The Annals of Statistics
1978, Vol. 6, No. 5, 1122-1137

LINEAR ESTIMATION FOR APPROXIMATELY LINEAR MODELS

BY JEROME SACKS[1] AND DONALD YLVISAKER[2]

Northwestern University and
University of California, Los Angeles

An approximate linear model is proposed to allow for deviations from an underlying ideal linear model as follows: If, in standard notation, $Y = A\beta + \varepsilon$ is the ideal model then $Y = A\beta + r + \varepsilon$ where $|r_i| \leqq M_i$ for M a given vector is an approximate linear model. The problem solved here is that of finding a linear estimate of a single linear function of β which minimaxes mean square error in the approximate model. The estimate obtained may be the standard one from the ideal model, but in general it is not. The estimate is calculated as a solution to a set of nonlinear equations (generalizing the usual normal equations) and an algorithm is given for obtaining the solution.

1. Introduction. Robust estimation of the parameters in a linear model has been the subject of intense inquiry in recent years (for example, [3], [5] and [6]) and considerable success has been achieved in dealing with those problems which arise when there is departure from the assumption of normality of errors. The issue of "model robustness," i.e., behavior of estimates when there is departure from the assumed linear model, has long been recognized to be of central importance but it has not received the concerted attention given to the issue of "distributional robustness." It is the aim of this paper to propose and discuss some approximately linear models which admit deviations from an ideal linear model as follows: if $Y = A\beta + \varepsilon$ is the model (standard notation is used here) then the approximate models to be studied have the form

$$Y = A\beta + r + \varepsilon \tag{1.1}$$

where r is an n-vector satisfying $|r_i| \leqq M_i$ for a given n-vector M. The case where M is the 0-vector reduces the approximate model to the linear one.

The approximate linear models of (1.1) seem flexible enough to admit common types of deviations from an ideal model and they are tractable enough to permit determination of that linear estimate of a single linear function of β which minimaxes the mean square error. The modification of the standard Gauss–Markov estimate achieved by this approach is, qualitatively, to

Received October 1976; revised October 1977.

[1] This author's research was supported in part by NSF Grants MPS 73-08523 and PMS 76-07066.

[2] This author's research was supported in part by NSF Grants MPS 72-4591, MPS 75-7120 and MCS 77-02121.

AMS 1970 *subject classifications.* Primary 62J05, 62J10; Secondary 62J35.

Key words and phrases. Approximately linear models, minimum mean square linear estimation, normal equations.

downgrade the contribution of those observations for which the departure from the linear model is great (technically, this means that the weight attached to Y_i is small when M_i is large).

As an illustration consider the approximately linear regression model where independent observations Y_i taken at locations $x_i \in R^1$ have mean $f(x_i)$ and variance σ^2, $i = 1, \cdots, n$, and f is approximately linear in the sense that, for some $m \geqq 0$,

$$(1.2) \qquad |f(x) - \beta_0 - \beta_1 x| \leqq mx^2 \qquad \text{for some } \beta_0 \text{ and } \beta_1 \text{ and all } x.$$

In the notation used at (1.1) $M_i = mx_i^2$ represents the possible departure from linearity of the ith observation. With $m = 0$ one has straight line regression but with $m > 0$ the collection of possible regressions is much larger. In any event, one may think of β_0 and β_1 as $\beta_0 = f(0)$, $\beta_1 = f'(0)$. Among the estimates of β_0 of the form $\sum_{i=1}^n c_i Y_i = \sum_{i=1}^n c(x_i) Y_i$, the one that minimaxes the mean square error over the model (1.2) when $m > 0$ has $c(x)$ defined by a positive quadratic loop covering the origin with the possible addition of a negative quadratic loop supported on an interval disjoint from the first. The particulars depend on m and on the design points $x_1, \cdots, x_n$. The case $m = 0$ is a degenerate one: the x^2 term is missing and the best c is, of course, a linear function. In the ideal linear model observations farthest from the origin are weighted most heavily but in the approximately linear model the observations closest to the origin have the most weight. Corresponding results hold for the estimation of β_1. The details are in Section 3 (Theorem 1 and Remark 3).

The calculation of an estimate of a single linear function of β requires solution of a set of nonlinear equations (see (3.8)) which reduce, as expected, to the normal equations when the ideal model holds, that is, when $M \equiv 0$. Fortunately, a uncomplicated algorithm can be provided to solve these nonlinear equations and the algorithm appears to be very efficient. Details of this appear in Section 5 with some computed examples in Section 6.

The problem of simultaneous estimation of several linear functions of the parameters can also be solved but the computability of the solutions is doubtful. This situation is mitigated somewhat by the fact that the use of single parameter estimates in the simultaneous estimation problem is fairly efficient (of course, it is perfectly efficient in the ideal model). A discussion of this appears in Section 4.

2. Approximately linear models. Approximately linear models are presented here as relaxed versions of the strict linear models employed in a variety of statistical contexts. The proposed models are parametric in nature, although the parametrizations necessarily possess some novel features. Approximate models are first introduced in a general setting in order to emphasize their flexibility. Afterwards, some concrete examples are given and the corresponding parametrizations identified.

Let f and σ be real-valued functions on some index set $\mathscr{X}$ with $\sigma > 0$.

Assume one has observations

$$Y_i = Y_{x_i} = f(x_i) + \sigma(x_i)\varepsilon_i\,, \qquad x_i \in \mathscr{X},\ i = 1, \cdots, n\,, \tag{2.1}$$

where the ε_i are uncorrelated random variables, each with mean zero and variance one so that $f(x_i) = EY_i$. Imagine f (and possibly σ) to be unknown.

Let $f_0, f_1, \cdots, f_k$ and M be real-valued functions on $\mathscr{X}$ with $M \geqq 0$. The observations are said to follow an *approximately linear model* if

$$\begin{aligned} &\text{There is a vector } \beta = (\beta_0, \cdots, \beta_k)' \text{ so that} \\ &|f(x) - \textstyle\sum_{j=0}^{k} \beta_j f_j(x)| \leqq M(x) \quad \text{for all } x \in \mathscr{X}\,. \end{aligned} \tag{2.2}$$

If $M \equiv 0$ then (2.2) gives the linear model determined by $\{f_0, \cdots, f_k\}$ while if $M \geqq 0$, the set of f's given by (2.2) contains the linear model.

If f satisfies (2.2) and if $r(x) = f(x) - \sum \beta_j f_j(x)$ then $|r(x)| \leqq M(x)$ and the model (2.1) becomes

$$Y_i = \sum \beta_j f_j(x_i) + r(x_i) + \sigma(x_i)\varepsilon_i$$

which has the form of (1.1). It is slightly more convenient to describe the model by (2.1) and (2.2).

Under (2.2) the parameters of the model are said to be *identified* if

$$\begin{aligned} &|\textstyle\sum_{j=0}^{k} \beta_j f_j(x)| \leqq M(x) \qquad \text{for all } x \in \mathscr{X} \text{ and} \\ &\qquad\qquad\qquad\qquad\quad \text{some } \beta \text{ implies } \beta = 0\,. \end{aligned} \tag{2.3}$$

If $M \equiv 0$, (2.3) is the assumption of linear independence of $f_0, \cdots, f_n$ over $\mathscr{X}$; when $M \geqq 0$ (2.3) is, evidently, a more stringent requirement than linear independence. The sense of (2.3) is this: one can readily show, via the triangle inequality, that when the parameters of the approximately linear model are identified (2.2) cannot hold for two distinct coefficient vectors. Thus, if (2.2) and (2.3) hold, it makes sense to refer to the guaranteed and unique coefficients $\beta_0, \cdots, \beta_k$ as the (regression) parameters of the function f.

Here are some examples.

EXAMPLE 1. Let $\mathscr{X} = R^d$ and take $f_0, \cdots, f_k$ to be the monomials of degree $\leqq \nu$ in d variables, i.e., each f_j has the form $f_j(x) = \prod_{i=1}^{d} x_{(i)}^{\nu_i}$, $\sum_{i=1}^{d} \nu_i \leqq \nu$. Take $M(x) = o(|x|^\nu)$ as $x \to 0$. It is easily checked that (2.3) holds. (2.2) means that f has a Taylor series expansion to order ν at the point 0, the β_j are the Taylor coefficients of f, and f differs from its Taylor series by no more than $M(x)$ at the point x. As a particular case, the choices $d = 1$, $f_0(x) = 1$, $f_1(x) = x$ and $M(x) = x^2$ yield the approximately linear regression model $|EY_x - \beta_0 - \beta_1 x| = |f(x) - f(0) - f'(0)x| \leqq x^2$ for all x.

EXAMPLE 2. Let $\mathscr{X} = R^1$ and suppose $z_0, \cdots, z_k$ are distinct real numbers. Take $f_j(x) = \prod_{i \neq j}^{k} (x - z_i) / \prod_{i \neq j}^{k} (z_j - z_i)$ and let M be any nonnegative function that vanishes at $z_0, \cdots, z_k$. The f_j are the Lagrange interpolating polynomials of degree k so (2.3) applies. Under (2.2), $\beta_j = f(z_j)$ and the function f differs

from its Lagrange interpolating polynomial by no more than $M(x)$ at the point x. As a special case, the choices $k = 1$, $z_0 = 0$, $z_1 = 1$ and $M(x) = x^2 \wedge (x-1)^2$ yield the approximately linear regression model $|EY_x - \beta_0(1-x) - \beta_1 x| = |f(x) - f(0)(1-x) - f(1)x| \leqq x^2 \wedge (x-1)^2$ for all x.

EXAMPLE 3. Let $f_0, \cdots, f_k$ be a Chebyshev system on an interval $\mathscr{X} \subset R^1$, i.e., $\sum_0^k a_j f_j(x)$ does not vanish more than k times (counting multiplicities) unless all the a_j's are zero. If M has $k+1$ zeroes (counting multiplicities) on $\mathscr{X}$ then (2.3) is satisfied and, if f satisfies (2.2), the coefficients β_j can be related to certain linear combinations of $\{f(x_l), f^{(1)}(x_l), \cdots, f^{(m_l)}(x_l)\}$ where $M(x_l) = 0$ with multiplicity m_l. For example, $\mathscr{X} = [0, \pi]$, $f_0(x) \equiv 1$, $f_1(x) = \sin x$, $f_2(x) = \cos x$, $M(x) = |\sin 2x|$ gives an approximate trigonometric regression model.

EXAMPLE 4. Begin with a standard linear model in the form $Y = A\beta + \varepsilon$ where $Y = (Y_1, \cdots, Y_n)'$ is a vector of observations, ε is a vector of uncorrelated random variables with mean zero, and A is a known $n \times (k+1)$ matrix with rank $k+1$. Now take $\mathscr{X} = \{1, \cdots, n\}$ and regard the columns $f_0, \cdots, f_k$ of A as functions on $\mathscr{X}$. If (2.3) applies for some $M \geqq 0$, one may take the approximately linear model to be $|EY - A\beta| \leqq M$ for some (unique) β. It is important to note that (2.3) can only apply if M vanishes often enough and, when this is so, the β_j are appropriate linear combinations of the mean values at the points where $M = 0$.

As a particular case, consider the $r \times c$ analysis of variance model without interactions and with one observation per cell. Take $M = 0$ in the first row and column with M arbitrary but nonnegative elsewhere. The approximately linear model which results (using standard notation) has

$$\begin{aligned} &|EY_{ij} - \mu - \mu_{i\cdot} - \mu_{\cdot j}| \\ &\quad = |EY_{ij} - E(Y_{1\cdot} + Y_{\cdot 1} - Y_{11}) - E(Y_{i1} - Y_{\cdot 1}) - E(Y_{1j} - Y_{1\cdot})| \\ &\quad = |E(Y_{ij} - Y_{i1} - Y_{1j} + Y_{11})| \leqq M_{ij} \qquad \text{for} \quad i \leqq r, \quad j \leqq c. \end{aligned}$$

Other approximate linear models in the same context are given by

$$|EY_{ij} - \mu| = |E(Y_{ij} - Y_{11})| \leqq M_i \qquad \text{with} \quad M_1 = 0$$

which arises as an approximation to the linear model $Y_{ij} = \mu + \varepsilon_{ij}$ and, similarly,

$$|EY_{ij} - \mu - \mu_{i\cdot}| = |E(Y_{ij} - Y_{i1})| \leqq M_{ij} \qquad \text{with} \quad M_{i1} = 0, \quad i = 1, \cdots, r,$$

which arises as an approximation to the linear model $EY_{ij} = \mu + \mu_{i\cdot} + \varepsilon_{ij}$.

3. Linear estimation of a single parameter. Suppose that observations are taken according to the model given by (2.1) and (2.2) for some specified functions $f_0, \cdots, f_k$ and $M \geqq 0$. The problem investigated here is the estimation of a single linear combination $L'\beta = \sum_{j=1}^k l_j \beta_j$. The estimates considered are linear in the observations; a comment on this restriction is made in Remark 5. The identifiability assumption (2.3) is not needed here; see Remark 6 for further

discussion about (2.3). In this section and the next σ^2 is to be regarded as known; the problem with unknown σ^2 is discussed in Section 6.

Consider the mean square error of estimation of $L'\beta$ by $\sum_{i=1}^n c_i Y_i$ when $f(x) = \sum_{j=0}^k \beta_j f_j(x) + r(x)$ and $|r(x)| \leqq M(x)$ for all x. One finds

$$
\begin{aligned}
&E(\textstyle\sum_{i=1}^n c_i Y_i - L'\beta)^2 \\
&\quad = E(\textstyle\sum_{i=1}^n c_i(Y_i - f(x_i)))^2 + (\sum_{i=1}^n c_i f(x_i) - L'\beta)^2 \\
&\quad = \textstyle\sum_{i=1}^n c_i^2\sigma^2(x_i) + (\sum_{i=1}^n c_i \sum_{j=0}^k \beta_j f_j(x_i) + \sum_{i=1}^n c_i r(x_i) - L'\beta)^2 .
\end{aligned} \tag{3.1}
$$

Now β is an arbitrary $(k+1)$-vector so (3.1) is unbounded unless

$$
\textstyle\sum_{i=1}^n c_i f_j(x_i) = l_j , \qquad j \leqq k . \tag{3.2}
$$

As in the linear model $L'\beta$ is said to be *estimable* if (3.2) admits some solution in the c_i's. Then, if $L'\beta$ is estimable and the c_i satisfy (3.2), (3.1) becomes

$$
\textstyle\sum_{i=1}^n c_i^2\sigma^2(x_i) + (\sum_{i=1}^n c_i r(x_i))^2 . \tag{3.3}
$$

The convexity of (3.3) in the c_i's reduces consideration to estimates with $c_i = c(x_i)$, i.e., to weights c_i that depend only on the location of the observations. Note also that the maximum of (3.3) as $r(x)$ varies subject to $|r(x)| \leqq M(x)$ is given by

$$
\textstyle\sum_{i=1}^n c^2(x_i)\sigma^2(x_i) + (\sum_{i=1}^n |c(x_i)|M(x_i))^2 . \tag{3.4}
$$

Thus, in order to minimax the mean square error, one must find a function c which minimizes (3.4) subject to the estimability conditions (3.2).

To facilitate the exposition, let ξ be the design (counting) measure, $\xi(\{x\}) = \#(x_i = x)$ and write $F(x) = (f_0(x), \cdots, f_k(x))'$. As described above the problem is to minimize

$$
J(c) = \textstyle\int c^2\sigma^2 \, d\xi + (\int |c|M \, d\xi)^2 \tag{3.5}
$$

subject to

$$
\textstyle\int cF \, d\xi = L . \tag{3.6}
$$

THEOREM 1. *Suppose $L'\beta$ is estimable. Then there is a unique c^0 which minimizes* (3.5) *subject to* (3.6) *and is given by*

$$
c^0(x) = \sigma^{-2}(x)[(b'F(x) - \lambda M(x))^+ - (b'F(x) + \lambda M(x))^-] \tag{3.7}
$$

where (b, λ) is any solution to

$$
\begin{aligned}
&\textstyle\int \sigma^{-2}[(b'F - \lambda M)^+ - (b'F + \lambda M)^-]F \, d\xi = L \\
&\textstyle\int \sigma^{-2}[(b'F - \lambda M)^+ + (b'F + \lambda M)^-]M \, d\xi = \lambda .
\end{aligned} \tag{3.8}
$$

(Note: (3.8) means c^0 satisfies (3.6) and $\int |c^0|M \, d\xi = \lambda$.)

PROOF. According to Theorem 3.9 of Whittle [11], there is a vector b of Lagrange multipliers such that

$$
K(c) = J(c) - 2 \textstyle\int c \cdot b'F \, d\xi \tag{3.9}
$$

is minimized at c^0 where c^0 minimizes (3.5) under (3.6). Because K is strictly convex, if c^0 minimizes K it does so uniquely and

$$\lim_{\varepsilon \downarrow 0} \frac{K(c^0 + \varepsilon\varphi) - K(c^0)}{\varepsilon} \geqq 0 \tag{3.10}$$

for any φ. Calculating the limit in (3.10) one finds

$$\begin{aligned} \int [c^0\sigma^2 + (\int |c^0|M\, d\xi)(\operatorname{sgn} c^0)M - b'F]\varphi\, d\xi \\ + (\int |c^0|M\, d\xi) \int (1 - |\operatorname{sgn} c^0|)|\varphi|M\, d\xi \geqq 0 \end{aligned} \tag{3.11}$$

for any φ.

If c^0 minimizes (3.5) subject to (3.6) and $\int |c^0|M\, d\xi = 0$ then, from (3.11), $c^0(x) = \sigma^{-2}(x) \cdot b'F(x)$ a.e. ξ and thus c^0 has the required form with $\lambda = 0$.

If c^0 minimizes (3.5) subject to (3.6) and $\int |c^0|M\, d\xi = \lambda > 0$, then (3.11) implies that on the set $\{c^0 = 0\}$, $|b'F| \leqq \lambda M$; on the set $\{c^0 > 0\}$, $c^0 = \sigma^{-2}(b'F - \lambda M)$; on the set $\{c^0 < 0\}$, $c^0 = \sigma^{-2}(b'F + \lambda M)$. Thus c^0 has the required form.

If (b, λ) satisfies (3.8) and c^0 is defined by (3.7) then, since $\lambda \geqq 0$, (3.11) holds for any φ. It follows that K is minimized at c^0 which, by uniqueness and the first sentence of the proof, is the solution. The theorem is proved.

REMARK 1. When $\lambda = 0$, c^0 gives the least squares estimate of $L'\beta$. This necessarily occurs when $M \equiv 0$ but it can also happen when $M \geqq 0$, in which case the least squares weight function c^* must satisfy $\int |c^*|\, M\, d\xi = 0$. The latter event is not a common occurrence and is generally precluded in settings like those of Examples 1, 2 and 3 of Section 2.

REMARK 2. It is useful to note that

$$\begin{aligned} \int (c^0)^2\sigma^2\, d\xi &= \int c^0(b'F - (\operatorname{sgn} c^0)\lambda M)\, d\xi \\ &= L'b - \lambda \int |c^0|M\, d\xi = L'b - \lambda^2 \end{aligned}$$

which implies that

$$J(c^0) = L'b\,. \tag{3.12}$$

In particular this means that if the problem is one of estimating β_0 then $L'b = b_0 > 0$.

REMARK 3. The qualitative difference between the estimate from the approximately linear model and the one from the ideal linear model is that the former estimate dampens the influence of those observations drawn at levels "far" from the linear structure, i.e., of observations at levels where M is large. In fact, there is also a truncation effect in operation. To see this, consider the particular linear regression model cited at the end of Example 1 in Section 2. For the ideal linear version the least squares weight function for estimating β_0 is linear after multiplication by σ^2, and observations far from $x = 0$ are given the largest weight. In the approximate model with $M(x) = x^2$, the optimum weight function (following multiplication by σ^2) consists of a positive quadratic loop of the form $(b_0 + b_1 x - \lambda x^2)$ covering the origin (since $b_0 > 0$ by the comment following

(3.12)) with the possible addition of a negative quadratic loop $(b_0 + b_1 x + \lambda x^2)$ on an interval disjoint from the first. Here observations far from $x = 0$ tend to be discarded. The specifics depend, of course, on the measure ξ as it gives rise to values of b_0, b_1 and λ. The presence of the negative loop is assured if ξ is concentrated on $(0, \infty)$ since in this case one cannot satisfy $\int (b_0 + b_1 x - \lambda x^2) x\sigma^{-2}(x)\, d\xi = 0$. The negative loop will usually not be present if 0 is in the "middle" of the support of ξ.

REMARK 4. When $M \equiv 0$ the optimum estimate of $L'\beta$ is $L'\hat{\beta}$ where $\hat{\beta}$ is the vector of optimum estimates of the individual β_i's. This is usually false when $M \not\equiv 0$.

REMARK 5. The restriction to linear estimates remains an open issue. In the case of the ideal model it is not a serious one (at least from the minimax view). In the case of no error (i.e., $\sigma^2 \equiv 0$) the use of linear estimates is no restriction (this result is credited to Smolyak by Michelli [10]). It is certain that when ε is nondegenerate normal there is something to be gained by use of nonlinear estimates, but how much is unclear. Of course, any discussion which attempts to handle both model robustness and distributional robustness would necessarily entail dealing with nonlinear estimates.

REMARK 6. As noted at the beginning of this section (2.3) has not been used in finding the optimum estimates. The function of (2.3) is to permit an unequivocal interpretation of the parameters and therefore of the estimates. One can always imbed a given problem in a new one for which identifiability holds; however, the new $\mathscr{X}$ may introduce fictitious treatments or locations and may be far from realistic so that interpretation of the parameters will remain elusive. The point of view taken here is that to each problem there is a natural $\mathscr{X}$ (certainly this is true of the examples in Section 2) and that identifiability with respect to this $\mathscr{X}$ is required to make clear the meaning of the parameters and the estimates. (It may be noted here that for identification of aparticular parametric function $L'\beta$, (2.3) may be weakened to

$$(3.13) \qquad |\sum \beta_j f_j(x)| \leqq M(x) \quad \text{for all } x \in \mathscr{X} \text{ and some } \beta \text{ implies } L'\beta = 0.$$

When (3.13) holds, one may have $\beta \neq \gamma$ so that $|f(x) - \sum_{j=0}^{k} \beta_j f_j(x)| \leqq M(x)$ for all $x \in \mathscr{X}$ and $|f(x) - \sum_{j=0}^{k} \gamma_j f_j(x)| \leqq M(x)$ for all $x \in \mathscr{X}$ but then necessarily $L'\beta = L'\gamma$.)

4. Several parameters. This section is devoted to a discussion of the simultaneous estimation of several linear functions of β under the model of (2.1) and (2.2).

Suppose Λ is an $s \times (k + 1)$ matrix ($s \leqq k + 1$) and the problem is to estimate $\Lambda\beta$ using linear estimates. If C is an s-vector of functions on $\mathscr{X}$ the estimability

restriction on the estimate $\sum_{i=1}^{n} C(x_i)Y_i$ becomes, as in Section 3,

$$\int CF'\,d\xi = \Lambda\,, \tag{4.1}$$

and the mean square error matrix is

$$\begin{aligned} E(\textstyle\sum_{i=1}^{n} C(x_i)Y_i - \Lambda\beta)(\sum_{i=1}^{n} C(x_i)Y_i - \Lambda\beta)' \\ = \int CC'\sigma^2\,d\xi + (\int Cr\,d\xi)(\int Cr\,d\xi)' = I_V(C) + I_B(r, C)\,, \end{aligned} \tag{4.2}$$

where the notation V, B refers to variance and bias, respectively.

In order to compare matrices consider criteria Φ, i.e., functions Φ on the nonnegative definite $s \times s$ matrices satisfying $\Phi(\Sigma_1) \geqq \Phi(\Sigma_2)$ whenever $\Sigma_1 - \Sigma_2$ is nonnegative definite, and the problem of minimizing

$$J_\Phi(C) = \max_{|r|\leqq M} \Phi(I_V(C) + I_B(r, C)) \tag{4.3}$$

subject to (4.1). This is not a tractable problem even in the case $\Phi(\Sigma) = \operatorname{tr}(\Sigma)$ where the problem is to minimize, subject to (4.1),

$$\begin{aligned} J_{\mathrm{tr}}(C) &= \int C'C\sigma^2\,d\xi + \max_{|r|\leqq M} (\int Cr\,d\xi)'(\int Cr\,d\xi) \\ &= \int C'C\sigma^2\,d\xi + \max_{|u|=1} (\int CMu\,d\xi)'(\int CMu\,d\xi)\,. \end{aligned} \tag{4.4}$$

Even though J_{tr} is strictly convex in C the Lagrange multiplier technique employed in Theorem 1 does not help much because of the nature of the second term on the right side of (4.4) under small perturbations of C.

Some useful bounds can be obtained which suggest that in a variety of cases use of the single linear combination estimates of Section 3 will give satisfactory results. To see this let L_α' denote the αth row of Λ and let $v_\alpha = J(c_\alpha^0)$ where c_α^0 is the solution of Theorem 1 for the problem with L replaced by L_α. Then, if c_α denotes the αth coordinate of C and (4.1) holds,

$$\begin{aligned} J_{\mathrm{tr}}(C) &\geqq \textstyle\sum_{j\neq\alpha} \int c_j^2\sigma^2\,d\xi + \int c_\alpha^2\sigma^2\,d\xi + (\int |c_\alpha| M\,d\xi)^2 \\ &\geqq \textstyle\sum_{j\neq\alpha} \int c_j^2\sigma^2\,d\xi + v_\alpha \\ &\geqq \textstyle\sum_{j\neq\alpha} v_j^* + v_\alpha \end{aligned}$$

where v_j^* is the variance of the least squares estimate of $L_j'\beta$. It follows that

$$\min J_{\mathrm{tr}}(C) \geqq \max_{1\leqq\alpha\leqq s} (\textstyle\sum_{j\neq\alpha} v_j^* + v_\alpha)\,. \tag{4.5}$$

Let $C^0 = (c_1^0, \cdots, c_s^0)'$. Then

$$J_{\mathrm{tr}}(C^0) \leqq \textstyle\sum_1^s v_\alpha\,. \tag{4.6}$$

If the ratio of the right sides of (4.5) and (4.6) is close to 1 then C^0 is a satisfactory estimate. Of course the upper bound of (4.6) is made unnecessary by a computation of $J_{\mathrm{tr}}(C^0)$.

Here is another criterion which can be discussed in a similar way. Let $\Phi(\Sigma) =$ max eigenvalue of $\Sigma = \max_{a'a=1} a'\Sigma a$. Then

$$J_{ME}(C) = \max_{a'a=1} (\int (a'C)^2\sigma^2\,d\xi + (\int |a'C| M\,d\xi)^2)$$

and minimizing J_{ME} subject to (4.1) again presents difficulties. A lower bound

is obtained by writing

$$\begin{aligned}\inf_C J_{ME}(C) &\geqq \max_{a'a=1} \inf_C [\int (a'C)^2\sigma^2\, d\xi + (\int |a'C|M\, d\xi)^2] \\ &\geqq \max_{a'a=1} \inf_{\gamma|\int \gamma F' = a'\Lambda} J(\gamma) \\ &= \max_{a'a=1} v(a'\Lambda) \geqq \max_{1\leqq\alpha\leqq s} v_\alpha\end{aligned}$$

where $v(a'\Lambda)$ is the minimum value of J for the problem of Section 3 when L' is replaced by $a'\Lambda$. A crude upper bound on $J_{ME}(C^0)$ is easily obtained from (4.6) since $J_{ME}(C^0) \leqq J_{tr}(C^0)$. This bound is useful if there is one β_α which dominates in the sense that it is much more difficult to estimate than the others.

For criteria of the form $\Phi(\Sigma) = \operatorname{tr} \Sigma Q$ where Q is positive definite, the discussion with J_{tr} is pertinent by changing Λ to $Q^{\frac{1}{2}}\Lambda$. Other criteria of the type discussed here can also be dealt with in the same fashion but a detailed analysis is yet to be done.

5. Computation of c^0. In this section an algorithm for computing the optimum solution of Theorem 1 is proposed and discussed. Although $\sigma^2\,(= \sigma^2(x))$ is assumed known here, the algorithm makes sense if an estimate of σ^2 is used; further discussion on this point is to be found in Section 6.

Some notation is required. To this end let S^+ and S^- be disjoint subsets of $\mathscr{X}$ and let $S^* = S^+ \cup S^-$, $S = (S^+, S^-)$. Define the function M_S by $M_S(x) = M(x)$ if $x \in S^+$, $M_S(x) = -M(x)$ if $x \in S^-$, and let H_S be the $(k+2) \times (k+2)$ matrix

$$H_S = \begin{pmatrix} \int_{S^*} FF'\sigma^{-2}\, d\xi & \int_{S^*} FM_S\sigma^{-2}\, d\xi \\ \int_{S^*} M_S F'\sigma^{-2}\, d\xi & \int_{S^*} M_S^2\sigma^{-2}\, d\xi + 1 \end{pmatrix}. \tag{5.1}$$

If $S_0^+ = \{c^0 > 0\}$ and $S_0^- = \{c^0 < 0\}$ then the equations (3.8) can be written as

$$H_S\binom{b}{-\lambda} = \binom{L}{0} \tag{5.2}$$

with $S = S_0$. On the other hand, if (b, λ) satisfies (5.2) and if $c(x)$ is defined by the right side of (3.7) for this (b, λ), then $c = c^0$ provided $\{c > 0\} = S^+$ and $\{c < 0\} = S^-$. Thus to find c^0, it suffices to find an $S = (S^+, S^-)$ so that when (5.2) is solved for (b, λ) and c is formed as at (3.7), S^+ and S^- are the positive and negative sets of c, respectively.

Based on the above observation, the following algorithm has proved useful. Let $S_1 = (S_1^+, S_1^-)$ be a starting pair of sets. If S_{p-1} is defined for $p \geqq 2$, let (b_p, λ_p) satisfy

$$H_{S_{p-1}}\binom{b_p}{\lambda_p} = \binom{L}{0} \tag{5.3}$$

and set

$$\begin{aligned} c_p(x) &= \sigma^{-2}\{(b_p'F - \lambda_p M)^+ - (b_p'F + \lambda_p M)^-\} \\ S_p^+ &= \{c_p > 0\}, \qquad S_p^- = \{c_p < 0\}. \end{aligned} \tag{5.4}$$

If $S_p = S_{p-1}$ (i.e., $S_p^+ = S_{p-1}^+$ and $S_p^- = S_{p-1}^-$), then stop because $c_p = c^0$. If $S_p \neq S_{p-1}$, continue through (5.3) to S_{p+1}.

There are two potential difficulties in reaching the optimum solution as outlined: the equations (5.3) may not be solvable at some stage, and cycling might occur. In the case of one regression function ($k = 0$) it is shown below that for any starting pair S_1 for which $\int_{S_1^*} f M_{S_1} \sigma^{-2}\, d\xi > 0$, the algorithm stops with $S_{p-1} = S_p$, some p. The corresponding result when $k > 0$ (and with S_1 subject only to minor restrictions) is not known, but in all examples thus far attempted the solution has been found by the present method, and in very few steps.

It has been pointed out to us by George Knafl that the algorithm just described is closely related to the Newton–Raphson algorithm. This can be seen as follows: let $v = \binom{b}{\lambda}$, let c be determined from (3.7) and take $S^+(v)$, $S^-(v)$ to be the sets where $c > 0$, $c < 0$, respectively. Let $g(v) = H_{S(v)} v$ and $\Delta = \binom{L}{0}$. The problem is to solve $g(v) = \Delta$ and, formally, the Newton–Raphson method produces recursively

$$v_p = v_{p-1} - (Jg(v_{p-1}))^{-1}(g(v_{p-1}) - \Delta)$$

where $(Jg)^{-1}$ is the inverse of the Jacobian of g. If $S(v)$ does not change in a neighborhood of v_{p-1} (as will often happen), $Jg(v_{p-1}) = H_{S(v_{p-1})}$ and then $v_p = H^{-1}_{S(v_{p-1})}\Delta$ as at (5.3). Jg does not exist everywhere and, even though directional derivatives exist so that Jg may be defined and is perhaps invertible, it is not clear how to produce a proof of convergence. Modifications of Newton–Raphson can be employed to assure convergence in most cases; e.g., let $v_p = v_{p-1} - d(g(v_{p-1}) - \Delta)$ for a positive constant d. These modifications are presently being investigated.

Here is the proof for the case $k = 0$. Write $f_0 = f$ and fix $L = 1$. Assume first that the least squares estimate is not optimum, as can be easily checked, so that the optimum (b_0, λ_0) satisfies $b_0 > 0$, $\lambda_0 > 0$. S_0 will denote the optimum pair (S_0^+, S_0^-).

Take S_1 to be a starting pair with $S_1^+ = \{b_1 f - \lambda_1 M > 0\}$ and $S_1^- = \{b_1 f + \lambda_1 M < 0\}$, $b_1 > 0$, $\lambda_1 > 0$. The function f is positive on S_1^+ and negative on S_1^-, insure that at least one of these sets is nonempty. Now solve (5.3) with $p = 2$ to produce (b_2, λ_2): this may be done since $|H_{S_1}| \geqq \int_{S_1^*} f^2\sigma^{-2}\, d\xi > 0$. It can be easily argued from (5.3) that both b_2 and λ_2 are positive. Denote b_i/λ_i by ρ_i, $i = 0, 1, \cdots$, and write the second equation of (5.3) in the form

$$\int_{S_1^+} (\rho_2 f - M)M\sigma^{-2}\, d\xi + \int_{S_1^-} (-\rho_2 f - M)M\sigma^{-2}\, d\xi = 1\,. \tag{5.5}$$

Suppose first that $\rho_1 < \rho_0$. If also $\rho_2 < \rho_0$, make the following replacements in (5.5): ρ_2 by ρ_0, S_1^+ by S_0^+ ($\supset S_1^+$) and S_1^- by S_0^- ($\supset S_1^-$). But then

$$\int_{S_0^+} (\rho_0 f - M)M\sigma^{-2}\, d\xi + \int_{S_0^-} (-\rho_0 f - M)M\sigma^{-2}\, d\xi > 1\,, \tag{5.6}$$

which contradicts the nature of S_0, (b_0, λ_0), since optimality necessitates equality in (5.6). Thus $\rho_1 < \rho_0$ implies $\rho_2 \geqq \rho_0$.

Suppose next that $\rho_1 > \rho_0 > \rho_2$ so, in particular, $S_0^+ \subset S_1^+$ and $S_0^- \subset S_1^-$.

Then

$$\int_{S_1^+}(\rho_2 f - M)M\sigma^{-2}\,d\xi < \int_{S_1^+}(\rho_0 f - M)M\sigma^{-2}\,d\xi \leqq \int_{S_0^+}(\rho_0 f - M)M\sigma^{-2}\,d\xi \tag{5.7}$$

because $\rho_0 f - M \leqq 0$ on $S_1^+ - S_0^+$. Similarly,

$$\int_{S_1^-}(-\rho_2 f - M)M\sigma^{-2}\,d\xi < \int_{S_0^-}(-\rho_0 f - M)M\sigma^{-2}\,d\xi\,. \tag{5.8}$$

Adding (5.7) to (5.8) leads to the contradiction at (5.6). Thus $\rho_0 < \rho_1$ implies $\rho_0 \leqq \rho_2$.

Suppose finally that $\rho_0 < \rho_1 < \rho_2$. The first equation at (5.3) with $p = 2$ can be written as

$$\int_{S_1^+}(f\rho_2 - M)f\sigma^{-2}\,d\xi + \int_{S_1^-}(f\rho_2 + M)f\sigma^{-2}\,d\xi = \lambda_2^{-1}\,.$$

Now f is positive on S_1^+ and negative on S_1^- so it cannot be that both $S_2^+ = \{b_2 f - \lambda_2 M > 0\}$ and $S_2^- = \{b_2 f + \lambda_2 M < 0\}$ are empty, since $\lambda_2 > 0$. Then solve (5.3) with $p = 3$ to get (b_3, λ_3). An argument above gives $\rho_3 \geqq \rho_0$ and then, if $\rho_3 > \rho_2$,

$$\int_{S_2^+}(\rho_3 f - M)M\sigma^{-2}\,d\xi > \int_{S_2^+}(\rho_2 f - M)M\sigma^{-2}\,d\xi \geqq \int_{S_1^+}(\rho_2 f - M)M\sigma^{-2}\,d\xi\,. \tag{5.9}$$

In the same way,

$$\int_{S_2^-}(-\rho_3 f - M)M\sigma^{-2}\,d\xi > \int_{S_1^-}(-\rho_2 f - M)M\sigma^{-2}\,d\xi\,. \tag{5.10}$$

Add (5.9) to (5.10) to get a contradiction. Thus when $\rho_0 < \rho_1 < \rho_2$, $\rho_3 \leqq \rho_2$.

The above facts together show that $b_2/\lambda_2 \geqq b_3/\lambda_3 \geqq \cdots \geqq b_0/\lambda_0$. In terms of the pairs S_i, this means $S_2^+ \supseteq S_3^+ \supseteq \cdots \supseteq S_0^+$, $S_2^- \supseteq S_3^- \supseteq \cdots \supseteq S_0^-$. Since only finitely many sets are available, $S_p = S_0$ for some p. The special starting pairs S_1 used above are not required for if $\int_{S_1^+} f M_{S_1}\sigma^{-2}d\xi > 0$, (5.3) can be solved for (b_2, λ_2) and, in fact, $b_2 > 0$, $\lambda_2 > 0$.

A convenient starting place for the algorithm is $S_1^+ = \{c^* > 0\}$ and $S_1^- = \{c_1^* < 0\}$, where c^* is the least squares weight function. This choice is efficient if M is small. If M is not small, one can proceed by solving the problem for εM with ε small and then using $c^0(\varepsilon)$ (the optimum solution for εM) to determine a starting pair for the problem with $2\varepsilon M$, etc. This procedure has been carried out and found to be efficient, because the number of iterations required to solve an auxiliary problem is usually 1 or 2, even if ε is moderate.

6. Examples; unknown σ^2. This section covers some specifics of the general examples given in Section 2. The results are suggestive of what might be expected in other cases. In Examples 1 — 6 below, the calculations are done as if σ^2 is known; the final comment of the section deals with the question of unknown σ^2.

EXAMPLE 1. Consider the model of the type discussed in Example 4 of Section

2 with

$$Y_{ij} = \mu + r_i + \sigma_i \varepsilon_{ij}\,; \qquad j = 1, \cdots, n_i,\ i = 1, 2\,,$$

and $|r_i| \leqq M_i$. Assume $M_1 \leqq M_2$. The problem is to estimate μ and here $F = f_0 \equiv 1$. If $v_1 = \sigma_1^2/n_1$, $v_2 = \sigma_2^2/n_2$ then this model can be reduced to

$$Y_{i\cdot} = \mu + r_i + v_i^{\frac{1}{2}}\bar{\varepsilon}_i \qquad i = 1, 2\,,$$

$\mathscr{X} = \{1, 2\}$. Since $L = 1$ and $L'b > 0$ (see (3.12)) it must be that $b > 0$ and, consequently, $c^0(i) = (b - \lambda M_i)^+/v_i$—the negative part never appears. Because $M_1 \leqq M_2$ the optimum S^+ is either $\{1\}$ or $\{1, 2\}$. In order for S^+ to be $\{1\}$ it must be, according to the discussion at the beginning of this section, that

$$\binom{b}{-\lambda} = \begin{pmatrix} \frac{1}{v_1} & \frac{M_1}{v_1} \\ \frac{M_1}{v_1} & \frac{M_1^2}{v_1} + 1 \end{pmatrix}^{-1} \binom{1}{0}$$

satisfies $b - \lambda M_1 > 0$ and $b - \lambda M_2 \leqq 0$. Since $b, -\lambda$ are the same positive multiple of $(M_1^2/v_1) + 1$ and $-M_1/v_1$ respectively, the first condition $b - \lambda M_1 > 0$ is assured, while $b - \lambda M_2 \leqq 0$ if and only if

$$\frac{M_1^2}{v_1} + 1 - \frac{M_1 M_2}{v_1} \leqq 0\,. \tag{6.1}$$

Thus if (6.1) holds the optimum estimate of μ is determined by $c^0(1) = 1$, $c^0(2) = 0$, i.e., the estimate is $Y_{1\cdot}$. If (6.1) does not hold then $S^+ = \{1, 2\}$ and

$$H_S = \begin{pmatrix} \frac{1}{v_1} + \frac{1}{v_1} & \frac{M_1}{v_1} + \frac{M_2}{v_2} \\ \frac{M_1}{v_1} + \frac{M_2}{v_2} & \frac{M_1^2}{v_1} + \frac{M_2^2}{v_2} + 1 \end{pmatrix}.$$

One then calculates

$$c^0(1) = \frac{(M_2 - M_1)\dfrac{M_2}{v_1 v_2} + \dfrac{1}{v_1}}{\dfrac{1}{v_1} + \dfrac{1}{v_2} + \dfrac{1}{v_1 v_2}(M_2 - M_1)^2}\,,$$

$$c^0(2) = \frac{(M_1 - M_2)\dfrac{M_1}{v_1 v_2} + \dfrac{1}{v_2}}{\dfrac{1}{v_1} + \dfrac{1}{v_2} + \dfrac{1}{v_1 v_2}(M_2 - M_1)^2}\,.$$

Note that if $M_1 = 0$ (which is required for identifiability) the result implies that information from the second population is always useful, but to a small extent if M_2 is large.

EXAMPLE 2. In the context of Example 1 of Section 2 take $k = 0$, $f_0 \equiv 1$, $M(x) = |x|$, and d arbitrary. Then $c^0 = \sigma^{-2}(b - \lambda|x|)^+$ since $L = 1$ implies $b > 0$

and then no negative loop appears. It follows that the observations which are used lie in some sphere around 0. The calculation of b, λ is easy and assured by the algorithm of Section 5.

EXAMPLE 3. Consider the case of approximately linear regression, i.e., the setup of Example 1 of Section 2 with $k = 1$, and take $M(x) = m|x|^2$. Let $L' = (1, 0, \cdots, 0)$. Then the c^0 which is optimum for estimating β_0 is given by

$$c^{0+} = \sigma^{-2}(b + \textstyle\sum_1^d b_j x^{(j)} - \lambda m|x|^2)^+$$

where $x^{(j)}$ is the jth coordinate of x, and c^{0-} is obtained by changing the sign of the coefficient of $|x|^2$ in c^{0+} and taking the negative part. Thus the support of c^{0+} is the set of all points x in the support of ξ which lie in a ball B^+ (say) and the negative part of c^0 has support on a ball whose center is diametrically opposite to that of B^+ but whose radius is smaller (or 0).

For the specific situations given below the calculations were done using the algorithm of Section 5. In all cases it is only necessary to describe c^{0+}.

EXAMPLE 3(a). Let $d = 1$, $m = \frac{1}{2}$, $\sigma^2 \equiv .525$, $n = 21$, and let the x_i's be uniformly spaced on $[0, 3]$ including the endpoints. For estimating β_0,

$$c^{0+}_{\beta_0} = (.413 - .555x - .173x^2)^+, \qquad J(c^0_{\beta_0}) = .217.$$

If λ is desired, note that $\sigma^{-2}\lambda m = .173$. For estimating β_1,

$$c^{0+}_{\beta_1} = (-.549 + 1.34x - .605x^2)^+, \qquad J(c^0_{\beta_1}) = .704.$$

The weight functions $c^0_{\beta_0}$ and $c^0_{\beta_1}$ differ markedly from their least squares versions. In particular, the positive loop of $c^0_{\beta_0}$ is over $[0, .6]$ while the negative loop covers $[1.20, 1.95]$. Here the truncation effect is manifest: observations at $x \geqq 2.10$ are not used. For $c^0_{\beta_1}$ the situation is similar: it is negative on $[0, .3]$ and positive on $[.6, 1.65]$.

EXAMPLE 3(b). Let $d = 1$, $m = \frac{1}{2}$, $\sigma^2 \equiv .525$, $n = 21$, and let the x_i's be uniformly spaced on $[-1, 2]$ including the endpoints. Then

$$c^{0+}_{\beta_0} = (.109 - .00022x - .102x^2)^+, \qquad J(c^0_{\beta_0}) = .057,$$
$$c^{0+}_{\beta_1} = (-.0037 + .454x - .354x^2)^+, \qquad J(c^0_{\beta_1}) = .239.$$

In this case $c^{0-}_{\beta_0} \equiv 0$, which typically happens when 0 lies near the middle of the support of ξ. The negative loop of $c^0_{\beta_1}$ covers the x_i's in $[-1, -.1]$ while the positive loop covers $[.05, 1.25]$. The marked reduction in mean square error from that of 3(a) is due to the change in design.

EXAMPLE 4. Consider the nearly quadratic model with $k = 2$, $f_j(x) = x^j$, $j = 0, 1, 2$, $M(x) = |x|^3/6$, $\sigma^2 \equiv .525$, $n = 21$, and the x_i's equally spaced in $[-1, 2]$. Then

$$c^{0+}_{\beta_0} = (.120 + .0271x - .107x^2 - .0279|x|^3)^+, \qquad J(c^0_{\beta_0}) = .063,$$
$$c^{0+}_{\beta_1} = (.0215 + .185x - .0771x^2 - .414|x|^3)^+, \qquad J(c^0_{\beta_1}) = .097.$$

Here $c^0_{\beta_0}$ has a large positive loop covering $[-.85, .9]$ and a small negative loop over $[1.7, 2]$. Similarly, $c^0_{\beta_1}$ has one positive and one negative loop. In other cases, there is the possibility of two positive or two negative loops. In fact, leaving other matters the same, the first possibility occurs for estimating β_0 when the design is on $[0, 3]$ and $M(x) = |x|^3/2$; two negative loops show up when the design is on $[-1, 2]$ and $M(x) = |x|^3/2$.

It is interesting to compare the computations in 3(b) with the present ones. Real differences in the models do not appear in comparing the $c^0_{\beta_0}$'s: these are fairly close to one another, with a small difference in mean square error. However, the effect of changing models on the estimation of β_1 is noticeable: there is a substantial drop in mean square error. It appears that an approximately linear model may be adequate for estimating $f(0)$, but that one requires at least an approximately quadratic model to get satisfactory estimates of $f'(0)$.

EXAMPLE 5. Consider the following specific case of Example 2, Section 2. Let $k = 1, f_0(x) = 1 - x, f_1(x) = x, M(x) = \min(x^2, (1 - x)^2)/2, \sigma^2 \equiv .525$, and suppose there are 21 observations at equally spaced points of $[0, 3]$. Then

$$c^{0+}_{\beta_0} = (.282(1 - x) + .054x - .262M(x))^+, \qquad J(c^0_{\beta_0}) = .282,$$
$$c^{0+}_{\beta_1} = (.06(1 - x) + .09x - .178M(x))^+, \qquad J(c^0_{\beta_1}) = .09.$$

Here $c^0_{\beta_1}$ is positive on $[0, 2.10]$ and is never negative, while $c^0_{\beta_0}$ is positive on $[0, 1, 2]$ and negative on $[1.35, 2.4]$.

EXAMPLE 6. The computation of the bounds (4.5) and (4.6) for the simultaneous estimation of β_0 and β_1 in 3, 4 and 5 produces the following table.

TABLE 1

Example	Lower Bound ((4.5))	Upper Bound ((4.6))
3(a)	.797	.921
3(b)	.270	.296
4	.145	.160
5	.342	.372

Unknown σ^2. When σ^2 is unknown it has to be estimated. An appropriate procedure in 1, for example, is to estimate σ_i^2 by the sample variance in the ith population, $i = 1, 2$. In regression problems, one would normally try an adaptive or iterative technique. One can take a starting value of σ_1^2, use it to estimate $\{f(x_i)\}$ by means of the algorithm of Section 5, where M at each x_i can be conveniently taken to be $M(|x - x_i|)$. The resulting estimates $\{f^{(1)}(x_i)\}$ can then be used to estimate σ^2 by σ_2^2, so that the process can be iterated with σ_2^2 as a new starting value (once should be enough). This has been done for σ^2 constant and found to be adequate.

When σ is not constant, a number of methods could be used in estimating it. Fortunately, the effect of moderate changes in σ^2 on c^0 appears to be small.

When estimating all $f(x_i)$'s, a good starting place for computing the estimate of $f(x_{i+1})$ is provided by the solution to the problem of estimating $f(x_i)$ where x_i is close to x_{i+1}. Then the number of iterations is small and the entire procedure is not costly unless n is large. Further work is required in order to be more definitive about this problem.

7. Miscellaneous remarks.

Designs. With a design element present, the issue of inaccuracies in ideal regression models has been raised by Box and Draper [4], and has received attention in the work of Karson, Manson and Hader [7] and Kiefer [8]. The types of inaccuracy investigated in [7] and [8] do not, however, bear much resemblance to the kind allowed by the approximately linear models discussed here. A design problem in the setting of (2.1) and (2.2) has been explored by Marcus and Sacks [9].

Asymptotic behavior. In regression models of the type described in Example 1 of Section 2 it is possible to obtain the asymptotic behavior of $J(c^0)$ as the number of observations gets large, provided the sequence of design measures $\{\xi_n\}$ behaves regularly enough. For example, consider the case of approximately νth degree polynomial regression in dimension 1 with $M(x) = |x|^{\nu+1}$, $\mathscr{X}$ an interval in R^1. If $P_n = (1/n)\xi_n$ is such that $\lim_{n\to\infty} n^{(\frac{1}{2})-\varepsilon} \sup_t |P_n(t) - P(t)| = 0$, and if P has a density p which is continuous and positive at 0, then for estimating the coefficient β_0 of the constant term,

$$J(c^0_{\beta_{0,n}}) = O(n^{-(2\nu+2)/(2\nu+3)}) . \tag{7.1}$$

In fact, the exact limiting behavior can be found: it depends on $p(0)$ and solutions to continuous versions of the equations at (3.8). In dimension d, the corresponding result is

$$J(c^0_{\beta_{0,n}}) = O(n^{-(2\nu+2)/(2\nu+2+d)}) . \tag{7.2}$$

Estimation of the other coefficient can be similarly handled. For example, in estimating the coefficient β_1 of x in the context of (7.1), one finds

$$J(c^0_{\beta_{1,n}}) = O(n^{-2\nu/(2\nu+3)}) . \tag{7.3}$$

Thus for $d = 1$, $\nu = 1$, $J(c^0_{\beta_{0,n}}) = O(n^{-\frac{4}{5}})$ and $J(c^0_{\beta_{1,n}}) = O(n^{-\frac{2}{5}})$; for $d = 1$, $\nu = 2$, $J(c^0_{\beta_{0,n}}) = O(n^{-\frac{6}{7}})$, $J(c^0_{\beta_{1,n}}) = O(n^{-\frac{4}{7}})$. This gives the asymptotic counterpart to the numbers obtained in Section 6, 3(b) and 4.

The details of (7.1) to (7.3) are a bit messy and so are not presented here.

In the context of Example 4 of Section 2 it would be useful to know the different types of asymptotic behavior which can occur when, for instance, the number of cells in a two-way layout becomes large. This is as yet unexplored.

Other approximate models. In regression models especially, it would seem natural to require, instead of (2.2), that $f(x) - \sum \beta_j f_j(x) = r(x)$ with r a smooth function. For example, in the context of Example 1 of Section 2 with $d = 1$,

$\mathscr{X} = [0, 1]$ and $\nu = 1$, the requirement that f'' (and therefore r'') be continuous with $|f''(x) \leqq |1$ gives rise to a class of f's which are included in those of (2.1), (2.2) when $M(x) = x^2/2$. This kind of modification leads to related problems which are, unfortunately, quite complicated. One such problem has been studied by Berkovitz and Pollard [1], [2]. The ease and generality of solution for the models of (2.1) and (2.2), and the fact that there is some numerical evidence indicating very little gain in going to the smoother models, suggest that the present setup might prove more useful in practice.

Acknowledgment. The authors are grateful to the referee for pointing out the reference to Whittle [11] which led to a considerably shortened proof of Theorem 1. Thanks are also due to George Knafl whose help with the algorithm of Section 5 and computations in Section 6 was indispensable.

REFERENCES

[1] BERKOVITZ, L. D. and POLLARD, HARRY (1967). A non-classical variational problem arising from an optimal filter problem. *Arch. Rational Mech. Anal.* **26** 281–304.

[2] BERKOVITZ, L. D. and POLLARD, HARRY (1970). A non-classical variational problem arising from an optimal filter problem, II. *Arch. Rational Mech. Anal.* **38** 161–172.

[3] BICKEL, P. J. (1973). On some analogues to linear combinations of order statistics in the linear model. *Ann. Statist.* **1** 597–616.

[4] BOX, G. E. P. and DRAPER, N. R. (1959). A basis for the selection of a response surface design. *J. Amer. Statist. Assoc.* **54** 622–654.

[5] HUBER, PETER J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *Ann. Statist.* **1** 799–821.

[6] JUREČKOVÁ, J. (1971). Nonparametric estimate of regression coefficients. *Ann. Math. Statist.* **42** 1328–1338.

[7] KARSON, M. J., MANSON, A. R. and HADER, R. J. (1969). Minimum bias estimation and experimental design for response surfaces. *Technometrics* **11** 461–475.

[8] KIEFER, J. (1973). Optimal designs for fitting biased multiresponse surfaces. In *Multivariate Analysis III*, 287–297. Academic Press, New York.

[9] MARCUS, M. B. and SACKS, J. (1976). Robust designs for regression problems. To appear in: *Proceedings of the Symposium on Statistical Design Theory and Related Topics.*

[10] MICCHELLI, C. A. (1977). Optimal estimation of linear functionals. IBM preprint.

[11] WHITTLE, PETER (1971). *Optimization under Constraints.* Wiley-Interscience, New York.

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF CALIFORNIA
LOS ANGELES, CALIFORNIA 90024