REGRESSION DIAGNOSTICS AND APPROXIMATE INFERENCE PROCEDURES FOR PENALIZED LEAST SQUARES ESTIMATORS

by

Richard F. Gunst and Randall L. Eubank

Technical Report No. SMU/DS/TR-181
Department of Statistics -- NASA Contract

September 1983

Research sponsored by the National Aeronautics Space Administration - Contract No. NASA NCC 9-9

Department of Statistics
Southern Methodist University
Dallas, Texas 75275

REGRESSION DIAGNOSTICS AND APPROXIMATE INFERENCE PROCEDURES FOR PENALIZED LEAST SQUARES ESTIMATORS

R. L. Eubank and R. F. Gunst*

ABSTRACT

Generalizations of least squares diagnostic techniques are presented for a class of penalized least squares estimators. Efficient computation of these diagnostics is afforded by expressions which relate coefficient estimates and residuals from fits to subsets of the data to the corresponding quantities from a fit to the complete data set. From these expressions approximate confidence intervals and test statistics can be obtained using jackknife and bootstrap procedures. Applications are discussed for the special cases of smoothing splines and ridge regression.

KEY WORDS

Bootstrap confidence intervals; Jackknife confidence intervals; Leverage values; Ridge regression; Smoothing splines; Studentized residuals.

AUTHOR'S FOOTNOTE

* R. L. Eubank is Assistant Professor and R. F. Gunst is Associate Professor and Chairman, Department of Statistics, Southern Methodist University. This research was supported in part by Office of Naval Research Contract No. N00014-82-K-0207 and by the National Aeronautics and Space Administration under Contract No. NCC 9-9. The authors would like to thank Mr. Tom Carmody for furnishing Figure 1 and Dr. Paul Speckman for helpful discussions on the material in this manuscript.

REGRESSION DIAGNOSTICS AND APPROXIMATE INFERENCE PROCEDURES FOR PENALIZED LEAST SQUARES ESTIMATORS

R. L. Eubank and R. F. Gunst

1. INTRODUCTION

Regression diagnostics are an integral component of comprehensive regression modeling efforts, in large part because of recent theoretical advances which lead to computational efficiency. With few exceptions (a notable one being Pregibon (1981)) these advances have been restricted to ordinary least squares (OLS) estimation for linear models. In this paper diagnostic techniques are extended to a class of penalized least squares estimators which include smoothing splines and ridge regression estimators as special cases. An additional benefit of these results is the ability to efficiently compute jackknife confidence intervals and other inferential statistics for model parameters.

Let $y = (y_1,\ldots,y_n)'$ be a vector of observed responses which follow the model

$$y = \eta + \varepsilon , \qquad (1.1)$$

where $\eta = (\eta_1,\ldots,\eta_n)'$ is a vector of unknown constants and $\varepsilon = (\varepsilon_1,\ldots,\varepsilon_n)'$ is a vector of zero mean, uncorrelated errors with common variance σ^2. It is assumed that η is to be approximated by a linear form $X\beta$ where X is a known nxp matrix of rank $p \leq n$ having ith row x_i' and $\beta = (\beta_1,\ldots,\beta_p)'$ is a vector of parameters which is to be estimated. The class of estimators which are investigated in this article are those obtained as the solution to

$$\min_{\beta}\{\Sigma_{j=1}^{n}(y_j-x_j'\beta)^2 + \lambda\beta'Q\beta\}, \quad \lambda \geq 0 \ , \qquad (1.2)$$

with Q denoting an arbitrary positive (semi-) definite matrix. For a given Q, X, and λ, expression (1.2) has a unique solution:

$$\tilde{\beta} = C(\lambda)y \ , \qquad (1.3)$$

where

$$C(\lambda) = (X'X + \lambda Q)^{-1}X' \ . \qquad (1.4)$$

The estimator $\tilde{\beta}$ is termed a <u>penalized least squares estimator</u> of β. Observe that when $\lambda = 0$, $\tilde{\beta}$ reduces to the OLS estimator

$$\hat{\beta} = (X'X)^{-1}X'y \ .$$

At the other extreme, if Q is positive definite $\tilde{\beta} \to \underline{0}$ as $\lambda \to \infty$. In many instances it is preferable to use a value of λ between these two extremes and a variety of methods are available for estimating its value from data. For example, Golub, Heath and Wahba (1979) discuss generalized cross-validation (GCV) as well as other data-driven methods for selecting λ.

It is often reasonable to make the stronger assumption that $\eta = X\beta$ under which model (1.1) becomes the linear regression model

$$y = X\beta + \varepsilon \ . \qquad (1.6)$$

When this model holds and no further assumptions are made, $\tilde{\beta}$ will be termed a generalized ridge regression estimator of β; however, the results presented below are of sufficient generality to include cases in which the η_j represent values from an unknown regression function, η, which is to be estimated nonparametrically. When appropriately formulated (see Section 5) the smoothing spline estimator of η is seen to be a special case of estimator (1.3).

As with ordinary least squares, the penalized least squares "hat matrix" (see Hoaglin and Welsch 1978) provides important diagnostic information about the influence of individual observations (y_i, x_i') on the associated prediction equation. The hat matrix corresponding to $\tilde{\beta}$ is defined to be

$$H(\lambda) = \{h_{ij}(\lambda)\} = XC(\lambda) \ . \tag{1.7}$$

This matrix transforms the response vector y to the vector of fitted values, $\tilde{y} = (\tilde{y}_1, \ldots, \tilde{y}_n)'$; i.e.,

$$\tilde{y} = H(\lambda)y \ .$$

The element $h_{ij}(\lambda)$ is a direct measure of the influence of y_j on the fit to y_i. In particular, the "leverage value" $h_{ii}(\lambda)$ measures the influence of y_i on its own prediction.

This study of the estimator class (1.3) begins with a derivation of some of the properties of $H(\lambda)$ in Section 2. In Section 3 techniques are presented for computing estimates and fitted values when observations are deleted from the data set. The results of this section are applied, in Section 4, to obtain approximate inference procedures for the parameter vector β and to derive diagnostic measures for detecting influential observations. Specific applications to nonparametric estimation by smoothing splines and to ridge regression estimators are detailed in Section 5. Concluding remarks are made in Section 6.

2. LEVERAGE VALUES FOR PENALIZED LEAST SQUARES

In this section certain properties of the hat matrix $H(\lambda)$ will be derived. It will be seen that the characteristics of its elements are closely related to those of the hat matrix H for the corresponding OLS estimator:

$$H = \{h_{ij}\} = X(X'X)^{-1}X' \ . \qquad (2.1)$$

Since H in equation (2.1) is a (orthogonal) projection operator, the following properties are easily proven:

$$\begin{aligned} &\text{i) } 0 \leq h_{ii} \leq 1 \\ &\text{ii) } -1 \leq h_{ij} \leq 1, \quad i \neq j \qquad (2.2) \\ &\text{iii) } h_{ij} = 1 \ \Rightarrow \ h_{ij} = 0, \quad i \neq j \ . \end{aligned}$$

When X contains a constant column, somewhat sharper results are provided by

$$\begin{aligned} &\text{i)' } \ n^{-1} \leq h_{ii} \leq 1 \\ &\text{ii)' } \ -(n-1)n^{-1} \leq h_{ij} \leq 1, \quad i \neq j \qquad (2.3) \\ &\text{iii)' } \ h_{ii} = 1 \iff h_{ij} = 0 \ , \quad i \neq j \ . \end{aligned}$$

Extreme rows of X result in large leverage values. The rough cutoff of $h_{ii} > 2p/n$ suggested by Hoaglin and Welsch (1978) is often used to identify such rows. Note from iii) and iii)' that, as $h_{ii} \to 1$, $h_{ij} \to 0$, $i \neq j$ and $\hat{y}_i = x_i'\hat{\beta} \to y_i$, indicating that an observation with a large leverage value will tend to dominate its own fit.

For $\lambda > 0$, $H(\lambda)$ is no longer a projection matrix. The following theorem establishes bounds for the elements of $H(\lambda)$ as a function of the corresponding elements of H, thereby providing an analog of properties i) and ii) in equation (2.2).

Theorem 2.1. The elements of $H(\lambda)$ satisfy

$$|h_{ij}(\lambda)| \leq (1 + \lambda d_1)^{-1}\{h_{ii}h_{jj}\}^{1/2} \qquad (2.4)$$

where d_1 is the smallest eigenvalue of $(X'X)^{-1}Q$.

Proof. Using the spectral decomposition (eg. Kshirsagar 1972, Chapter 7) of X write $X = UL^{1/2}Z'$, where $L = \text{diag}(\ell_1,\ldots,\ell_p)$ is a diagonal matrix containing the nonzero eigenvalues of XX' (and X'X) in ascending order, and $U = [u_1,\ldots,u_p]$ and Z are the corresponding matrices of eigenvectors of XX' and X'X, respectively. $H(\lambda)$ can now be expressed as

$$H(\lambda) = U(I + \lambda L^{-1/2}Z'QZL^{-1/2})^{-1}U' \ . \qquad (2.6)$$

Let $0 \leq d_1 \leq d_2 \leq \ldots \leq d_p$ denote the eigenvalues of $L^{-1/2}Z'QZL^{-1/2}$ (which are also the eigenvalues of $(X'X)^{-1}Q$). Using $\Gamma = [\gamma_1,\ldots,\gamma_p]$ to denote the corresponding matrix of eigenvectors, individual elements of $H(\lambda)$ can now be represented as

$$h_{ij}(\lambda) = \Sigma_{r=1}^{p} b_{ir}b_{jr}(1+\lambda d_r)^{-1}, \quad b_{kr} = u_k'\gamma_r. \qquad (2.7)$$

Application of the Cauchy-Schwarz inequality in equation (2.7) along with the ordering of the d_r completes the proof. □

Theorem 2.1 and its proof have several important consequences. First, it furnishes tighter bounds for the elements of $H(\lambda)$ than the inequalities in equation (2.2); i.e.,

$$\begin{aligned} &\text{i) } 0 \leq h_{ii}(\lambda) \leq (1 + \lambda d_1)^{-1} \\ &\text{ii) } -(1 + \lambda d_1)^{-1} \leq h_{ij}(\lambda) \leq (1 + \lambda d_1)^{-1},\ i \neq j \ . \end{aligned} \qquad (2.8)$$

In addition, from equation (2.7), it is apparent that $h_{ii}(\lambda)$ is monotonically decreasing with λ from $h_{ii}(0) = h_{ii}$ to $h_{ii}(\infty)$. Note that in general $h_{ii}(\infty) > 0$ unless $d_1 > 0$; when $d_1 > 0$, $h_{ii}(\infty) = 0$. Since $h_{ij}(\lambda)$ is continuous in λ, standard results from calculus can be used to show that for λ sufficiently small (large) $h_{ij}(\lambda)$ will have the same sign as h_{ij} $(h_{ij}(\infty))$ provided that $h_{ij} \neq 0$ $(h_{ij}(\infty)\neq 0)$.

Two important special cases occur when (i) $0 = d_1 = \dots = d_m < d_{m+1} \leq \dots \leq d_p$ and (ii) $Q = I$. These special cases have applications to smoothing splines and ridge regression, respectively, which will be explored in Section 5. The important details are summarized in the following two corollaries.

Corollary 1. Suppose $0 = d_1 = \dots = d_m < d_{m+1} \leq \dots \leq d_p$ and define $h_{ij}(\infty) = \Sigma_{r=1}^{m} b_{ir} b_{jr}$, where the b_{kr} are as in equation (2.7). Then

$$h_{ii}(\infty) + (1+\lambda d_p)^{-1} \sum_{r=m+1}^{p} b_{ir}^2 \leq h_{ii}(\lambda) \leq h_{ii}(\infty) + (1+\lambda d_{m+1})^{-1} \sum_{r=m+1}^{p} b_{ir}^2 . \tag{2.9}$$

Corollary 2. If $\tilde{\beta} = (X'X + \lambda I)^{-1} X'y$ then

$$|h_{ij}(\lambda)| \leq \ell_p (\ell_p + \lambda)^{-1} \{h_{ii} h_{jj}\}^{1/2}$$

where ℓ_p is the largest eigenvalue of $X'X$. The upper bound for the ith leverage value, viz. $\ell_p(\ell_p+\lambda)^{-1}$, is obtained when $x_j' = \ell_p^{1/2} z_p'$ where z_p is the eigenvector corresponding to ℓ_p.

3. DELETING OBSERVATIONS FROM AN ESTIMATOR

The development of exact tests and interval estimates for β using the penalized least squares estimator $\tilde{\beta}$ is a difficult, and as yet unresolved, problem. In contrast, approximate techniques based on nonparametric procedures such as the jackknife and bootstrap are easy to propose but their practicality depends on the ability to efficiently perform the necessary calculations. In this section a simple method of deleting observations from $\tilde{\beta}$ is derived which requires no refitting of the data. This is found, in Section 4, to make the use of inference techniques such as jackknife confidence regions for $\tilde{\beta}$ a practical alternative and to allow a generalization of several types of regression diagnostic measures to the penalized least squares setting.

For $q \leq n-p$ let $J = \{j_1,\ldots,j_q\}$ be a subset of the indices $\{1,\ldots,n\}$ and let $\tilde{\beta}^{(J)}$ represent the coefficient estimates obtained using only those (y_j, x_j') with $j \notin J$. The following theorem provides a partial characterization of $\tilde{\beta}^{(J)}$.

<u>Theorem 3.1</u>. Let $\tilde{\beta}^{[J]}(w_{j_1},\ldots,w_{j_q})$ solve

$$\min_{\beta}\{\sum_{j\notin J}(y_j-x_j'\beta)^2+\sum_{j\varepsilon J}(w_j-x_j'\beta)^2+\lambda\beta'Q\beta\} \qquad (3.1)$$

and define $\tilde{y}_i^{(J)} = x_i'\tilde{\beta}^{(J)}$, $i = 1,\ldots,n$. Then,

$$\tilde{\beta}^{[J]}(\tilde{y}_{j_1}^{(J)},\ldots,\tilde{y}_{j_q}^{(J)}) = \tilde{\beta}^{(J)} \ . \qquad (3.2)$$

Theorem 3.1 has the consequence that $\tilde{\beta}^{(J)}$ can be obtained by applying $C(\lambda)$ to a "new data vector" wherein y_j has been replaced by $\tilde{y}_j^{(J)}$ for all $j\varepsilon J$. This would seem to presuppose knowledge of $\tilde{\beta}^{(J)}$; however, such is not the case and in many cases of interest it is possible to compute the $\tilde{y}_j^{(J)}$ without explicit computation of $\tilde{\beta}^{(J)}$. This property follows by application of the next theorem.

Theorem 3.2. The values $\tilde{y}_j^{(J)}$, $j\varepsilon J$, satisfy the linear equation system

$$\tilde{y}_i^{(J)} - \Sigma_{j\varepsilon J} h_{ij}(\lambda)\tilde{y}_j^{(J)} = \tilde{y}_i - \Sigma_{j\varepsilon J} h_{ij}(\lambda) y_j$$

$$= \Sigma_{j\not\in J} h_{ij}(\lambda) y_j \ , \quad i \,\varepsilon\, J \ . \qquad (3.3)$$

Proof of Theorems 3.1-3.2. Set $w_j = \tilde{y}_j^{(J)}$. Proof of Theorem 3.1 is provided by the following inequalities:

$$\Sigma_{j\not\in J}(y_j - x_j'\tilde{\beta}^{(J)})^2 + \Sigma_{j\varepsilon J}(w_j - x_j'\tilde{\beta}^{(J)})^2 + \lambda\tilde{\beta}^{(J)'}Q\tilde{\beta}^{(J)}$$

$$= \Sigma_{j\not\in J}(y_j - x_j'\tilde{\beta}^{(J)})^2 + \lambda\tilde{\beta}^{(J)'}Q\tilde{\beta}^{(J)} \leq \Sigma_{j\not\in J}(y_j - x_j'\beta)^2 + \lambda\beta'Q\beta$$

$$\leq \Sigma_{j\not\in J}(y_j - x_j'\beta)^2 + \Sigma_{j\varepsilon J}(w_j - x_j'\beta)^2 + \lambda\beta'Q\beta \ .$$

To verify equation (3.3) note that $x_i'\tilde{\beta}^{[J]}(w_{j_1},\ldots,w_{j_q})$ is linear in w_j, $j\varepsilon J$, and can, therefore, be written as

$$x_i'\tilde{\beta}^{[J]}(w_{j_1},\ldots,w_{j_q}) = x_i'\tilde{\beta} + \Sigma_{j\varepsilon J} h_{ij}(\lambda)(w_j - y_j) \ . \qquad (3.4)$$

Letting $w_j = x_j'\tilde{\beta}^{(J)}$ gives the desired result. □

To illustrate uses for Theorem 3.1-3.2 confine attention, for the moment, to the instance $q = 1$, $J = \{j\}$ for some $j \,\varepsilon\, \{1,\ldots,n\}$. To distinguish this important special case the notation

$$\tilde{\beta}^{[j]} = \tilde{\beta}^{(J)} \qquad (3.5)$$

and

$$\tilde{y}_i^{[j]} = x_i'\tilde{\beta}^{[j]} \qquad (3.6)$$

is utilized. Application of Theorem 3.2 to this special case yields the following expression for $\tilde{y}_j^{[j]}$:

$$\tilde{y}_j^{[j]} = (\tilde{y}_j - h_{jj}(\lambda)y_j)/(1-h_{jj}(\lambda)). \qquad (3.7)$$

This relationship explicitly demonstrates the ability to obtain each of the $\tilde{y}_j^{[j]}$ without refitting the model.

The term "deleted residual" will be used to designate the difference $y_j - \tilde{y}_j^{[j]}$. Equation (3.7) provides an efficient computational form for the deleted residual; viz.,

$$e_{[j]} \equiv y_j - \tilde{y}_j^{[j]} = e_j/(1-h_{jj}(\lambda)), \quad j = 1,\ldots,n, \quad (3.8)$$

where e_j is the jth residual from the fit to the entire data set:

$$e_j = y_j - \tilde{y}_j, \quad j = 1,\ldots,n. \quad (3.9)$$

Substituting equation (3.8) into equation (3.2) yields

$$\tilde{\beta}^{[j]} = \tilde{\beta} - c_j(\lambda)e_{[j]}, \quad j = 1,\ldots,n, \quad (3.10)$$

where $c_j(\lambda)$ is the jth column of $C(\lambda)$.

Formulas (3.8) and (3.10) include as special cases the equivalent expressions for ordinary least squares, $\lambda = 0$ (e.g., Beckman and Trussel 1974; Hoaglin and Welsch 1978). In the case of smoothing splines equation (3.8) was established by Craven and Wahba (1979) using a method of proof similar to the one employed here.

4. INFERENCE AND DIAGNOSTICS

Equation (3.8) provides a fundamental expression for the derivation of approximate confidence intervals to complement the point estimator $\tilde{\beta}$. Define the jth vector of pseudo-values by

$$\begin{aligned}\tilde{b}_j &= n\tilde{\beta} - (n-1)\tilde{\beta}^{[j]} \\ &= \tilde{\beta} + (n-1)c_j(\lambda)e_{[j]} \ . \quad (4.1)\end{aligned}$$

Then the jackknife estimator of β based on $\tilde{\beta}$ is $\tilde{b} = n^{-1}\Sigma_{j=1}^{n}\tilde{b}_j$

and the variance-covariance matrix of $\tilde{\beta}$ or $\tilde{b}$ can be estimated by

$$\tilde{V} = \Sigma_{j=1}^{n}(\tilde{b}_j - \tilde{b})(\tilde{b}_j - \tilde{b})'/n(n-1) \ . \tag{4.2}$$

For a linear functional $a'\beta$, an approximate $100(1-\alpha)\%$ confidence interval is provided by

$$a'\tilde{\beta} \pm Z_{\alpha/2}(a'\tilde{V}a)^{1/2} \quad \text{or} \quad a'\tilde{b} \pm Z_{\alpha/2}(a'\tilde{V}a)^{1/2} \tag{4.3}$$

where $Z_{\alpha/2}$ is the $100(1-\alpha/2)$ percentage point of the standard normal distribution (critical values from a Student's t distribution with n-1 degrees of freedom could be used in place of $Z_{\alpha/2}$ in expression (4.3)). Notice that the interval estimates (4.3) can be computed using information available entirely from the original fit. When $\lambda = 0$, equations (4.1)-(4.2) reduce to formulae given in Miller (1974), Hinkley (1977a), and Efron (1982, Chapter 3) for jackknifing $\hat{\beta}$.

Diagnostic measures which parallel those utilized for ordinary least squares can also be derived as a result of (3.8) and (3.10). To do so first note that a natural estimator of σ^2 associated with the penalized least squares estimator $\tilde{\beta}$ is

$$\tilde{\sigma}^2 = \Sigma_{i=1}^{n}e_i^2/tr(I-H(\lambda)) \tag{4.4}$$

where tr denotes the matrix trace. This estimator reduces to the usual estimator of σ^2 associated with $\hat{\beta}$, namely $\hat{\sigma}^2 = \Sigma_{i=1}^{n}e_i^2/(n-p)$, when $\lambda = 0$. The estimator (4.4) has been found to be quite effective for spline smoothing by Wahba (1983). Studentized (deleted) residuals can then be defined as

$$t_{[j]} = e_j/\tilde{\sigma}_{[j]}(1-h_{jj}(\lambda))^{1/2} \tag{4.5}$$

where $\tilde{\sigma}_{[j]}^2$ is the estimator (4.4) computed from the reduced data set wherein (y_j, x_j') has been excluded. An explicit formula for $\tilde{\sigma}^2_{[j]}$ is

$$\tilde{\sigma}^2_{[j]} = \sum_{\substack{i=1\\ i\neq j}}^{n} (e_i + h_{ij}(\lambda)e_{[j]})^2/\mathrm{tr}(I-H^{[j]}(\lambda)) \qquad (4.6)$$

with

$$\mathrm{tr}(I-H^{[j]}(\lambda)) = n-1 - \sum_{\substack{i=1\\ i\neq j}}^{n} [h_{ii}(\lambda) + h_{ij}(\lambda)^2/(1-h_{jj}(\lambda))]. \quad (4.7)$$

To prove formulas (4.6)-(4.7) observe that $\tilde{y}_i^{[j]}$ can be written as $\Sigma_{r\neq j} a_{ir} y_r$. The coefficients a_{ir} can be deduced from equation (3.2) and used to establish equation (4.7). The form of the numerator follows easily from expression (3.10).

The studentized residuals along with formulas (4.6)-(4.7) are generalizations of relations which hold when $\lambda = 0$ (e.g., Gunst and Mason 1980, Chapter 7). These residuals provide a scaled measure of how the fit to y_j changes when its value is not used to estimate β. They can, therefore, be used to detect overly influential data values. The value of $t_{[j]}$ might be compared to values from a Student's t distribution with approximately $\mathrm{tr}(I-H^{[j]}(\lambda))$ degrees of freedom. Simulation results discussed in Section 5 indicate that Student's t critical values provide a reasonably good approximation for 5% cutoff values for the $t_{[j]}$. Through similar considerations a variety of other diagnostic measures can also be suggested. One such example is

$$\mathrm{DFFITS}_j = (\tilde{y}_j - x_j'\tilde{\beta}^{[j]})/\tilde{\sigma}_{[j]}\cdot h_{jj}(\lambda)^{1/2}$$

$$= [h_{jj}(\lambda)/(1-h_{jj}(\lambda))]^{1/2} t_{[j]}\ , \ j = 1,\ldots,n,$$

(see Velleman and Welsch 1981 or Belsley, Kuh and Welsch 1980).

Deleting $q \geq 2$ observations is somewhat more complicated than the case $q = 1$. When $q \geq 2$ it is no longer obvious that equations (3.3) always uniquely determine the $\tilde{y}_j^{(J)}$. This will be true if and only if $(I-H(\lambda))_J$, the submatrix of $I-H(\lambda)$ corresponding to those indices in J, has rank q. For example, when $q = 2$, $J = \{i,j\}$ this condition is equivalent to $(1-h_{ii}(\lambda))(1-h_{jj}(\lambda)) - h_{ij}(\lambda)^2 \neq 0$. Instances where this is not satisfied would seem rare in practice.

Now suppose that one obtains m random samples of q indices each, $J_1,\ldots,J_m$, by sampling with replacement from $\{1,\ldots,n\}$. A bootstrap estimator of the variance-covariance matrix of $\tilde{\beta}$ is provided by

$$\tilde{W} = \Sigma_{r=1}^{m}(\tilde{\beta}^{(J_r)} - \tilde{\beta}^*)(\tilde{\beta}^{(J_r)} - \tilde{\beta}^*)'/(m-1) \tag{4.8}$$

where $\tilde{\beta}^* = m^{-1}\Sigma_{r=1}^{m}\tilde{\beta}^{(J_r)}$. If the matrices $(I-H(\lambda))_{J_r}$ all have rank q, W can be computed using equations (3.2)-(3.3) and its elements can then be used to obtain bootstrap analogs of the jackknife confidence intervals (4.3). A similar approach when all possible subsets of size q are used leads to the development of grouped jackknife interval estimates of β (see Efron 1982, Chapter 2).

To conclude note that when $\lambda = 0$ Theorems 3.1 - 3.2 can be used to establish "leave-q-out" identities such as equation (7) of Draper and John (1981). It is, therefore, possible to generalize leave-q-out diagnostics such as those discussed in Gentleman and Wilk (1975a, b) and Draper and John (1978, 1981) to the case of penalized least squares estimation.

5. EXAMPLES

In this section the application of results in Sections 3 and 4 to the special cases of smoothing splines and ridge regression will be illustrated.

5.1 Smoothing Splines

Suppose η is a smooth response function and that $\eta_j = \eta(t_j)$, $0 \leq t_1 < \dots < t_n \leq 1$, in model (1.1). For $n \geq m$ the smoothing spline estimator of η, denoted by $\tilde{\eta}$, is obtained by minimizing

$$\Sigma_{j=1}^{n}(y_j - f(t_j))^2 + \lambda\int_0^1 f^{(m)}(t)^2 dt \tag{5.1}$$

over all functions f having m-1 absolutely continuous derivatives and a square integrable mth derivative. Schoenberg (1964) proposed this type of nonparametric estimator for η and showed that $\tilde{\eta}$ was a spline function of order 2m with knots at the t_j. General discussions of smoothing splines can be found in Wahba (1977), Wegman and Wright (1983) and Eubank (1983).

Demmler and Reinsch (1975) (see also Speckman 1983) develop a basis for spline smoothing which consists of functions $x_1, \dots, x_n$ and constants $0 = q_1 = \dots = q_m < q_{m+1} < \dots < q_n$ which satisfy

$$\Sigma_{r=1}^{n} x_i(t_r)x_j(t_r) = \delta_{ij} \tag{5.2}$$

and

$$\int_0^1 x_i^{(m)}(t)x_j^{(m)}(t)dt = q_j\delta_{ij} \ , \tag{5.3}$$

where δ_{ij} is the Kronecker delta. They show that the minimizer of criterion (5.1) is necessarily of the form

$$f(t) = \Sigma_{j=1}^{n}\beta_j x_j(t) \quad ; \tag{5.4}$$

hence, it sufficies to minimize criterion (5.1) over functions of this type. Substituting f(t) from (5.4) into (5.1) and invoking the relationships in equation (5.3) gives the equivalent criterion

$$\min_{\beta}\{\Sigma_{j=1}^{n}(y_j-\Sigma_{r=1}^{n}\beta_r x_r(t_j))^2 + \lambda\Sigma_{j=1}^{n}\beta_j^2 q_j\} \ . \qquad (5.5)$$

Comparison with (1.2) reveals this to be a special case of penalized least squares estimation with $p = n$, $x_j' = (x_1(t_j),\ldots,x_n(t_j))$ and $Q = \mathrm{diag}\ (q_1,\ldots,q_n)$. Therefore,

$$\tilde{\beta} = D(\lambda)X'y \qquad (5.6)$$

where $D(\lambda) = \mathrm{diag}((1 + \lambda q_1)^{-1},\ldots,(1 + \lambda q_n)^{-1})$.

The hat matrix corresponding to the estimator (5.6) is $H(\lambda) = XD(\lambda)\hat{X}'$; moreover, since $X'X = I$ the eigenvalues of $(X'X)^{-1}Q$ are simply the q_j. Applying Corollary 1 of Section 2 the following bounds are obtained for $h_{ii}(\lambda)$:

$$h_{ii}(\infty) + (1 + \lambda q_n)^{-1}\Sigma_{r=m+1}^{n}x_r(t_i)^2 \leq h_{ii}(\lambda) \leq h_{ii}(\infty)$$

$$+ (1 + \lambda q_{m+1})^{-1}\ \Sigma_{r=m+1}^{n}x_r(t_i)^2\ , \qquad (5.7)$$

where $h_{ii}(\infty) = \Sigma_{r=1}^{m}x_r(t_i)^2$. It follows from Demmler and Reinsch (1975) that $h_{ii}(\infty)$ is the ith leverage value for regression on polynomials of order m. Equation (5.7) therefore establishes a connection between the leverage values for spline smoothing and those for polynomial regression. These results generalize to multivariate "Thin Plate" or Laplacian smoothing splines (e.g., Wahba 1981; Wahba and Wendleberger 1980; and Wendelberger 1981) where the $h_{ii}(\lambda)$ may be partilarly useful in the detection of sensitive points in the design.

To illustrate the behaviour of some of the diagnostic and inferential methods proposed in Section 4, a small scale simulation was conducted. Data sets were generated from model (1.1) with

$$\eta_i = \eta(t_i) = 4.26\{\exp(-3.25t_i)-4\exp(-6.5t_i)+3\exp(-9.75t_i)\},$$
$$t_i = (i-1)/n, \qquad n = 80 ,$$

and normal errors with σ values of .05, .1, .2 and .4. This function is a rescaled version of one studied by Wahba and Wold (1975). The basic experiment was replicated $r = 50$ times (i.e., 50 data sets of size 80) with each replicate being "treated" by all four values of σ. A cubic smoothing spline ($m = 2$) was fitted to each data set with λ selected via GCV.

Approximate 95% jackknife confidence intervals for the η_i, centered at $\tilde{\eta}_i$, were computed by taking $a_i' = (x_1(t_i),\ldots,x_n(t_i))$ in equation (4.3). The proportion of times the true function value was contained in its interval estimate was recorded along with the value of $\tilde{\sigma}^2$ and the proportion of times $|t_{[j]}|$ exceeded the 5% (two-tailed) critical value for the Student t distribution. Summary statistics for the simulation are given in Table 1. A typical example of these results, for $\sigma = .1$, appears in Figure 1.

[Insert Table 1, Figure 1]

The empirical confidence levels in Table 1 are somewhat lower than might be desired. However, by using 99% rather than 95% intervals, confidence levels in excess of 94% were obtained in all cases. This is typical of simulations performed with other function types and other configurations for the values of r, n, and σ. These results will appear elsewhere. As illustrated in Table 1, the

Student's t approximation to $t_{[j]}$ and the estimator $\tilde{\sigma}^2$ performed well.

5.2 Ridge Regression

Ridge regression estimators (Hoerl and Kennard 1970; Marquardt 1970) are solutions to the criterion function (1.2) when (i) only the nonconstant predictor variables from model (1.1) are included in X, (ii) the predictor variables are standardized so that X'X is in correlation form, and (iii) $Q = I$. Much controversy persists over automated selection of λ, the effect of standardization on ridge estimation, and the assumptions underlying the validity of the ridge estimator (e.g., Draper and Van Nostrand 1979; Smith and Campbell 1980, with discussion). In order to demonstrate the application of the results of Section 2-4, assume that for a specific regression analysis the criticisms noted above are satisfactorily answered and that a ridge regression analysis is deemed appropriate.

Ridge regression diagnostics can be obtained from the results of Sections 2-4 under the conditions stated above; however, the efficient computational expressions for deleted estimators (i.e., $\tilde{\beta}^{[j]}$ and $\tilde{\beta}^{(J)}$) and deleted residuals (i.e., $e_{[j]}$) are exact only if the reduced X matrix is not restandardized when rows are deleted. Hinkley (1977a) noted a similar restriction when he cautioned against obtaining (least squares) jackknife estimates of the constant term of a regression model using centered predictor variables. Since the major benefits of centering and standardization cited by Marquardt (1980) are essentially maintained when one (or a small number) of the rows of the standardized X matrix is (are) deleted, only the original matrix of predictor variables is standardized in the following example.

Gunst and Mason (1980, Appendix A) contains a data set on the gross national product (GNP) of 49 countries of the world along with the six additional socioeconomic indices: an infant death rate (INFD), a physician/population ratio (PHYS), population density (DENS), population density measured in terms of agricultural land area (AGDS), a literacy measure (LIT), and an index of higher education (HIED). Table 2 displays regression diagnostics for the fit of $\ell n(\text{GNP})$ by the six socioeconomic indices.

[Insert Table 2]

The relatively small value of $\lambda(0.08)$ which was chosen for this illustration has little effect on the bounds for ridge leverage values given by Corollary 2 since $\ell_6/(\ell_6+0.08)=0.97$. With the exception of Malta, least squares leverage values which exceed $2(p+1)/n = 0.286$ are also large with the ridge estimator using the analogous bound $2(\text{tr}[H(\lambda)]+1)/n = 0.271$. Although the ridge DFFITS values appear to be slightly more uniform than those of least squares (e.g., none of the former are greater than 1.0 in magnitude), four of the five observations which exceed $2\{(p+1)/n\}^{1/2}=0.756$ for least squares also exceed $2\{(\text{tr}[H(\lambda)]+1)/n\}^{1/2}=0.736$ for ridge regression--Malta is again the exception--and a similar comment can be made about the $t_{[j]}$.

Malta is obviously affecting the two estimation procedures differently. It has high leverage and is influential on the least squares fit but has neither high leverage nor an influential impact on the ridge regression fit. A scatterplot of DENS and AGDS reveals that Malta lies well off the concentrated linear scatter (r = 0.97) between these two variates. Thus by lessening the effect of the strong pairwise correlation between DENS and AGDS on the estimation of the regression coefficients, the ridge estimator is also lessening the influence of Malta on the fit. Although the other least squares and ridge diagnostics identify equally important characteristics of this data set, comparison of the two sets of diagnostics has provided important insight about Malta which might have gone unappreciated had only the least squares diagnostics been examined.

Table 3 displays least squares, ridge ($\lambda = .08$), and jackknifed ridge ($\tilde{b}$) coefficient estimates and confidence intervals. The purpose of presenting the ridge and jackknifed ridge estimates is to highlight typical characteristics of these estimators, not to draw definitive conclusions relative to this data set. Note in particular that, while similar, the ridge and jackknifed ridge estimates are somewhat different. In addition, both of these latter two estimators produce jackknife confidence intervals (using expressions (4.3)) which are shorter than least squares. In view of the simulation results in Section 5.1, it might be advisable to adjust these confidence intervals (not done here) by using a larger Student t critical point. If one uses 99% nominal coverage, the ridge confidence interval for the coefficient of DENS includes the origin.

Obviously a more complete analysis of this data set is needed in order to resolve questions which remain about influential observations and the significance of the predictor variables. Any thorough analysis must incorporate prior knowledge about the regression coefficients and information concerning the intended use of the conclusions which are to be drawn from the fitted model. These topics are beyond the scope of this paper; nevertheless, this example illustrates some important characteristics of penalized least squares diagnostics and approximate inference procedures.

6. CONCLUDING REMARKS

The results of this paper generalize least squares regression diagnostics and certain approximate inference procedures to a class of (quadratic) penalized least squares estimators for linear models. Theorems 3.1 and 3.2 produce expressions for deleted estimators and residuals which provide exact, computationally efficient, calculation of quantities such as pseudo values and Studentized residuals. These results have wide application, two specific illustrations being nonparametric estimation with smoothing splines and ridge regression.

Much research remains to be conducted regarding the properties and usage of the procedures proposed in this paper. For example, the jackknife confidence intervals do not achieve the nominal confidence level, although they are well-known to be insensitive to a variety of unimodal error distributions. Corrections for the jackknife such as those proposed in Hinkley (1977b, 1978) may alleviate coverage difficulties and the behavior of jackknife intervals under nonnormal errors merits further investigation. Likewise, the sensitivity of jackknife confidence intervals to the choice of λ warrants further study. For instance, in the ridge regression example increasing λ from 0.08 to 0.20 decreases the estimated standard errors of the individual coefficients between 5 percent (HIED) and 50 percent (AGDS). On the other hand, the Studentized residuals and the estimator of σ^2 performed well in the simulation in Section 5.1. Similarly, the ridge regression diagnostics highlighted an important characteristic of the presence of Malta which could have been overlooked if only the least squares diagnostics were examined.

REFERENCES

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.

BELSLEY, D. A., KUH, E. and WELSCH, R. E. (1980), Regression Diagnostics, New York: John Wiley and Sons, Inc.

CRAVEN, P. and WAHBA, G. (1979), "Smoothing Noisy Data with Spline Functions," Numerische Mathematik, 31, 377-403.

DEMMLER, A. and REINSCH, C. (1975), "Oscillation Matrices with Spline Smoothing," Numerische Mathematik, 24, 375-382.

DRAPER, N. and JOHN, J. A. (1978), "On Testing for Two Outliers or One Outlier in Two-Way Tables," Technometrics, 20, 69-78.

DRAPER, N. R. and JOHN, J. A. (1981), "Influential Observations and Outliers in Regression," Technometrics, 23, 21-26.

DRAPER, N. R. and VAN NOSTRAND, R. C. (1979), "Ridge Regression and James-Stein Estimation: Review and Comments," Technometrics, 21, 451-466.

EFRON, B. (1982), The Jackknife, the Bootstrap, and Other Resampling Plans, Philadelphia: Society for Industrial and Applied Mathematics, Monograph No. 38.

EUBANK, R. L. (1983), "Approximate Regression Models and Splines," Communications in Statistics: Statistical Reviews (to appear).

GENTLEMAN, J. F. and WILK, M. B. (1975a), "Detecting Outliers in a Two-Way Table: I. Statistical Behavior of Residuals," Technometrics, 17, 1-14.

GENTLEMAN, J. F. and WILK, M. B. (1975b), "Detecting Outliers. II. Supplementing the Direct Analysis of Residuals," Biometrics, 31, 387-410.

GOLUB, G. H., HEATH, M., and WAHBA, G. (1979), "Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter," Technometrics, 21, 215-223.

GUNST, R. F. and MASON, R. L. (1980), Regression Analysis and Its Application: A Data-Oriented Approach, New York: Marcel-Dekker, Inc.

HINKLEY, D. V. (1977a), "Jackknifing in Unbalanced Situations," Technometrics, 19, 285-292.

HINKLEY, D. V. (1977b), "Jackknife Confidence Limits Using Student t Approximations," Biometrika, 64, 21-28.

HINKLEY, D. V. (1978), "Improving the Jackknife with Special Reference to Correlation Estimation," Biometrika, 65, 13-21.

HOAGLIN, D. C. and WELSCH, R. E. (1978), "The Hat Matrix in Regression and ANOVA," American Statistician, 32, 17-22.

HOERL, A. E. and KENNARD, R. W. (1970), "Ridge Regression: Biased Estimation for Nonorthogonal Problems," Technometrics, 12, 55-67.

KSHIRSAGAR, A. M. (1972), Multivariate Analysis, New York: Marcel Dekker, Inc.

MARQUARDT, D. W. (1970), "Generalized Inverses, Ridge Regression, Biased Linear Estimation and Nonlinear Estimation," Technometrics, 12, 591-612.

MARQUARDT, D. W. (1980). "You Should Standardize the Predictor Variables in Your Regression Models," Journal of the American Statistical Association, 75, 87-91.

MILLER, R. G. (1974), "An Unbalanced Jackknife," Annals of Statistics, 2, 880-891.

PREGIBON, D. (1981), "Logistic Regression Diagnostics," The Annals of Statistics, 9, 705-724.

SCHOENBERG, I. J. (1964), "Spline Functions and the Problem of Graduation," Proceedings of the National Academy of Sciences (USA), 52, 947-950.

SMITH, G. and CAMPBELL F. (1980), "A Critique of Some Ridge Regression Methods," Journal of the American Statistical Association, 75, 74-81.

SPECKMAN, P. (1983), "The Asymptotic Integrated Error for Smoothing Noisy Data by Splines," Numerische Mathematik (to appear).

VELLEMAN, P. and WELSCH, R. E. (1981), "Efficient Computing of Regression Diagnostics," The American Statistician, 35, 234-242.

WAHBA, G. (1977), "A Survey of Some Smoothing Problems and the Method of Generalized Cross-Validation for Solving Them," In Applications of Statistics (P. R. Krishnaiah, ed.) 507-523. Amsterdam: North Holland.

WAHBA, G. (1981), "Numerical Experiments with the Thin Plate Histospline," Communications in Statistics, A10, 2475-2514.

WAHBA, G. (1983), "Bayesian Confidence Intervals for the Cross-Validated Smoothing Spline," Journal of the Royal Statistical Society, Series B, 45, 133-150.

WAHBA, G. and WENDELBERGER, J. (1980), "Some New Mathematical Methods for Variational Objective Analysis Using Splines and Cross Validation," Monthly Weather Review, 108, 1122-1143.

WAHBA, G. and WOLD, S. (1975), "A Completely Automatic French Curve: Fitting Spline Functions by Cross Validation," Communications in Statistics, A4, 1-17.

WEGMAN, E. J. and WRIGHT, I. W. (1983), "Splines in Statistics," Journal of the American Statistical Association, 78, 351-365.

WENDELBERGER, J. (1981), "The Computation of Laplacian Smoothing Splines with Examples," Technical Report No. 648, Department of Statistics, University of Wisconsin-Madison.

TABLE AND FIGURE TITLES

Tables

Figure

TABLE 1. Summary Statistics for the Simulation

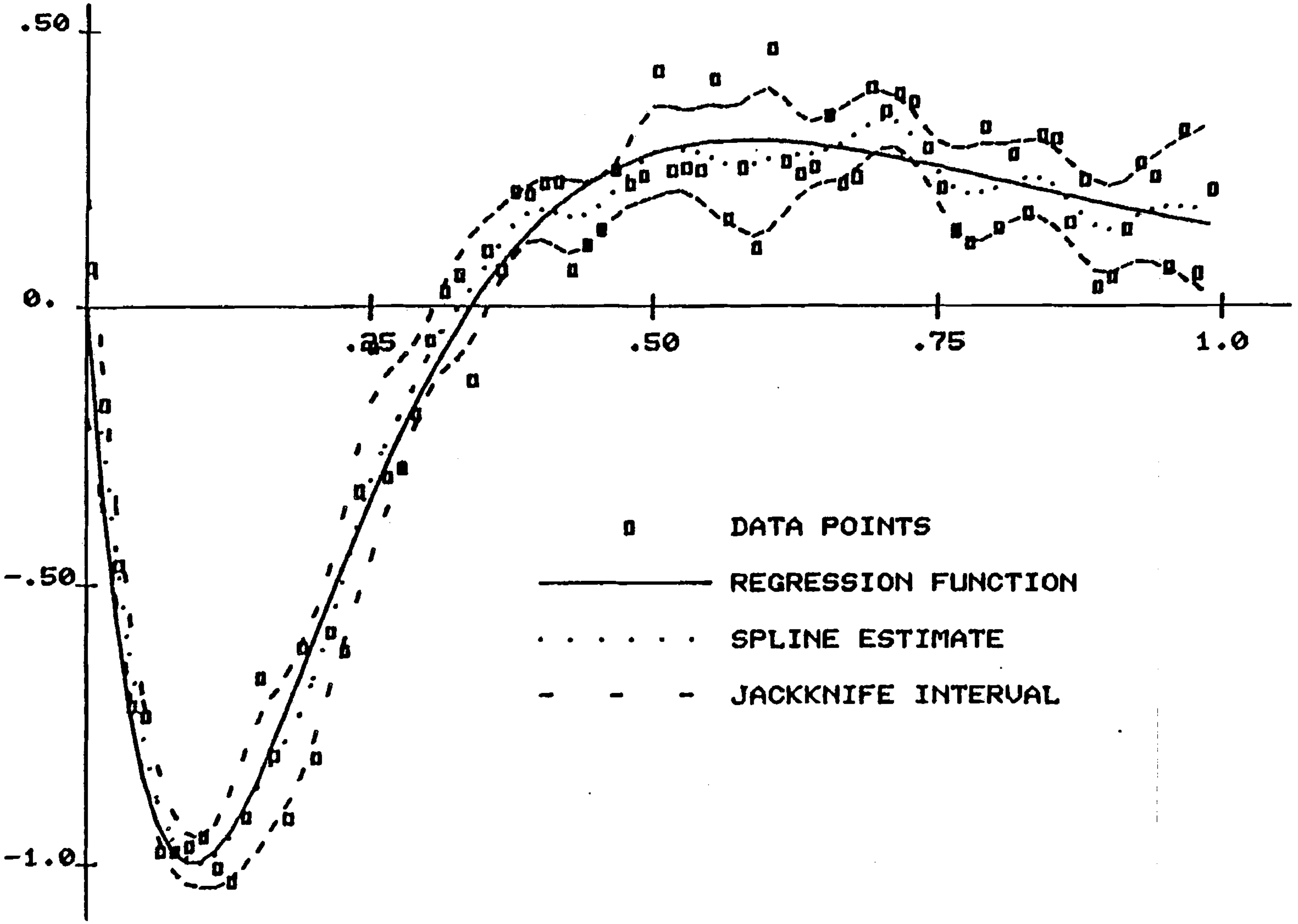
	Empirical Confidence Levels		Empirical Significance Levels		Estimated Variance	
σ	Average	Std. Error	Average	Std. Error	Avg.	MSE
.05	.8838	.0084	.0508	.0025	.0023	$2x10^{-7}$
.10	.8868	.0087	.0510	.0024	.0091	$3x10^{-6}$
.20	.8863	.0102	.0493	.0021	.0366	$5x10^{-5}$
.40	.8843	.0149	.0490	.0023	.1490	$6x10^{-4}$

TABLE 2. Regression Diagnostics for GNP Data, Selected Observations

	Least Squares			Ridge (λ=.08)		
Obsn.	h_{jj}	$t_{[j]}$	DFFITS$_j$	$h_{jj}(.08)$	$t_{[j]}$	DFFITS$_j$
BARBADOS	.238	-2.026	-1.131	.137	-1.929	-.769
CANADA	.042	2.011	.419	.039	2.111	.423
HONG KONG	.511	-.107	-.109	.471	-.138	-.130
INDIA	.558	1.337	1.502	.507	.903	.917
JAPAN	.049	-2.799	-.633	.046	-2.743	-.602
LUXEMBOURG	.084	2.356	.713	.077	2.391	.690
MALTA	.688	1.506	2.236	.262	.426	.254
SINGAPORE	.632	.562	.736	.516	.632	.653
TAIWAN	.178	-2.401	-1.119	.129	-2.475	-.953
U.S.	.490	.804	.787	.447	.951	.855

TABLE 3. Coefficient Estimates and Nominal 95% (Individual) Confidence Intervals

Predictor Variable	Least Squares Estimates	Ridge Regression (λ = .08)	Jackknifed Ridge
	(a) Coefficient Estimates		
INFD	-1.870	-1.772	-1.695
PHYS	.171	- .125	.113
DENS	-1.094	- .410	- .606
AGDS	.862	.151	.453
LIT	2.298	1.985	2.163
HIED	1.454	1.411	1.662
	(b) Confidence Intervals		
INFD	(-3.012,- .729)	(-2.218,-1.326)	(-2.142,-1.250)
PHYS	(-1.192, 1.535)	(- .524, .274)	(- .286, .512)
DENS	(-4.718, 2.530)	(- .767,- .053)	(- .963,- .249)
AGDS	(-2.738, 4.462)	(- .188, .490)	(.114, .792)
LIT	(.748 3.848)	(1.408, 2.562)	(1.586, 2.740)
HIED	(.528, 2.380)	(.994, 1.828)	(1.245, 2.079)


.50
0.
-.50
-1.0
.25
.50
.75
1.0
DATA POINTS
REGRESSION FUNCTION
SPLINE ESTIMATE
JACKKNIFE INTERVAL

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER SMU/DS/TR-181	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) REGRESSION DIAGNOSTICS AND APPROXIMATE INFERENCE PROCEDURES FOR PEANLIZED LEAST SQUARES ESTIMATORS		5. TYPE OF REPORT & PERIOD COVERED Technical Report
		6. PERFORMING ORG. REPORT NUMBER SMU/DS/TR-181
7. AUTHOR(s) Richard F. Gunst and Randall L. Eubank		8. CONTRACT OR GRANT NUMBER(s) NASA NCC9-9
9. PERFORMING ORGANIZATION NAME AND ADDRESS Southern Methodist University Dallas, Texas 75275		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS National Aeronautics and Space Administration Houston, Texas		12. REPORT DATE September 1983
		13. NUMBER OF PAGES 28
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report)
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

This document has been approved for public release and sale; its distribution is unlimited. Reproduction in whole or in part is permitted for any putpose of The United States Government.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Bootstrap confidence intervals, Jackknife confidence intervals; Leverage values; Ridge regression; Smoothing Splines; Studentized residuals.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Generalizations of least squares diagnostic techniques are presented for a class of penalized least squares estimators. Efficient computation of these diagnostics is afforded by expressions which relate coefficient estimates and residuals from fits to subsets of the data to the corresponding quantities from a fit to the complete data set. From these expressions approximate confidence intervals and test statistics can be obtained using jackknife and bootstrap procedures. Applications are discussed for the special cases of smoothing splines and ridge regression.