Diagnostics for Penalized Least-Squares Estimators

by
R. L. Eubank and R. F. Gunst

Technical Report No. SMU-DS-TR-196
Department of Statistics ONR Contract

December 1985

Research sponsored by the Office of Naval Research
Contract N00014-85-K-0340

Department of Statistics
Southern Methodist University
Dallas, Texas 75275

Diagnostics for Penalized Least-Squares Estimators

R. L. Eubank and R. F. Gunst[1]

Summary. Diagnostic methods for a class of penalized least-squares estimators are derived from a Bayesian perspective. The class of estimators considered includes generalized ridge estimators, partial splines and thin plate smoothing splines. The proposed diagnostics include scaled residuals, leverage values and various measures of influence.

Key words and phrases: Bayes estimators, influence, leverage, partial splines, residuals, ridge regression, smoothing splines, thin plate splines.

AMS 1980 subject classification: Primary 62J99, Secondary 62J07, 62G05, 65D07.

[1] The authors' research was supported by the Office of Naval Research, under Contract No. N00014-82-K-0207, and by the National Aeronautics and Space Administration under Contract No. NCC 9-9.

1. **Introduction.** The development of diagnostic methods to accompany statistical estimation techniques represents an important area of statistical research. A focal point for the study of such techniques is least-squares estimators for linear regression models for which a plethora of diagnostic measures are now available. The objective of this note is to show that there are parallels of the diagnostic indicators currently used for linear regression models which are applicable to a more general class of estimators that includes ordinary least-squares estimators, ridge regression estimators, and several variants of smoothing splines.

The estimators to be considered can be described as follows. Let $(\underline{u}_1', y_1)$, ..., $(\underline{u}_n', y_n)$ represent n observations on a response variable y and a q-vector of independent variables $\underline{u}' = (u_1, \ldots, u_q)$. The y_i and $\underline{u}_i$ are related by

$$y_i = \mu(\underline{u}_i) + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{1}$$

where μ is an unknown regression function and the ε_i are zero mean, uncorrelated, random errors having common variance σ^2. The objective is to estimate μ. To do so we suppose that associated with each $\underline{u}_i$ there is a known p-vector, $\underline{x}_i$, and coefficients, $\underline{\beta}' = (\beta_1, \ldots, \beta_p)$, such that $\underline{x}_i'\underline{\beta}$ either equals or represents an approximation to $\mu(\underline{u}_i)$. Penalized least-squares estimators are then obtained by minimizing the criterion

$$n^{-1} \sum_{i=1}^{n} (y_i - \underline{x}_i'\underline{\beta})^2 + \lambda \underline{\beta}' G \underline{\beta}, \qquad \lambda \geq 0, \tag{2}$$

for some specified positive semi-definite matrix G. If $X' = [\underline{x}_1, \ldots, \underline{x}_n]$ has full row rank the unique solution is

$$\tilde{\underline{\beta}}(\lambda) = C(\lambda)\underline{y} \tag{3}$$

with

$$C(\lambda) = (X'X + n\lambda G)^{-1}X'$$

and $\underline{y}' = (y_1, \ldots, y_n)$.

The estimation framework from which (2) and (3) derive is sufficiently general to encompass many estimators used in practice. For example, if $\mu(\underline{u}_i) = \underline{u}_i'\underline{\beta}$ and $\underline{x}_i = \underline{u}_i$, (3) is a generalized ridge estimator. Thus, the estimators under consideration also include ordinary least-squares estimators for linear models. Other examples are provided by several variants of smoothing splines. The latter case is somewhat complicated and, therefore, the details are relegated to Section 5. The reader wishing more motivation may wish to glance ahead to that section.

To develop diagnostics appropriate for use with $\tilde{\underline{\beta}}(\lambda)$ we propose a Bayesian approach which begins by recognizing that the estimator is a Bayes estimator (see, e.g., Lindley and Smith 1982 and Leamer 1973) when, conditional on $\underline{\beta}$, $\underline{y} \sim N_n(X\underline{\beta}, \sigma^2 I)$, and the log of the prior density for $\underline{\beta}$ is, apart from an additive constant, $- n\lambda\underline{\beta}'G\underline{\beta}/2\sigma^2$. (Note that G may be singular in which case the prior is partially improper.) Thus $\tilde{\underline{\beta}}(\lambda)$ is the posterior mean of $\underline{\beta}$ for this model and

$$V(\underline{\beta}|\underline{y}) = E[(\underline{\beta} - \tilde{\underline{\beta}}(\lambda))(\underline{\beta} - \tilde{\underline{\beta}}(\lambda))'|\underline{y}] = \sigma^2(X'X + n\lambda G)^{-1}. \qquad (4)$$

Identity (4) represents an extension of the usual form ($\lambda = 0$) for variances and covariances for least-squares estimators. It plays an important role in the development of influence measures in Section 4.

To study residuals we use the predictive or unconditional distribution of $\underline{y}$ (see Box 1980 and references therein). Using this approach it is found that the vector of residuals,

$$\underline{e}(\lambda) = (e_1(\lambda), \ldots, e_n(\lambda))' = \underline{y} - X\tilde{\underline{\beta}}(\lambda),$$

has variance-covariance matrix

$$V(\underline{e}(\lambda)) = \sigma^2(I - H(\lambda)), \tag{5}$$

where

$$H(\lambda) = X(X'X + n\lambda G)^{-1}X' \tag{6}$$

is the hat matrix for estimator (3). When $\lambda = 0$ identity (5) reduces to the usual form for variances and covariances of residuals for least-squares estimators for linear regression models.

Our treatment of $\tilde{\underline{\beta}}(\lambda)$ as a Bayes estimator raises questions about the interpretation of (4) and (5) when $\underline{\beta}$ cannot be assumed stochastic. First we should note that there are cases, such as the smoothing spline setting, where the Bayesian formulation can be regarded as an approximation to the truth even when μ is not random (see Blight and Ott 1975, Wahba 1978, Wecker and Ansley 1983, Steinberg 1983 and Eubank 1984a). Draper and Van Nostrand (1979) conclude that ridge regression should only be used if the Bayesian model is appropriate or it is known that $\underline{\beta}'G\underline{\beta} \leq c^2$ for some known constant c. The latter case can be well approximated by the Bayesian model with λ chosen appropriately. Another interpretation can be obtained through consideration of the risk, $R(\lambda) = n^{-1}\Sigma_{i=1}^{n}E(\mu(\underline{u}_i) - \underline{x}_i'\,\tilde{\underline{\beta}}(\lambda))^2$. For smoothing splines Wahba (1983) argues that when λ^* is the minimizer of $R(\lambda)$ we should have $R(\lambda^*) \doteq \sigma^2\Sigma_{i=1}^{n}h_{ii}(\lambda^*)/n$, where $h_{ii}(\lambda)$ is the ith diagonal element of (6). A heuristic argument leading to a similar conclusion can also be advanced for ridge estimators. This suggests that when the Bayes assumptions are not valid and λ is selected in an optimal fashion (i.e., to minimize the risk) the use of (4) or (5) is in the spirit of using mean squared error rather than variance to assess the accuracy of a biased estimator. It should be emphasized that the arguments for this viewpoint are heuristic and more research is required before theoretically exact statements can be made. At present we merely view this as an indication that when using our methods where the Bayesian assumptions are not tenable one should choose λ in an optimal fashion.

In what follows we assume that λ in (2) has been specified a priori. Thus, our objective is detection of difficulties in the fit for a given value of λ. From the Bayesian viewpoint selection of λ is tantamount to selection of a prior. Thus it is necessary to have a specified value for λ before diagnostic (or inferential) analysis can be conducted. In practice, λ may require estimation from the data which can be viewed as an empirical Bayes approach. Various methods for selecting an optimal value of λ are discussed in Golub, Heath and Wahba (1979) and the references they cite.

Two further points should be made before proceeding. The first is that we are concerned with the detection of influential data rather than how to rectify any problems that might be discovered. Possible solutions to fit difficulties include M-estimation techniques which parallel those in Anderssen, Bloomfield and McNeil (1974), Huber (1979), Utreras (1981) and Cox (1983). The other point is that the diagnostics derived from our Bayesian approach are certainly not the only possibilities. For example, Wendelberger (1982) has found normal probability plots of residuals to be useful with smoothing splines.

2. Residual Diagnostics. In view of (5) a natural scaling of $e_i(\lambda)$ is provided by

$$t_i(\lambda,\sigma) = e_i(\lambda)/\sigma \ \{1 - h_{ii}(\lambda)\}^{\frac{1}{2}},$$

where $h_{ii}(\lambda)$ is the ith diagonal element of $H(\lambda)$ in (6). In practice σ will usually be unknown and require estimation from the data. Two possible estimators are discussed below.

An estimator of σ^2 which is unbiased under the Bayesian model is provided by

$$\sigma^2(\lambda) = \Sigma_{i=1}^{n} e_i(\lambda)^2/\mathrm{tr}(I - H(\lambda)) \ , \tag{7}$$

where tr denotes the matrix trace. This estimator generalizes the usual estimator of σ^2 for least-squares estimation from linear models with $\mathrm{tr}(I - H(\lambda))$ now assuming the role of degrees of freedom. It has been used by Wahba (1983) with good results in the context of spline smoothing.

In scaling residuals it is often advisable to remove the influence of the residual under study from the estimator of σ^2. An estimator which accomplishes this is

$$\sigma^2_{(i)}(\lambda) = \sum_{j\neq i} (e_j(\lambda) + h_{ji}(\lambda)e_i(\lambda)/(1-h_{ii}(\lambda)))^2/(n - 1 - \mathrm{tr}[H_{(i)}(\lambda)]), \tag{8}$$

where

$$\mathrm{tr}[H_{(i)}(\lambda)] = \sum_{j\neq i} [h_{jj}(\lambda) + h_{ij}(\lambda)^2/(1 - h_{ii}(\lambda))].$$

It follows from the Deletion Theorem in Section 4 that $\sigma^2_{(i)}(\lambda)$ is the estimator $\sigma^2(\lambda)$ computed from the data when the observation $(\underline{u}_i', y_i)$ has been deleted.

Using estimators (7) and (8) one obtains studentized residuals and studentized deleted residuals $t_i(\lambda, \sigma(\lambda))$ and $t_i(\lambda, \sigma_{(i)}(\lambda))$. By analogy with ordinary linear regression the $t_i(\lambda, \sigma(\lambda))$ and $t_i(\lambda, \sigma_{(i)}(\lambda))$ might be compared to critical values from Student's t distributions to detect points where the fit is inadequate. Approximate degrees of freedom for these statistics are provided by $\mathrm{tr}(I - H(\lambda))$ and $n - 1 - \mathrm{tr}[H_{(i)}(\lambda)]$, respectively.

3. Leverage. The diagonal elements of the hat matrix (leverage values) provide important information about the presence of extreme observations for the independent variables when least-squares estimators are used. This is due, in part, to the fact that leverage values provide measures of distance between the rows of the design matrix and the center of the data (assuming a constant term is included in the model). In this section we show that the elements of $H(\lambda) = \{h_{ij}(\lambda)\}_{i,j=1,n}$ possess similar properties as their linear regression counterparts and, hence, provide useful diagnostic information.

If $\underline{1} = (1, \ldots, 1)'$ is an eigenvector of $H(\lambda)$, i.e., $H(\lambda)\underline{1} = \delta\underline{1}$, then $h_{ii}(\lambda)$ has a distance interpretation since

$$h_{ii}(\lambda) = (\underline{x}_i - \bar{\underline{x}})'(X'X + n\lambda G)^{-1}(\underline{x}_i - \bar{\underline{x}}) + \delta n^{-1}, \tag{9}$$

where $\bar{\underline{x}} = n^{-1}X'\underline{1}$. Thus, in this case, $h_{ii}(\lambda)$ represents an assessment of the departure of $\underline{x}_i$ from the centroid $(\bar{\underline{x}})$ of the data. Although $\underline{x}_i$ is a function of $\underline{u}_i$, a large $h_{ii}(\lambda)$ may or may not correspond to an extreme $\underline{u}_i$. This presents no conceptual difficulties because $\mu(\underline{u}_i)$ is being approximated by $\underline{x}_i'\underline{\beta}$ and, hence, it is extreme values of the $\underline{x}_i$, rather than the $\underline{u}_i$, which will cause estimation difficulties.

Even when (9) is not satisfied the leverage values $h_{ii}(\lambda)$, $i = 1, \ldots, n$, still provide diagnostic information as a consequence of the following theorem whose proof follows by use of the singular-value decomposition for X and the Cauchy-Schwarz inequality.

Theorem 1. Let ℓ_1 denote the smallest eigenvalue of $(X'X)^{-1}G$ and let $h_{ii}(\infty) = \lim_{\lambda\to\infty} h_{ii}(\lambda)$. Then

$$|h_{ij}(\lambda)| \le (1 + n\lambda\ell_1)^{-1}[h_{ii}(0)h_{jj}(0)]^{\frac{1}{2}} \tag{10}$$

and

$$h_{ii}(\infty) \le h_{ii}(\lambda) \le h_{ii}(0) \ . \tag{11}$$

Also, $h_{ii}(\lambda) = 1$ if and only if $\underline{x}_i'\tilde{\underline{\beta}}(\lambda) = y_i$.

Theorem 1 has the implication that the elements of $H(\lambda)$ satisfy $-1 \le h_{ij}(\lambda) \le 1$ and $0 \le h_{ii}(\lambda) \le 1$, just as with least-squares estimators. Since the extreme case $h_{ii}(\lambda) = 1$ corresponds to estimation of $\mu(\underline{u}_i)$ by y_i we may infer that large leverage values are indicative of sensitive points in the design where an observation will tend to dominate its own fit. The bound (11) provides a useful benchmark for the determination of large leverage values for penalized least-squares estimators.

4. Influence measures. One approach to the assessment of influence involves the use of various summary statistics computed from an estimator's sample influence curve (SIC) (e.g., Cook and Weisberg 1980). The SIC for $\tilde{\underline{\beta}}(\lambda)$ is defined by

$$\underline{\mathrm{SIC}}_i = (n - 1)(\tilde{\underline{\beta}}(\lambda) - \tilde{\underline{\beta}}_{(i)}(\lambda)) \ , \tag{12}$$

where $\tilde{\underline{\beta}}_{(i)}(\lambda)$ is the estimator of $\underline{\beta}$ computed from the data set with $(\underline{u}_i', y_i)$ deleted. A closed form expression for (12) is provided by the following theorem.

<u>Theorem 2</u> (Deletion Theorem). For fixed λ and z let $\tilde{\underline{\beta}}(\lambda,z)$ denote the minimizer of $n^{-1}\{\Sigma_{j\neq i}(y_j - \underline{x}_j'\underline{\beta})^2 + (z - \underline{x}_i'\underline{\beta})^2\} + \lambda\underline{\beta}'G\underline{\beta}$. Then,

$$\tilde{\underline{\beta}}_{(i)}(\lambda) = \tilde{\underline{\beta}}(\lambda, \underline{x}_i'\tilde{\underline{\beta}}_{(i)}(\lambda)) \tag{13}$$

and

$$\underline{x}_i'\tilde{\underline{\beta}}_{(i)}(\lambda) = y_i - e_i(\lambda)/(1 - h_{ii}(\lambda)) \ . \tag{14}$$

The proof of the Deletion Theorem parallels that of Lemma 3.1 in Craven and Wahba (1979). One consequence of the theorem is that the estimator $\tilde{\underline{\beta}}_{(i)}(\lambda)$ can be obtained by applying $C(\lambda)$ to the original response vector $\underline{y}$ with y_i replaced by (14). Thus,

$$\underline{SIC}_i = (n - 1)\ \underline{c}_i(\lambda)e_i(\lambda)/(1 - h_{ii}(\lambda)) \ ,$$

where $\underline{c}_i(\lambda)$ is the ith column of $C(\lambda)$.

To develop a general framework for assessing influence we now proceed as in Cook and Weisberg (1980, 1982) and consider the subset of $\mathbb{R}^p$ defined, for a given positive definite matrix M and constant $m > 0$, by

$$L(M,m) = \{\underline{\ell} \in \mathbb{R}^p\colon \underline{\ell}'M^{-1}\underline{\ell} \le [(n-1)^2 m]^{-1}\}.$$

To each $\underline{\ell} \in L(M,m)$ there corresponds a linear functional $\underline{\ell}'\tilde{\underline{\beta}}(\lambda)$. The maximum impact of the ith observation on the functionals in L(M,m) is then measured by

$$\begin{aligned} \rho_i(\lambda,M,m) &= \sup_{\underline{\ell}\in L(M,m)} [\underline{\ell}'\underline{SIC}_i]^2 \\ &= t_i(\lambda,\sigma)^2\ \frac{\sigma^2}{m}\ P_i(\lambda,M) \ , \end{aligned} \tag{15}$$

where $P_i(\lambda,M)$ is the potential defined by

$$P_i(\lambda,M) = \underline{x}_i' (X'X + n\lambda G)^{-1}M(X'X + n\lambda G)^{-1}\underline{x}_i/(1 - h_{ii}(\lambda)) .$$

By choosing M and m appropriately in (15) a variety of useful diagnostics can be obtained. In view of (4) one obvious choice for M is $M = (X'X + n\lambda G)$. By then taking $m = \sigma^2(\lambda)\ \mathrm{tr}[H(\lambda)]$, $m = \sigma^2_{(i)}(\lambda)\ \mathrm{tr}[H(\lambda)]/\mathrm{tr}[I - H(\lambda)]$ or $m = \sigma^2_{(i)}(\lambda)$ one obtains penalized least squares versions of influence measures discussed by Cook (1977), Atkinson (1981) and Belsley, Kuh and Welsch (1980). Alternative choices for M and m can be proposed through similar arguments to those in Cook and Weisberg (1982) (see, e.g., their Table 3.5.4). Still other choices may be suggested by the specific estimator under study and the type of diagnostic information that is desired.

We note in passing that the close connection between the SIC and jackknife standard error estimates (see Efron 1982) has the implication that measures of accuracy associated with $\tilde{\underline{\beta}}(\lambda)$ can be developed using results in this section. This point will not be pursued further here.

5. **Examples.** In this final section we apply the results of Sections 2-4 to several classes of penalized least-squares estimators. The first example represents an extension of work by Eubank (1984b, 1985) and gives a partial answer to a question posed by Wahba (1984a) regarding the applicability of regression type diagnostics to partial splines.

5.1. Partial splines. Suppose $\underline{u}_i' = (s_{i1}, \ldots, s_{id}, z_{i1}, \ldots, z_{ik},) = (\underline{s}_i', \underline{z}_i')$ and $\underline{y}$ follows the partially linear model $\underline{y} = \underline{f} + Z\underline{\gamma} + \underline{\varepsilon}$, where $\underline{f}' = (f(\underline{s}_1), \ldots, f(\underline{s}_n))$ for some unknown function f, $Z' = [\underline{z}_1, \ldots, \underline{z}_n]$ and $\underline{\varepsilon}' = (\varepsilon_1, \ldots, \varepsilon_n)$. If f is sufficiently smooth an estimator of $\mu(\underline{u})$ can be obtained by minimization with respect to f and $\underline{\gamma}$ of

$$n^{-1} \sum_{i=1}^{n} (y_i - f(\underline{s}_i) - \underline{z}_i' \, \underline{\gamma})^2 + \lambda J(f,f), \tag{16}$$

where

$$J(f,g) = \sum_{\alpha_1+\ldots+\alpha_d=m} \frac{m!}{\alpha_1!\ldots\alpha_d!} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{\partial^m f(\underline{s})}{\partial s_1^{\alpha_1} \ldots \partial s_d^{\alpha_d}} \, \underline{ds} \, .$$

When formulated in the appropriate function space and under certain restrictions on $\underline{s}_1, \ldots, \underline{s}_n$ and Z there is a unique solution to this problem which provides an estimator of μ that is known as a partial spline. Details can be found in Wahba (1984a,b, 1985) (see also Shiller 1984). Special cases of these estimators include univariate smoothing splines (d = 1, k = 0) and multivariate thin plate smoothing splines (d > 1, k = 0).

It follows from results in Wahba (1984b) that there are functions $B_1(\underline{s}), \ldots, B_n(\underline{s})$ such that the minimizer of (16) has the form $\mu_\lambda(\underline{u}) = \Sigma_{j=1}^{n} \theta_j B_j(\underline{s}) + \underline{z}'\underline{\gamma}$. General expressions for the B_j can be deduced from Wahba (1984b). We merely note that the linear span of the B_j includes polynomials of order m in $\underline{s}$. Substituting the form for μ_λ into (16) and minimizing with respect to $\underline{\beta}' = (\underline{\theta}', \underline{\gamma}')$ reveals that the minimizer coincides with (3) when $X = [B,Z]$, for $B = \{B_j(\underline{s}_i)\}_{i,j=1,n}$, and

$$G = \begin{bmatrix} \Omega & 0 \\ 0 & 0 \end{bmatrix} ,$$

where Ω has typical element $J(B_i, B_j)$.

From a Bayesian perspective the form of G has the implication that a non-informative prior is being used for the coefficient vector $\underline{\gamma}$. In addition, Ω has rank $n - \binom{d+m-1}{d}$, which can be shown to mean that a diffuse prior has been placed on the "polynomial portion" of μ.

Since the estimator of μ derived from (16) fits into the framework of Section 1 this suggests that the residual and influence diagnostics discussed in Sections 2 and 4 are appropriate for use with partial splines. With regard to leverage values, since $B(B'B + n\lambda\Omega)^{-1}B'\underline{1} = \underline{1}$ one can show that $H(\lambda)\underline{1} = \underline{1}$. Hence the $h_{ii}(\lambda)$ have a distance interpretation and, of course, Theorem 1 holds. Leverage values should therefore provide useful design diagnostics for partial spline estimators.

Examples of the use of the diagnostics in Sections 2-4 to detect difficulties in multivariate thin plate smoothing spline and univariate smoothing spline fits to data can be found in Carmody (1985) and Eubank (1985). For other examples involving some closely related measures see Silverman (1985).

5.2 Ridge regression. A generalized ridge regression estimator of μ results from setting $x_{i1} = 1$ and letting x_{ij} $(j > 1)$ be standardized predictor variable values $(\sum_i x_{ij} = 0, \sum_i x_{ij}^2 = 1)$. Ridge regression is usually used to estimate the standardized regression coefficients for the (p-1) nonconstant predictor variables but minimization of (2) does not require that the constant term be estimated separately from the other regression coefficients.

The generalized ridge estimators are

$$\tilde{\beta}_1 = \bar{y} \qquad \tilde{\underline{\beta}}_2(k) = (X_2'X_2 + kG_2)^{-1}X_2' \underline{y} \ ,$$

where $\underline{\beta}' = (\beta_1, \underline{\beta}_2')$, $X = [\underline{1}, X_2]$, $G = \text{diag}(0,G_2)$ with G_2 $(p-1) \times (p-1)$ nonnegative definite, and $k = n\lambda$. Ordinary ridge estimators are obtained by setting $G_2 = I$:

$$\tilde{\beta}_1 = \bar{y} \qquad \tilde{\underline{\beta}}_2(k) = (X_2'X_2 + kI)^{-1}X_2' \underline{y} \ .$$

Leverage values for both the generalized and the ordinary ridge estimators have a distance interpretation as in (2).

It is beyond the scope of this paper to discuss selection of the ridge parameter k. We note in passing the controversy over automated and stochastic selection of k, the role of standardization, and assumptions underlying theoretical properties of the ridge estimator (e.g., Draper and Van Nostrand 1979; Smith and Campbell 1980, with discussion). Our interest here is on ridge diagnostics and their interpretation for a fixed value, k, of the ridge parameter regardless of how it is chosen.

Efficient computation of the ridge estimates and residuals when observations are deleted (e.g., equations (13) and (14)) is possible only if the reduced X_2 matrix is not restandardized when the ith row is deleted. Since the major benefits of centering and standardization cited by Marquardt (1980) are essentially retained when one of the rows of the standardized X_2 matrix is 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 produce (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 1 displays regression diagnostics for the fit of ℓn(GNP) by the six socioeconomic indices. We now present an analysis of this data using the diagnostics discussed in Sections 2 - 4.

The ridge parameter was estimated as $(p-1)\hat{\sigma}^2/\tilde{\underline{\beta}}'(0)\tilde{\underline{\beta}}(0) = .08$ as suggested by Hoerl, Kennard and Baldwin (1976). This value of k also corresponds to a relatively stable portion of the ridge trace. Table 1 provides a summary of the diagnostics which result for the least-squares and ridge regression estimators for this data. For detection of influence we use the measure DFITS_i which is (±) the square root of $\rho_i(\lambda,\ X'X + kI,\ \sigma^2_{(i)}(\lambda))$ (see Belsley, Kuh and Welsch 1980).

TABLE 1. Regression Diagnostics for GNP Data, Selected Observations

	Least Squares			Ridge (k = .08)		
Obsn.	h_{ii}	t_i	DFITS_i	h_{ii}	t_i	DFITS_i
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	1.377	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

With the exception of Malta, least-squares leverage values which exceed $2(p+1)/n = .286$ are also large with the ridge estimator using the analogous bound $2(\mathrm{tr}[H(\lambda)]+1)/n = 0.271$. Although the ridge DFITS 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 the bound suggested by Belsley, Kuh and Welsch (1980) $2\{(p+1)/n\}^{\frac{1}{2}} = 0.756$ for least squares also exceed the parallel bound $2\{(\mathrm{tr}[H(\lambda)]+1)/n\}^{\frac{1}{2}} = 0.736$ for ridge regression--Malta is again the exception--and a similar comment can be made about the t_i.

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.

Obviously a more complete analysis of this data set is needed in order to resolve questions which remain about influential observations. 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 the usefulness of penalized least-squares diagnostics.

5.3. Other examples. There appear to be many other estimators to which the work in Sections 1-4 is applicable. These include various estimators derived from the context of minimax estimation, discrete smoothness priors and discretization of smoothing spline estimators. See, for example, Rice (1982), Shiller (1973, 1984) and Engle et al. (1983). Other illustrations can be found in work by Allen (1974) and Blight and Ott (1975).

REFERENCES

Allen, D. M. (1974). The relationship between variable selection and data augmentation and a method for prediction. Technometrics 16, 125-127.

Anderssen, R. S., Bloomfield, P. and McNeil, D. R. (1974). Spline functions in data analysis. Tech. Rep. 69, Series 2, Dept. of Statistics, Princeton University.

Atkinson, A. C. (1981). Two graphical displays for outlying and influential observations in regression. Biometrika 68, 13-20.

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

Blight, B. J. N. and Ott, L. (1975). A Bayesian approach to model inadequacy for polynomial regression. Biometrika 62, 79-88.

Box , G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness. J. R. Statist. Soc. A 143, 383-430.

Carmody, T. J. (1985). Diagnostics for Multivariate Smoothing Splines. Unpublished Ph.D. dissertation, Dept. of Statist., Southern Methodist Univ.

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

Cook, R. D. and Weisberg, S. (1980). Characterizations of an empirical influence function for detecting influential cases in regression. Technometrics 22, 495-508.

Cook, R. D. and Weisberg, S. (1982). Residuals and Influence in Regression. New York: Chapman and Hall.

Cox, D. (1981). Asymptotics for M-type smoothing splines. Ann. Statist. 11, 530-551.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions. Numer. Math. 31, 377-403.

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. SIAM, Monograph No. 38, CBMS-NSF.

Engle, R., Granger, C., Rice, J. and Weiss, A. (1983). Nonparametric estimates of the relationship between weather and electricity demand. Discussion paper no. 83-17, Department of Economics, University of California, San Diego.

Eubank, R. L. (1984a). Approximate regression models and splines. Commun. Statist. - Theor. Meth. A13(4), 433-484.

Eubank, R. L. (1984b). The hat matrix for smoothing splines. Statist. and Prob. Letters 2, 9-14.

Eubank, R. L. (1985). Diagnostics for smoothing splines. J. R. Statist. Soc. B, to appear.

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 Applications: A Data-Oriented Approach. Marcel-Dekker: New York.

Hoerl, A. E., Kennard, R. W. and Baldwin, K. F. (1975). Ridge regression: some simulations. Comm. Statist. 4, 105-123.

Huber, P. J. (1979). Robust smoothing. In Robustness in Statistics (Launer and Wilkinson, eds.), 33-47. Academic Press: New York.

Leamer, E. E. (1973). Multicollinearity: a Bayesian interpretation. Rev. Econ. Statist. 55, 371-380.

Lindley, D. V. and Smith, A. F. M. (1972). Bayes estimates for the linear model. J. R. Statist, Soc. B 34, 1-41.

Marquardt, D. W. (1980). You should standardize the predictor variables in your regression models. J. Amer. Statist. Assoc. 75, 87-91.

Rice, J. (1982). An approach to peak area estimation. J. Research National Bureau of Standards 87, 53-65.

Shiller, R. J. (1973). A distributed lag estimator derived from smoothness priors. Econometrika 41, 775-788.

Shiller, R. J. (1984). Smoothness priors and nonlinear regression. J. Amer. Statist. Assoc. 79, 609-615.

Silverman, B. W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting (with discussion). J. Roy. Statist. Soc. B, to appear.

Smith, G. and Campbell, F. (1980). A critique of some ridge regression methods. J. Amer. Statist. Assoc. 75, 74-81.

Steinberg, D. M. (1983). Bayesian models for response surfaces of uncertain functional form. MRC Technical Summary Report No. 2474, Univ. Wisconsin-Madison.

Utreras, F. (1981). On computing robust splines and applications. SIAM J. Sci. Stat. Comput. 2, 153-163.

Wahba, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression. *J. Roy. Statist. Soc. B* *40*, 364-372.

Wahba, G. (1983). Bayesian "confidence intervals" for the cross-validated smoothing spline. *J. R. Statist. Soc. B* *45*, 133-150.

Wahba, G. (1984a). Partial spline models for the semiparametric estimation of functions of several variables. In *Statistical Analysis of Time Series*, Tokyo: Institute of Statistical Mathematics, 319-329.

Wahba, G. (1984b). Cross validated spline methods for the estimation of multivariate functions from data on functionals. In *Statistics: An Appraisal, Proceedings 50th Anniversary Conference Iowa State Statistical Laboratory*, eds. H. T. David, H. A. David, Iowa State Univ. Press, 205-235.

Wahba, G. (1985). Comments to Peter J. Huber, Projection Pursuit, *Ann. Statist.*, to appear.

Wecker, W. E. and Ansley, C. F. (1983). The signal extraction approach to nonlinear regression and spline smoothing. *J. Amer. Statist. Assoc.* *78*, 81-89.

Wendelberger, J. G. (1981). The computation of Laplacian smoothing splines with examples. Tech. Rep. No. 648, Dept. of Statist., Univ. Wisconsin-Madison.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (*When Data Entered*)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER SMU-DS-TR-196	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (*and Subtitle*) Diagnostics for Penalized Least-Squares Estimators		5. TYPE OF REPORT & PERIOD COVERED Technical Report
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(*s*) R. L. Eubank and R. F. Gunst		8. CONTRACT OR GRANT NUMBER(*s*) ONR-N00014-85-K-0340
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Statistics Southern Methodist University Dallas, Texas 75275		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS NR 042-479
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research Arlington, VA 22217		12. REPORT DATE December 1985
		13. NUMBER OF PAGES 19
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*)

Bayes Estimators, Influence, Leverage, Partial Splines, Residuals, Ridge Regression, Smoothing Splines, Thin Plate Splines.

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

Diagnostic methods for a class of penalized least-squares estimators are derived from a Bayesian perspective. The class of estimators considered includes generalized ridge estimators, partial splines and thin plate smoothing splines. The proposed diagnostics include scaled residuals, leverage values and various measures of influence.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (*When Data Entered*)