DIAGNOSTICS FOR SMOOTHING SPLINES

BY

R. L. EUBANK

Technical Report No. SMU-DS-TR-187
Department of Statistics ONR Contract

September 1984

DIAGNOSTICS FOR SMOOTHING SPLINES

by R. L. Eubank
Southern Methodist University, USA

SUMMARY

Diagnostic measures appropriate for use with smoothing splines are derived and their properties are investigated. The proposed measures focus on detection of observations which substantially influence the fit and provide additional information over that obtained from examination of residuals alone. A numerical example illustrates the technique.

Keywords: CROSS-VALIDATION; DIAGNOSTICS; INFLUENCE; LEVERAGE; STUDENTIZED RESIDUALS; SPLINE SMOOTHING

Present address: Professor R. L. Eubank, Department of Statistics
Southern Methodist University, Dallas, TX. 75275, USA.

1. INTRODUCTION

Consider the situation where responses $y_1,\ldots,y_n$ are observed corresponding to values $t_1,\ldots,t_n$ of an independent variable which, for convenience, are assumed to satisfy

$$0 \leq t_1 < t_2 < \ldots < t_n \leq 1.$$

The y_j and t_j are related by the model

$$y_j = \eta(t_j) + \varepsilon_j \quad , \quad j = 1,\ldots,n \; , \tag{1.1}$$

where η is some unknown response function and the ε_j are zero mean, uncorrelated random variables. In this paper we consider the problem of nonparametric estimation of the regression function η. Diagnostic methods, similar to those used in ordinary regression analysis, are proposed and their properties investigated for a particular nonparametric estimator known as the cross-validated smoothing spline.

It will be assumed throughout that η is smooth in the sense that, for some positive integer m, η belongs to the function class

$$W_2^m[0,1] = \{f{:}f^{(j)} \text{ absolutely continuous, } j = 0,\ldots,m-1, \int_0^1 f^{(m)}(t)^2 dt < \infty\} \; .$$

Under this restriction a natural estimator is the function minimizing

$$n^{-1} \sum_{j=1}^{n} (y_j - f(t_j))^2 + \lambda \int_0^1 f^{(m)}(t)^2 dt, \quad \lambda > 0, \tag{1.2}$$

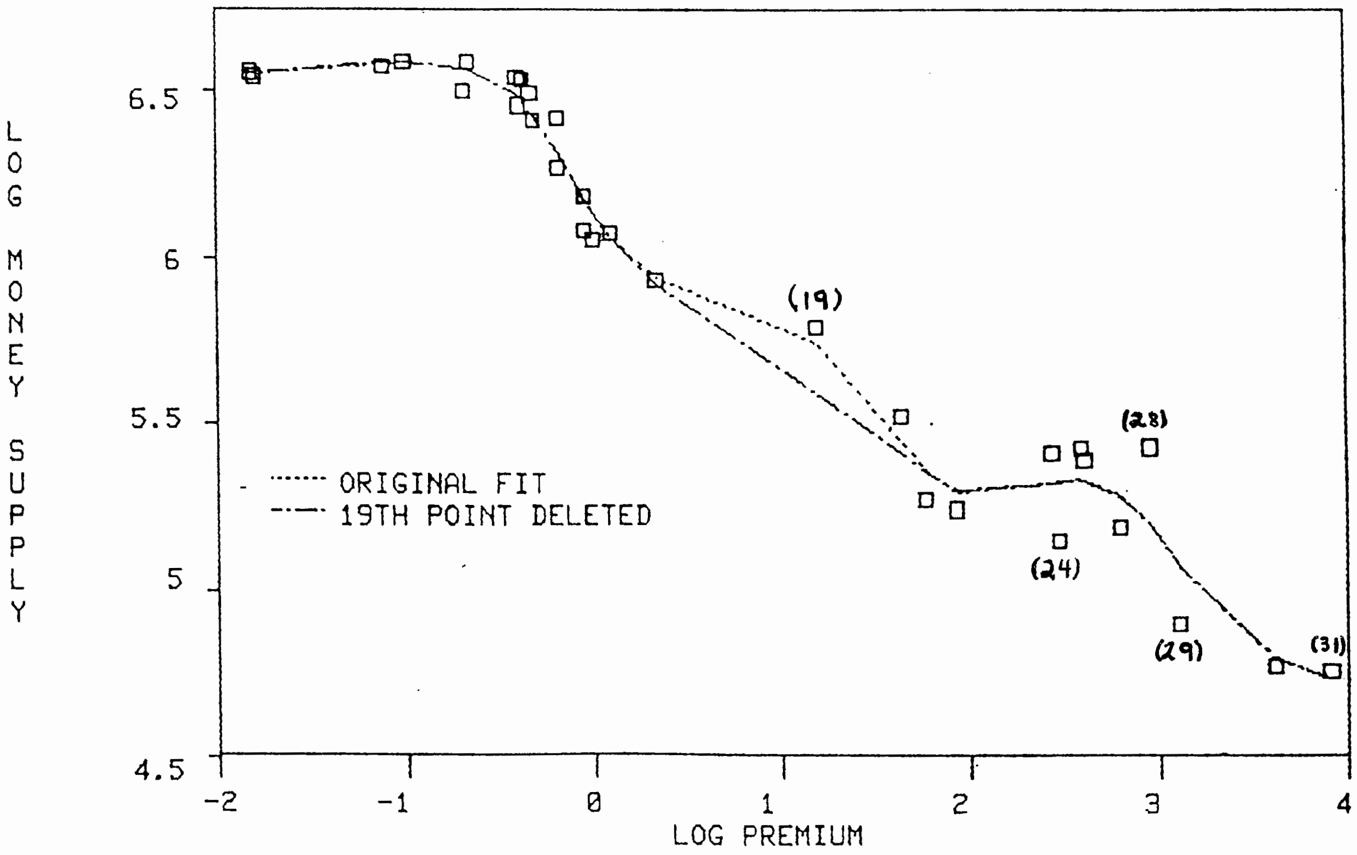
over all $f \in W_2^m[0,1]$. If $n \geq m$ there is a unique solution to this problem which we denote by $\hat{\eta}_\lambda$. The estimator $\hat{\eta}_\lambda$ is a polynomial spline of order 2m with knots at the t_j that is usually refered to as a smoothing spline (see, e.g., Schoenberg 1964 or Reinsch 1967). The parameter λ in (1.2) governs the balance between fidelity to the data (as measured by $n^{-1}\Sigma_{j=1}^{n}(y_j-f(t_j))^2$) and smoothness (as gauged by $\int_0^1 f^{(m)}(t)^2 dt$). The extreme cases, $\lambda = 0$ and ∞, corresponded to interpolation of the data and regression on polynomials or order m, respectively. The latter instance illustrates one of the many connections between smoothing splines and polynomial regression. Expository discussions of the statistical properties of smoothing splines are provided in Wegman and Wright (1983) and Eubank (1984a).

Several procedures are available for estimation of the smoothing parameter, λ, from data. The most popular of these seems to be the method of generalized cross-validation (GCV) developed by Craven and Wahba (1979). GCV has been shown to have various efficiency and consistency properties (Craven and Wahba 1979, Speckman 1982 and Li 1983) and, perhaps more importantly, tends to work well in practice. Other methods include maximum likelihood type procedures such as that of Wecker and Ansley (1983). However, results in Wahba (1983b) suggest that GCV should, on the average, work as well (or better) than maximum likelihood methods. Consequently, attention will be restricted to the estimator $\hat{\eta}_\lambda$ with λ estimated by GCV. This estimator is termed the cross-validated smoothing spline.

Regression analysts have long recognized that estimators obtained through minimization of a squared error criterion can be adversely effected by influential data points. This fact has lead to the development of an extensive literature on types and uses of diagnostic measures for regression modeling. Just as in the regression setting, the presence of the squared error term in (1.2) has the consequence that influential observations can (locally) dominate a smoothing spline fit. A dramatic illustration of this fact is provided by data for the logarithm of the money supply versus the logarithm of the premium, or discount, on a forward contract for foreign exchange during the German hyperinflation (see Frenkel 1977). This data is plotted along with a cross-validated cubic (m=2) smoothing spline fit in Figure 1 and a listing of the data is provided in the Appendix. Also plotted is a smoothing spline fit where the response for August 1922 has been deleted. The new fit clearly reveals the influence the deleted response had on the original spline estimate.

The adverse influence that outliers can have on smoothing spline estimates has been recognized by Anderssen, Bloomfield and McNeil (1974), Huber (1979) and Cox (1983). Their solution to such difficulties is the use of robust smoothing splines. In contrast, the objective of this paper is the development of techniques for detection of influential data, observations which significantly impact upon the fit. As noted by Beckman and Cook (1983) an outlier need not be influential. However, as they point out, influential observations should be regarded as special types of outliers. Andrews and Pregibon (1978) call these "the outliers that matter" and observe that if an outlier is not influ-

FIGURE 1. GERMAN HYPERINFLATION DATA



ential "there is little point in agonizing over how deviant it appears". Thus, the results in this paper can be regarded as providing data screening methodology that focuses our attention on influential data points which merit closer scrutiny. Further examination may, or may not, lead to remedial action such as the use of robust smoothing techniques.

In the next section background material on smoothing splines is developed that will be required for the sequel. Then, in Section 3, a variety of diagnostic indicators are proposed for use with smoothing splines. The use of these measures is illustrated in Section 4 through a numerical example.

2. NOTATION AND PRELIMINARIES

Smoothing splines are linear estimators. Thus, there is an $n \times n$ matrix $H(\lambda) = \{h_{ij}(\lambda)\}$ which transforms the vector of responses, $\underline{y} = (y_1, \ldots, y_n)'$, to the vector of fitted values, $\underline{\hat{\eta}}_\lambda = (\hat{\eta}_\lambda(t_1), \ldots, \hat{\eta}_\lambda(t_n))'$, i.e.,

$$\underline{\hat{\eta}}_\lambda = H(\lambda)\underline{y} \ .$$

$H(\lambda)$ is known as the hat matrix (Eubank 1984b) and its typical element, $h_{ij}(\lambda)$, determines how much influence y_j has on the fit to y_i. By analogy with the linear regression setting (Hoaglin and Welsch 1978) the diagonal elements of $H(\lambda)$ are termed <u>leverage values</u>. The properties of $H(\lambda)$ have been studied by Eubank (1984b) where it is shown, for example, that $0 \leq h_{ii}(\lambda) \leq 1$.

A specific form for $H(\lambda)$ can be derived from the work of Wahba (1978). Let T denote the n×m matrix with ijth element t_i^{j-1}, $i=1,\ldots,n$, $j=1,\ldots,m$, let I_r denote the r×r identity matrix and define the covariance kernel

$$K(s,t) = \int_0^s \frac{(t-u)^{m-1}(s-u)^{m-1}}{(m-1)!^2}\,du, \quad s \leq t. \tag{2.1}$$

Then, it is shown in Wahba (1978) that

$$I_n - H(\lambda) = n\lambda U(U'K_nU + n\lambda U'U)^{-1}U' , \tag{2.2}$$

where U is any $n\times(n-m)$ matrix of rank $n-m$ such that $U'T = \phi$, a null matrix, and K_n is the n×n matrix with ijth entry $K(t_i,t_j)$. Since $U'K_nU$ and $U'U$ are positive definite there is a nonsingular matrix B such that $B'[U'K_nU + n\lambda U'U]B = \Delta + n\lambda I_{n-m}$, where $\Delta=\mathrm{diag}(\delta_1,\ldots,\delta_{n-m})$ and the δ_j are the eigenvalues, arranged in ascending order, of $(U'U)^{-1}(U'K_nU)$. Set $X_2 = UB$ and notice that $X_2'X_2 = I_{n-m}$ and $X_2'T = \phi$. If we now define $X = [X_1 \vdots X_2]$ as an appropriately augmented unitary matrix, $H(\lambda)$ is seen to admit the representation

$$H(\lambda) = XD(\lambda)X' , \tag{2.3}$$

where $D(\lambda) = \mathrm{diag}(d_1(\lambda),\ldots,d_n(\lambda))$ for $d_j(\lambda) = 1$, $j = 1,\ldots,m$, $d_j(\lambda) = \delta_{j-m}/(\delta_{j-m} + n\lambda)$, $j=m+1,\ldots,n$.

A representation for $\hat{\eta}_\lambda$ also follows from (2.3). Let x_j denote the unique function in $W_2^m[0,1]$ which minimizes $\int_0^1 x^{(m)}(t)^2 dt$ over all functions satisfying $x(t_i) = x_{ij}$, the ijth element of X. Functions $x_1,\ldots,x_m$ are necessarily polynomials of order m which span the set of mth order polynomials. The remaining functions are splines of order 2m with knots at the t_i. It can then be shown that

$$\hat{\eta}_\lambda(t) = \Sigma_{j=1}^n \hat{\beta}_j(\lambda)x_j(t) \qquad (2.4)$$

where

$$\hat{\underline{\beta}}_\lambda = (\hat{\beta}_1(\lambda),\ldots,\hat{\beta}_n(\lambda))' = D(\lambda)X'\underline{y} \qquad (2.5)$$

(see also Demmler and Reinsch 1975 for discussion of this representation). The elements of $\hat{\underline{\beta}}(\lambda)$ are smoothing spline counterparts of the coefficient estimates in ordinary regression analysis.

Smoothing splines can be derived and motivated from other perspectives than minimization of criterion (1.2). For example, suppose that instead of following model (1.1) the observations are obtained at "time points" $t_1,\ldots,t_n$ from the stochastic process

$$y(t) = \sum_{j=0}^{m-1} \alpha_j t^j + \sigma_s Z(t) + \varepsilon(t),\ t\varepsilon[0,1],\ \sigma_s>0, \qquad (2.6)$$

where $Z(\cdot)$ is a zero mean process with covariance kernel (2.1) that is uncorrelated with the white noise process $\{\varepsilon(t);\ 0 \leq t \leq 1\}$. It then follows from Kimeldorf and Wahba (1970) that $\hat{\eta}_\lambda(t)$ with $\lambda=\sigma^2/n\sigma_s^2$ is the best linear unbiased predictor of $y(t)$ based on $y_1,\ldots,y_n$.

A closely related Bayesian model assumes that η has the same prior distribution as the process

$$\xi(t) = \sum_{j=0}^{m-1} \alpha_j t^j + \sigma_s Z(t), \quad t\varepsilon[0,1]\ , \qquad (2.7)$$

where the Z process is now assumed to be normal with the same mean and covariance kernel as before. The vector of polynomial coefficients, $\underline{\alpha} = (\alpha_0,\ldots,\alpha_{m-1})'$, is assumed to be independent of the Z process and have an m-variate normal distribution with mean zero and variance-covariance matrix γI_m. In the case of a partially improper prior, $\gamma \to \infty$, it is shown in Wahba (1978) that $\hat{\eta}_\lambda(t)$, $\lambda = \sigma^2/n\sigma_s^2$, is the best linear unbiased predictor of $\eta(t)$.

Models (2.6)-(2.7) do not coincide with the deterministic response function model (1.1) (see Wahba 1983a for discussion of this point). They can, nonetheless, be regarded as approximations to the truth in the sense that the response function is approximated as far as possible by a polynomial with a stochastic approximation used for the remainder. This is essentially the philosophy discussed in Blight and Ott (1975), Wahba (1978) and Steinberg (1983) and again illustrates the close connection between smoothing splines and polynomial regression.

It is possible to utilize models (2.6) - (2.7) to obtain certain identities which indicate some of the formal similarities between smoothing splines and ordinary regression analysis. Let

$$\underline{r}_\lambda = (r_1(\lambda),\ldots,r_n(\lambda))' = [I_n - H(\lambda)]\underline{y}$$

denote the vector of residuals and define

$$\underline{\eta} = (\eta(t_1),\ldots,\eta(t_n))'.$$

Then, under model (2.6), $\underline{y}$ has variance-covariance matrix $V(\underline{y}) = \sigma_s^2 K_n + \sigma^2 I_n$. Using (2.2) and recalling that $n\lambda = \sigma^2/\sigma_s^2$, we find that

$$V(\underline{r}_\lambda) = \sigma^2(I_n - H(\lambda)). \qquad (2.8)$$

This parallels the form for variances and covariances of residuals from linear regression.

Another identity of interest stems from the Bayesian model (2.7). It was shown by Wahba (1983a) that

$$V(\hat{\underline{\eta}}_\lambda|\underline{y}) = E[(\hat{\underline{\eta}}_\lambda-\underline{\eta})(\hat{\underline{\eta}}_\lambda-\underline{\eta})'|\underline{y}] = \sigma^2H(\lambda) \quad . \tag{2.9}$$

More conventional notation would use $V(\underline{\eta}|\underline{y})$ rather than $V(\hat{\underline{\eta}}_\lambda|\underline{y})$ in (2.9). However, we prefer to follow the lead of Wahba (1983a) and think of $\sigma^2H(\lambda)$ as the "variance-covariance" matrix for the vector of fitted values. From this same perspective we define

$$V(\hat{\underline{\beta}}_\lambda|\underline{y}) = \sigma^2D(\lambda) \quad , \tag{2.10}$$

by using the fact that $\hat{\underline{\beta}}_\lambda = X'\hat{\underline{\eta}}_\lambda$.

To conclude this section we point out certain formulae for deleting observations from smoothing spline fits that will be needed in the next section. Let $\hat{\eta}_\lambda^{[j]}$ denote the smoothing spline fit, for fixed λ, when the observation (t_j,y_j) has been excluded from the data. Using Lemmas 3.1 and 3.2 of Craven and Wahba (1979) it is seen that

$$\hat{\eta}_\lambda^{[j]}(t) = \sum_{i=1}^{n} \hat{\beta}_i^{[j]}(\lambda)x_i(t) \quad , \tag{2.11}$$

where

$$\hat{\underline{\beta}}_\lambda^{[j]}=(\hat{\beta}_1^{[j]}(\lambda),\ldots,\hat{\beta}_n^{[j]}(\lambda))=D(\lambda)X'[\underline{y}-(r_j(\lambda)/(1-h_{jj}(\lambda)))\underline{e}_j], \tag{2.12}$$

with $\underline{e}_j$ denoting the jth column of I_n.

3. DIAGNOSTIC MEASURES FOR SMOOTHING SPLINES

At present diagnostic analysis of smoothing spline fits consists primarily of examination of plots or normal plots (Wendelberger 1981) involving the residuals. These forms of analysis are important and should be included in any diagnostic package. However, such methods, by themselves, are inadequate as can be seen from the example in the introduction (see also the analysis in Section 4). The residual corresponding to the 19th observation is obviously quite small in the original fit and, hence, its impact on the fit, as indicated by the refitted curve with this case deleted, would have been missed in plots or examinations of raw (or even scaled) residuals alone. In this section we derive diagnostic tools designed to highlight such influential data points. Our philosophy is to justify the form of a diagnostic by using a convenient model, such as one of those in Section 2, and then see how it works for a model of interest, such as (1.1). The close connection between smoothing splines and polynomial regression makes us believe the "right choices" for diagnostics should be similar to those commonly used for linear regression modeling. This viewpoint guides the search for appropriate measures and is reflected in much of the notation and terminology which follows. For discussions of diagnostic techniques used in linear regression which impact on the present study the reader is refered to Gunst and Mason (1980, Chap. 7), Belsley, Kuh and Welsch (1980, Chap. 2), Cook and Weisberg (1982, Chap. 3), Hoaglin and Welsch (1978) and Velleman and Welsch (1981).

Throughout this section, unless stated otherwise, the value used for λ is taken as $\hat{\lambda}$ the GCV estimate. The concepts we develop (although not our simulations) will, of course, apply to other methods of selecting λ. For notational convenience we adopt the conventions $\hat{\underline{\eta}}_{\hat{\lambda}} = \hat{\underline{\eta}}$, $\hat{\underline{\beta}}_{\hat{\lambda}} = \hat{\underline{\beta}}$, $\underline{r}_{\hat{\lambda}} = \underline{r}$, $H(\hat{\lambda}) = H$, $D(\hat{\lambda}) = D$ with analogous notation used for the elements of these vectors and matrices.

Assuming for the moment that σ is known, identity (2.8) suggests using the standardized residuals

$$T_j(\sigma) = r_j/\sigma(1-h_{jj})^{1/2}, \ j = 1,\dots,n, \qquad (3.1)$$

for detection of responses which do not conform to the fit. The measures which stem from the $T_j(\sigma)$, when σ is unknown, will be seen to occupy a central role in subsequent development of diagnostic indicators.

When σ is unknown it must be estimated in (3.1). A natural estimator, proposed by Wahba (1983a) is

$$s^2 = \Sigma_{j=1}^n r_j^2/tr(I_n-H) \ , \qquad (3.2)$$

with tr denoting the matrix trace. This estimator can be motivated from the viewpoint of ordinary least squares regression in which case $tr(I_n-H)$ corresponds to degrees of freedom. For smoothing splines I_n-H is not idempotent and, hence, $tr(I_n-H)$ will not usually be integral in value. However, as in Wahba (1983a), we regard $tr(I_n-H)$ as the equivalent degrees of freedom (EDF) for smoothing splines.

An indication of the properties of s^2 as an estimator of σ^2 is provided through study of estimators of the form

$$s^2(\lambda) = \Sigma_{j=1}^n r_j(\lambda)^2/tr(I_n-H(\lambda)).$$

Note that under model (2.6) with $\lambda = \sigma^2/n\sigma_s^2$,

$$E[\Sigma_{j=1}^n r_j(\lambda)^2] = tr[(\sigma_s^2 K_n + \sigma^2 I_n)(I_n-H)^2] + \underline{\alpha}'T'(I_n-H)^2 T\underline{\alpha}$$
$$= \sigma^2 tr(I_n-H),$$

as $(I_n-H)T$ is a null matrix. Consequently, $s^2(\lambda)$ is an unbiased estimator under this model and could be justified from this perspective. Alternatively, the following proposition states that, for an appropriately chosen deterministic sequence of λ's, $s^2(\lambda)$ is asymptotically unbiased for σ^2 when the data derive from model (1.1).

Proposition. Suppose that $m \geq 2$, $\eta \in W_2^m[0,1]$ and the sequence of sampling points $\{t_{1,n},\ldots,t_{n,n}\}$ satisfies $(2j-1)/(2n)=\int_0^{t_{j,n}} p(t)dt$ for some continuous nonvanishing density, p, on [0,1]. Then, if $\lambda \to 0$ as $n \to \infty$ in such a way that $n\lambda^{1/2m} \to \infty$,

$$E[s^2(\lambda)] = \sigma^2(1 + o(1))$$

where $o(1) \to 0$ as $n \to \infty$.

Proof: First observe that, under model (1.1),

$$E[s^2(\lambda)] = \{\sigma^2 tr[(I_n-H(\lambda))^2]+\underline{\eta}'(I_n-H(\lambda))^2\underline{\eta}\}/tr(I_n-H(\lambda)).$$

It follows from Speckman (1981) that, under the assumptions on η, λ and the t_j, $n^{-1}tr[H(\lambda)^2]$ and $n^{-1}trH(\lambda)$ are both o(1) which implies that $tr[(I_n-H(\lambda))^2]/tr(I_n-H(\lambda)) = 1 + o(1)$. Using Lemma 4.1 of Craven and Wahba (1979) we see that $n^{-1}\underline{\eta}'(I_n-H(\lambda))^2 \underline{\eta} \leq \lambda\int_0^1 \eta^{(m)}(t)^2dt$ which establishes the proposition.

It is of somewhat more interest to have a result of this type for $s^2 = s^2(\hat{\lambda})$ where $\hat{\lambda}$ is the GCV estimate of λ. At present no such result is available. However, simulations in Wahba (1983a) indicate satisfactory behaviour for s^2 as an estimator of σ^2.

If σ is replaced by s in (3.1) the resulting statistics are the studentized residuals

$$T_j = r_j/s(1-h_{jj})^{1/2}, \quad j = 1,\ldots,n. \qquad (3.3)$$

Motivated by practices in ordinary regression analysis, one might wish to examine an observation for which $|T_j|$ exceeds an appropriate critical value from a Student's t distribution with $\mathrm{tr}(I_n-H)$ EDF.

To ascertain how a Student's t bound for the studentized residuals might work in practice, a small scale simulation was conducted. Data was generated from model (1.1) with normal errors, equally spaced t_j and response functions

$$\eta_1(t) = 4.26\{\exp(-3.25t)-4\exp(-6.5t)+3\exp(-9.75t)\}$$

and

$$\eta_2(t) = \{B_{10,5}(t) + B_{7,7}(t) + B_{5,10}(t)\}/3$$

where

$$B_{p,q}(t) = \Gamma(p+q)t^{p-1}(1-t)^{q-1}/\Gamma(p)\Gamma(q), \quad 0 \le t \le 1,$$

and $\Gamma(\cdot)$ is the gamma function. The function η_1 is a rescaled version of a function considered by Wahba and Wold (1975) whereas η_2 was utilized in the simulation study in Wahba (1983a). Two sample size configurations were used: m = 50 replicates of sample size n = 80 and m = 100 replicates of n = 40. Four values were chosen for σ, σ = .2, .4, .6, and .8. For a given function and sample size configuration the same samples were used for all four values of σ. Different samples were generated for different sample sizes and/or functions. For each sample the cross-validated cubic smoothing spline fit was computed and the proportion of times the $|T_j|$ exceeded the 5% (two-tailed) critical value for a Student's t distribution with approximate degrees of freedom $tr(I_n-H)$ was recorded. The results are summarized in Table 1. Approximate standard errors can be obtained using the usual formula for the estimated standard error of a proportion estimate.

Examination of the values in Table 1 suggest that the proposed Student's t bound is satisfactory and, perhaps, somewhat conservative in small samples. We have obtained similar results in simulations using other significance levels and several other functions.

Table 1. Empirical Significance Levels for Studentized and Studentized Deleted Residuals

	Studentized Residuals		Studentized Deleted Residuals	
η_1	m=100,n=40	m=50,n=80	m=100,n=40	m=50,n=80
σ=.2	.0433	.0473	.0515	.0523
σ=.4	.0390	.0460	.0493	.0518
σ=.6	.0398	.0455	.0475	.0505
σ=.8	.0410	.0455	.0500	.0495
η_2				
σ=.2	.0383	.0458	.0495	.0493
σ=.4	.0408	.0458	.0510	.0488
σ=.6	.0418	.0448	.0490	.0478
σ=.8	.0418	.0448	.0495	.0488

The use of s as an estimator of σ in (3.1) has the disadvantage that the estimate involves the influence of (t_j,y_j), the observation under inspection. It may, therefore, be preferable to use $s_{(j)}$, the estimator s computed with (t_j,y_j) deleted from the data, as an estimator of σ. A closed form expression for $s^2_{(j)}$, obtained using (2.11), is

$$s^2_{(j)} = \sum_{\substack{i=1\\i\neq j}}^{n} (r_i + h_{ij}r_j/(1-h_{jj}))^2/tr(I_{n-1}-H^{[j]}) \qquad (3.4)$$

where

$$trH^{[j]} = \sum_{\substack{i=1\\i\neq j}}^{n} (h_{ii} + h^2_{ij}/(1-h_{jj})) \ . \qquad (3.5)$$

Using $s_{(j)}$ to estimate σ in (3.1) gives the diagnostic measures

$$T_{(j)} = r_j/s_{(j)}(1-h_{jj})^{1/2}, \quad j = 1,\ldots,n \ . \qquad (3.6)$$

The $T_{(j)}$ are referred to as studentized deleted residuals and might be compared to percentage points from a Student's t distribution with $tr(I_{n-1}-H^{[j]})$ for its EDF. An indication of the efficacy of such a bound is provided by simulation results in Table 1, obtained in the same manner as those for studentized residuals by comparison of the $|T_{(j)}|$ to 5% (two-tailed) critical values. The approximation seems quite good with the observed significance levels somewhat closer to the nominal 5% level than those of studentized residuals.

The T_j and $T_{(j)}$ are useful in the detection of fit inadequacies. However, for assessing influence a standard approach is to consider some aspect of the fit both with and without a particular observation. A class of diagnostic indicators derived from this perspective can be described as follows. Let $\mathcal{L}$ be the set of all bounded linear functionals on $W_2^m[0,1]$. A common practice is to use $\ell(\hat{\eta})$ to estimate $\ell(\eta)$ (see Wahba 1983a and Wahba and Wold 1975). Now to each $\ell \varepsilon \mathcal{L}$ there corresponds an n-vector $\underline{\ell} = (\ell(x_1),\ldots,\ell(x_n))'$ which dictates how ℓ acts on $\hat{\eta}$. Thus, to assess the influence of (t_j, y_j) on $\ell(\hat{\eta})$ we could examine $(\ell(\hat{\eta})-\ell(\hat{\eta}^{[j]}))^2 = [\underline{\ell}'(\underline{\hat{\beta}}-\underline{\hat{\beta}}^{[j]})]^2$. Motivated by this discussion we define, for a given positive definite matrix Q and positive constant c, the set

$$\mathcal{L}(Q,c) = \{\ell \varepsilon \mathcal{L}\colon \underline{\ell}'Q^{-1}\underline{\ell} \le c^{-1}\}$$

and the corresponding diagnostic measures

$$\rho_j(Q,c) = \sup_{\ell\varepsilon\mathcal{L}(Q,c)} [\ell(\hat{\eta})-\ell(\hat{\eta}^{[j]})]^2,\ j = 1,\ldots,n.$$

By use of identities (2.11)-(2.12) and results on the extrema of quadratic forms, $\rho_j(Q,c)$ can be expressed as

$$\rho_j(Q,c) = \sigma^2|T_j(\sigma)|^2\underline{x}_j'DQD\underline{x}_j/c(1-h_{jj}), \qquad (3.7)$$

where $\underline{x}_j'$ is the jth row of X. The $\rho_j(Q,c)$, j = 1,...,n, provide an entire class of diagnostics indexed by both Q and c. We now address the question of choices for these index parameters.

A convenient choice for Q in (3.7) is $Q = D^{-1}$. In this case, $\rho(D^{-1},c) = \sigma^2|T_j(\sigma)|^2h_{jj}/c(1-h_{jj})$ which is obviously invariant under change of basis. This particular measure can also be motivated from the Bayesian model (2.7). Using (2.10) we obtain $V(\ell(\hat{\eta})|\underline{y}) = \sigma^2\underline{\ell}'D\underline{\ell}$ and it then follows that

$$\sup_{\ell\in\mathcal{L}}(\ell(\hat\eta)-\ell(\hat\eta^{[j]}))^2/V(\ell(\hat\eta)|\underline{y}) = (\hat\eta(t_j)-\hat\eta^{[j]}(t_j))^2/\sigma^2 h_{jj}$$

$$= \rho_j(D^{-1},\sigma^2).$$

Thus, apart from a constant multiple, $\rho_j(D^{-1},c)$ has the interpretation of being the maximum "scaled" change in $\ell(\hat\eta)$ due to deletion of (t_j,y_j) over all bounded linear functionals. In particular, since point evaluation is a bounded linear functional we have

$$(\hat\eta(t_i)-\hat\eta^{[j]}(t_i))^2/V(\hat\eta(t_i)|\underline{y}) \le (\hat\eta(t_j)-\hat\eta^{[j]}(t_j))^2/\sigma^2 h_{jj}.$$

Consequently, to assess the impact of (t_j,y_j) on the fit we need only examine, initially, its influence on the fit at y_j.

Another option is to take $Q = D^{-2}$ which gives the measures $\rho_j(D^{-2},c) = \sigma^2|T_j(\sigma)|^2/c(1-h_{jj})$, $j=1,\dots,n$. These diagnostics can be motivated directly from model (1.1) since it can be shown that

$$\rho_j(D^{-2},\sigma^2) = \sup_{\ell\in\mathcal{L}}(\ell(\hat\eta)-\ell(\hat\eta^{[j]}))^2/V(\ell(\hat\eta)).$$

We prefer measures such as $\rho_j(D^{-1},c)$, however, due to their similarity to those commonly used in ordinary regression analysis.

Several choices can be suggested for c. Of particular interest are $c = s^2$ and $c = s^2_{(j)}$. These correspond to the use of measures such as

$$\text{DFITS}_j = T_j(h_{jj}/(1-h_{jj}))^{1/2}, \quad j = 1,\dots,n\,, \tag{3.8}$$

and

$$\text{DFITS}_{(j)} = T_{(j)}(h_{jj}/(1-h_{jj}))^{1/2}, \quad j =1,\dots,n\,. \tag{3.9}$$

Rough bounds for these indicators can be obtained by first approximating h_{jj} and $(1-h_{jj})$ by their averages $n^{-1}\mathrm{tr}H$ and $n^{-1}\mathrm{tr}(I_n-H)$ to obtain the approximation $\mathrm{tr}H/\mathrm{tr}(I_n-H)$ for $h_{jj}/(1-h_{jj})$. In view of the results in Table 1, one might then set aside for closer inspection those observations with values for DFITS_j or $\mathrm{DFITS}_{(j)}$ which exceed $2(\mathrm{tr}H/\mathrm{tr}(I_n-H))^{1/2}$.

Other choices for c include $c = s^2\mathrm{tr}H$ and $c = s^2_{(j)}\mathrm{tr}H/\mathrm{tr}(I_n-H)$. These lead to smoothing spline analogs of Cook's distance measure (Cook 1977) and a measure due to Atkinson (1981).

The reader may have observed the ubiquitous role played by the leverage values, h_{jj}, in the diagnostics proposed in this section. Since the h_{jj} provide diagnostics pertaining to the independent variable (Eubank 1984b) their appearance is a natural consequence of the fact that (t_j,y_j) can be extreme in t_j and/or y_j. Thus, for example, we see that the measure $\rho_j(Q,\hat{\sigma}^2)^{1/2}$ consists of two components, $[\underline{x}_j'DQD\underline{x}_j/(1-h_{jj})]^{1/2}$ and $|T_j(\hat{\sigma})|$ which reflect diagnostics for the value of t_j and fit to y_j, respectively. This fact can be used to suggest various graphical methods for displaying the $\rho_j(Q,\hat{\sigma}^2)$ such as plots of $([\underline{x}_j'DQD\underline{x}_j/(1-h_{jj})]^{1/2}, |T_j(\hat{\sigma})|)$, $j=1,\ldots,n$.

The leverage values for smoothing splines have a distance interpretation similar to that of their counterparts in the least squares regression setting. To see this, set $\underline{\bar{x}} = n^{-1}X'\underline{1}$ where $\underline{1}$ is an $n\times 1$ vector of all unit elements. Then, since $\underline{\bar{x}}'=[\underline{1}'X_1 \vdots \underline{\phi}']$ and $X_1'X_1\underline{1} = \underline{1}$, we have

$$(\underline{x}_j-\underline{\bar{x}})'D(\underline{x}_j-\underline{\bar{x}}) = h_{jj} - n^{-1},$$

where $\underline{x}_j$ is the jth row of X. Thus $h_{jj}-n^{-1}$ can be regarded as measuring how extreme the jth row of X is relative to the average $\underline{\bar{x}}$. It follows from Eubank (1984b) that $0 \leq h_{jj} \leq 1$ and $h_{jj} = 1$ if and only if $h_{ij} = 0$ for all $i \neq j$. Consequently, h_{jj} which are close to one indicate extreme rows in the X matrix and the corresponding response will tend to dominate its own fit. Since the ith row of X consists of the values $x_j(t_i)$, $j=1,\ldots,n$, extreme rows of X typically correspond to extreme values for the t_i.

In concluding this section we note that, in some instances, it may be useful to consider extensions of measures such as (3.3), (3.6), and (3.7) where $k>1$ observations are under inspection. Development of techniques for addressing such problems are beyond the scope of the present paper. However, measures applicable to this purpose can be readily derived through appropriate generalizations of identities (2.11)-(2.12).

4. NUMERICAL ILLUSTRATION

In this section the German hyperinflation data is examined using the methodology developed in the previous section. The objective is to illustrate the additional insight that can be obtained from the measures proposed in Section 3, over the use of tools such as residual plots alone, rather than to provide a complete, in depth, analysis of the data.

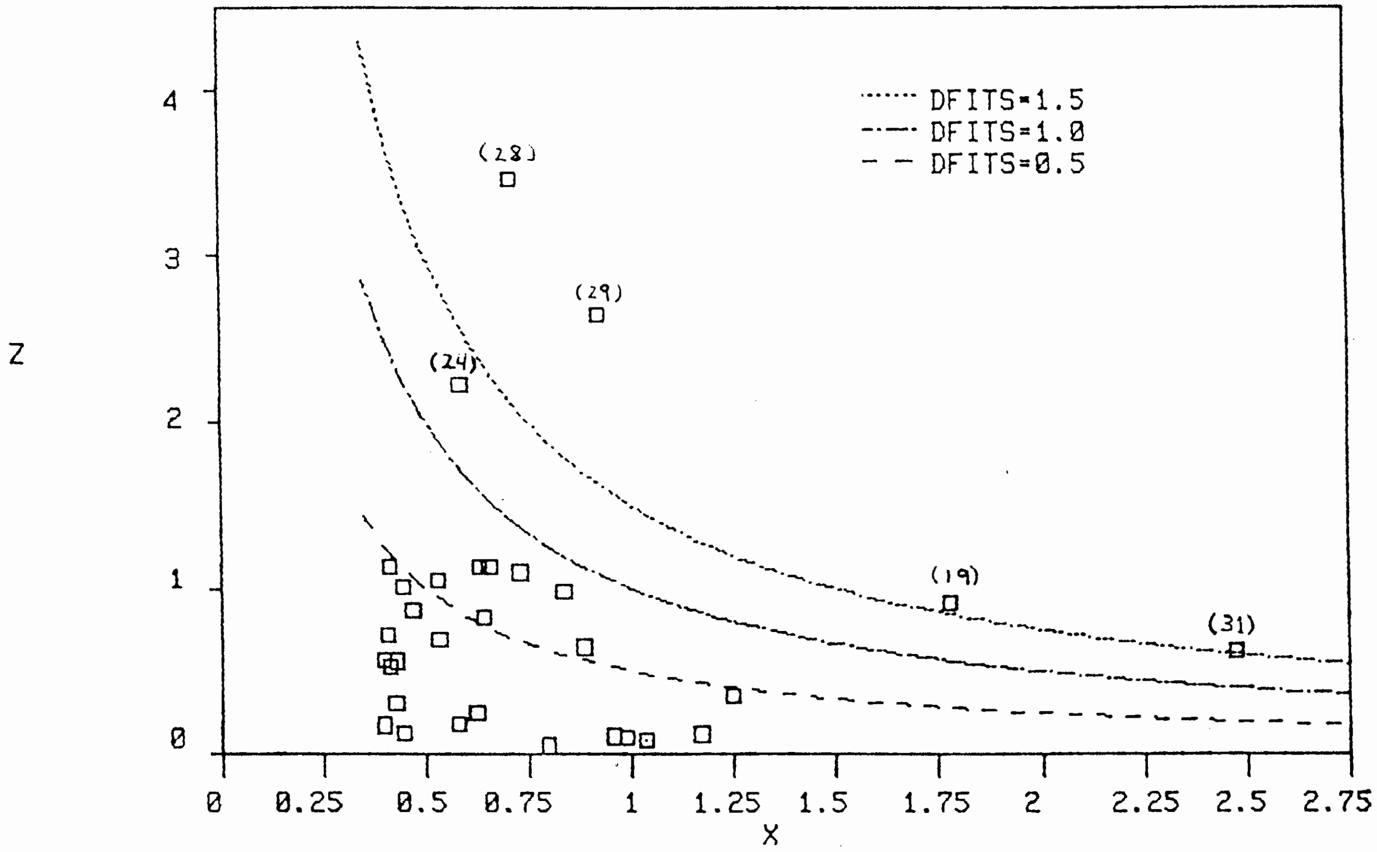
For ease of exposition, the German hyperinflation data, shown in Figure 1, is given in the Appendix. Throughout this section specific cases will be referred to by the observation numbers assigned there. This data set has also been studied in the context of smoothing splines by Wecker and Ansley (1981).

Many of the important aspects of the cross-validated smoothing spline fit to the hyperinflation data are revealed by a plot of the points $(x_i, z_i) = ([h_{ii}/(1-h_{ii})]^{1/2}, |T_{(i)}|)$, i=1,...,n, shown in Figure 2. The curves shown in this figure are the contours |DFITS| = constant, for constant values .5, 1 and 1.5. For this data a rough bound is $2[trH/tr(I_n-H)]^{1/2} = 1.41$, which provides us with a quantitative measure of which observations should be judged influential and indicates where our attention should be focused on the plot.

Examination of Figure 2 indicates that observations 19, 28, 29, 31 and, to a lesser extent, 24 are influential data points. However, they are influential for different reasons. These are readily deduced by the position of their respective points on the plot.

Observations 19 and 31 are influential because of high leverage. From examination of the data in the Appendix, we see that case 19 has high leverage resulting from inadequate information about values of the log exchange premium over a portion of its range. Recognition of this fact provides the explanation for the marked change in fit, illustrated in the introduction, that occured when this value was deleted from the data. This also indicates that conclusions drawn about predictions made in this interval will be predicted largely

FIGURE 2. DFITS FOR HYPERINFLATION DATA

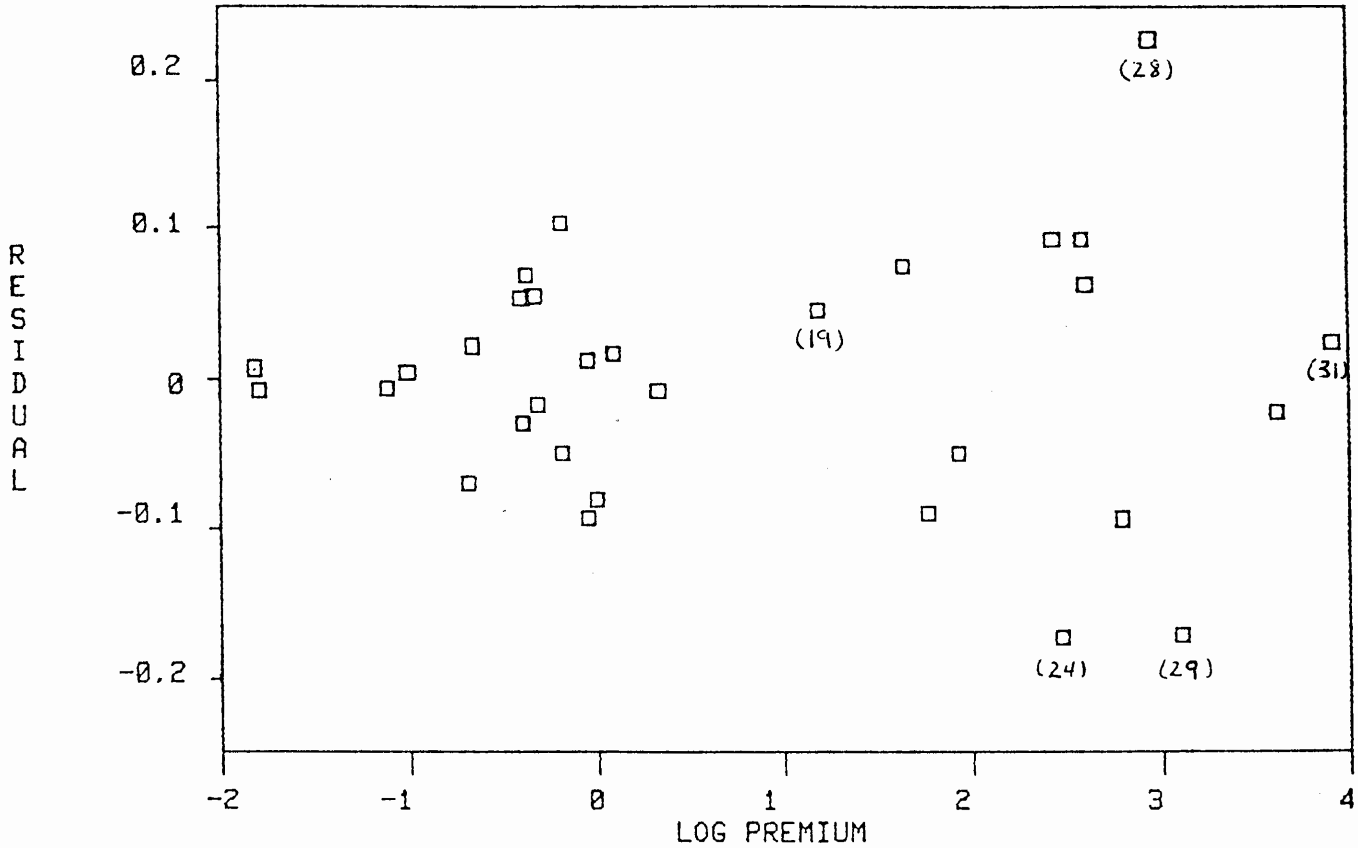


on the information in the 19th observation. The high leverage for observation 31 is reflective of the position of its value for the log exchange premium at the extreme of the data set.

Observations 28 and 29 are seen to be influential because they lie significantly far from the fitted curve as measured by their values for $|T_{(i)}|$. Observation 24 also has a significantly large studentized deleted residual. However, it is not a high leverage point and, as indicated by its lower value for DFITS, is less influential on the fit. This latter observation enforces the point that not all outliers will severely influence the fit.

To demonstrate the additional utility of the diagnostics in Section 3 over examination of residuals alone, a plot of the residuals has been presented in Figure 3. Cases 24, 28 and 29 stand out as outliers. However, examination of their studentized deleted residual values, all of which exceed 2, provides a quantitative confirmation of this intuitive designation for these points. In addition, by consideration of their values for DFITS, we recognize that, of these three points, observations 28 and 29 have the most influence on the fit. Also note that, from examination of the residual plot, no special importance would be attributed to observations 19 and 31. This is not surprising since high leverage points will always have small residuals. Their influence was clearly revealed, however, by examination of the DFITS values.

FIGURE 3. RESIDUALS FOR HYPERINFLATION DATA



In summary, the analysis in this section has pointed out several sources of difficulty for the cross-validated smoothing spline fit to the hyperinflation data which would not have been recognized through examination of residuals alone. Specific remedies would depend on the subject matter area and might include consideration of alternative transformations for the data, downweighting certain points in the original criterion (1.2) or the use of robust smoothing methodology which takes leverage into account. These possibilities will not be pursued here.

ACKNOWLEDGEMENTS

The author would like to thank Tom Carmody and Lori Thombs for their assistance with some of the computations. Helpful conversations with Dick Gunst and Bill Schucany are also gratefully acknowledged.

This work was supported by the ONR under Contract No. N00014-82-K-0207.

APPENDIX: GERMAN HYPERINFLATION DATA

The data below represent the (natural) logarithm of real money supply (Y) and the logarithm of the premium, or discount, on a forward contract for foreign exchange (t) during the German hyperinflation from February 1921 to August 1923.

Observation	Month/		
No.	Year	Y	t
1	03/21	6.5605	-1.8202
2	02/21	6.5474	-1.7958
3	04/21	6.5802	-1.1087
4	05/21	6.5927	- .9927
5	08/21	6.5019	- .6832
6	06/21	6.5896	- .6539
7	07/21	6.5414	- .3960
8	12/21	6.4580	- .3930
9	09/21	6.5381	- .3653
10	10/21	6.4977	- .3271
11	01/22	6.4129	- .3093
12	11/21	6.4225	- .1863
13	02/22	6.2669	- .1839
14	04/22	6.0839	- .0429
15	03/22	6.1841	- .0837
16	05/22	6.0578	0.0
17	06/22	6.0774	.0999
18	07/22	5.9321	.3343
19	08/22	5.7858	1.1845
20	09/22	5.5203	1.6369
21	03/23	5.2718	1.7630
22	12/22	5.2421	1.9243
23	04/23	5.4116	2.4336
24	11/22	5.1504	2.4774
25	05/23	5.4239	2.5908
26	10/22	5.3290	2.6053
27	01/23	5.1921	2.7955
28	06/23	5.4269	2.9565
29	02/23	4.9010	3.1122
30	08/23	4.7712	3.6169
31	07/23	4.7589	3.9176

Source: Wecker and Ansley (1981).

REFERENCES

Anderssen, R. S., Bloomfield, P. and McNeil, D. R. (1974). Spline functions in data analysis. Tech. Rep. No. 69, Series 2, Department of Statistics, Princeton Univ.

Andrews, D. F. and Pregibon, D. (1978). Finding the outliers that matter. J. R. Statist. Soc. B, 40, 85-93.

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

Beckman, R. J. and Cook, R. D. (1983). Outlier..........s. Technometrics, 25, 119-163.

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.

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

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

Cox, D. D. (1983). 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.

Demmler, A. and Reinsch, C. (1975). Oscillation matrices with spline smoothing. Numer. Math., 24, 375-382.

Eubank, R. L. (1984a). Approximate regression models and splines. Commun. Statist.-Statist. Reviews, to appear.

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

Frenkel, J. A. (1977). The forward exchange rate, expectations, and the demand for money: the German hyperinflation. Amer. Econ. Rev., 67, 653-670.

Gunst, R. F. and Mason, R. L. (1980). *Regression Analysis and Its Application: A Data-Oriented Approach*. New York: Marcel Dekker.

Hoaglin, D. C. and Welsch, R. F. (1978). The hat matrix in regression and ANOVA. *Amer. Statist.*, *32*, 17-22.

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

Kimeldorf, G. S. and Wahba, G. (1970). Spline functions and stochastic processes. *Sankhya Ser. A*, *32*, 173-180.

Li, K. C. (1983). From Stein's unbiased risk estimates to the method of generalized cross-validation. Tech. Rep. No. 83-34, Purdue University.

Reinsch, C. H. (1967). Smoothing by spline functions, *Numer. Math.*, *10*, 177-183.

Schoenberg, I. J. (1964). Spline functions and the problem of graduation. *Proc. Nat. Acad. Sci. USA*, *52*, 947-950.

Speckman, P. (1981). The asymptotic integrated error for smoothing noisy data by splines. *Numer. Math.*, to appear.

Speckman, P. (1982). Efficient nonparametric regression with cross-validated smoothing splines. *Ann. Statist.*, to appear.

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

Velleman, P. and Welsch, R. E. (1981). Efficient computing of regression diagnostics. *Amer. Statist.*, *35*, 234-242.

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

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

Wahba, G. (1983b). A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. Tech. Rep. No. 712, Dept. of Statistics, Univ. of Wisconsin-Madison.

Wahba, G. and Wold, S. (1975). A completely automatic French curve: fitting spline functions by cross validation. Comm. Statist.-Theor. Meth., A4(1), 1-17.

Wecker, W. E. and Ansley, C. F. (1981). Extensions and examples of the signal extraction approach to regression. To appear in Applied Time Series Analysis of Economic Data (A Zellner, ed.)

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.

Wegman, E. J. and Wright, I. W. (1983). Splines in statistics. J. Amer. Statist. Assoc., 78, 351-365.

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

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER SMU/DS/TR-187	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Diagnostics for Smoothing Splines		5. TYPE OF REPORT & PERIOD COVERED Technical Report 6. PERFORMING ORG. REPORT NUMBER SMU/DS/TR-187
7. AUTHOR(s) R. L. Eubank		8. CONTRACT OR GRANT NUMBER(s) N00014-82-K-0207
9. PERFORMING ORGANIZATION NAME AND ADDRESS Southern Methodist University Dallas, Texas 75275		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS NR042 - 479
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research Arlington, VA 22217		12. REPORT DATE September 1984 13. NUMBER OF PAGES 29
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)

Cross-validation; diagnostics; influence; leverage; studentized residuals; spline smoothing

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

Diagnostic measures appropriate for use with smoothing splines are derived and their properties are investigated. The proposed measures focus on detection of observations which substantially influence the fit and provide additional information over that obtained from examination of residuals alone. A numerical example illustrates the technique.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)