditions on β would make the assumption that

$$\beta \sim \text{Gau}_p(\delta, \tau^2 I) \tag{3-5}$$

a reasonable one? By scaling the x's as we have, we have made them dimensionless so that the β^* parameters reflect only the relative slopes of the regression plane and not merely differences in the units in which the x's are measured. The assumption (3-5) asserts that the $\beta_i - \delta_i$ behave like a sample from a Gaussian distribution with unknown variance and zero mean. This is more plausible for the β^*'s than the original β's which may have

differing units. Finally, the constant term in a regression is usually of quite a different character than the regression parameters. It merely centers the regression plane to pass through the "middle" of the point cloud. Hence we have centered each variable at its weighted mean and chosen $\hat{\beta}_0$ so that the fitted regression plane passes through the point $(\bar{y}, \bar{x}_1, \ldots, \bar{x}_p)$. Thus β_0 is not included in the parameters that have been given priors.

Under assumption (3-5) the posterior distribtion of β is

$$\beta|y \sim \text{Gau}_p(\delta + (X^T\langle w\rangle X + kI)^{-1}X^T\langle w\rangle(y - X\delta), \sigma^2(X^T\langle w\rangle X + kI)^{-1}) \quad (3\text{-}6)$$

and the marginal distribution of y is

$$y \sim \text{Gau}_p(X\delta, \sigma^2\langle w\rangle^{-1} + \tau^2 XX^T) \ . \quad (3\text{-}7)$$

Interpreting Ridge Regression

The Bayesian motivation for ridge regression may be satisfactory for many purposes, but the following interpretation shows that it also has close ties with regression on principal components.

We may rewrite (2-8) in the following form:

$$\hat{\beta}^*_R = \delta^* + (I + k(X^{*T}\langle w\rangle X^*)^{-1})^{-1}(\hat{\beta}^*_{LS} - \delta^*) \quad (3\text{-}8)$$

where

$$\hat{\beta}^*_{LS} = \langle s\rangle \hat{\beta}_{LS} \quad (3\text{-}9)$$

A particularly revealing form of ridge regression appears when we transform to the "principle component axes." The usual orthogonal diagonalization of $X^{*T}\langle w\rangle X^*$ is given by

$$X^{*T}\langle w\rangle X^* = V\langle\lambda\rangle V^T \quad (3\text{-}10)$$

where V is p×p orthogonal and $\langle\lambda\rangle$ is the system of eigenvalues of $X^{*T}\langle w\rangle X^*$.

We define the "principal component parameters" by

$$\gamma^* = V^T\beta^* \tag{3-11}$$

and the corresponding transformed prior mean by

$$\nu^* = V^T\delta^* \tag{3-12}$$

One property of least squares is that

$$\hat{\gamma}^*_{LS} = V^T\hat{\beta}^*_{LS} \quad . \tag{3-13}$$

We may define $\hat{\gamma}^*_R$ so that this is also true for ridge regression, i.e.,

$$\hat{\gamma}^*_R = V^T\hat{\beta}^*_R \quad . \tag{3-14}$$

Starting with (3-8) and (3-14) we then obtain

$$\hat{\gamma}^*_R = \nu^* + \left\langle \frac{\lambda}{\lambda+k} \right\rangle (\hat{\gamma}^*_{LS} - \nu^*) \quad . \tag{3-15}$$

Hence we see that the ridge regression estimators of the principal component parameters are found by shrinking the least squares estimators of γ^*_i towards ν^*_i by an amount that reflects the size of λ_i relative to k. If λ_i large, then $(\hat{\gamma}^*_{LS})_i$ is shrunk very little; when λ_i is small, then it is shrunk a lot. Thus when $\delta = 0$, $\hat{\gamma}^*_R$ may be viewed as a type of selection of variables technique using the principle components as the variables and the size of the eigenvalues as the selection criterion.

4. The Choice of k

There are two types of choices of k which we shall discuss here. The first type is in the spirit of empirical Bayes methods because prior parameters are estimated from the data. The second type is based on estimates of certain optimum values of k.

In all cases σ^2 is treated as a known constant and set equal to its estimated value $\hat{\sigma}^2$ from (3-3). Furthermore, while the theoretical analysis uses y, in the actual computations the centered values, $\tilde{y}$ are used. This introduces an error of order N^{-1} into the analysis, but this is more than overcome by the resulting simplification in the resulting formulas.

Empirical Bayes Choices of k

From (3-7) we have that the marginal distribution of y is given by

$$y \sim \text{Gau}_N(X^*\delta^*,\ \sigma^2 \langle w \rangle^{-1} + \tau^2 X^* X^{*T}) \tag{4-1}$$

From (4-1) it follows that

$$\langle w^{\frac{1}{2}} \rangle\, y \sim \text{Gau}_N(\langle w^{\frac{1}{2}} \rangle X\delta,\ \sigma^2 I + \tau^2 \langle w^{\frac{1}{2}} \rangle X^* X^{*T} \langle w^{\frac{1}{2}} \rangle) \tag{4-2}$$

and hence that

$$E[(y - X\delta)^T \langle w \rangle (y - X\delta)] = N\sigma^2 + \tau^2\ \text{trace}(X^{*T} \langle w \rangle X^*). \tag{4-3}$$

If we let

$$U^T \langle w \rangle U = \| U \|_w^2 \tag{4-4}$$

then (4-3) may be expressed as

$$E[\| y - \tilde{X}\delta \|_w^2] = N\sigma^2 + p\tau^2 \tag{4-5}$$

since $\text{trace}(X^{*T} \langle w \rangle X^*) = p$.

Therefore an unbiased estimate of τ^2 is given by

$$\hat{\tau}^2 = (\| \tilde{y} - \tilde{X}\delta \|_w^2 - N\hat{\sigma}^2)/p\ . \tag{4-6}$$

Thus the ratio of $\hat{\sigma}^2$ to $\hat{\tau}^2$ yields a plausible though biased estimate of $k = \sigma^2/\tau^2$. We call this

$$k_a = \frac{p\,\hat{\sigma}^2}{\left\| \tilde{y} - \tilde{X}\delta \right\|_w^2 - N\,\hat{\sigma}^2} \tag{4-7}$$

Sclove [1973] suggests keeping this estimate of k positive by replacing N by N-p. This yields

$$k_{a1} = \frac{p\,\hat{\sigma}^2}{\left\| \tilde{y} - \tilde{X}\delta \right\|_w^2 - (N-p)\hat{\sigma}^2} \tag{4-8}$$

Alternatively we might use a "positive part" estimator of the form

$$k_{a2} = \max(0, k_a) \tag{4-9}$$

to keep k from being negative.

Another set of empirical Bayes choices of k stem from the following observation. If σ^2 is regarded as known, then $\hat{\beta}^*_{LS}$ is a sufficient statistic (marginally) for τ^2 so that we may reduce by sufficiency to the marginal distribution of

$$\hat{\beta}^*_{LS} = (X^{*T}\langle w\rangle X^*)^{-1}X^{*T}\langle w\rangle\, y \sim \mathrm{Gau}_p(\delta^*, \sigma^2[(X^{*T}\langle w\rangle X^*)^{-1} + k^{-1}I]). \tag{4-10}$$

Equivalently, we may use the marginal distribution of

$$\begin{aligned}\hat{\gamma}^*_{LS} &= V^T\hat{\beta}^*_{LS} \\ &\sim \mathrm{Gau}_p(\nu^*, \sigma^2\langle k^{-1} + \lambda^{-1}\rangle)\end{aligned} \tag{4-11}$$

If we set

$$\hat{\xi}^*_{LS} = \langle (k^{-1} + \lambda^{-1})^{-\frac{1}{2}}\rangle\,(\hat{\gamma}^*_{LS} - \nu^*) \tag{4-12}$$

then

$$\hat{\xi}^*_{LS} \sim \mathrm{Gau}_p(0, \sigma^2 I)\ . \tag{4-13}$$

and we see that

$$\left\| \hat{\xi}^*_{LS} \right\|^2 = \sum_{i=1}^{p} \frac{(\hat{\gamma}^*_i - \nu^*_i)^2}{k^{-1} + \lambda_i^{-1}}$$

$$\sim \sigma^2 \text{ [chi-square on p d.f.]} . \qquad (4\text{-}14)$$

Dempster (Wermuth [1972]) suggests setting $\left\| \hat{\xi}^*_{LS} \right\|^2$ equal to its expected value and estimating σ^2 by $\hat{\sigma}^2$. This yields the following equation for k

$$\sum_i \frac{(\hat{\gamma}^*_i - \nu^*_i)^2}{k^{-1} + \lambda_i^{-1}} = p\,\hat{\sigma}^2 . \qquad (4\text{-}15)$$

We shall call the solution to (4-15) (if it exists) k_d. Sclove [1973] suggests a method that is equivalent to the following observation. $\left\| \hat{\xi}^*_{LS} \right\|^2$ and $(N-p)\,\hat{\sigma}^2$ are independent with $\sigma^2\,\chi^2_p$ and $\sigma^2\,\chi^2_{N-p}$ distributions respectively. Thus set

$$F_{p,\,N-p} = \frac{\left\| \hat{\xi}^*_{LS} \right\|^2}{(N-p)\,\hat{\sigma}^2} . \qquad (4\text{-}16)$$

Sclove then suggest setting $F_{p,\,N-p}$ equal to its expected value, i.e.,

$$E(F_{p,\,N-p}) = \frac{N-p}{N-p-2} \qquad (\text{if } N-p \geq 3) .$$

This yields the following equation for k

$$\sum_i \frac{(\hat{\gamma}^*_i - \nu^*)^2}{k^{-1} + \lambda_i^{-1}} = p\,\hat{\sigma}^2 \frac{N-p}{N-p-2} . \qquad (4\text{-}17)$$

We shall call the solution to (4-17) (if it exists) k_s.

If we regard $\sigma^2 = \hat{\sigma}^2$ as known, then the likelihood function for k based on $\hat{\gamma}^*_{LS}$ is

$$L(\hat{\gamma}^*_{LS}, k) = (2\pi)^{-p/2}\,\sigma^{-p} \prod_{i=1}^{p} (k^{-1} + \lambda_i^{-1})^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{p} \frac{(\hat{\gamma}^*_i - \nu^*_i)^2}{k^{-1} + \lambda_i^{-1}} \right\} \qquad (4\text{-}18)$$

Differentiating (4-18) in k yields the following equation for k

$$\sum_i \frac{(\hat{\gamma}_i^* - \nu_i^*)^2}{(k^{-1} + \lambda_i^{-1})^2} = \hat{\sigma}^2 \sum_i \frac{1}{k^{-1} + \lambda_i^{-1}} \quad . \tag{4-19}$$

We shall call the solution to (4-19) (if it exists) k_m.

The expressions, k_a, k_{a1}, k_{a2}, k_d, k_s, and k_m, are all of the empirical Bayes choices for k that we will consider, except for some minor modifications we shall make later in this section.

Estimating Optimal k-values

Consider the expected squared distance between $\hat{\beta}_R^*$ and the true parameter value β^* (where expectations are now computed relative to the condition distribution of y given β^* and σ^2). There is a value of k that minimizes this squared distance, but it depends on the unknown parameters. If we use the least squares estimates of β^* in the resulting formula for the optimal k-value we obtain a data-dependent choice of k that estimates this best choice of k. This is the spirit in which we present the next two choices of k. There are actually two meaningful "distances" in this problem and we consider them in sequence. The first is the simple Euclidean distance given by

$$E\left\| \hat{\beta}_R^* - \beta^* \right\|^2 = \sum_i E[(\hat{\beta}_R^*)_i - \beta_i^*]^2$$

$$= E\left\| \hat{\gamma}_R^* - \gamma^* \right\|^2 = \sum_i E[(\hat{\gamma}_R^*)_i - \gamma_i^*]^2 \tag{4-20}$$

But

$$E[(\hat{\gamma}_R^*)_i - \gamma_i^*]^2 = E[\nu_i^* + d_i((\hat{\gamma}_{LS}^*)_i - \nu_i^*) - \gamma_i^*]^2$$

$$= E[d_i((\hat{\gamma}_{LS}^*)_i - \nu_i^*) - (\gamma_i^* - \nu_i^*)]^2$$

$$= d_i^2 E((\hat{\gamma}_{LS}^*)_i - \gamma_i^*)^2 + (1-d_i)^2(\gamma_i^* - \nu_i^*)^2$$

where $d_i = \lambda_i/(\lambda_i + k)$.

However, because $\hat{\beta}^*_{LS} \sim \text{Gau}_p(\beta^*, \sigma^2(X^{*T}\langle w\rangle X^*)^{-1})$

we have

$$\hat{\gamma}^*_{LS} \sim \text{Gau}_p(\gamma^*, \sigma^2 \langle \lambda \rangle^{-1}) \ , \tag{4-21}$$

and hence

$$E[(\hat{\gamma}^*_R)_i - \gamma^*_i]^2 = d_i^2 \ \sigma^2 \ \lambda_i^{-1} + (1-d_i)^2(\gamma^*_i - \nu^*_i)^2 \ .$$

Therefore, we have

$$E \left\| \hat{\beta}^*_R - \beta^* \right\|^2 = \sum_i \sigma^2 \ d_i^2 \ \lambda_i^{-1} + (1-d_i)^2(\gamma^*_i - \nu^*_i)^2. \tag{4-22}$$

If we differentiate (4-23) in k to find that value which minimizes the expected squared (Euclidean) distance of $\hat{\beta}^*_R$ to β^* we obtain the following equation for k (after substituting $\hat{\gamma}^*_{LS}$ for γ^* and $\hat{\sigma}^2$ for σ^2)

$$\sum_i \frac{k\lambda_i(\hat{\gamma}^*_i - \nu^*_i)^2}{(\lambda_i + k)^3} = \hat{\sigma}^2 \sum_i \frac{\lambda_i}{(\lambda_i + k)^3} \ . \tag{4-23}$$

We shall denote the solution to (4-23) by k_{Ob}.

The other notion of choseness that is relevant to this problem is the expected squared weighted distance from $\hat{y}_R = \tilde{X}\hat{\beta}_R$ to $\tilde{X}\beta$. We now examine the result of minimizing this quantity. We have

$$\begin{aligned} E \left\| \hat{y}_R - \tilde{X}\beta \right\|^2_w &= E[(\hat{\beta}_R - \beta)^T \tilde{X}^T \langle w \rangle \tilde{X}(\hat{\beta}_R - \beta)] \\ &= E[(\hat{\beta}^*_R - \beta^*)^T X^{*T} \langle w \rangle X^*(\hat{\beta}^*_R - \beta^*)] \\ &= E(\hat{\gamma}^*_R - \gamma^*)^T \langle \lambda \rangle (\hat{\gamma}^*_R - \gamma^*) \\ &= \sum_i \lambda_i \ E[(\hat{\gamma}^*_R)_i - \gamma^*_i]^2 \ . \end{aligned} \tag{4-24}$$

Hence we may write

$$E\left\| \hat{y}_R - X\beta \right\|_w^2 = \sum_i [d_i^2 \sigma^2 + (1-d_i)^2 \lambda_i (\gamma_i^* - \nu_i^*)^2] \qquad (4\text{-}25)$$

where $d_i = \lambda_i/(\lambda_i + k)$. Minimizing in k produces the following equation analogous to (4-23)

$$\sum_i \frac{k\lambda_i^2(\hat{\gamma}_i^* - \nu_i^*)^2}{(\lambda_i + k)^3} = \hat{\sigma}^2 \sum_i \frac{\lambda_i^2}{(\lambda_i + k)^3} . \qquad (4\text{-}26)$$

We shall denote the solution to (4-26) by k_{Oy}.

Further Study of These Choices of k

We now have eight possible methods for choosing k in ridge regression. In order to thin down the candidates we begin by considering what they look like in the important special case when the x's are orthogonal. By this we mean that $\lambda_i \equiv 1$ which implies that

$$X^{*T} \langle w \rangle X^* = I . \qquad (4\text{-}27)$$

When this happens, all of the equations for determining k have easy solutions. They are given by:

$$\frac{1}{1 + k_m} = \frac{1}{1 + k_d} = 1 - \frac{p\hat{\sigma}^2}{\left\| \hat{\beta}_{LS}^* - \delta^* \right\|^2} \qquad (4\text{-}28)$$

$$\frac{1}{1 + k_s} = 1 - \frac{p\hat{\sigma}^2}{\left\| \hat{\beta}_{LS}^* - \delta^* \right\|^2} \left(\frac{N-p}{N-p-2}\right) \qquad (4\text{-}29)$$

$$\frac{1}{1 + k_{Ob}} = \frac{1}{1 + k_{Oy}} = 1 - \frac{p\hat{\sigma}^2}{\left\| \hat{\beta}_{LS}^* - \delta^* \right\|^2 + p\hat{\sigma}^2} . \qquad (4\text{-}30)$$

Turning now to k_a we see that

$$\frac{1}{1 + k_a} = 1 - \frac{p\hat{\sigma}^2}{\left\| \tilde{y} - \tilde{X}\delta \right\|_w^2 - (N-p)\hat{\sigma}^2}$$

But the following identity holds true

$$\left\| \tilde{\tilde{y}} - \tilde{\tilde{X}}\delta \right\|_w^2 = \left\| \tilde{\tilde{y}} - \hat{\tilde{\tilde{y}}}_{LS} \right\|_w^2 + \left\| \hat{\tilde{\tilde{y}}}_{LS} - \tilde{\tilde{X}}\delta \right\|_w^2 \qquad (4\text{-}31)$$

$$= (N-p)\hat{\sigma}^2 + (\hat{\beta}_{LS} - \delta)^T X^T \langle w \rangle X(\hat{\beta}_{LS} - \delta)$$

So that we have

$$\left\| \tilde{\tilde{y}} - \tilde{\tilde{X}}\delta \right\|_w^2 = (N-p)\hat{\sigma}^2 + \left\| \hat{\beta}^*_{LS} - \delta^* \right\|^2 \qquad (4\text{-}32)$$

and hence we may express k_a as

$$\frac{1}{1 + k_a} = 1 - \frac{p\,\hat{\sigma}^2}{\left\| \hat{\beta}^*_{LS} - \delta^* \right\|^2} \qquad (4\text{-}33)$$

Similarly we have

$$\frac{1}{1 + k_{a1}} = 1 - \frac{p\,\hat{\sigma}^2}{\left\| \hat{\beta}^*_{LS} - \delta^* \right\|^2 + p\,\hat{\sigma}^2} \qquad (4\text{-}34)$$

and

$$\frac{1}{1 + k_{a2}} = 1 - \max\left(1, \frac{p\,\hat{\sigma}^2}{\left\| \hat{\beta}^*_{LS} - \delta^* \right\|^2}\right) . \qquad (4\text{-}35)$$

Now in this case (i.e., $\lambda_i \equiv 1$) we have

$$\hat{\beta}^*_{LS} \sim \text{Gau}_p(\beta^*, \sigma^2 I)$$

with $\hat{\sigma}^2$ an independent, chi-square distributed estimate of the common variance σ^2.

James and Stein [1961] showed that in this type of situation $\hat{\beta}^*_{LS}$ can be uniformly improved upon (in the sense of lowering the value of $E\left\| \hat{\beta}^* - \beta^* \right\|^2$) by an estimator of the form

$$\hat{\beta}^*_{JS} = \delta^* + \frac{1}{1 + k_{JS}} (\hat{\beta}^*_{LS} - \delta^*) \qquad (4\text{-}36)$$

where k_{JS} is given by

$$\frac{1}{1 + k_{JS}} = 1 - \frac{(p-2)\,\hat{\sigma}^2}{\left\| \hat{\beta}^*_{LS} - \delta^* \right\|^2} \tag{4-37}$$

providing that p exceeds 2. Further slight improvements can be achieved if $\hat{\sigma}^2$ is replaced by $\frac{N-p}{N-p+2}\,\hat{\sigma}^2$ and if $(1+k_{JS})^{-1}$ is prevented from going negative by a device like that used in (4-35). However, the bulk of the improvement stems from the use of the factor (4-37). Comparing the corresponding values for our proposed choices of k we see that k_a, k_d, and k_m agree with k_{JS} except for a "p" replacing the correct "p-2". The extra factor in k_S appears to go in the wrong direction. k_{Ob}, k_{Oy} and k_{al} all agree on a shrinking factor that is too small in general. If we use the value of k_{JS} to calibrate the performance of our choices of k in the orthogonal case, then we are motivated to alter the definitions of k_a, k_d and k_m so that they agree with k_{JS} when $\lambda_i \equiv 1$. Because they fail to agree with k_{JS}, we will drop k_{Ob}, k_{Oy}, k_{al}, and k_S from further discussion.

It is easy to change the definition of k_a so that it agrees with k_{JS} in the orthogonal case. We shall use

$$k'_a = \frac{(p-2)\,\hat{\sigma}^2}{\left\| \tilde{y} - \tilde{X}\delta \right\|^2_w - (N-2)\,\hat{\sigma}^2} \quad . \tag{4-38}$$

It is also obvious how to change (4-15) so that k_d agrees with k_{JS} in in the orthogonal case. We propose the following simple modification. Instead of (4-15) use

$$\sum_i \frac{(\hat{\gamma}^*_i - \nu^*_i)^2}{(k^{-1} + \lambda_i^{-1})} = (p-2)\,\hat{\sigma}^2 \quad . \tag{4-39}$$

We shall call the solution to (4-39) k'_d.

It is less obvious how to change (4-19) so that k_m agrees with k_{JS} in the orthogonal case. We suggest the following slight change in (4-19), others might be better.

$$\sum_i \frac{(\hat{\gamma}_i^* - \nu_i^*)^2}{(k^{-1} + \lambda_i^{-1})^2} = (p-2)\,\hat{\sigma}^2\,\frac{1}{p}\sum_i \frac{1}{k^{-1} + \lambda_i^{-1}} \qquad (4\text{-}40)$$

We shall call the solution to (4-40) k'_m.

We have now reduced our eight choices of k to 3, k'_a, k'_d and k'_m. It should be understood that none of this applies when $p \leq 2$ and that k is never allowed to be negative for any of these choices. Equations (4-39) and (4-40) may easily be solved (if solutions exist) by Newton's method starting at $k = 0$. When $w_i \equiv 1$ and $\delta = 0$, it is easy to show that (4-39) has a unique solution if and only if the usual R^2 exceeds $(p-2)/N$. Conditions for the existence of solutions to (4-40) are similar in spirit but more complicated.

In order to distinguish further between k'_a, k'_d and k'_m we need comparisons of their respective performances. k'_a is appealing since it does not require as much work to compute as the other two do.

5. Combining Ridge and Robust Regression Methods

Ridge regression was invented to deal with the problem of near multicollinearity in regression. Another problem that besets the user of regression methods is outliers and other forms of non-Gaussian errors. Robust regression methods have been proposed to deal with such problems (see Huber [1972], Bickel [1973], Andrews [1973] and Tukey [1973] for reviews of methods and discussions of current research). Considerable attention

has been given to various versions of Huber's M-estimators. Tukey proposes using iteratively reweighted least squares as a device for computing M-estimators and other related robust estimators. The weights are computed sequentially from the residuals of the previous iteration. The end product of Tukey's method is a set of weights $\{w_i\}$ such that the robust estimator $\hat{\beta}_w$ is computed by

$$\hat{\beta}_w = (X^T \langle w \rangle X)^{-1} X^T \langle w \rangle y \qquad (5\text{-}1)$$

In view of the analysis and development given in the previous sections we propose here to use weighted ridge regression to combine ridge and robust methods. The specific proposal is to take the weights found for the robust method and do weighted ridge regression using formula (2-9) and (2-1). The choice of k is still problematic but two alternatives present themselves.

(a) Use several k's in the diagnostic mode proposed by Hoerl and Kennard. This will help identify unstable parameter estimates.

(b) Use k'_a, k'_d or k'_m computed from the data to obtain point estimates of β. Further work is necessary to see if these choices of k differ substantially.

The expectation is that a ridgified robust estimator will combine the benefits of both approaches and be no more difficult to compute than $\hat{\beta}_w$ or $\hat{\beta}_R$ separately.

REFERENCES

Andrews, D. [1973]. "Some Monte Carlo Results on Robust/Resistant Regression," (unpublished manuscript).

Baranchik, A. [1970]. "A family of minimax estimators of the mean of a multivariate normal distribution," Annals of Math Statist 41, 642-645.

Bickel, P. [1973]. "On some analogues to linear combinations of order statistics in the linear model," Annals of Statist 1, 597-616.

Hoerl, A., and R. Kennard [1970]. "Ridge regression. Biased estimation for nonorthogonal problems," Technometrics 12, 55-68.

Huber, P. [1972]. "Robust statistics: a review," Annals of Math Statist 43, 1041-1067.

James, W., and C. Stein [1961]. "Estimation with Quadratic loss," Proceedings of the Fourth Berkeley Symposium 1, U. of Calif. Press, 361-379.

Marquardt, D. [1970]. "Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation, Technometrics 12, 591-611.

Mayer, L., and T. Willke. "On biased estimation in linear models," Technometrics 15, 497-508.

Sclove, S. [1968]. "Improved estimators for coefficients in linear regression," JASA 63, 596-606.

Sclove, S. [1973]. "Least squares with random regression coefficient," Technical Report 87, Economic series, Stanford University.

Tukey, J. [1973]. "A way forward for robust regression," (unpublished m.s.).

Wermuth, N. [1972]. An Empirical Comparison of Regression Methods. Unpublished doctoral dissertation, Harvard Univeristy, Department of Statistics.