HATLINK: A Link Between Least Squares Regression and Nonparametric Curve Estimation

by

Richard L. Einspom

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in
Statistics

APPROVED:

Jeffrey B. Birch, Chairman

Klaus H. Hinkelmann

Raymond H. Myers

Eric P. Smith

George R. Terrell

August 25, 1987

Blacksburg, Virginia

HATLINK: A Link Between Least Squares Regression and Nonparametric Curve Estimation

by

Richard L. Einsporn

Jeffrey B. Birch, Chairman

Statistics

(ABSTRACT)

For both least squares and nonparametric kernel regression, prediction at a given regressor location is obtained as a weighted average of the observed responses. For least squares, the weights used in this average are a direct consequence of the form of the parametric model prescribed by the user. If the prescribed model is not exactly correct, then the resulting predictions and subsequent inferences may be misleading. On the other hand, nonparametric curve estimation techniques, such as kernel regression, obtain prediction weights solely on the basis of the distance of the regressor coordinates of an observation to the point of prediction. These methods therefore ignore information that the researcher may have concerning a reasonable approximate model. In overlooking such information, the nonparametric curve fitting methods often fit to anomolous patterns in the data.

This paper presents a method for obtaining an improved set of prediction weights by striking the proper balance between the least squares and kernel weighting schemes. The method is called "HATLINK," since the appropriate balance is achieved through a mixture of the hat matrices corresponding to the least squares and kernel fits. The mixing parameter is determined adaptively through cross-validation (PRESS) or by a version of the Cp statistic. Predictions obtained through the HATLINK procedure are shown through simulation studies to be robust to model misspecification by the researcher. It is also demonstrated that the HATLINK procedure can be used to perform many of the usual tasks of regression analysis, such as estimate the error variance, provide confidence intervals, test for lack of fit of the user's prescribed model, and assist in the variable selection process. In accomplishing all of these tasks, the HATLINK procedure provides a model-robust alternative to the standard model-based approach to regression.

Acknowledgements

I would like to express my appreciation to the many persons in the Department of Statistics at Virginia Tech who have helped me during my five years here. In particular, I would like to thank my advisor, Professor Jeffrey B. Birch, for his excellent editing work and for his friendship. Professor Klaus H. Hinkelmann and the faculty of the Department of Statistics are to be commended for providing a solid graduate program. Professor Raymond H. Myers is especially appreciated for his inspiring teaching in the seven courses I have taken with him. I would also like to express gratitude to my fellow graduate students, particularly Dr. David D. Morris and Dr. Ann Giovannitti - Jensen, my comrades through all five years of graduate work. Special thanks is due Mrs. Sharon L. Myers and Mrs. Mollissa Faye Roop for frequently providing computing advice.

I am especially indebted to my wife Lea for her steady support and understanding, and to the members of the Circle of Life prayer group for supporting me through their prayers. Most of all, I am thankful to the Lord, in Whom all things are possible.

Table of Contents

List of Illustrations

List of Tables

Chapter I

I. INTRODUCTION

In the usual regression framework, a response variable Y is viewed as being related to one or more regressor variables X_1, X_2, ..., X_k according to a model of the form

$$Y = f(X_1, X_2, \ldots, X_k) + \varepsilon ,$$

where ε represents a random error. It is typically assumed that a specific parametric form for the function f is known, with the parameters of f to be estimated by the data. Subsequent predictions, estimates, and hypothesis tests, are all based on the assumption that the correct function has been specified. Therefore, the usual regression inferences may be misleading when the user has incorrectly specified the function f.

In applications of regression, it is often the case that the user does not know the exact form for the true underlying f. Except in situations where physical or biological laws dictate the form of the true model, users are generally only able to prescribe models that are approximations to the true relationships. Techniques such as data transformation and nonlinear regression can be used in some cases to provide models that are better approximations to the true f, but these methods

do not guarantee that the exact form of f will be obtained. Model specification is particularly problematic in the multiple regression setting, where f is often automatically taken to be a linear combination of a set of X's. Many of the usual regression inferences are especially questionable in such circumstances. For a given situation, it would therefore be advantageous to be able to:

(i) determine whether f has been misspecified and measure the degree to which misspecification has occurred, and

(ii) provide a model-robust method for performing many of the usual tasks of standard regression analysis.

The procedure outlined in this paper is designed to accomplish both of these goals. This procedure should therefore be of use whenever the true form of f is in doubt, especially in the multiple regression setting.

Closely related to the proposed procedure is the method of nonparametric kernel regression. In this approach, the true function f is viewed as being unknown, and no parametric form for f is specified by the user. Let $\underline{X}_0 = (X_1, \ldots, X_k)'$ be a location at which one would like to predict Y. If Y is related to the variables $X_1, \ldots, X_k$ according to some relatively smooth (yet unknown) function, then observations near $\underline{X}_0$ should provide information about $f(\underline{X}_0)$, the true value of the underlying function f at the regressor location $\underline{X}_0$. Observations remote from $\underline{X}_0$ may have little relationship to $f(\underline{X}_0)$. This is precisely the logic behind kernel regression, in which prediction of $f(\underline{X}_0)$ is a weighted average of the observations (Y's) where the weight assigned to an observation decreases as the distance from its $\underline{X}$ location to $\underline{X}_0$ increases. Note that the kernel regression method provides a way for predicting Y at any given $\underline{X}$ location, thereby producing an estimated (nonlinear) regression curve. No closed form expression for this estimated curve is obtained, but the graph of the kernel regression curve may suggest the true functional form of f.

This paper establishes a bridge, or link, between nonparametric kernel regression and the standard model-based approach. As in the standard approach, a function f is provided by the user. However, adjustments are made to the usual least squares predictions to the extent that the data

suggest that the true f differs from the form specified. The adjusting occurs through mixing the least squares "hat" matrix with the hat matrix corresponding to a version of nonparametric kernel regression. For this reason, the proposed method will be entitled "HATLINK". The proper balance between least squares and kernel regression may be determined from the data by cross-validation (PRESS) or by a variation on Mallows' C_p statistic. If the data do not indicate a departure from the specified f, then HATLINK will produce the least squares model-based predictions. If the prescribed model appears to differ from the true model, greater emphasis will be placed on local information in making predictions. That is, prediction at a given location X_0 will be more heavily influenced by the observations which are close to X_0 than would be the case under least squares.

It is demonstrated through simulation studies (Chapter IV) that HATLINK does provide improved prediction compared to both the model-based approach (using least squares) and the "model-free" approach (kernel regression). This advantage is shown to hold for varying degrees and types of model misspecification, both in the single and multiple regressor cases.

In addition to being a method for making predictions in the face of possible model misspecification, HATLINK yields model-robust alternatives to many of the usual regression analytics. For example, in contrast to the least squares approach, HATLINK provides an estimate of error variance that is relatively stable whether or not the model specified by the user is correct. Similarly, confidence intervals resulting from HATLINK are better able to maintain coverage of the true underlying function when an incorrect model is specified. The method also gives diagnostic information regarding possible lack of fit of the user's model, and serves as a model-robust means for variable selection.

The following chapter provides relevant background information on the model-based least squares approach to regression, plus details of the kernel regression method. Included in the next chapter is a review of the literature on kernel regression and other nonparametric regression techniques. In Chapter III the HATLINK procedure is described for both the single and multiple regressor cases. Also developed in Chapter III are methods for estimating error variance, setting confidence intervals, and detecting lack of fit through the HATLINK procedure.

In order to demonstrate the effectiveness of the HATLINK method and its related analytics, the results of simulations of a variety of regression situations are given in Chapter IV. The application of HATLINK to specific examples is presented in Chapter V. Chapter VI contains a summary of the findings and a discussion of areas for future investigation. Certain mathematical developments, to be referred to in Chapter III, are included in the Appendix, along with excerpts of the computer program used in this research.

Chapter II

II. LEAST SQUARES AND NONPARAMETRIC REGRESSION METHODS

This chapter contains a brief review of least squares regression, followed by a description of the kernel method and its variations. For completeness, other nonparametric methods which are comparable to kernel regression are discussed as well. The review of background material will focus on the situation where only one regressor variable is used, since that is the primary setting for which nonparametric regression methods have been developed in the literature. The few references which discuss nonparametric regression in the multiple regression setting are mentioned at the end of the chapter.

II.1 Ordinary Least Squares

First consider the model-based least squares approach to regression. This brief review will emphasize the role of the hat matrix and its entries, as this information is crucial to the development of the HATLINK procedure. Let $Y_i = f(X_{1i}, X_{2i}, \ldots X_{ki}) + \varepsilon_i$, $i = 1, \ldots, n$, where the ε_i are iid with mean zero and constant variance σ^2, k is the number of regressor variables, and n is the number of observations. In linear regression it is asumed that the correct parametric form for f can be specified by the user. The vector form for this model is $\underline{Y} = X\underline{\beta} + \underline{\varepsilon}$. Here, $X = [\underline{1}\ \underline{X}_1\ \underline{X}_2 \ldots \underline{X}_k]$, where $\underline{1}' = [1\ 1 \ldots 1]$ and $\underline{X}_j' = [X_{j1}\ X_{j2} \ldots X_{jn}]$, $j = 1, \ldots, k$, and $\underline{\beta}' = [\beta_0\ \beta_1 \ldots \beta_k]$. The estimate of the parameter vector $\underline{\beta}$ obtained by the least squares method is

$$\hat{\underline{\beta}}_{ols} = (X'X)^{-1}X'\underline{Y}\ , \tag{II.1.1}$$

and predictions at the data locations are obtained as

$$\hat{\underline{Y}}_{ols} = X\hat{\underline{\beta}}_{ols} = X(X'X)^{-1}X'\underline{Y} = H_{ols}\underline{Y}\ . \tag{II.1.2}$$

The matrix $H_{ols} = X(X'X)^{-1}X' = (h_{ij}^{(ols)})$ is commonly known as the "hat" matrix, due to its role in obtaining the predictions ("y-hats"). The importance of the matrix was stressed by Hoaglin and Welsch (1978), who presented a number of its mathematical properties. Some of these properties are as follows:

(i) H_{ols} is symmetric idempotent,

(ii) $-1 \le h_{ij}^{(ols)} \le 1$,

(iii) $\sum_{i=1}^{n} h_{ii}^{(ols)} = p$, where $p = k + 1$ is the number of parameters to be estimated,

(iv) $\sum_{j=1}^{n} h_{ij}^{(ols)} = 1$ for each i.

A number of results in linear regression can be expressed in terms of the hat matrix or its elements. Several such results, which will be used later in this proposal, are listed below.

$$\text{(i)} \quad Var(\hat{\underline{Y}}_{ols}) = Var(H_{ols}\underline{Y}) = \sigma^2 H_{ols} \ , \tag{II.1.3}$$

$$\text{(ii)} \quad Var(\hat{Y}_i^{(ols)}) = \sigma^2 (H_{ols})_{ii} = \sigma^2 h_{ii}^{(ols)} \ , \tag{II.1.4}$$

$$\text{(iii)} \quad \underline{e}_{ols} = \underline{Y} - \hat{\underline{Y}}_{ols} = \underline{Y}(I - H_{ols}) \ , \tag{II.1.5}$$

$$\text{(iv)} \quad Var(\underline{e}_{ols}) = \sigma^2 (I - H_{ols}) \ , \tag{II.1.6}$$

$$\text{(v)} \quad Var(e_i^{(ols)}) = \sigma^2 (I - H_{ols})_{ii} = \sigma^2 (1 - h_{ii}^{(ols)}) \ , \tag{II.1.7}$$

$$\text{(vi)} \quad \hat{\sigma}^2_{ols} = \frac{\sum e_i^{2(ols)}}{n-2} = \frac{\sum e_i^{2(ols)}}{tr\,[(I - H_{ols})(I - H_{ols})']} \ , \tag{II.1.8}$$

where I in the above expressions represents the n by n identity matrix. For the present development, it is important to note that equation II.1.2 reduces to

$$\hat{Y}_i^{(ols)} = \sum_{j=1}^{n} h_{ij}^{(ols)} Y_j \ , \tag{II.1.9}$$

for predicting Y at a particular data location X_i. Thus, prediction at X_i is a weighted average of the observations Y_j, with weights given by the i^{th} row of the hat matrix. From expression II.1.9, it can be seen that an observation at X_j with a relatively large (positive or negative) $h_{ij}^{(ols)}$ would have a very heavy influence on the prediction $\hat{Y}_i^{(ols)}$. Such an observation is often referred to as a "high leverage point".

For simple linear regression the elements of the hat matrix can be expressed as

$$h_{ij}^{(ols)} = (\frac{1}{n}) + \frac{(X_i - \overline{X})(X_j - \overline{X})}{\sum_{k=1}^{n} (X_k - \overline{X})^2} \ . \tag{II.1.10}$$

Note that $h_{ij}^{(ols)}$ tends to be relatively large in magnitude if X_j is greatly removed from $\overline{X}$. Thus, in simple linear regression, prediction at X_i is generally most heavily influenced by those points at the most extreme X locations. Conversely, some observations taken relatively close to X_i may be essentially ignored in predicting $f(X_i)$, since the corresponding $h_{ij}^{(ols)}$'s can be near zero.

In order to illustrate the nature of the hat matrix entries when a least squares fit is obtained for the user's model, the following example is presented. Suppose the model $Y = \beta_0 + \beta_1 X + \varepsilon$ is prescribed by the user. Let ten observations be taken, one at each of the locations X = -10, -3, -1.5, -1, -.5, .5, 1, 1.5, 3, and 10. Then the least squares weights $h_{0j}^{(ols)}$ associated with predicting Y at locations X_0 = 0, 3, and 6 are shown in the following chart.

Observation at X =	-10	-3	-1.5	-1	-.5	.5	1	1.5	3	10
Weight at X_0 = 0 :	.10	.10	.10	.10	.10	.10	.10	.10	. 10	.10
Weight at X_0 = 3 :	-.033	.060	.080	.087	.093	.107	.113	.120	.140	.233
Weight at X_0 = 6 :	-.167	.020	.060	.073	.087	.113	.126	.140	.180	.367

Considering the least squares prediction at location X_0 = 0 , the mean of the X's, the chart shows that this prediction is the simple mean of all ten observations. Specifically, the remote observations at X = -10 and X = 10 are given the same weight as the nearby observations taken at X = -.5 and X = .5. For predicting at location X_0 = 3 , the least squares method gives weight .233 to the observation taken at X = 10. This weight is substantially greater than the weight .140 given to the observation taken right at location X = 3. The observation taken at X = 10 is given weight .367 in making the least squares prediction at X_0 = 6 . This is in contrast to the much smaller weight (.180) given to the observation at location X = 3, which is closer to X_0 = 6 . Further, the weight for this prediction given to the remote observation at X = -10, namely -.167, is nearly as great in magnitude as for X = 3. Moreover, the magnitude of the weight at X = -10, and therefore the importance placed on this in the least squares prediction, is larger than the weight given to any of the observations from X = -3 to X = 1.5. Such a weighting scheme is a direct ramification of the simple linear regression model prescribed by the user. In general, the particular weighting scheme used for making predictions by the least squares method will be strongly tied to the form of the user's parametric model.

If the user is certain that the prescribed model is correct throughout the entire range of the observed X's, then the model-based least squares weighting scheme is desirable. In that case the various optimality theorems for least squares regression apply. A problem with the least squares

weighting scheme arises when the user's specified model differs from the true model. If the straight-line model is not correct, then the least squares weighting scheme could produce very poor predictions at many X locations. For prediction of $f(X_0)$ in such cases, it would no longer make sense to give heavy prediction weights to observations Y_j taken at X locations remote from X_0. Rather, it would be more logical to emphasize observations taken closer to X_0.

This local emphasis is accomplished by the method of nonparametric kernel regression. For this approach and for other nonparametric regression methods the model $Y = f(X) + \varepsilon$ is used, where the function is taken to be entirely unknown. If f is well-behaved then observations Y_j taken at locations nearest to X_0 should be related to $f(X_0)$ more strongly than observations taken at locations far from X_0. This suggests an alternative weighting scheme for predicting $f(X_0)$ in which the Y_j's are weighted according to some decreasing function of the distance of their X locations from X_0. This type of weighting scheme is obtained through the use of kernel regression. Details of this procedure will be discussed in the following section.

II.2 Kernel Regression

Watson(1964) and Nadaraya(1964) proposed an estimator of $f(X_0)$ that can be expressed as

$$\hat{Y}_{\text{ker}}(X_i) = \sum_{j=1}^{n} h_{ij}^{(\text{ker})} Y_j \,, \tag{II.2.1}$$

where

$$h_{ij}^{(\text{ker})} = \frac{K\left(\frac{X_i - X_j}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{X_i - X_j}{h}\right)} \,, \quad \text{for some } h > 0 \,. \tag{II.2.2}$$

That is, the weights $h_{ij}^{(ker)}$ on the observations Y_j for predicting at location X_i are obtained through a function K( ) of the distance from X_i to X_j . The nature of the function K(u) and the constant h will be discussed on the following page. Note that there is a type of hat matrix $H_{ker} = (h_{ij}^{(ker)})$ associated with kernel regression, since II.2.1 implies $\hat{\underline{Y}}_{ker} = H_{ker}\underline{Y}$. Further, the denominator in expression II.2.2 is present for the purpose of making the rows of H_{ker} sum to one. That is , $\sum_{j=1}^{n} h_{ij}^{(ker)} = 1$, as is true for the least squares hat matrix. However, H_{ker} will not generally be symmetric (or idempotent).

Prediction at any non-data point X_0 can be obtained by

$$\hat{Y}_{ker}(X_0) = \sum_{j=1}^{n} h_{0j}^{(ker)} Y_j \ , \qquad \text{(II.2.3)}$$

where

$$h_{0j}^{(ker)} = \frac{K\left(\frac{X_0 - X_j}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{X_0 - X_j}{h}\right)} \ , \quad \text{for some } h > 0 \, . \qquad \text{(II.2.4)}$$

Since it is possible to obtain predictions at any location, one can think of kernel regression as producing an estimated regression curve. A plot of this curve can provide valuable information about the true relationship between Y and X, and the graph may suggest a reasonable parametric model for f (Cleveland, 1979, Denby, 1987).

A drawback of kernel regression and the other related nonparametric regression procedures is that no closed form expression for $\hat{f}(X)$ is obtained. For that reason, application of kernel regression is somewhat limited in comparison to the usual model-based methods. Nevertheless, the kernel method is being used for its graphical and predictive benefits. For example, dozens of articles in fields such as medicine, forestry, and environmental sciences have referenced Cleveland's 1979 paper, which presented a nonparametric regression procedure similar to the kernel method.

The function K(u) which appears in expressions II.2.2 and II.2.4 is called the kernel function and is often taken to be a probability density function. According to a number of authors, the exact form of K does not appear to be crucial. Butler(1975) recommends that K and its first derivative

be smooth and continuous, and that the first derivative of K(u) be zero at u = 0. In recent literature, the kernel functions used are generally nonnegative and symmetric about zero. Some are strictly decreasing in $|u|$, while others decrease to zero at some finite value of $|u|$ and remain zero beyond that point. Typical of kernel functions used in the literature is one considered by Butler:

$$K(u) = \frac{1}{1 + cu^2}, \quad c > 0 . \tag{II.2.5}$$

The term h in expressions II.2.2 and II.2.4 is a positive constant called the bandwidth, and it plays an important role in kernel regression. This value controls the extent to which local observations are emphasized in making predictions. A large value of h, for instance, will lead to a prediction at location X_0 that is approximately the average of all the observations. That is, $\hat{Y}_{\text{ker}}(X_0) \cong \overline{Y}$. Conversely, a small h will result in a prediction at X_0 that is based only on the very closest observations, ignoring the rest of the data.

In order to select an appropriate value for the bandwidth in a particular regression problem, the PRESS procedure of Allen(1974) is commonly used. The "leave-one-out" approach is usually referred to as cross-validation in the present context. (Stone, 1974, and Geisser, 1975, introduced the notion of PRESS in a very general framework, and Stone's title of cross-validation seems to have gained preeminence.) Wong(1983) and Rice(1984b) show consistency results for kernel regression estimates with bandwidths obtained by cross-validation. This method selects the bandwidth h so that the quantity

$$PRESS = \sum_{i=1}^{n} (Y_i - \hat{Y}_{i,-i}^{(\text{ker})})^2 \tag{II.2.6}$$

is minimized, where $\hat{Y}_{i,-i}^{(\text{ker})}$ denotes the predicted value of Y at location X_i based on kernel regression, where the prediction is made without the i^{th} data point.

Determination of $\hat{Y}_{i,-i}^{(\text{ker})}$ for a given bandwidth is relatively straightforward. "Minus i" prediction at location X_i is given by $\hat{Y}_{i,-i}^{(\text{ker})} = \sum_{j \neq i} h_{ij,-i}^{(\text{ker})} Y_j$, where $h_{ij,-i}^{(\text{ker})} = h_{ij}^{(\text{ker})} / (1 - h_{ii}^{(\text{ker})})$. This is true since, for a given bandwidth h, the numerator, $K((X_i - X_j)/h)$, of expression II.2.2 for $h_{ij}^{(\text{ker})}$ does

not change when the i^{th} point is deleted. Multiplying the denominator of II.2.2 by $1 - h_{ii}^{(ker)}$ gives the proper scaling so that the new set of weights $h_{ij,-i}^{(ker)}$ sum over j to one.

Cross-validation is used in this setting in an attempt to avoid overfitting to the specific data set. Clearly, minimization of the usual sum of squared errors, $SSE = \Sigma(Y_i - \hat{Y}_{i}^{(ker)})^2$, is not appropriate as a criterion for bandwidth selection, since this would always result in extreme overfitting. In practice, using cross-validation to select h does lead to estimated curves that are less heavily influenced by the individual observations. However, some amount of overfitting to the given data set may still occur. For that reason, it makes sense to use the information contained in the model postulated by the user to temper the kernel regression results. That is part of the motivation behind the proposed HATLINK method.

One difficulty with using PRESS (eq. II.2.6) to select the bandwidth is that a point with an extreme Y coordinate will dominate this sum. Minimization of PRESS in this case will essentially reduce to minimization of the PRESS residual for that observation. In this case bandwidth selection will be dominated by that one point, leading to a value of h that is not generally suitable for the rest of the data locations. Further, PRESS may appear very large because of just one or two observations. Since the value of PRESS is a useful indicator of ability to make predictions, an inflated PRESS may be misleading. It is therefore necessary to either alter the kernel regression in some way to make it robust to outliers, or to provide diagnostics that indicate the presence of outlying data points. Several proposed outlier diagnostics are presented in the Appendix, and Cleveland's robust version of kernel regression is briefly outlined in Section II.3.C.

II.3 Variations of Kernel Regression

II.3.A Local bandwidth kernel regression

The usual kernel regression estimate defined by equations II.2.1 and II.2.2 can be considered a global bandwidth procedure. That is, for a given data set, the same bandwidth applies throughout the X-space. In many cases, however, it may be very beneficial to allow the bandwidth to vary at different X locations. Cleveland(1979) and Georgiev and Greblicki(1986) use variable bandwidths that are adapted to the local density of the X's. Prediction at a point where there are a large number of observations is made with a small bandwidth, whereas a bigger bandwidth would be used at an isolated X-location. Density adjusted local bandwidths are obtained in the following way. Let X_0 be a point where prediction is to be made and h_0 be the bandwidth to be used at that location. Then h_0 is based on the distance to the r^{th} nearest neighbor to X_0 among the X_i. Cleveland uses cross-validation to select r rather than h, and employs a kernel function K which gives zero weight to all but the r-1 nearest observations to X_0.

Georgiev and Greblicki present a simulation study that shows for one particular true model that locally adjusted bandwidths provide somewhat better prediction than the corresponding global bandwidth procedure. The study evaluated both approaches on the basis of mean integrated absolute error (from the true underlying model), $MIAE = \mathrm{E} \int |\hat{Y}(X) - f(X)|\, dx$, considering a range of values for h and r. For this particular underlying model (a type of sine wave), the results indicate that the local bandwidth estimate with the best choice of r obtains lower MIAE for small to moderate sample sizes than the global bandwidth estimate with the best choice of h. This simulation did not incorporate cross-validation for the selection of h or r, but the results do suggest that MIAE is more sensitive to the choice of h than of r.

Another local bandwidth procedure is recommended by Müller and Stadtmüller (1985, 1987). They adapt the bandwidth according to the local curvature of f by estimating the first

through j^{th} derivatives of f at each X location. In regions of high estimated curvature, relatively small bandwidths would be used. Regions with little apparent curvature would call for large bandwidths. Müller and Stadtmüller present asymptotic and simulation results that demonstrate the effectiveness of the derivative approach to locally adjusting the bandwidth. More work is needed to determine whether locally adjusted bandwidths are generally superior to global bandwidths, and whether the superiority is great enough to merit the extra computation and complication.

II.3.B Local linear regression

Another variation on the basic kernel regression method was introduced by Cleveland(1979). Recall that in the usual approach, prediction at location X_0 is simply a weighted average of the observations (Y's). Rather than just averaging the observations (with weights based on distances from X_0), Cleveland uses weighted least squares regression to predict $f(X_0)$, where the weights are the $h_{0j}^{(ker)}$ from a preliminary nonparametric regression. The resulting prediction of $f(X_0)$ is obtained as a new weighted average of the Y's which now emphasizes the localized regression trend in the vicinity of X_0 . Specifically, the method works as follows. First, $\hat{\beta}_0$ and $\hat{\beta}_1$ are found to minimize the quantity $\sum_{j=1}^{n} h_{0j}^{(ker)}(Y_j - \beta_0 - \beta_1 X_j)^2$. Then $\hat{Y}_0 = \underline{X}_0'\hat{\underline{\beta}} = \underline{X}_0'(X'WX)^{-1}X'W\underline{Y} = \sum_{j=1}^{n} h_{0j}^{**} Y_j$, where $\underline{X}_0' = [1 \quad X_0]$ and W is a diagonal matrix with entries $h_{01}^{(ker)}$, ... , $h_{0n}^{(ker)}$. Cleveland also considers the use of local polynomial regression in this context, but recommends local linear regression due to computational considerations. Little evidence is presented, however, to show that it is worth the extra computation and complication to use local linear regression rather than ordinary kernel regression. An article by Müller (1987) demonstrates the close relationship between local linear regression and the kernel method.

II.3.C Robust kernel regression

Cleveland(1979) also provides a way to make kernel regression robust to outliers among the Y's. He uses an iterative scheme in which the weights h_{0j} are adjusted downward for points whose residuals $e_j = \hat{Y}_j - Y_j$ from the previous iteration are large in magnitude. This reweighting is accomplished through the use of a robust weight function such as the bisquare. (See, for example, Montgomery and Peck, 1982.)

II.4 Other Nonparametric Regression Methods

A number of other mehods have been suggested for the regression problem where the true model is completely unknown. Since these methods compete in some ways with the kernel method, they will be briefly described in this section. These procedures are not directly associated with the proposed HATLINK method, but are discussed here for the sake of completeness.

II.4.A The Priestley-Chao and Gasser-Müller estimates

An approach similar to the usual kernel regression method was proposed by Priestley and Chao(1972). For this estimate,

$$\begin{aligned} h_{0j} &= \left(\frac{X_{j+1} - X_j}{h}\right) K\left(\frac{X_0 - X_j}{h}\right) \quad , \; j = 1, \ldots, n-1 \\ h_{0n} &= \quad 0 \; , \end{aligned} \tag{II.4.1}$$

where $X_1 < X_2 < \ldots < X_n$ is assumed. The Priestley-Chao estimate requires an ordering of the X's and behaves rather poorly near the data boundaries. In several examples, Benedetti(1975) found the Priestley-Chao estimate to have inferior performance to the usual kernel estimate.

An improved version of the Priestley-Chao estimate was suggested by Gasser and Müller(1979). (See also Cheng and Lin, 1981.) Their estimate can be defined by

$$\begin{aligned} h_{0j} &= \frac{1}{h} \int_{X_i}^{X_{i+1}} K\left(\frac{X_0 - y}{h}\right) dy \ , \quad i = 1, \ldots, n-1 \\ h_{0n} &= 0 \ , \end{aligned} \tag{II.4.2}$$

with $X_1 < X_2 < \ldots < X_n$. It is clear that this estimate requires much more complicated computations, particularly if extended to the multiple regression case.

II.4.B Spline regression

The method of nonparametric spline regression has received much attention in the literature. The procedure evidently was developed as a means for obtaining a smooth curve through a set of data points, and is still referred to by some authors as "spline smoothing". However, this method can be used to provide good predictions in the regression setting where the true model is unknown, and therefore is an alternative to the kernel method. Silverman(1985) presents a review (followed by 42 discussants) of spline regression, and shows a relationship between the spline and kernel methods. Silverman defines the spline regression estimate to be the function $\hat{g}$ which minimizes

$$S(g) = \sum_{i=1}^{n} (Y_i - g(X_i))^2 \ + \ \delta \int (g''(x))^2 dx \tag{II.4.3}$$

over the class of all twice-differentiable functions. The resulting estimate is designed to give a good fit to the data while avoiding too much rapid local variation. The first term on the right-hand side of expression II.4.3 is the residual sum of squares for the estimating function g, while the integral term is considered a "roughness penalty", since the integral will be large for a function g which

fluctuates rapidly. The constant $\delta > 0$ is called the "smoothing parameter", and its value dictates the balance between the dual goals of obtaining a smooth curve and acheiving a good fit to the data. For example, if δ is near zero, then S(g) would be minimized by a possibly very erratic $\hat{g}$ which overfits the data. On the other hand, a large value of δ would lead to the minimization of S(g) by a very smooth (nearly linear) function $\hat{g}$.

In general, the function $\hat{g}$ obtained by minimizing S(g) can be shown (Reinsch, 1967) to have the following properties:

(i) $\hat{g}$ is a cubic polynomial in each interval (X_i, X_{i+1}). (Assume the X's are ordered.)

(ii) At the data points X_i, the curve $\hat{g}$ and its first two derivatives are continuous, but may be discontinuous in the third derivative.

(iii) $\hat{g}$ is linear outside of the range of the data.

Curves satisfying (i) and (ii) are called "cubic splines" with "knots" at the points X_i. Computational schemes have been developed for obtaining $\hat{g}$, some allowing for the smoothing parameter δ to be selected by cross-validation. (Craven and Wahba, 1979; Silverman, 1984.)

Silverman(1985) indicates the connection between spline regression and variable kernel regression. For predicting by spline regression at a location X_0 that is not too close to the data perimeter,

$$h_{0j} \doteq \frac{1}{n}\frac{1}{e(X_0)}\frac{1}{h(X_0)}K\left(\frac{X_0 - X_j}{h(X_0)}\right), \tag{II.4.4}$$

holds for spline regression when n is large. In this expression, $e(X_0)$ represents the local density of the X's, and would have to be estimated in an application. The bandwidth is given by

$$h(X_0) = \delta^{1/4} n^{-1/4} e(X_0)^{-1/4}, \tag{II.4.5}$$

and the "effective" kernel function is

$$K(u) = \frac{1}{2} e^{-|u|/\sqrt{2}} \sin\left(\frac{|u|}{\sqrt{2}} + \frac{\pi}{4}\right). \tag{II.4.6}$$

The function K(u) can take on small negative values, so that h_{0j} may be slightly negative in some cases. Negative h_{0j} values are common in ordinary least squares and therefore may also be present in the proposed HATLINK procedure, which combines least squares and kernel regression.

Note that the bandwidth $h(X_0)$ for the spline estimate varies across the sample and is related to the local density $e(X_0)$ at any given location. This corresponds to the variable bandwidth kernel estimate mentioned previously. Silverman (1985) claims that a low negative power dependency of the bandwidth on the local density is desirable due to some theoretical considerations.

Müller and Stadtmüller(1985) and Marriot(1985) list a number of reasons for their preference of kernel regression to spline smoothing. In particular, these authors feel that kernel regression is much more suitable for generalization to the multiple regression case. Because of this issue and the fact that spline regression adjusts the hat matrix in such an indirect fashion, it is clearly more appropriate to incorporate the kernel method into the HATLINK procedure.

II.4.C Nearest neighbor regression

The nearest neighbor regression method has been developed by Cover and Hart(1967), Cover(1968), Stone(1977), and Li(1984). In this approach, prediction at location X_0 is also obtained as a weighted average of the observations (Y's). Like kernel regression, the nearest neighbor method also uses a local weighting scheme, but the weights are obtained as a function of the distances of the X_i's from X_0 only through the *order* of those distances. This is in contrast to kernel regression, where the weights are based on the actual distances from the prediction location. For example, Stone(1977) suggests a triangular weighting scheme defined by

$$h_{0j} = \begin{cases} \dfrac{2(k-j+1)}{k(k+1)}, & j = 1,\ldots,k \\ 0, & j = k+1,\ldots,n, \end{cases} \qquad \text{(II.4.7)}$$

where the indices are ordered such that $|X_0 - X_1| \leq |X_0 - X_2| \leq \ldots \leq |X_0 - X_n|$. For this weight function, only the k observations closest to X_0 receive nonzero weight in the prediction of $f(X_0)$. For example, if k = 4 , then the weights .4, .3, .2, and .1 , are applied to the 4 observations nearest to X_0 , in order of increasing distance from X_0 . The nearest neighbor approach to regression seems to have been developed as a way to smooth time series data, since the method makes predictions as a moving average of the observations. When the X's are not evenly spaced, it is not clear that this is the proper approach.

II.5 Application of Nonparametric Regression Methods

II.5.A The void

Nearly all the examples used in the literature to illustrate the use of nonparametric methods are for situations where there is a *single regressor* and *numerous observations* . Cleveland (1979), for example, presents three single regressor data sets of sizes 30, 50, and 158. Silverman (1985) shows spline regression fits to two data sets, one with 272 observations and the other with about 130 observations. The examples presented deal with situations where any of the typical parametric model forms, such as low-order polynomials, would clearly be inadequate. Moreover, nothing appears in the literature in the way of comparing the prediction performances of parametric (model-based) and nonparametric regression methods. A definite void exists in the literature for the regression situations where a reasonable, but not exactly correct parametric model can be specified.

II.5.B Error variance

Cleveland (1979) proposed an estimate of error variance based on his local linear regression procedure. This estimate may be adapted to apply for kernel regression as well. It is defined by

$$\hat{\sigma}^2_{\text{ker}} = \frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i^{\text{ker}})^2}{tr[(I - H_{\text{ker}})'(I - H_{\text{ker}})]} \quad . \tag{II.5.1}$$

This form is suggested by the least squares estimate (eq. II.1.8), where the denominator was chosen to make the estimator unbiased for σ^2 . In this case,

$$E\left[\sum_{i=1}^{n} (Y_i - \hat{Y}_i^{(\text{ker})})^2\right] = E\left[\underline{Y}'(I - H_{\text{ker}})'(I - H_{\text{ker}})\underline{Y}\right]$$
$$= \sigma^2 \, tr[(I - H_{\text{ker}})'(I - H_{\text{ker}})] + \underline{\mu}_0' \, (I - H_{\text{ker}})'(I - H_{\text{ker}}) \, \underline{\mu}_0 \quad .$$

Here $\underline{\mu}_0 = (\mu_{01}, \ldots, \mu_{0n})$, where $\mu_{0i} = f(X_i)$ is the true value of f at location X_i . Thus,

$$E[\hat{\sigma}^2_{\text{ker}}] = \sigma^2 + \frac{\underline{\mu}_0'(I - H_{\text{ker}})'(I - H_{\text{ker}})\underline{\mu}_0}{tr[(I - H_{\text{ker}})'(I - H_{\text{ker}})]} \quad . \tag{II.5.2}$$

Note that $H_{\text{ker}} \underline{\mu}_0$ represents the predictions which would be obtained through kernel regression using the true values μ_{0i} as the observations. That is, the kernel method is fitting the observations $Y_i = f(X_i) + \varepsilon_i$ with ε_i equal to zero. For nearly any reasonable underlying function f with relatively few bends compared to the number of observations, kernel regression will predict $\underline{\mu}_0$ nearly perfectly. So $H_{\text{ker}} \underline{\mu}_0 \cong \underline{\mu}_0$, and the second term in eq. III.5.2 is approximately zero. Thus, $\hat{\sigma}^2_{\text{ker}}$ is approximately unbiased for σ^2 for any reasonable underlying function f .

Note here that eq. III.5.2 indicates that the least squares estimate of σ^2 , based on H_{ols} , will be biased upward when the user's model is biased. By the user's model being biased it is meant that the user has specified a form for the function f in the model $Y = f(X) + \varepsilon$, that leads to a

distribution of ε which does not have mean zero. In this case, $H_{ols}\underline{\mu}_0$ will not equal $\underline{\mu}_0$, so the second term in III.5.2 will be positive. This upward bias in the least squares estimate is demonstrated in Chapter IV, where various regression settings are examined through simulations.

The estimate of error variance for kernel regression given by by equation III.5.2 was found in preliminary simulation studies to generally overestimate the true value of σ^2 . The following variation on III.5.2 has proven through simulation studies to be more successful in estimating σ^2 :

$$s^2_{\text{ker}} = \frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i^{(\text{ker})})^2}{n - tr[H_{\text{ker}}]} \quad . \tag{II.5.3}$$

Silverman (1985) and other authors use the quantity $n - tr[H_{\text{ker}}]$ as the denominator for their nonparametric regression variance estimates. The argument used earlier for $\hat{\sigma}^2_{\text{ker}}$ can be applied to demonstrate that s^2_{ker} is also approximately unbiased for σ^2 . This follows since $tr[(I - H_{\text{ker}})'(I - H_{\text{ker}})] \cong n - tr[H_{\text{ker}}]$.

II.5.C Degrees of freedom

By analogy to least squares, we may interpret the denominator of either of the variance estimates defined above as representing the degrees of freedom (df) for error for the kernel regression fit. Note that these denominators, $tr[(I - H_{\text{ker}})'(I - H_{\text{ker}})]$ and $n - tr[H_{\text{ker}}]$ are not typically integers. Each of these measures of error degrees of freedom corresponds to a representation for the model degrees of freedom. That is, either $n - tr[(I - H_{\text{ker}})'(I - H_{\text{ker}})]$ or $tr[H_{\text{ker}}]$ may be considered to be the model degrees of freedom associated with the kernel regression. Cleveland (1979) refers to this quantity as the "equivalent" degrees of freedom for the nonparametric regression fit. Tibshirani and Hastie (1987) relate this quantity to the number of parameters that would have been used to fit a comparable parmetric model. If $tr[H_{\text{ker}}] = 4.3$, for example, one would interpret the kernel fit as being equivalent to fitting some model having 4.3 parameters. In the present work,

the forms $tr[H_{ker}]$ and $(n - tr[H_{ker}])$ will be used to represent the equivalent model and error df, respectively.

II.5.D. Confidence intervals

Several authors present methods for obtaining confidence limits for the true value of f at any given location. In the context of spline regression, Silverman (1985) forms an approximate 95 percent confidence interval for $f(X_i)$ as

$$\hat{Y}_i \pm 2\hat{\sigma}\sqrt{h_{ii}} \quad . \tag{II.5.4}$$

Similar intervals are used by Hastie and Tibshirani (1987). In each of these articles, the intervals are developed heuristically, by analogy to the least squares intervals. The behavior of such intervals has not been carefully examined in the literature. However, the simulation studies in Chapter IV will consider the performance of confidence intervals based on kernel regression. Also, similar intervals will be constructed, and their behavior examined, for the regression method developed in Chapter III.

II.5.E. Multiple regression

Very little has appeared in the literature regarding the use of nonparametric techniques in the multiple regression framework. This lack of development is partly due to the fact that some of the procedures, such as spline regression and nearest neighbor regression, do not easily extend to the multiple regressor case. The basic kernel approach can be extended to multiple regression in a straightforward manner (section III.4.E), but requires very large data sets in order to accurately predict Y over the multidimensional regressor space. Cleveland and Devlin (1986) extend the local

linear method to multiple regression by fitting locally and in a moving fashion that is analogous to how a moving average is computed for a time series.

Butler (1975) and Hastie and Tibshirani (1986, 1987) consider the problem of variable selection in multiple regression. Butler uses the sum of squared residuals in conjunction with kernel regression as a criterion for selecting a best subset of regressor variables from a list of candidate regressors. The work of the Hastie and Tibshirani concentrates on the restricted class of "additive models." This class includes models of the form $Y = f_1(X_1) + f_2(X_2) + \ldots + f_r(X_r) + \varepsilon$. These authors employ a forward selection procedure to determine whether candidate regressor variables should be added to the regression model. These methods are nonparametric in the sense that each variable is considered for addition to the model according to any reasonable functional form. That is, variable selection is not restricted to adding variables linearly into the model.

Staniswalis, McCrady, Carter, Campbell, and Carchman (1987) use a multiple regression version of the kernel method to analyze a response surface. They recommend the kernel approach for exploring the shape of the surface and for identifying a region where a parametric model, such as a second order model, may be adequate. For situations where the user is not prepared to adopt a particular parametric model, the authors have developed a method for estimating the location of the optimal response and provide a confidence region for this location. The method is applied to a data set where 3 to 4 observations were taken at each of the 104 possible combinations of 8 levels of one regressor and 13 levels of a second regressor. The authors refer to this situation as involving a "small number of levels of the independent variables," and imply that this is a small to moderate sized data set for applying the kernel method with two regressors. They note that they are pleased that the kernel approach works reasonably well for this size of data set. It is hoped that the HATLINK method would provide further improvement in response surface applications for cases where some approximately correct model can be specified by the user.

Chapter III

III. THE HATLINK PROCEDURE

III.1 Description

In many regression situations the user does not know the exact form of the true underlying function f , but may be able to specify a parametric model that reasonably approximates the true function. When the specified model differs from the true model, predictions and other inferences obtained through least squares can be quite poor. It will be observed in Chapter IV that the least squares inferences are often seriously affected by fairly slight discrepancies between the user's model and the true model. On the other hand, kernel regression and the other nonparametric regression methods do not make use of the fact that the user is able to put forth a model form that is at least a rough approximation of the true underlying function. With small to moderate data sets, and without the basis of a reasonable model, the kernel method too often fits to anomolous patterns in the data. Therefore, in cases where the user is able to specify an approximately correct model, it may be beneficial to compromise between the extreme positions held by least squares, where prediction is completely dominated by the choice of the model, and by kernel regression, where the

prediction process ignores any knowledge the user may have concerning the true model. The HATLINK procedure described in this chapter is one way of achieving such a compromise.

As the name of the proposed method suggests, the link between least squares and kernel regression occurs with the hat matrices associated with these two procedures. Let $H_{ols} = X(X'X)^{-1}X' = (h_{ij}^{(ols)})$ be the usual hat matrix that would be used in least squares regression for the postulated model. Let $H_{\text{ker}} = (h_{ij}^{(\text{ker})})$ be the hat matrix associated with kernel regression as discussed in Section II.2. The least squares predictions at the data locations would be given by $\hat{\underline{Y}}_{ols} = H_{ols}\underline{Y}$, whereas the kernel method would give predictions as $\hat{\underline{Y}}_{\text{ker}} = H_{\text{ker}}\underline{Y}$. Prediction at a data point X_i would be obtained as $\hat{Y}_i^{(ols)} = \sum_{j=1}^{n} h_{ij}^{(ols)} Y_j$ and $\hat{Y}_i^{(\text{ker})} = \sum_{j=1}^{n} h_{ij}^{(\text{ker})} Y_j$ for least squares and kernel regression, respectively. Recall that in least squares, observations Y_j taken at locations remote from X_i tend to have weights $h_{ij}^{(ols)}$ that are large in magnitude. If the prescribed model is not an accurate approximation of the true f in the range of the data, then such a weighting scheme would not be desirable. In these situations, prediction at X_i may be improved by de-emphasizing the more remote observations and giving more weight to observations closer to X_i. Therefore, it may be advantageous to adjust the least squares weights $h_{ij}^{(ols)}$ toward the local distance-based weights $h_{ij}^{(\text{ker})}$ of kernel regression. This adjustment is accomplished as follows by the HATLINK method.

Let a new hat matrix $H(\lambda)$ be defined for the HATLINK method by

$$H(\lambda) = \lambda H_{\text{ker}} + (1 - \lambda) H_{ols}, \quad 0 \le \lambda \le 1. \tag{III.1.1}$$

That is, $H(\lambda)$ is a convex combination of the hat matrices for the model-based and local distance (kernel) methods. Prediction at the vector of data locations is obtained through the HATLINK method by

$$\hat{\underline{Y}}(\lambda) = H(\lambda)\underline{Y}, \tag{III.1.2}$$

and at single data points by

$$\hat{Y}(X_i, \lambda) = \sum_{j=1}^{n} h_{ij} Y_j = \sum_{j=1}^{n} [\lambda h_{ij}^{(\ker)} + (1 - \lambda) h_{ij}^{(ols)}] Y_j , \qquad \text{(III.1.3)}$$

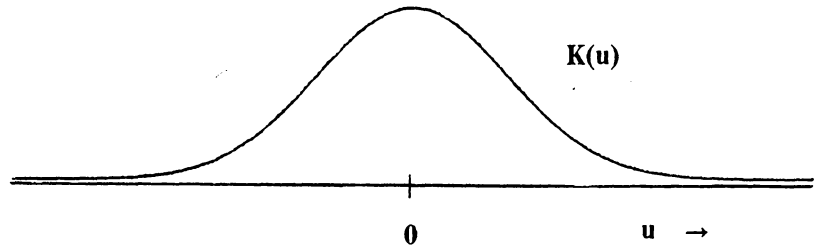
where h_{ij} will be used to denote the ij entry of $H(\lambda)$. For ease of notation, the term $\hat{Y}(X_i, \lambda)$ will be written as $\hat{Y}_i(\lambda)$ in subsequent material.

The value of the mixing parameter λ determines the degree to which the least squares hat matrix entries are adjusted, with $\lambda = 0$ corresponding to least squares, and $\lambda = 1$ resulting in pure kernel regression. In order to make predictions in a given regression problem, an appropriate value of the mixing parameter must be chosen. One method for making this choice is to select λ adaptively by cross-validation. That is, λ may be chosen so that the quantity $\text{PRESS}(\lambda) = \sum_{i=1}^{n} [Y_i - \hat{Y}_{i,-i}(\lambda)]^2$ is a minimum. In this expression, $\hat{Y}_{i,-i}(\lambda)$ represents the prediction at location X_i obtained by the HATLINK method with mixing parameter λ without using the observation taken at X_i . This and other methods for selecting λ are considered in detail later in this chapter.

Application of the HATLINK method also requires the selection of a specific version of kernel regression. In particular, a kernel function K(u) must be chosen, along with a method for selecting the bandwidth. As previously noted, the choice of kernel function K(u) does not appear to be crucial. The function used throughout this paper is

$$K(u) = e^{-u^2} . \qquad \text{(III.1.4)}$$

A sketch of this particular kernel function appears below.



For the kernel function K(u) used here, prediction of $f(X_i)$ is given by $\hat{Y}_i^{(\ker)} = \sum_{j=1}^{n} h_{ij}^{(\ker)} Y_j$, where

$$h_{ij}^{(\ker)} = \frac{e^{-(X_i - X_j)^2/h^2}}{\sum_{j=1}^{n} e^{-(X_i - X_j)^2/h^2}} \quad . \tag{III.1.5}$$

The kernel function K(u) defined in III.1.4 has a form similar to the equivalent kernel function of spline regression (eq.II.4.6), and therefore should give good predictions. Other kernel functions could be considered in future research.

A method for selecting the bandwidth h is also needed in order to define the kernel portion of the HATLINK procedure. Because of its ease of use and conceptual simplicity, the global bandwidth method (Section II.2) has been employed in the present research. In order to select the bandwidth for a given set of data, a variation of the PRESS statistic is used. The bandwidth h is chosen so that the quantity

$$PRESS^*(h) = \frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_{i,-i}(h))^2}{n - tr[H_{\ker}(h)]} \tag{III.1.6}$$

is a minimum. In expression III.1.6, the term $\hat{Y}_{i,-i}(h)$ represents the prediction at location X_i obtained by the kernel method with bandwidth h without using the observation taken at X_i. The denominator in the above expression provides additional protection against overfitting, since $tr[H_{\ker}(h)]$ becomes large for very small bandwidths. In preliminary simulation studies, use of this version of PRESS to select h was found to lead to improved kernel prediction compared to the original PRESS, defined in II.2.6.

Additionally, for the single regressor situation, $PRESS^*(h)$ was further revised so that observations taken at the largest and smallest X locations were not included in the summation. Again, this adjustment was found to lead to improved prediction by the kernel method. The motivation for this revision is that $PRESS^*(h)$ might otherwise lead to an overly small bandwidth in order to accomodate prediction at the endpoints. (See the last paragraph of Section II.2.)

III.2 Examples

Three basic examples are presented in order to illustrate the application of the kernel and HATLINK methods. In all three cases it is assumed that the user has prescribed a simple linear regression model, $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$. Twenty observations were taken, two each at ten evenly spaced X locations from X = 1 to X = 10. These observations were based on true models of the form $Y_i = c\,(X_i - 5.5)^3 + X_i + \varepsilon_i$, with the ε_i generated as independent N(0,16) random variates. The same set of errors ε_i were used in each example. The constant c was varied in the following manner:

Example #1: c = 0,

Example #2: c = .12,

Example #3: c = .25 .

Scatterplots of the raw data for these examples are displayed in Figures III.2.1-3.

Prediction performance of the least squares fit to the user's model, and the fits obtained through the kernel and HATLINK methods, are evaluated here by two criteria. These are defined as

$$SSEP = \sum_{i=3}^{18} [\hat{Y}_i - f(X_i)]^2 \quad , \tag{III.2.1}$$

and

$$SAEP = \sum_{i=3}^{18} | \hat{Y}_i - f(X_i) | \quad . \tag{III.2.2}$$

Here, "SSEP" denotes the sum of squared errors of prediction at the data locations, while "SAEP" stands for the sum of absolute errors of prediction. Note that predictions obtained at the endpoints of the X region are excluded from these sums. This step was taken in order to provide a fairer comparison for the kernel method, which (in the form of kernel used here) is biased toward $\overline{Y}$ at the endpoints. (A similar evaluation of kernel appears in Georgiev and Greblicki, 1986).

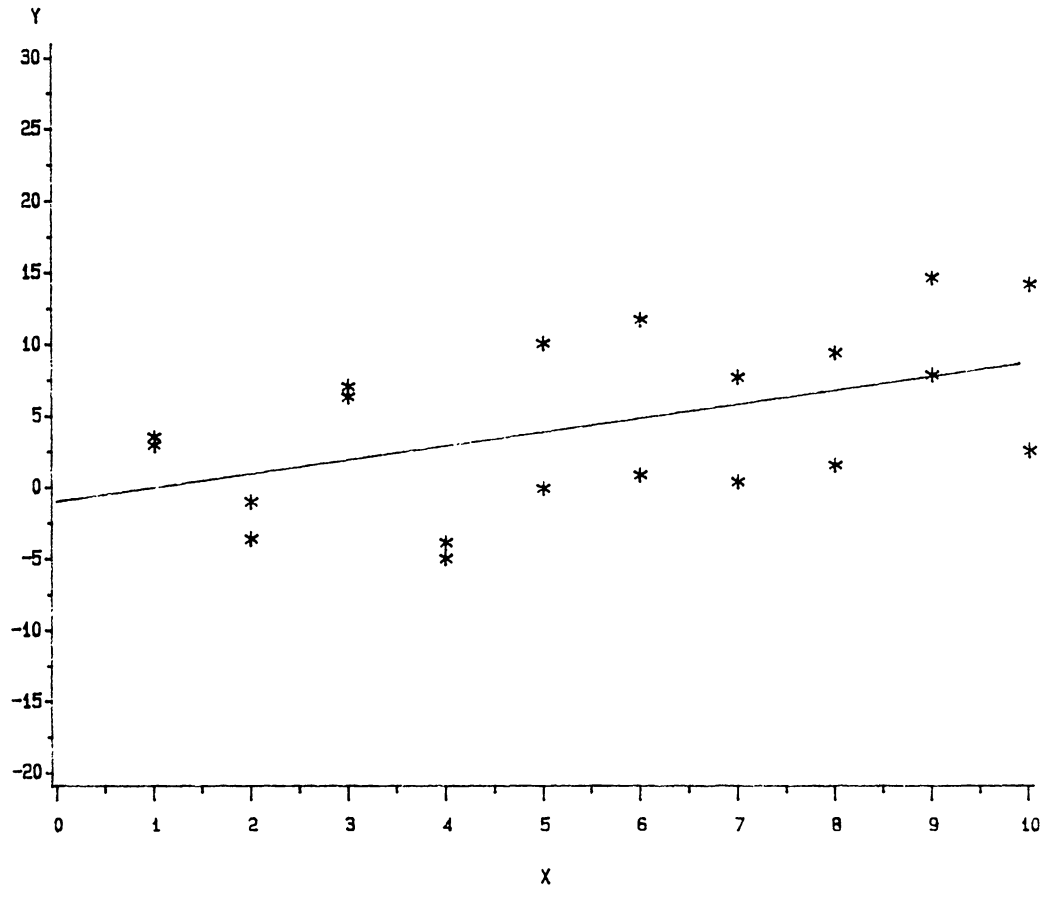


Figure III.2.1. Data and Linear Fit by L.S. for Example #1.

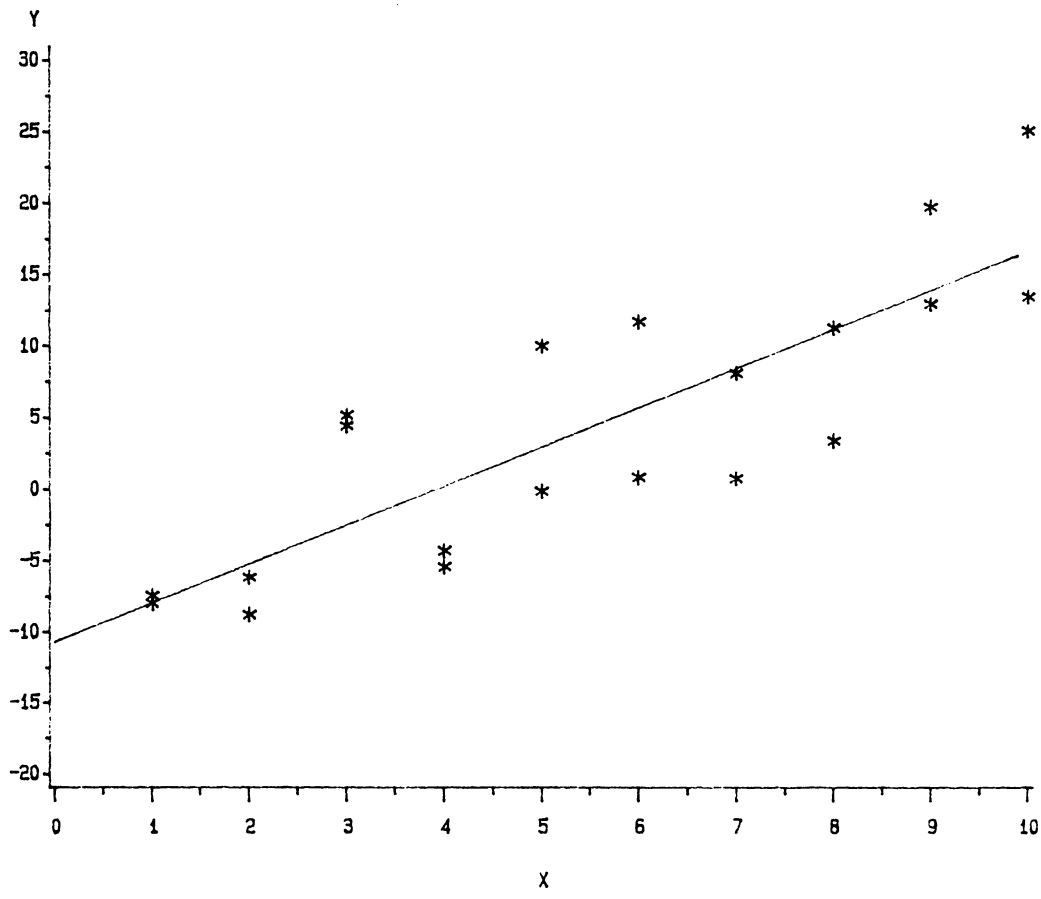


Figure III.2.2. Data and Linear Fit by L.S. for Example #2.

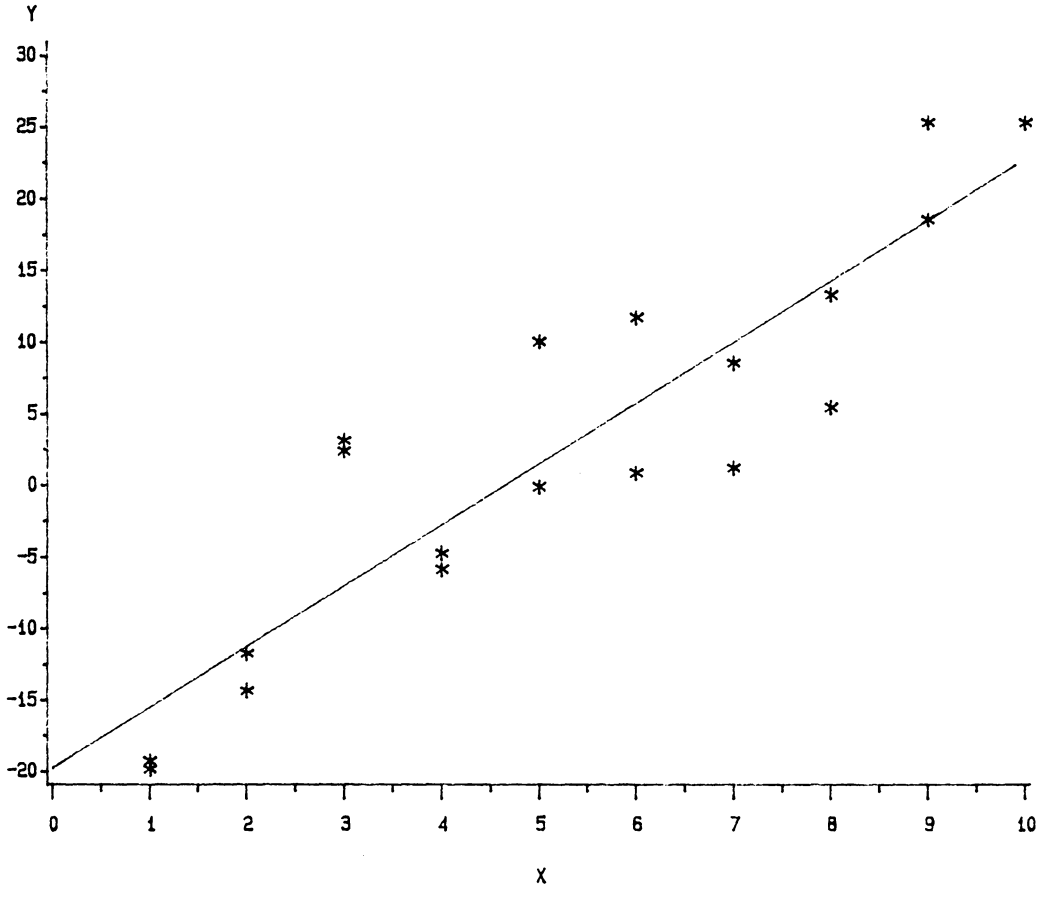


Figure III.2.3. Data and Linear Fit by L.S. for Example #3.

III.2.A. Example #1 User's model is correct (c = 0)

Since c = 0 , the true model reduces to $Y_i = X_i + \varepsilon_i$, so the user's simple linear regression model is correct. Using the $PRESS^*(h)$ criterion to select the kernel bandwidth, the kernel method procuced a slightly curvilinear fit to the data, as shown in Figure III.2.4. The degrees of freedom for this fit are $tr[H_{ker}] = 2.15$, compared to 2 df for the user's postulated model. To give a better indication of how the kernel fit is obtained, the row of the kernel hat matrix corresponding to location X = 6 is presented in Table III.2.1, along with the X = 6 row of H_{ols} . The HATLINK method here reproduced the simple linear regression fit, as $PRESS(\lambda)$ was minimized at a value of $\lambda = 0$. As is nearly always the case when the user has correctly specified the model, the kernel method resulted in higher values for SSEP and SAEP than for the least squares fit. A comparison of SSEP and SAEP for the three methods is given in Table III.2.2.

III.2.B. Example #2 User's model slightly off (c = .12)

This is an example in which the true model departs from the user's model by a very modest amount. The usual F test for detecting lack of fit for the user's model has power of only .18 at level $\alpha = .05$. For this particular data set, the F statistic was not significant at the .10 level. Furthermore, since the scatterplot (Figure III.2.2.) reveals no nonlinear pattern, the user should have no reason to question the correctness of the prescribed model.

In this case the kernel method produces an overfit (Figure III.2.5.), with $tr[H_{ker}] = 6.81$ equivalent model degrees of freedom. The overfit is largely due to the way in which the $PRESS^*(h)$ criterion handles the two close pairs of observations occurring at data locations X = 3 and X = 4. A very small bandwidth for the kernel would fit close to the mean of the replicates at each of these two locations, making the squared PRESS residuals for these points as low as possible.

Table III.2.1. Portion of the X = 6 Row of the Kernel and Least Squares Hat Matrices for Example #1. Entries are for one of the two observations taken at each of the 10 regressor locations.

X Location	1	2	3	4	5	6	7	8	9	10
Least Squares	.036	.039	.042	.045	.048	.051	.055	.058	.061	.064
Kernel	.009	.021	.039	.061	.080	.088	.080	.061	.039	.021

Table III.2.2. Prediction Performance of the Least Squares, Kernel and HATLINK Regression Methods for Example #1. In this example PRESS(λ) was minimized at $\lambda = 0$, so the HATLINK and least squares predictions are identical. The model degrees of freedom, $tr[H]$, is also listed for each regression method.

Method	Mean SAEP	Mean SSEP	Model df
Least Squares	18.1	20.6	2.00
HATLINK	18.1	20.6	2.00
Kernel	21.4	33.8	2.15

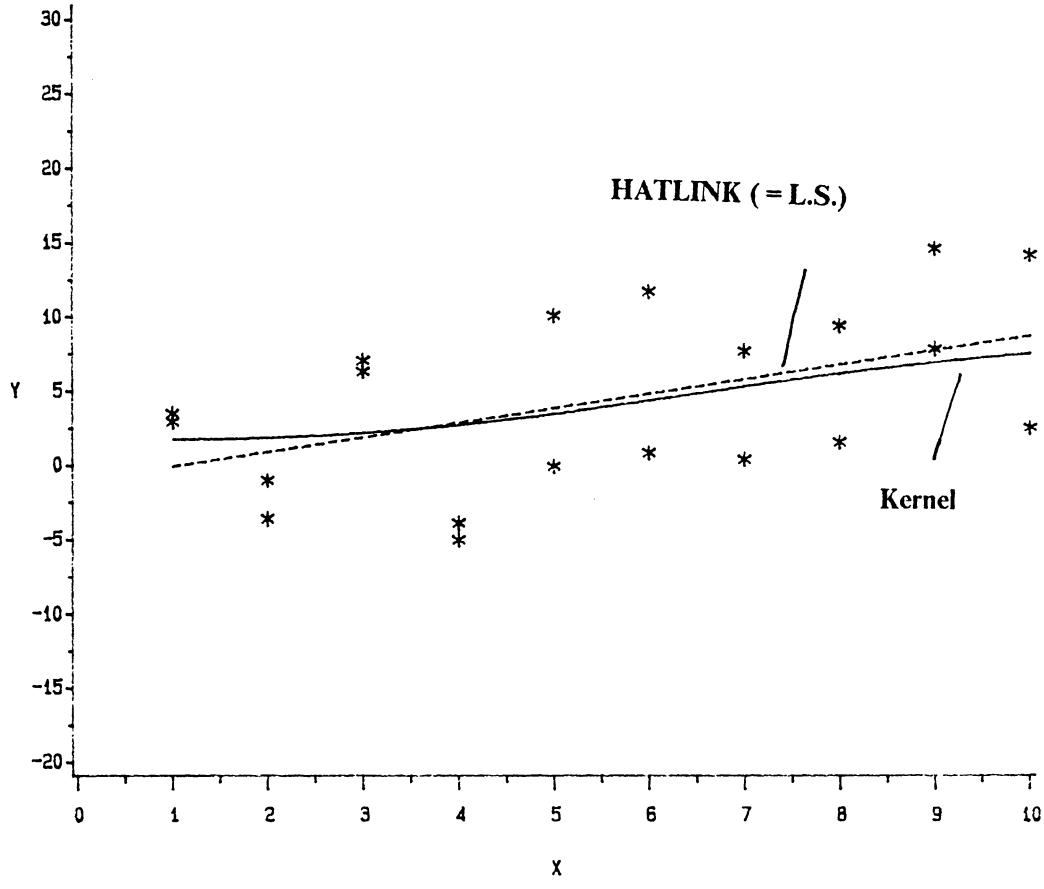


Figure III.2.4. Kernel and HATLINK fits for Example #1.

Table III.2.3. Portion of the X = 6 Row of the Kernel and Least Squares Hat Matrices for Example #2. Entries are for one of the two observations taken at each of the 10 regressor locations.

X Location	1	2	3	4	5	6	7	8	9	10
Least Squares	.036	.039	.042	.045	.048	.051	.055	.058	.061	.064
HATLINK	.032	.035	.038	.041	.053	.082	.058	.051	.054	.057
Kernel	.000	.000	.000	.002	.085	.327	.085	.002	.000	.000

Table III.2.4. Prediction Performance of the Least Squares, Kernel, and HATLINK Regression Methods for Example #2. In this example PRESS(λ) was minimized at $\lambda = 0.11$. The model degrees of freedom, $tr[H]$, is also listed for each regression method.

Method	Mean SAEP	Mean SSEP	Model df
Least Squares	27.8	71.0	2.00
HATLINK	25.7	65.2	2.53
Kernel	27.7	81.9	6.81

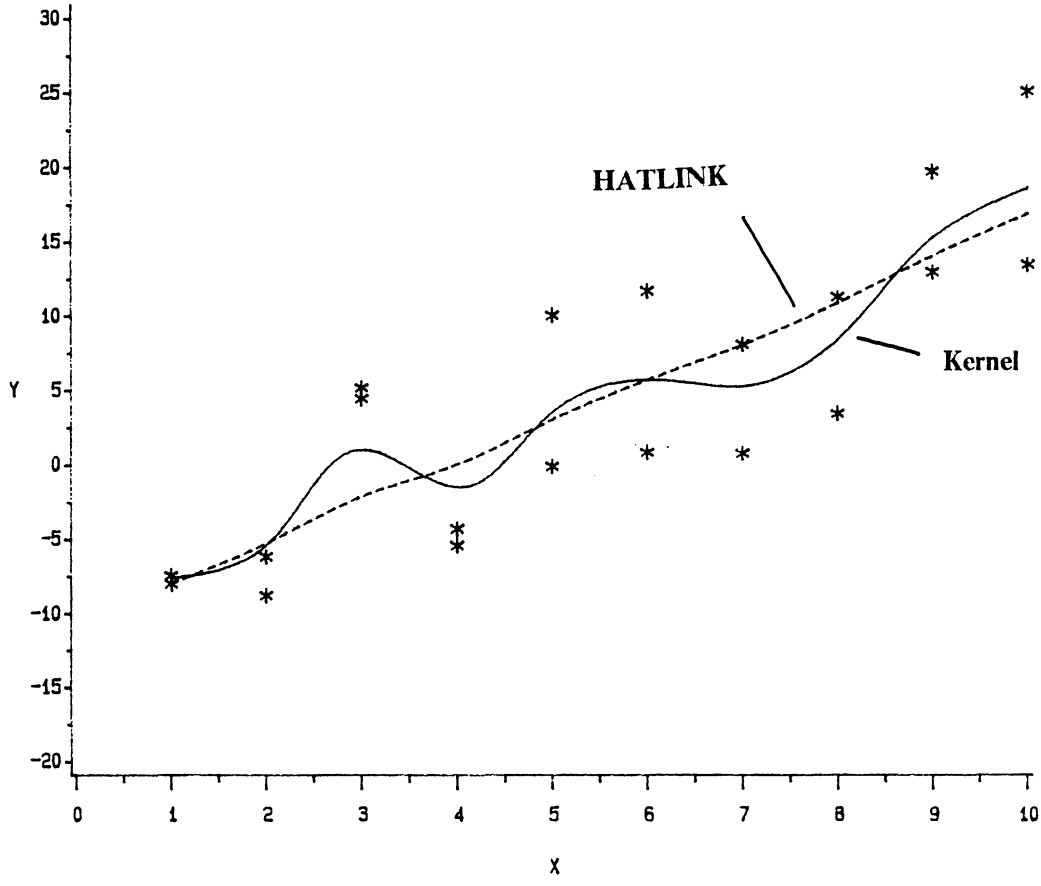


Figure III.2.5. Kernel and HATLINK fits for Example #2.

Table III.2.5. Prediction Performance of the Least Squares, Kernel and HATLINK Regression Methods for Example #3. In this example PRESS(λ) was minimized at $\lambda = 0.79$. The model degrees of freedom, $tr[H]$, is also listed for each regression method.

Method	Mean SAEP	Mean SSEP	Model df
Least Squares	54.6	243.8	2.00
HATLINK	28.0	92.5	5.89
Kernel	29.6	92.8	6.90

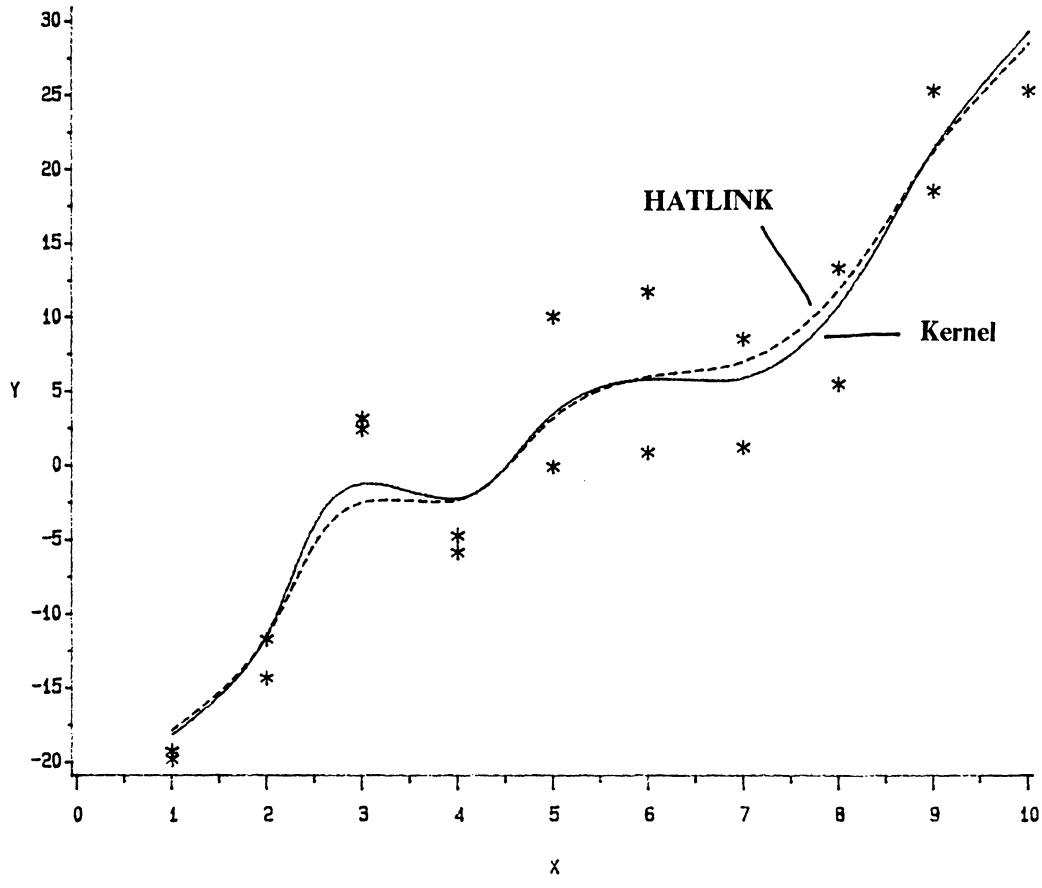


Figure III.2.6. Kernel and HATLINK fits for Example #3.

The fit obtained by the user's model provides better prediction than the kernel method here, according to the SSEP evaluation, while the two methods are about the same with regard to SAEP. (See Table III.2.4.) The HATLINK procedure gives improved prediction in this example compared to both of the other methods. A value of $\lambda = .11$ was selected by the PRESS criterion, and this value corresponds to a fit having $tr[H(\lambda = .11)] = 2.53$ equivalent degrees of freedom. Notice the contrast among the three methods in the regression weights they use for predicting Y at X = 6. (Table III.2.3.) Figure III.2.5 shows the kernel, least squares, and HATLINK fits to this data set.

III.2.C. Example #3 User's model off by a moderate amount (c = .25)

In this example there is a fairly wide gap between the user's model and the true model. An $\alpha = .05$ level lack of fit test at this value of c has power of .60 . The test for lack of fit for this data set was significant at $\alpha = .05$ and the scatterplot (Figure III.2.3.) does reveal a nonlinear pattern, so the user would likely discard the simple linear regression model in favor of a more complex model. Thus, comparisons of the least squares simple linear regression fit to the kernel and HATLINK fits on the basis of SSEP or SAEP are not entirely fair. Such comparisons do, however, give an indication of what might happen in multiple regression, when a similar degree of misspecification in one of the regressors would probably go undetected. Further, the example is useful in that it provides insight into the performance of HATLINK when the specified model is off to a greater extent.

As expected, the kernel fit resulted in lower SSEP and SAEP than for the simple linear regression fit. The equivalent model degrees of freedom are $tr[H_{ker}] = 6.90$, with the kernel overfitting as before because of the observations at X = 3 and X = 4. HATLINK, with $\lambda = .79$, had $tr[H(\lambda = .79)] = 5.89$ model df, and gave slightly better predictions than kernel. The fits are displayed in Figure III.2.6 and the evaluations of these fits are shown in Table III.2.5.

III.2.D. Summary of examples

These three examples show what are typically the prediction performances of the usual regression approach, the kernel method, and the HATLINK method, under conditions where (i) the user's model is correct, (ii) the user's model is slightly but not noticeably off, and (iii) the user's model is off by a moderate amount. It is generally the case that when the user's model is correct, the corresponding least squares predictions are slightly better than those obtained through HATLINK and much better than for the kernel fit. When the specified model is far off, then kernel regression and HATLINK are about the same in their ability to predict, while predictions obtained by least squares for the user's model are terrible. In the much more realistic situation where the user's model is off only slightly, HATLINK tends to do a better job of prediction than both of the other methods. However, under any of these types of conditions, the comparative performance of the three methods varies rather widely from data set to data set. The simulation results presented in Chapter IV will establish that the above examples are truly representative of the typical prediction performances of the three procedures.

III.3 Variations on the HATLINK Procedure

Many adjustments to the basic HATLINK method, as oulined in Section III.1, are possible. These adjustments generally fall into two categories: (i) variations on the kernel portion of HATLINK, and (ii) variations on the method for selecting λ .

III.3.A Variations on the kernel portion of HATLINK

Results obtained by the HATLINK procedure will depend to some extent on the particular version of kernel regression employed. For instance, the use of a kernel function other than $K(u) = e^{-u^2}$ will lead to a different H_{ker} matrix. Similarly, using a method other than cross-validation to choose the bandwidth would also result in a different H_{ker} matrix. However, neither the choice of kernel function nor the method used for bandwidth selection should have a huge impact on HATLINK. All that is needed is a reasonable local distance-based hat matrix to serve as a counterpart to the model-based least squares hat matrix. Any of the kernel functions K(u) used in the recent literature and any sensible method for bandwidth selection will provide a suitable matrix H_{ker} for most practical situations. However, some improvement of the kernel portion of HATLINK may be gained through the use of one or more of the variations on kernel regression mentioned in Section II.3, such as local bandwidth kernel regression or robust kernel regression. Each of these variations introduces an additional degree of complexity to the computation and interpretation of H_{ker} . Further work is needed to determine whether the potential gains achieved through these methods are worth the additional complication.

Two less complicated bandwidth selection methods have been given some consideration in the present research. Both are global bandwidth procedures and both have been investigated through empirical studies in conjunction with HATLINK. The first method is to allow the user to select a desirable value for the model df for the kernel. For use in HATLINK it is often advantageous to let the kernel fit have several more degrees of freedom than the parametric model that has been specified. Once a value has been chosen for the kernel df, the corresponding bandwidth is determined through a search routine. It is useful to incorporate this nonstochastic kernel approach into certain tests for lack of fit that are developed Section III.4.D. An empirical evaluation of the nonstochastic kernel method is presented in Appendix A. It will be seen that certain versions of the HATLINK procedure tend to provide improved predictions when used in conjunction with the nonstochastic kernel method.

A second variation on kernel regression will be referred to as the bounded kernel method. In this approach, the user selects upper and/or lower bounds for the kernel degrees of freedom. If the PRESS bandwidth selection procedure gives a kernel fit with df beyond these boundaries, then a kernel with df equal to the nearest boundary value is used. Setting an upper bound can be used to limit the complexity of the kernel fit, and thereby avoiding serious overfitting. Setting the lower bound equal to the df for the user's model will insure that the kernel fit is at least as complex as the parametric model. Results of an empirical investigation of the bounded kernel approach are also presented in Appendix A.

III.3.B Methods for selecting the mixing parameter

It is of vital importance to the success of the HATLINK procedure to be able to select an appropriate value of the mixing parameter λ for any given set of data. Therefore, much of the present research effort has been devoted to evaluating and comparing the performance of a number of different methods for choosing λ . The λ selection methods considered fall into two categories -- those based on versions of the PRESS statistic, and those which are related to Mallows' *Cp* statistic (Mallows, 1973).

PRESS methods

The PRESS statistic,

$$PRESS(\lambda) = \sum_{i=1}^{n} (Y_i - \hat{Y}_{i,-i}(\lambda))^2 \quad ; \qquad \text{(III.3.1)}$$

is a concave quadratic function in λ with minimum often occurring in the interval from 0 to 1 . In such cases, HATLINK is formed using the minimizing value of λ . For some data sets the

minimum of PRESS may occur at $\lambda \leq 0$ or $\lambda \geq 1$, so that HATLINK reverts to least squares ($\lambda = 0$) or kernel regression ($\lambda = 1$) , respectively.

Choosing λ by cross-validation (PRESS) relies on the ability of the PRESS criterion to strike the proper balance between overfitting and underfitting. However, it has been shown empirically that minimizing the PRESS statistic as defined above often produces a value of λ that is too large. In cases where the kernel has overfit to spurious trends in the data, the usual PRESS criterion may yield a fit for HATLINK that is the same as or nearly the same as the kernel fit. A more conservative approach is to adjust the PRESS statistic, as was done in bandwidth selection (eq.III.1.6), by dividing PRESS by $n - tr[H(\lambda)]$. That is, select the value of λ which minimizes

$$PRESS^{*}(\lambda) = \frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_{i,-i}(\lambda))^2}{n - tr[H(\lambda)]} . \qquad \text{(III.3.2)}$$

Since $PRESS^{*}(\lambda)$ is no longer a quadratic function in λ , a search routine is employed to locate its minimum. Simulation studies (Chapter IV) indicate that the $PRESS^{*}(\lambda)$ criterion performs quite well in providing suitable λ values for various sorts of regression settings.

Cp methods

Since we are dealing with situations where the user's specified model may be inaccurate, it is natural to be concerned with the bias in this model. A good model would be one for which the bias is reasonably small across the range of interest in the regressor variable X, while simultaneously maintaining a small prediction variance. That is essentially the motivation behind the use of Mallows' *Cp* statistic (Mallows, 1973) for model selection. In the following section the formulation of the *Cp* statistic will be reviewed, and it will be shown how these principles apply to the selection of λ .

The *Cp* statistic is based on the quantity,

$$MSE[\hat{Y}(\underline{X}_0)] = E[\hat{Y}(\underline{X}_0) - E\,Y(\underline{X}_0)]^2 \quad , \tag{III.3.3}$$

the mean squared error of prediction at any regressor location $\underline{X}_0$. (See, for example, Myers, 1986.) This quantity can be re-expressed as the sum of a prediction variance component and a squared bias component:

$$MSE[\hat{Y}(\underline{X}_0)] = Var(\hat{Y}(\underline{X}_0)) + [E\,\hat{Y}(\underline{X}_0) - E\,Y(\underline{X}_0)]^2 \quad . \tag{III.3.4}$$

The usual *Cp* statistic is defined as an estimate of

$$\frac{1}{\sigma^2}\sum_{i=1}^{n} MSE\,\hat{Y}(X_i) = \frac{1}{\sigma^2}\sum_{i=1}^{n}[\,Var\,\hat{Y}(X_i) + (\,Bias\hat{Y}(X_i)\,)^2\,] \quad , \tag{III.3.5}$$

the standardized MSE of a fitted value summed over all data locations. For predictions obtained through the hat matrix H = (h_{ij}) , the variance component in expression III.3.5 can be written as

$$\begin{aligned}
\frac{1}{\sigma^2}\sum_{i=1}^{n} Var(\hat{Y}(X_i)) &= \frac{1}{\sigma^2}\sum_{i=1}^{n} Var(\sum_{j=1}^{n} h_{ij}\,Y_j) \\
&= \frac{1}{\sigma^2}\sum_{i=1}^{n}(\sigma^2\sum_{j=1}^{n} h_{ij}^2) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} h_{ij}^2 \\
&= tr[HH'] \quad .
\end{aligned} \tag{III.3.6}$$

It can be shown (Appendix B) that the bias component in III.3.5 can be expressed in the following form.

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}[\,Bias\,\hat{Y}(X_i)\,]^2 = \frac{1}{\sigma^2}\{\,E(SSE) - \sigma^2\,tr[(I-H)'(I-H)]\,\} \quad , \tag{III.3.7}$$

where $SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ is the usual sum of squared errors for the regression. Combining the expressions for variance and squared bias, equation III.3.5 becomes

$$\frac{1}{\sigma^2}\sum_{i=1}^{n} MSE\,\hat{Y}(X_i) = tr[HH'] + \frac{1}{\sigma^2}\{\,E(SSE) - \sigma^2\,tr[(I-H)'(I-H)]\,\} \tag{III.3.8}$$

If the regression fit is obtained through least squares, then equation III.3.8 reduces to

$$\frac{1}{\sigma^2} \sum_{i=1}^{n} MSE\, \hat{Y}(X_i) = p + \frac{(n-p)}{\sigma^2} [\, E(s^2) - \sigma^2] \quad , \tag{III.3.9}$$

where p is the number of parameters in the user's model, and s^2 is the usual MSE for the user's model.

The usual *Cp* statistic is defined by

$$Cp = p + \frac{(n-p)}{\hat{\sigma}^2} (s^2 - \hat{\sigma}^2) \quad , \tag{III.3.10}$$

where $\hat{\sigma}^2$ is some appropriate estimate of σ^2 , and s^2 (for the model in question) is used to estimate $E(s^2)$. An increase in the number of parameters p in the user's model will increase the variance portion (p) of the *Cp* statistic, while possibly reducing the bias component. If the resulting *Cp* for the extra-parameters model is lower than for the user's original model, one may have reason to believe that the reduction in bias has more than compensated for the variance increase due to adding these parameters.

This same rationale can be applied to the use of the *Cp* criterion to select the mixing parameter λ . One may think of the kernel method as generally fitting a higher order model than the one the user prescribes, so that choosing a λ greater than zero essentially amounts to adding to the complexity of the user's model. This leads to an increase in prediction variance while potentially reducing the bias. Moreover, if λ is allowed to increase from $\lambda = 0$ (model) to $\lambda = 1$ (kernel) , prediction variance will steadily increase while the bias may decrease. Recall that under certain conditions the kernel is consistent, providing asymptotically unbiased fits. Thus, the statistic *Cp* , viewed as a function of λ (denoted $Cp(\lambda)$), measures the trade-off between variance increase and bias decrease as λ ranges from zero to one. The value of λ which minimizes $Cp(\lambda)$ over $0 \le \lambda \le 1$ provides the HATLINK fit with the lowest esimated sum of mean squared errors at the data locations.

Several versions of $Cp(\lambda)$ are possible, depending on the ways in which $E[SSE(\lambda)]$ and σ^2 are estimated in equation III.3.8. Since $\lambda = 1$ (kernel) is analogous to the full model in the usual variable selection process, it makes sense to estimate σ^2 using the kernel fit. That is, estimate $\hat{\sigma}^2$ by

$$s^2_{\text{ker}} = \frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i^{(\text{ker})})^2}{\text{error df}} = \frac{SSE(\text{ker})}{\text{error df}} . \qquad \text{(III.3.11)}$$

The appropriate value for error degrees of freedom may reasonably be represented by any of the following quantities.

(i) error df $= n - tr[H_{\text{ker}}]$

(ii) error df $= tr[(I - H_{\text{ker}})'(I - H_{\text{ker}})]$

(iii) error df $= n - tr[H_{\text{ker}}H_{\text{ker}}']$

When the quantity (i) is used for error df, then III.3.11 is the kernel variance estimate presented in equation II.5.3. If the error degrees of freedom are represented by (ii), then III.3.11 is the kernel estimate of variance (equation II.5.1) formed by analogy to Cleveland's estimate. (The issues of variance estimation, model degrees of freedom, and error degrees of freedom will be discussed in more detail in Section III.4.)

Similarly, the quantity $E[SSE(\lambda)]$ in equation II.3.8 may be estimated by $SSE(\lambda) = \Sigma(Y_i - \hat{Y}_i(\lambda))^2 = s^2(\lambda)(\text{error df}(\lambda))$, where $s^2(\lambda)$ is some estimate of error variance based on the HATLINK fit. (See Section III.4.A.) Using the quantity $(n - tr[H(\lambda)H'(\lambda)])$ to represent error df for the HATLINK fit, we may form a *Cp* statistic to estimate III.3.8 as follows.

$$Cp(\lambda) = tr[H(\lambda)H'(\lambda)] + \frac{(s^2(\lambda) - s^2_{\text{ker}})tr[(I - H(\lambda))'(I - H(\lambda))]}{s^2_{\text{ker}}} . \qquad \text{(III.3.12)}$$

The value of λ which minimizes *Cp* would then be selected for use in HATLINK. Unfortunately, empirical studies have indicated that the prediction performance is relatively poor for the HATLINK method based on λ chosen through eq.III.3.12. However, several variations on

eq.III.3.12 have resulted in improved prediction. Four of these *Cp* versions which provide better selection of the mixing parameter are listed here.

$$Cp1(\lambda) = tr[H(\lambda)] + \frac{(s^2(\lambda) - s^2_{\ker})(n - tr[H(\lambda)])}{s^2_{\ker}}, \quad \text{(III.3.13)}$$

where $s^2_{\ker} = \frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i^{(\ker)})^2}{n - tr[H_{\ker}]}$ and $s^2(\lambda) = \frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i(\lambda))^2}{n - tr[H(\lambda)]}$. (See Section III.4.A.)

$$Cp2(\lambda) = n - tr[(I - H(\lambda))'(I - H(\lambda))] + \frac{(s^2(\lambda) - s^2_{\ker})(tr[(I - H(\lambda))'(I - H(\lambda))])}{s^2_{\ker}}, \quad \text{(III.3.14)}$$

where $s^2_{\ker}$ and $s^2(\lambda)$ have denominators $tr[(I - H_{\ker})'(I - H_{\ker})]$ and $tr[(I - H(\lambda))'(I - H(\lambda))]$, respectively.

$$Cp3(\lambda) = tr[H(\lambda)] + \frac{(s^2(\lambda) - s^2_{ols})(n - tr[H(\lambda)])}{s^2_{ols}}, \quad \text{(III.3.15)}$$

where s^2_{ols} is the least squares estimate of variance for the user's model, and $s^2(\lambda)$ is as in $Cp1(\lambda)$.

$$Cp4(\lambda) = 2(tr[H(\lambda)] - p) + \frac{(SSE(\lambda) - SSE_{ols})(n - tr[H_{\ker}])}{SSE_{\ker}}, \quad \text{(III.3.16)}$$

where p is the number of parameters in the parametric model.

$Cp1(\lambda)$ and $Cp2(\lambda)$ are based on the alternative definitions of effective model and error degrees of freedom. Both employ $s^2_{\ker}$ as the estimate of σ^2, whereas the third version, $Cp3(\lambda)$, uses s^2_{ols} to estimate σ^2. When the user's model is nearly correct, then one may expect s^2_{ols} to estimate σ^2 better than $s^2_{\ker}$. Consequently, $Cp3(\lambda)$ should, in such cases, be a better means for selecting the

mixing parameter. On the other hand, when the user's model is off, then s^2_{ker} more accurately estimates the variance than does s^2_{ols} , so that *Cp*3 should not work as well as Cp1 or Cp2. In fact, these insights have been borne out in the various simulations that have been conducted. (Chapter IV.)

The fourth Cp version, *Cp*4(λ) , is formulated as an estimated *change* in the sum of mean squared errors as λ is allowed to increase from zero (least squares). *Cp*4 is designed to perform well in situations where a fairly small bandwidth is required for the kernel fit. (See Appendix C for a discussion of the limiting behavior of the Cp4 statistic.)

These four versions of *Cp* and the two versions of PRESS have been investigated extensively through simulation studies. The results of these simulations are presented in Chapter IV. Numerous other methods for selecting the value of λ for a given data set are possible. Several of the possible altenatives are mentioned below.

Other methods for selecting the mixing parameter

One variation on the HATLINK method is to select the mixing parameter λ and the bandwidth h jointly by cross-validation. That is, PRESS(λ, *h*) or PRESS*(λ, *h*) would be minimized over the set of pairs (λ , h), where $0 \le \lambda \le 1$ and $h > 0$. This approach is based on a slightly differently philosophy than HATLINK, which attempts to find the optimal balance between the best model-based fit and the best model-free (kernel) fit. That is, HATLINK's choice of λ is conditional on the optimal bandwidth. The joint (λ , h) approach seeks to optimize prediction over all mixtures of least squares and all possible bandwidths simultaneously. Since the optimization is done over a larger data set, this second method will, in fact, lead to a lower value of PRESS than the original HATLINK method. However, this does not guarantee that the joint (λ , h) approach will provide for better prediction of the true f, since the joint method is more susceptible to overfitting. Empirical investigations of a few cases have shown no general improvement in SSEP for the joint method. Further, the joint minimization approach is based on one

optimization criterion, and is therefore less flexible than the HATLINK method, which allows different criteria to be used at the two stages of optimization.

It is also possible to take a subjective approach to choosing the mixing parameter. For example, rather than automatically choosing the value of λ which minimizes PRESS (or one of the other statistics listed in the preceding section), the HATLINK user may want to consider how much reduction in PRESS is actually achieved as λ is gradually increased from zero. A plot of PRESS versus λ could be used here in a manner similar to the way in which a plot of PRESS versus the the ridge parameter is used in ridge regression (Myers, 1986). That is, the mixing parameter for HATLINK could be selected to be as small a value as possible so that no great reduction in PRESS would be gained by using a λ higher than this value.

Another approach is to use a predetermined value for λ in the HATLINK regression. When this is done in conjunction with a predetermined bandwidth kernel fit (Section III.3.A) , certain pseudo-tests discussed in the next section have nicer properties. (See Section III.4.D.)

III.4 Development of the HATLINK regression method

Several developments which are intended to enhance the usefulness of the HATLINK method are presented in this section. With these developments, most of the usual tasks of standard regression analysis can be accomplished through HATLINK. (Of course, when $\lambda > 0$ there will be no estimation, interpretation, and testing of model parameters.) First, an estimate of error variance will be given. This estimate will then be used in the formation of regression confidence intervals. Other common regression items, such as R^2 , and the F-test for model utility, have analogous counterparts in HATLINK. All of these quantities are, like HATLINK predictions, robust to possible model misspecification. Additionally, HATLINK provides a number of ways for diagnosing lack of fit of the user's model.

A further development of HATLINK is its extension to the multiple regression setting. All of the items mentioned above for the univariate case apply for multiple regression by means of HATLINK. Other tasks, such as variable selection, which are specific to the multiple regressor situation, may be addressed through the HATLINK procedure.

III.4.A. Estimation of error variance

In order for HATLINK to be useful in ways similar to the usual regression method, it is imperative that an accurate estimate of error variance be established. Further, the estimator should remain accurate when the functional form of the user's model is incorrect. These requirements seem to be met by the estimator,

$$s^2(\lambda) = \frac{\sum_{i=1}^{n} [Y_i - \hat{Y}_i(\lambda)]^2}{n - tr[H(\lambda)]} . \qquad \text{(III.4.1)}$$

The simulation studies in Chapter IV demonstrate that this estimator performs well for $\lambda's$ chosen by several of the selection procedures outlined in the previous section.

The denominator of $s^2(\lambda)$ is interpreted as the equivalent error degrees of freedom, while $tr[H(\lambda)]$ represents the equivalent model degrees of freedom for the HATLINK fit. (See Section II.5.C.) Using the quantity $tr[(I - H(\lambda))'(I - H(\lambda))]$ for error df in the denominator was found to generally produce overestimates of σ^2 , as was noted earlier for the kernel version of s^2 with this denominator.

III.4.B. Confidence intervals

Confidence intervals for the value of the true underlying function f at any given location X_i are formed as follows for HATLINK.

$$\hat{Y}(X_i) \pm t_{n-tr[H(\lambda)]}\, s(\lambda)\sqrt{h_{ii}} \quad . \tag{III.4.2}$$

Note that the variance in the estimate $\hat{Y}_i(\lambda)$ is

$$Var\hat{Y}_i(\lambda) = Var(\sum_{j=1}^{n} h_{ij}\, Y_j) = \sigma^2 \sum_{j=1}^{n} h_{ij}^2 \cong \sigma^2\, h_{ii}$$

So the standard error of $\hat{Y}_i(\lambda)$ is estimated by $s(\lambda)\sqrt{h_{ii}}$ in (III.4.2). Since the error df, $n - tr[H(\lambda)]$, is not usually an integer, the value for t must be found by interpolating the t-table.

Compared to the usual intervals for the user's model by least squares, those formed for HATLINK using III.4.2 tend to be somewhat wider at X locations in the interior of the X-range. This happens since $h_{ii}^{(ker)} > h_{ii}^{(ols)}$ generally holds for locations X_i away from the extremes of the X-range because of the nature of kernel to give heavier prediction weight to local observations. See, for example, the entries of the hat matrices for Example #2 in Table III.2.3. Intuitively, one might expect the HATLINK intervals to be somewhat wider, as they are attempting to cover the true value of f no matter what functional form f might have. The usual ordinary least squares intervals are designed to accomplish the simpler task of covering the function f, whose form is supposedly known to be in the restricted class of functions indexed by the parameters of the prescribed model. It will be seen in the next chapter that the usual least squares intervals often fail to contain the true $f(X_i)$ when the user's model is only slightly misspecified. The HATLINK intervals, however, are somewhat robust to model misspecification.

III.4.C. Measures of regression utility

The usual coefficient of determination, R^2 , measures how much of the fluctuation in the response variable Y can be "explained" by fitting the regression on X *according to the prescribed model* . This version of R^2 is restrictive in the sense that its interpretation is limited to the particular type of model given by the user. If Y is related to X by some other funcional form, the usual R^2 will understate the degree to which X and Y are truly related. However, an R^2 statistic based on HATLINK will indicate how much of the variation in Y can be attributed to its relationship to X through *any* smooth continuous function. The R^2 statistic for HATLINK is defined in the natural way as follows.

$$R^2 = \frac{SSY - SSE(\lambda)}{SSY} . \qquad \text{(III.4.3)}$$

In the above expression, $SSY = \Sigma(Y_i - \overline{Y})^2$ and $SSE(\lambda) = \Sigma(Y_i - \hat{Y}_i(\lambda))^2$.

Since an overfit to a specific data set occasionally occurs through HATLINK, it seems reasonable to adjust R^2 to protect against this possibility. Let

$$R^2 adjusted = \frac{SSY - \dfrac{(n-1)\,SSE(\lambda)}{n - tr[H(\lambda)]}}{SSY} . \qquad \text{(III.4.4)}$$

An overfit would result in a misleadingly small value for $SSE(\lambda)$, so that R^2 (eq. III.4.3) would be inflated. However, the factor, $(n-1)/(n - tr[H(\lambda)])$, in $R^2 adjusted$ (eq.III.4.4) is increasing in $tr[H(\lambda)]$, and therefore acts as a penalty for overfitting.

Another measure of the overall usefulness of the postulated model in standard regression analysis is the test statistic,

$$F^{O}_{tr[H_{ols}]-1,\,n-tr[H_{ols}]} = \frac{SSY - SSE_{ols}}{(tr[H_{ols}] - 1)\,s^2_{ols}} . \qquad \text{(III.4.5)}$$

Again, this statistic is limited in that it only provides information as to whether the *specified model* is useful for predicting Y. The HATLINK version of the overall F test, defined by

$$F^{O}_{tr[H(\lambda)]-1,\,n-tr[H(\lambda)]} = \frac{SSY - SSE(\lambda)}{(tr[H(\lambda)] - 1)\,s^2(\lambda)}, \qquad \text{(III.4.6)}$$

is more general, as it indicates whether there is *any* (reasonable) function of X that can be used to predict Y. Of course, this HATLINK version of the F^{O} statistic does not exactly follow an F distribution. It will be demonstrated in Chapter IV, however, that this pseudo-F statistic does provide useful diagnostic information when compared to percentage points of an F-distribution.

III.4.D. Diagnostics for the lack of fit of the user's model

Information regarding the adequacy of the model specified by the user is provided naturally through the use of HATLINK. For example, the value of λ found to minimize PRESS is indicative of whether or not the user's model is correct. Also, F-like statistics may be constructed which are analogous to the usual F formed using the reduction in sum of squares principle and to the usual F for testing lack of fit.

Methods based on the value of the mixing parameter selected

If one of the λ selection criteria (Section III.3.B) , such as $PRESS(\lambda)$, is minimized at $\lambda = 0$ for a given set of data, then that would imply that the user's model is correct. If the minimum occurs at a high value of λ , then the criterion would be suggesting that the model has been incorrectly specified. One may therefore think of the value of λ obtained through minimizing $PRESS(\lambda)$ as being a test statistic for testing the hypotheses,

$$H_0: \lambda_{true} = 0 \text{ (model correct)}$$
$$H_A: \lambda_{true} > 0 \text{ (model incorrect)} .$$

Here, λ_{true} denotes the value of λ that would, in theory, provide the best possible fit. That is, for the particular set of X values and the true underlying model, $Y = f(X) + \varepsilon$, there is some value λ_{true} which, on the average, provides the best estimation of f through HATLINK.

It will be seen in Chapter IV that the criteria $PRESS^*(\lambda)$, $Cp1$, and $Cp3$, are the best suited for the purpose of diagnosing lack of fit by means of the λ values they select. Unfortunately, the distribution of the statistic λ found by minimizing $PRESS^*$, for example, does not have a common form, such as a chi-squared distribution. Difficulty in deriving the distribution of such a statistic arises from the non-idempotency and asymmetry of H_{ker} . Formal tests could only be formed through empirical methods, wherein one would obtain a critical value for λ through a number of simulations under the null hypothesis. (This is done, for example, in the "KANOVA" procedure of Krutchkoff, 1987, where simulations under the null hypothesis are conducted in order to set critical values for one-way and two-way analysis of variance type tests.) Such a critical value will depend not only on the form of the user's model, but also on the error variance and the particular X locations for the data set in question. Applications of this empirical testing method based on λ are presented in Sections IV.3.A and IV.4.C.

Reduction in sum of squares F test

The following F-like diagnostic is suggested by analogy to the reduction in sum of squares principle of standard regression analysis (for example, Draper and Smith, 1981).

$$F^*_{tr[H(\lambda)] - tr[H_{ols}],\, n - tr[H(\lambda)]} = \frac{SSE_{ols} - SSE(\lambda)}{(tr[H(\lambda)] - tr[H_{ols}])\, s^2(\lambda)} . \qquad \text{(III.4.7)}$$

One may think of the user's model as the "reduced model" and the HATLINK fit as being associated with some "full model." The difference $SSE_{ols} - SSE(\lambda)$ then corresponds to the reduction in sum of squares achieved by replacing the user's model with the HATLINK model. The effective increase in degrees of freedom in making this replacement is given by $tr[H(\lambda)] - tr[H_{ols}]$. Though F^* does not truly follow an F distribution, the value of F^* may be used for diagnostic purposes by comparison with critical values for an F with $tr[H(\lambda)] - tr[H_{ols}]$ and $n - tr[H(\lambda)]$ numerator and denominator df.

When a value of $\lambda = 0$ is selected for HATLINK, both the numerator and denominator of F^* are zero. We will define F^* to be zero in this situation, since $\lambda = 0$ itself implies no lack of fit. A problem with F^* arises, however, when a λ slightly above zero is selected. This will cause a very small value for $tr[H(\lambda)] - tr[H_{ols}]$ to appear in the denominator of F^* . This, in turn, may lead to a large F^* when there is only a very modest reduction in sum of squares achieved. Further, there is the problem of extrapolating the F table when the numerator df is between 0 and 1. Under the null hypothesis that the user's model is correct, small values of λ are selected quite often in practice, occasionally resulting in high values of F^* . Therefore, due to the instability of F^* under the null hypothesis, it is not recommended that F^* be used alone as a diagnostic for lack of fit under the usual HATLINK formulation.

It does make sense to consider F^* as the second stage of a two-stage diagnostic procedure. First, a value greater than one for the numerator degrees of freedom will be required. That is, if the numerator df is below one then F^* will be taken to be zero. Thus, the determination of the model degrees of freedom for the HATLINK fit, through the selection of a kernel bandwidth and a value for λ , serves as a first stage for detecting lack of fit for the user's model. If the degrees of freedom for the HATLINK fit is small, so that the numerator df, $tr[H(\lambda)] - tr[H_{ols}]$, is below one, then this information alone is indicative that the user's model is specified correctly. When the numerator degrees of freedom is large, this suggests that the user's model is incorrect. However, for such cases, there may or may not be a substantial improvement in the fit obtained by the HATLINK model compared to the fit obtained through the user's parametric model. By incorporating the reduction in sum of squared errors associated with moving from the fit for the parametric model to the

HATLINK fit, the F^* statistic will indicate whether such improvement has occurred. The use of this two-stage diagnostic procedure involving F^* is demonstrated through simulations in Chapter IV.

Lack of fit F test

The standard test statistic for lack of fit in regression can be espressed in the form,

$$F_{tr[H_m]-tr[H_{ols}],\, n-tr[H_m]} = \frac{\underline{Y}'(H_m - H_{ols})\,\underline{Y} \;/\; (tr[H_m] - tr[H_{ols}])}{\underline{Y}'(I - H_m)\,\underline{Y} \;/\; (n - tr[H_m])} \quad , \qquad \text{(III.4.8)}$$

where H_m is the hat matrix that would be used to fit to the mean of the observed Y's at each individual X location. That is, H_m is the hat matrix associated with fitting the means model $Y_i = \mu_i + \varepsilon_i$, where μ_i denotes the mean response at location X_i. Here, the null hypothesis that the postulated model is adequate is tested by considering the difference between the least squares fitted value and the mean of the responses at each data location. The quantities, $\underline{Y}' H_m \underline{Y}$ and $\underline{Y}' H_{ols} \underline{Y}$ are the regression sums of squares for the means model fit and the least squares fit to the user's model, respectively. When there is only one response at each X location, then H_m reduces to I, resulting in zero degrees of freedom for the denominator, so that the test cannot be performed.

Now consider a variation of this F test based on the kernel or HATLINK fit. Since these methods make predictions based on local observations, the corresponding hat matrices play a similar role to that of the matrix H_m. In particular, note that the kernel hat matrix H_{ker} is approximately equal to H_m when the kernel bandwidth is extremely small. That is, $H_{\text{ker}}(h) \rightarrow H_m$ as $h \rightarrow 0$. Thus, the following revision of the standard F statistic (eq. III.4.8) is reasonable.

$$F^{**}_{\nu_1,\,\nu_2} = \frac{\underline{Y}'(H(\lambda) - H_{ols})'(H(\lambda) - H_{ols})\,\underline{Y} \;/\; \nu_1}{\underline{Y}'(I - H(\lambda))'(I - H(\lambda))\,\underline{Y} \;/\; \nu_2} \quad , \qquad \text{(III.4.9)}$$

where $\nu_1 = tr[(H(\lambda) - H_{ols})'(H(\lambda) - H_{ols})]$, and $\nu_2 = tr[(I - H(\lambda))'(I - H(\lambda))]$. The value of F^{**} for a given regression should be compared to the appropriate percentile of the usual F distribution, interpolating for ν_1 and ν_2 degrees of freedom. Unlike the usual F test for lack of fit, the F^{**} test does not require replicates.

As is the case for the F^{*} diagnostic, it has been shown empirically that the F^{**} statistic is unstable under the null hypothesis that the user's model is correct. When $\lambda = 0$ is selected for HATLINK, then $\nu_1 = 0$ and $H(\lambda) - H_{ols} = O$, a matrix of zeros. We define $F^{**} = 0$ in this situation. However, for very small nonzero $\lambda's$, F^{**} may be unreasonably large due to a small fractional value for ν_1 appearing in the denominator. Therefore, it is recommended that F^{**} be used as the second stage of a lack of fit diagnosis, being employed only for cases where the numerator degrees of freedom is at least one. When the numerator df falls below one, then F^{**} is given the value zero. An empirical investigation of the two-stage version of F^{**} appears in Chapter IV.

F tests based on a nonstochastic choice of h and lambda

Both F^{*}andF^{**} can be applied alone as diagnostics if the HATLINK procedure is conducted in the following manner. Instead of allowing the kernel bandwidth to be selected adaptively for a given regression, let the bandwidth be determined so that the model degrees of freedom $tr[H_{ker}]$ is equal to a prescribed value. (See Section III.3.A..) Then form HATLINK using a predetermined value for λ . (See the end of Section III.3.B.) The numerator degrees of freedom in F^{*} , $tr[H(\lambda)] - tr[H_{ols}]$, can be set to any desired level through the designation of λ and $tr[H_{ker}]$. In particular, it can be guaranteed that the numerator degrees of freedom for either F^{*} or F^{**} will be above one. For example, suppose there are $p = tr[H_{ols}]$ parameters in the user's model. Then taking $tr[H_{ker}] = p + 4$ and $\lambda = .25$ would make $tr[H(\lambda)] = \lambda tr[H_{ker}] + tr[H_{ols}] = p + 1$, so that F^{*} will have 1 and n-p-1 numerator and denominator degrees of freedom.

Similarly, F^{**} may be used directly, not just as part of a two-stage procedure, if the statistic is based on a nonstochastic HATLINK fit. Again, through the proper specification of kernel df and λ , cases where the numerator df falls below one can be avoided. Further, since $H(\lambda)$ will be more like the matrix H_m of the usual lack of fit F test (eq. III.4.8) whenever $tr[H(\lambda)]$ is relatively large, it is recommended that fairly high values for $tr[H_{ker}]$ and λ be prescribed. An empirical investigation of F^{**} has been done for the particular version of nonstochastic HATLINK with $\lambda = 1$ and $tr[H_{ker}] = 6$, where a simple linear regression model ($p = 2$) has been specified by the user. The results of this investigation are presented in Section IV.3.A

III.4.E. Multiple Regression

In the multiple regression setting, it is rarely the case that one will know the exact form of the function which relates Y to the set of regressors. Thus, the model-based least squares approach may result in poor prediction and misleading inferences. On the other hand, the kernel method suffers greatly due to the relative sparcity of the data in the multidimensional regressor space. In light of the shortcomings of the model-based and local-distance approaches, the multiple regression setting is the place where HATLINK has its greatest potential.

Extension of HATLINK

Extension of HATLINK to multiple regression first requires extension of the kernel portion of HATLINK. This can be accomplished in the following way. Let r denote the number of regressor variables and $\underline{X}_i$ be an $r \times 1$ vector representing the i^{th} data point. The goal is to predict Y at location $\underline{X}_i = (X_{i1}, \ldots, X_{ir})'$ for the model $Y_i = f(X_{i1}, \ldots, X_{ir}) + \varepsilon_i$ using a local distance-based set of weights. One such set of weights can be expressed as

$$h_{ij}^{(\ker)} = \frac{e^{-(\|\underline{X}_i - \underline{X}_j\| / h)^2}}{\sum_{j=1}^{n} e^{-(\|\underline{X}_i - \underline{X}_j\| / h)^2}}, \qquad \text{(III.4.10)}$$

where $\|\underline{X}_i - \underline{X}_j\|$ is some appropriate measure of distance and $h > 0$ is the bandwidth. So $\hat{Y}_{\ker}(\underline{X}_i) = \sum_{j=1}^{n} h_{ij}^{(\ker)} Y_j$ gives the local distance-based (kernel) prediction at $\underline{X}_i$. Let $H_{\ker} = (h_{ij}^{ker})$ be the corresponding hat matrix. In the present investigation the Euclidean distance measure is used, with the X-variables first being rescaled to have common variance. The bandwidth is selected by minimizing $PRESS^*(h)$ as in the univariate case. (Equation III.1.6.) We define the hat matrix for the HATLINK method as before:

$$H(\lambda) = \lambda H_{\ker} + (1 - \lambda) H_{ols}, \quad 0 \le \lambda \le 1. \qquad \text{(III.4.11)}$$

The mixing parameter λ is selected by minimizing any of the versions of the PRESS or Cp statistics listed in Section III.3.B for the single regressor case. All of the developments of HATLINK, such as confidence intervals, R^2, and lack of fit tests, extend to multiple regression. In addition, the problem of variable selection can be addressed through HATLINK.

Variable selection

In standard regression analysis, statistics such as R^2, s^2, PRESS, and Cp, are used to compare competing models. The HATLINK counterparts of these quantities may also be used for this purpose. For instance, if HATLINK were used to fit the competing models, $Y = f_1(X_1, X_2, X_3) + \varepsilon$ and $Y = f_2(X_3, X_5, X_6) + \varepsilon$, one would compare the two on the basis of $s^2(\lambda)$, for example. If $s^2(\lambda)$ is lower for the first model, this would suggest that Y is more strongly related to the set of variables $\{X_1, X_2, X_3\}$ than to $\{X_4, X_5, X_6\}$.

Note that such comparisons via HATLINK are robust to possible model misspecification. The standard approach to the preceding model comparison would be to consider the values of s_{ols}^2 for the models $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$ and $Y = \beta_0^* + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \varepsilon$. A

lower s^2_{ols} for the first model would suggest that the variables X_1, X_2, X_3, do a better job of explaining Y than a second model using the variables X_4, X_5, X_6. However, the second set of regressors may truly provide for better prediction of Y, only through some function f other than a simple linear combination. In this sense, model comparisons based on the usual regression techniques are not robust to model misspecification. This lack of model robustness is an inherent part of the various stepwise regression routines, which consider linear combination models exclusively.

In addition to using $R^2(\lambda)$, $R^2\ adjusted(\lambda)$, $PRESS(\lambda)$, $Cp(\lambda)$, and $s^2(\lambda)$, to compare models, there are HATLINK analogs to the sequential F test. In order to determine whether variable X_r should be added to a set of r-1 regressors, the following procedure based on the full versus reduced model approach of the F^* statistic (equation III.4.7) could be used. In this case, the HATLINK fit to the full set of r regressors would be compared to the HATLINK fit to the reduced set of r-1 regressors. This comparison would be accomplished through the test statistic,

$$F^* = \frac{SSE(\lambda, X_{1,}, , , X_{r-1}) - SSE(\lambda, X_{1,}, , , X_r)}{tr[H(\lambda, X_{1,}, , , X_r)] - tr[H(\lambda, X_{1,}, , , X_{r-1})]} . \qquad \text{(III.4.12)}$$

If the value of F^* exceeds the F table value for $tr[H(\lambda, X_{1,}, , , X_r)] - tr[H(\lambda, X_{1,}, , , X_{r-1})]$ numerator and $n - tr[H(\lambda, X_{1,}, , , X_r)]$ denominator degrees of freedom, then it would be concluded that variable X_r does contribute to the model, and therefore should be included with the other regressors. As was noted earlier in Section III.4.D, there is a problem with determining the proper table value of F for comparison with F^* when the numerator degrees of freedom falls below 1. In such cases, the recommended conclusion is that variable X_r does not contribute significantly to the model. This approach is evaluated empirically in Section IV.4.D.

A second method for determining whether a variable should be included in the model is the following stepwise approach. First, perform a HATLINK regression of Y on $X_1, X_2, \ldots, X_{r-1}$ and obtain the residuals. Then regress the residuals on $X_1, X_2, \ldots, X_{r-1}, X_r$, where X_r is now included in the set of regressors. The user's model incorporated in HATLINK during this second stage could be $e = \beta_0 + \beta_r X_r + \varepsilon$. The kernel could be allowed to fit $e = f(X_1, \ldots, X_r) + \varepsilon$. If X_r can

itself, *or interactively* , enhance the prediction of Y, then this improvement should be detected by any of the measures of regression utility developed in Section III.4.C. The lack of fit diagnostics given Section III.4.D would provide information as to whether the X_i term should enter the model in some form other than as an additive linear term.

A variation on this approach is to use a kernel regression which takes the residuals from the first stage and fits these to a function of X_i only. In fitting the model $e = f(X_i) + \varepsilon$ at the second stage, we would be working in the framework of "additive models" (Hastie and Tibshirani, 1986, 1987) discussed in Section II.5.E. In this approach, statistics such as R^2 , $R^2_{adjusted}$, and the F test for regression utlity (Section III.4.C) would detect whether the model is improved when some function of the variable X_i alone is added to the existing model. That is, the methods would not detect the need for X_i to enter interactively into the model. The lack of fit diagnostics (Section III.4.D) in this case would be used to indicate whether X_i should be *added* to the model in some functional form other than as a linear term. Application of this additive models approach to variable selection is demonstrated through an empirical study in Section IV.4.D.

Chapter IV

IV. RESULTS OF SIMULATIONS

IV.1. Direction and Scope of the Study

Chapter II reviewed the ways in which the regression fits are obtained through the model-based and local distance approaches. The drawbacks of both of these methods were highlighted, and it was argued that a compromise between the two would be an improvement as a general method for prediction. To demonstrate that such improvement does, in fact, occur for the HATLINK method, simulation results for a variety of regression settings are presented in this chapter. Analysis of these simulations will focus primarily on the accuracy of predictions, but consideration will also be given to the performance of the various statistics that were developed in Chapter III.

In order to define a particular regression setting for simulation, the following items must be specified:

(i) The sample size n.

(ii) The locations of the regressors.

(iii) The distribution of the error random variable.

(iv) The user's parametric model.

(v) The functional form of the true underlying model.

(vi) The degree to which the true model departs from the user's model.

Since there are countless ways to specify each of the above items, the number of possible regression settings is limitless. The present study has been kept to a manageable size by considering only a few particular specifications of each item. So that the investigation will proceed in an orderly fashion, the analysis will focus on certain families of regressions, for which items (i) through (v) above are held constant, while the degree of misspecification, item (vi), is allowed to vary. In studying such a family of regressions, the behavior of the least squares, kernel, and HATLINK methods will be observed when a parameter of the true model is varied from where the user's model is correct to a point where the user's model becomes clearly questionable.

The effects of having a different sample size, error variance, or set of regressor locations, items (i), (ii), and (iii), are investigated by comparing the results for related families which differ in one of these items. Several different functional forms for the true underlying model are considered. These include quadratic and sinusoidal functions in a single regressor (Sections IV.2 and IV.3), plus various second order models in two regressor variables (Section IV.3). For the single regressor cases, it will most often be assumed that the user has specified a simple linear regression model. However, the situation where the user's model is quadratic will be given some attention in Section IV.3.B. In the two regressor situations, both first order and second order user's models will be considered.

For each of the families of regressions studied in this chapter, an analysis of the prediction performance of the least squares, kernel, and HATLINK methods will be presented. The study of prediction performance is emphasized here since the primary goal of HATLINK is to provide improved predictions. In several instances a comparison of the six different criteria for selecting λ will be included. Additionally, for some of the families of regressions presented here, consideration

will be given to the behavior of the various statistics developed in Chapter III. The issue of variable selection will be addressed through an empirical investigation in Section IV.3.D.

IV.2 Quadratic Underlying Models in a Single Regressor

IV.2.A. Basic quadratic family of models

Consideration is first given to a sequence of nine situations where the user's model was assumed to be the simple linear model, $Y = \beta_0 + \beta_1 X + \varepsilon$, while the true models were of form, $Y = Q(X - 5.5)^2 + 5X + \varepsilon$. The quadratic coefficient Q was allowed to vary over nine different values, ranging from Q = 0 to Q = .55. In each case 20 observations were taken, two each at ten evenly spaced regressor locations from X = 1 to X = 10. Errors were generated at random from the N(0,16) distribution. Two hundred simulations were performed at level Q = .00, 100 simulations were carried out at levels Q = .20, .30, and .40, and 25 simulations were done at the remaining levels of Q. The same initial seed was used to generate random errors, so the results for the 100 runs at levels Q = .20, .30, and .40, for example, are based on the same 100 sets of random errors.

Since duplicate observations were taken at every X location, it is possible to perform the usual lack of fit F test in this situation. The power (P) of this test at $\alpha = .05$ varies from P = .05 at Q = 0 up to P = .61 at Q = .55, but most of the values of Q considered here represent only slight to moderate degrees of misspecification. Table IV.2.1 shows the nine levels of Q that were used, along with the corresponding powers of the lack of fit test. To emphasize how close the user's model is to being correct over most of this range of Q's, consider Figure IV.2.1. Here, the true values of the underlying function $f(X) = Q(X - 5.5)^2 + 5$ are plotted against X for the Q = .35 case. Observe that this underlying function is itself very nearly a straight line. Therefore, when a

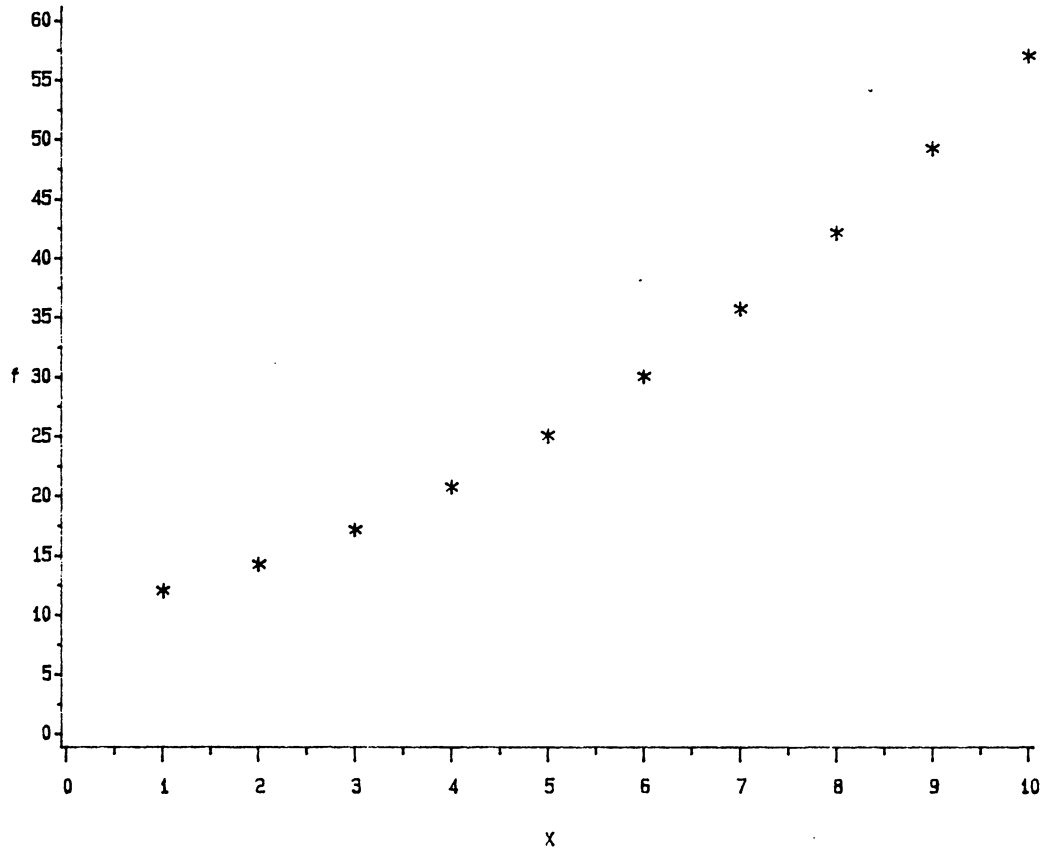


Figure IV.2.1. Plot of True Values of f(X) for the Quadratic Model with Quadratic Coefficient Q=.35.

random error term is added to each value of f(X), it will be quite difficult for a user to detect either visually, or through the test for lack of fit, that the straight line model is incorrect.

Table IV.2.1. Power (P) of the lack of fit test at $\alpha = .05$ for the 9 levels of the quadratic coefficient Q used in the basic quadratic family of models.

Q	.00	.10	.20	.25	.30	.35	.40	.45	.55
P	.05	.06	.11	.14	.19	.26	.33	.42	.61

Prediction results

In Table IV.2.2 the prediction performances of the least squares, kernel, and HATLINK methods are summarized in terms of SSEP, the sum of square errors of prediction at the data locations. (See Equation III.2.1.) Results are shown for HATLINK as constructed using the Cp3 criterion for selecting λ developed in Section III.3.B. The values presented in the table are the mean SSEP's for the set of simulations carried out for every value of the coefficient Q. At every level of Q, the various prediction methods were compared on the basis of mean SSEP through a randomized block analysis with the individual runs as blocks. The least significant difference procedure was then used to compare mean SSEP's for pairs of methods. Cases where the Cp3 version of HATLINK was found through the least significant difference procedure at level $\alpha = .05$ to be significantly better than the least squares method are noted by "+L" in the table. Cases where the Cp3 version of HATLINK was significantly worse than the least squares method are noted by "-L". Similarly, "+K" and "-K" are used to indicate situations where the Cp3 HATLINK predictions were significantly better or worse than the predictions obtained through the kernel method. Values of the least significant difference (lsd) are listed in Table IV.2.3 for the randomized block type analysis conducted at each of the levels of Q.

Table IV.2.2. Prediction Performance of the Least Squares, Kernel, and HATLINK methods for the Basic Quadratic Family of Regressions. Results are based on 200 simulations at level Q = .00 of the quadratic coefficient, 100 simulations at levels Q = .20, Q = .30, and Q = .40, and 25 simulations at the remaining levels of Q. Results shown for HATLINK are for the Cp3 version.

		Mean SSEP				Relative Prediction Efficiency		
Q	POWER	LS	KER	H		RPE(LS,K)	RPE(H,K)	RPE(H,LS)
.00	.05	19.6	56.4	22.9	(+K,-L)	2.87	2.46	0.86
.10	.06	26.3	61.7	29.3	(+K)	2.34	2.10	0.90
.20	.11	40.2	57.2	36.8	(+K)	1.42	1.55	1.09
.25	.14	52.7	63.7	47.0	(+K,+L)	1.21	1.35	1.12
.30	.19	64.7	58.5	47.6	(+K,+L)	0.90	1.23	1.36
.35	.26	82.2	64.3	56.5	(+L)	0.78	1.14	1.45
.40	.33	98.8	59.5	57.5	(+L)	0.60	1.04	1.72
.45	.42	121.4	65.3	63.9	(+L)	0.54	1.02	1.90
.55	.61	170.1	65.5	67.9	(+L)	0.39	0.96	2.51

Also included in Table IV.2.2 is a comparison of the relative prediction performances of the three regression methods. We make these comparisons on the basis of the relative prediction efficiency, RPE, defined as the ratio of the mean sum of squared errors of prediction for the two methods being compared. That is, the relative prediction efficiency of method A compared to method B is given by

$$RPE(A,B) = \frac{\overline{SSEP(B)}}{\overline{SSEP(A)}} , \qquad \text{(IV.2.1)}$$

where $\overline{SSEP(A)}$ is the mean value of SSEP for method A for the set of simulations at a particular level of Q. The RPE's for comparing the HATLINK procedure to the least squares and kernel methods are plotted in Figure IV.2.2 as a function of the power of the lack of fit F test for the nine Q values. Again, the results for HATLINK are obtained using the Cp3 criterion.

The results displayed in Table IV.2.2 and Figure IV.2.2 are quite favorable for the HATLINK procedure. As expected, the least squares method performs the best when the user's model happens to be exactly correct. In this situation, the kernel predictions are extremely poor, while the HATLINK predictions are closer to those obtained by least squares. Once the quadratic coefficient of the true model is increased to $Q = .20$, where the power of the lack of fit test is only .10, the least squares method is outperformed by the HATLINK procedure based on Cp3. By the time Q reaches .30, predictions obtained using the kernel method are better than those obtained by least squares. The HATLINK predictions are superior to the kernel predictions throughout this family of models, except for the extreme $Q = .55$ case, where the kernel slightly outpredicts the conservative Cp3 version of HATLINK.

Tables IV.2.3 and IV.2.4 provide a comparison of the six criteria developed in Section III.3.B for selecting λ. The prediction performances of the six versions of HATLINK corresponding to these criteria are evaluated in Table IV.2.3 for each of the nine levels of the quadratic coefficient Q. Again, this evaluation is done on the basis of SSEP, the sum of squared errors of prediction at the data locations. It is observed that the Cp3 criterion leads to the lowest mean SSEP's for all levels of Q except for $Q = .55$. The PRESS* criterion also performs relatively well

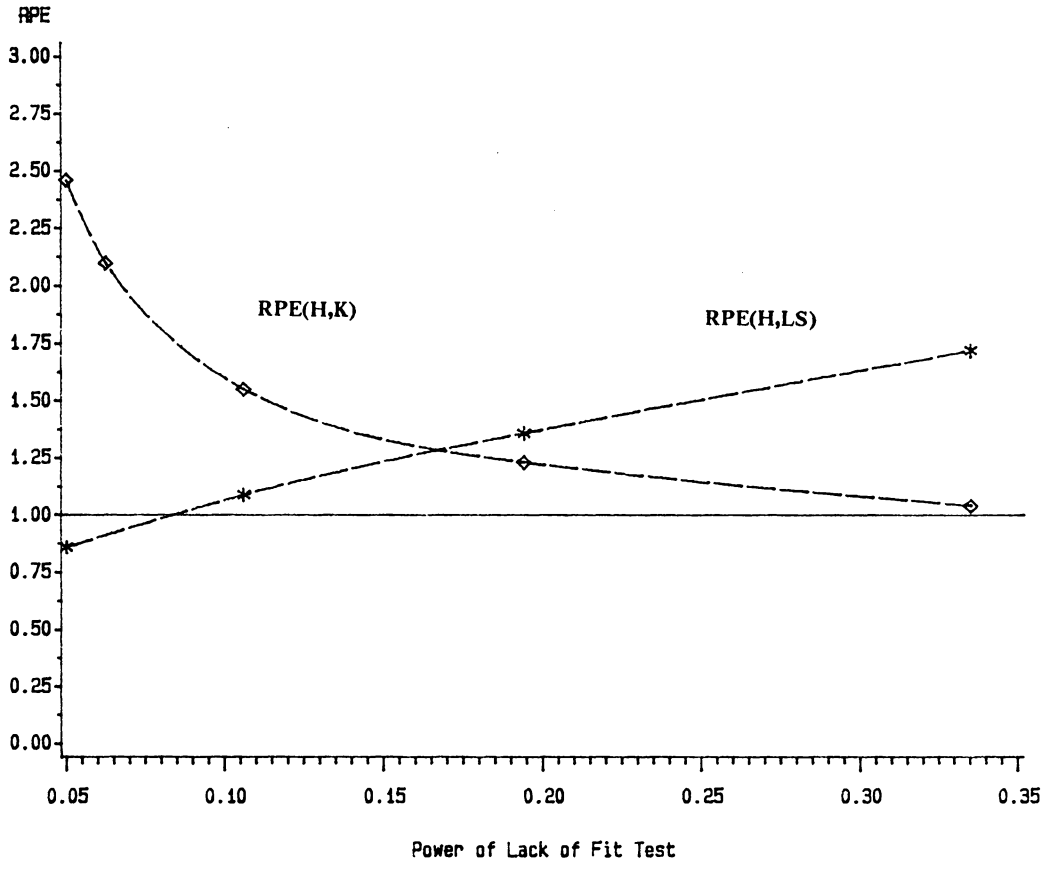


Figure IV.2.2. Relative Prediction Efficiencies (RPE).
Results are for the basic quadratic family of regressions.

Table IV.2.3. Prediction Performance of the HATLINK Method for the Six Criterion for Selecting the Mixing Parameter. Results are for the basic quadratic family of regressions.

		Mean Sum of Squared Errors of Prediction									
Q	POWER	LS	KER	PRESS	PRESS*	Cp1	Cp2	Cp3	Cp4	OPT	lsd
.00	.05	19.6	56.4	36.0	24.5	30.6	40.6	22.9	38.6	18.1	3.2
.10	.06	26.3	61.7	39.4	30.1	35.0	49.9	29.3	41.7	23.8	9.2
.20	.11	40.2	57.2	45.4	38.8	42.5	47.7	36.8	47.3	29.4	4.3
.25	.14	52.7	63.7	54.0	49.8	52.0	58.3	47.0	55.6	35.5	9.9
.30	.19	64.7	58.5	53.6	50.8	52.1	54.2	47.6	55.0	37.4	4.5
.35	.26	82.2	64.3	62.6	59.9	60.2	74.1	56.5	62.6	43.2	10.7
.40	.33	98.8	59.5	59.0	60.8	59.3	59.6	57.4	59.7	43.8	5.0
.45	.42	121.4	65.3	65.3	67.2	65.1	74.0	63.9	65.8	48.5	11.8
.55	.61	170.1	65.5	65.4	68.6	65.6	63.0	67.9	65.9	51.0	12.2

Table IV.2.4. Mean Values of the Mixing Parameter λ for the Six Selection Criteria. Results are for the basic quadratic family of regressions.

Q	POWER	PRESS	PRESS*	Cp1	Cp2	Cp3	Cp4	OPT
.00	.05	.37	.14	.26	.65	.14	.41	.13
.20	.11	.56	.28	.44	.66	.24	.63	.37
.30	.19	.68	.41	.57	.72	.34	.74	.54
.40	.33	.79	.54	.70	.80	.43	.84	.67
.55	.61	.91	.76	.86	.91	.63	.93	.84

for this family of regressions. Predictions obtained using the PRESS, Cp1, and Cp4 criteria are not as good as for Cp3 and PRESS* at the lower levels of Q, but they do provide better preditions than the kernel method at these levels. The Cp2 criterion resulted in the lowest mean SSEP of all methods at the Q = .55 level, but the method performs poorly at all other levels of Q.

The column labeled "OPT" in Table IV.2.3 shows prediction results for HATLINK using the optimal value for λ in each run. That is, for a given data set, the value of λ is chosen which minimizes the sum of squared errors of prediction, $SSEP(\lambda) = \Sigma(\hat{Y}_i(\lambda) - f(X_i))^2$, over the interval $[0, 1]$. In practice, of course, this minimizing value cannot be obtained, since the true underlying function f is not known. However, it is useful in the present research to compare the six practical methods for obtaining λ to this optimal method. For example, in Table IV.2.3, it is seen that predictions obtained through HATLINK based on the optimal λ are substantially better than those obtained by the other versions of HATLINK. This indicates that there is hope of improving the HATLINK procedure by developing improved methods for selecting λ .

Table IV.2.4 shows the mean λ values obtained by six criteria at several levels of the quadratic coefficient. The last column of this table shows the average values of λ chosen to minimize SSEP(λ) , as discussed in the preceding paragraph. In this table it is observed that the PRESS* and Cp3 criteria produce lower average values for λ than the other methods. Except for the Q = 0 case, the mean λ'*s* obtained by the PRESS* and Cp3 methods are lower at each level of Q than the mean optimal λ . The PRESS, Cp1, Cp2, and Cp4 criteria generally lead to higher values of λ , while the Cp2 method is evidently quite poor for selecting λ when the user's model is correct or nearly correct.

Variance estimates and confidence intervals

The behavior of the variance estimates developed for the kernel and HATLINK methods in Sections II.5.B and III.4.A is now reviewed. Table IV.2.5 displays the mean and standard deviation of the estimates of σ^2 based on the least squares, kernel, and HATLINK regressions for the

Table IV.2.5. Mean and Standard Deviations of the Estimates of Error Variance Obtained by the Least Squares and Kernel Methods and the Cp3 and Cp4 Versions of HATLINK. Also included are the mean degrees of freedom associated with these regression methods. Results are for 200 runs at level Q = .00 and 100 runs at each of the other three levels of the quadratic coefficient Q for the basic quadratic family. The true value of σ^2 is 16.0. For reference, the standard error of these mean estimates of σ^2 ranges from .41 (LS) to .49 (Kernel) at Q = .00, and from .72 (Cp3 and Cp4) to .75 (LS and Kernel) at Q = .40.

	Least Squares			Kernel			HATLINK -- Cp3			HATLINK -- Cp4		
Q	mean s^2	s.d. s^2	mean df	mean s^2	s.d. s^2	mean df	mean s^2	s.d. s^2	mean df	mean s^2	s.d. s^2	mean df
.00	16.5	5.8	2	15.9	6.9	4.9	16.5	6.4	2.4	15.5	6.6	3.5
.20	18.8	6.1	2	15.7	7.2	4.9	16.7	7.0	2.7	15.5	7.0	3.8
.30	21.7	6.7	2	15.7	7.2	5.0	17.1	7.1	3.0	15.6	7.1	4.2
.40	25.9	7.5	2	15.7	7.2	5.0	17.5	7.2	3.3	15.7	7.2	4.5

Table IV.2.6. Mean 95% Confidence Interval Widths at Location X = 6 and Percent Coverage of f(6). Intervals are formed by the least squares and kernel methods and by the Cp3 and Cp4 Versions of HATLINK. Results are for 200 simulations at level Q = .00 of the quadratic coefficient Q and 100 simulations at Q = .20, .30, and .40, for the basic quadratic family of regressions. The estimated standard error of these percentage estimates ranges from 1.5 (for all methods at Q = .00) to 4.8 (for the least squares intervals at Q = .30).

	Least Squares		Kernel		HATLINK -- Cp3		HATLINK -- Cp4	
Q	mean width	% cov.	mean width	% cov.	mean width	% cov.	mean width	% cov.
.00	3.8	94.5	7.7	97.5	4.5	94	5.6	94.5
.20	4.1	66	7.7	96	5.0	82	6.0	86
.30	4.4	37	7.8	96	5.5	75	6.6	86
.40	4.8	18	7.8	96	6.1	76	7.1	87

Table IV.2.7. Mean 95% Confidence Interval Widths at Location X = 8 and Percent Coverage of f(8). Intervals are formed by the least squares and kernel methods and by the Cp3 and Cp4 Versions of HATLINK. Results are for 200 simulations at level Q = .00 of the quadratic coefficient Q and 100 simulations at Q = .20, .30, and .40, for the basic quadratic family of regressions.

	Least Squares		Kernel		HATLINK -- Cp3		HATLINK -- Cp4	
Q	mean width	% cov.	mean width	% cov.	mean width	% cov.	mean width	% cov.
.00	5.0	95.5	7.7	95	5.4	96.5	6.2	94.5
.20	5.3	94	7.7	96	5.8	97	6.5	95
.30	5.7	97	7.8	97	6.2	98	7.0	97
.40	6.3	98	7.9	97	6.6	96	7.3	97

simulations performed at several of the levels of Q. Also shown are the mean values for the model degrees of freedom, $tr[H]$, associated with the kernel fits, and with the HATLINK fits based on the Cp3 and Cp4 methods for selecting λ . For simplicity, comparisons will be limited here to one method (Cp3) which tends to select small $\lambda's$, and one method (Cp4) that selects higher $\lambda's$.

Table IV.2.5 shows that the variance estimates obtained through the kernel and two HATLINK methods are fairly accurate when the user's model is correct (Q = 0), and remain accurate as the quadratic coefficient is increased. However, the usual least squares estimate of variance is biased upward when the user's model is incorrect. These results were foreseen in Section II.5.B, where it was demonstrated that the kernel variance estimate is approximately unbiased, provided that the true underlying function is not too erratic. In favor of the least squares estimator, it should be noted that this estimator was found to be more precise than the kernel and HATLINK estimators when the quadratic coefficient Q was .20 or less. That is, the standard deviation of the least squares variance estimates was lower at these levels of Q.

A method for constructing confidence intervals based on HATLINK was developed in Section III.4.B, where it was claimed that the resulting intervals should maintain coverage of the true underlying function when the user's model is incorrect. Further, it was stated that the usual intervals based on least squares could suffer under model misspecification. These claims are at least partly substantiated by the results shown in Table IV.2.6. This table shows the percentage of times that 95% confidence intervals formed at location X = 6 actually covered the true value, f(6). Note that this location is close to the point X = 5.5, where the bend in the quadratic underlying function is most severe. We observe that the usual least squares confidence intervals do, in fact, suffer when the model is incorrectly specified. What may be surprising is the poor coverage obtained by these intervals when the degree of model misspecification is very slight. For instance, the least squares 95% confidence intervals gave only 37% coverage of f(6) when Q = .30. Recall that the power (P) of the lack of fit test is only P = .19 at this level of Q.

On the other hand the interval based on kernel regression maintained excellent coverage of f(6) throughout the entire range of quadratic departures from the straight line. The two HATLINK intervals suffered some loss of coverage, but not to the extent that this occured for the least squares

intervals. This loss in coverage for the HATLINK intervals is due to the fact that the methods for selecting λ sometimes result in $\lambda = 0$. In such cases the HATLINK intervals are the same as those formed by the least squares, and therefore have reduced chances for covering the true function in such instances. Note that the intervals based on the conservative Cp3 criterion show slightly lower percent coverage for many levels of Q than the intervals based on the Cp4 criterion, which tends to select higher values of λ. In achieving improved coverage of f(6) through the kernel and HATLINK methods, a penalty is paid in the form of wider intervals. In particular, the kernel intervals are up to twice as wide, on the average, as the least squares intervals. Table IV.2.6 includes the mean interval widths for the various methods at each level of Q. Note that the intervals formed through HATLINK also tend to be wider than the least squares intervals, but narower than the kernel intervals. Thus, the HATLINK intervals present a compromise between validity of coverage and sensitivity, or width.

Table IV.4.7 shows the percent coverage and mean widths for intervals formed at location X = 8. This is a point where the least squares fitted value is generally close to the true value f(8), even when the quadratic coefficient Q is large. Therefore, the least squares intervals maintain the proper coverage of f(8) at all values of Q. However, since X = 8 is farther from the center of the data, the mean widths of the least squares intervals are greater than in the X = 6 case and not that much below the mean widths of the HATLINK intervals based on Cp3.

IV.2.B. Variations on the basic quadratic family

Different error variance

In the basic quadratic family of runs discussed in the preceding section, the error standard deviation was $\sigma_\varepsilon^2 = 16$. To consider the effect of having a different error variance, some additional runs were done using $\sigma_\varepsilon^2 = 4$ and $\sigma_\varepsilon^2 = 36$. So that the results for the differing values of error

Table IV.2.8. Comparison of Prediction Performances for Quadratic Models with Differing Error Variances. The results for HATLINK are for the Cp3 method for selecting the mixing parameter. The results presented are for 25 simulations at each level of Q considered.

			Mean SSEP			Relative Prediction Efficiency		
Q	POWER	σ^2	LS	KER	H	RPE(LS,K)	RPE(H,K)	RPE(H,LS)
0	.05	4	5.2	15.6	5.6	3.01	2.79	0.93
0	.05	16	20.8	60.9	24.0	2.93	2.54	0.86
0	.05	36	46.7	132.4	59.4	2.83	2.23	0.79
.125	.14	4	13.2	15.8	11.7	1.20	1.35	1.12
.25	.14	16	52.7	63.7	47.0	1.21	1.35	1.12
.38	.14	36	120.3	144.5	108.0	1.20	1.34	1.11
.175	.26	4	20.6	16.2	14.6	0.79	1.11	1.41
.35	.26	16	82.2	64.3	56.5	0.78	1.14	1.45
.52	.26	36	182.4	145.7	126.0	0.80	1.16	1.45

variance could be fairly compared, the same 25 sets of random errors were used for all cases. The results are presented in Table IV.2.8. As expected, the sums of squared errors of prediction for all methods are lower for $\sigma_\varepsilon^2 = 4$ and higher when $\sigma_\varepsilon^2 = 36$. Considering the values of RPE(LS,K) and RPE(H,K) for the three levels of variance at $Q = 0$, there appears to be a slight improvement in the kernel predictions relative to least squares and HATLINK as σ_ε^2 is increased. On the other hand, the performance at $Q = 0$ of HATLINK relative to least squares, measured by RPE(H,LS), improves somewhat as $\sigma^2{}_\varepsilon$ decreases, reaching RPE(H,LS) = .93 at $\sigma_\varepsilon^2 = 4$.

In order to fairly compare the results for different values of σ_ε^2 when Q is not equal to 0, it is necessary to match different levels of Q to the three levels of σ_ε^2 , so that the powers for the lack of fit tests are roughly equal. For example, a power (P) of approximately $P = .14$ is obtained for the following three combinations of Q and σ_ε^2 :

$Q = .125$ and $\sigma^2{}_\varepsilon = 4$

$Q = .25$ and $\sigma^2{}_\varepsilon = 16$

$Q = .38$ and $\sigma^2{}_\varepsilon = 36$.

Prediction results for these three combinations of Q and error variance are shown in Table IV.2.8., as are results for three additional combinations which make the power of the lack of fit test approximately $P = .26$. The values for the relative prediction efficiencies when the power is $P = .14$ are nearly constant over the three levels of σ_ε^2 . This is also true when the lack of fit test has power of $P = .26$. Thus, when the RPE's are viewed as a function of the power of the lack of fit test, the relative prediction efficiencies of the least squares, kernel, and HATLINK methods appear to remain nearly constant over differing levels of σ_ε^2 . An exception to this statement occurs at when there is no lack of fit ($Q = 0$), where, as noted previously, there are small changes in the RPE's for differing error variances.

Different regressor locations

Several alternative sets of X locations have been explored for the quadratic family of underlying models. The sets of regressor locations considered here are as follows.

(i) Twenty evenly spaced observations from X = 1 to X = 10 (with no replicates).

(ii) Two observations at each of ten locations, where the ten locations are obtained as the normal scores scaled to cover the interval from X = 1 to X = 10. Specifically, the X locations were X = 1.41, 2.94, 3.83, 4.55, 5.19, 5.81, 6.45, 7.17, 8.06, and 9.59.

(iii) Twenty observations at distinct locations between X = 1 and X = 10 as determined by the normal scores.

Table IV.2.9 compares the prediction results for the original ten location example to the results for twenty evenly spaced locations at several levels of Q. In this table it is observed that the kernel predictions, relative to least squares and HATLINK, are clearly worse for the twenty location situation. Apparently, the kernel method, with bandwidth selected by the PRESS*(h) is more likely to overfit to spurious trends in the data when there are no replicates. In the Q = .2 case, for example, the mean model degrees of freedom for the kernel increased from 4.84 in the ten location example to 5.94 in the twenty location example. For the latter example, there were five out of twenty-five simulations for which the kernel df exceeded 11. In these five simulations the SSEP's for the kernel procedure were extremely high.

The relative prediction efficiency of HATLINK compared to least squares changed only slightly for the twenty location example. In the case where the user's model was correct, RPE(H,LS) increased by a small amount for the twenty location, but decreased slightly for twenty locations at the Q = .35 level of misspecification. Thus, for the twenty location case, the least squares method holds a smaller advantage over HATLINK when the user's model is correct, but HATLINK does not outperform least squares by quite as wide a margin when the misspecification is moderate. Generally, though, the same basic pattern holds for RPR(H,LS) as a function of Q as was observed in the ten location case.

Table IV.2.9. Comparison of Prediction Performances for the Quadratic Family of Models for the 10 and 20 Location Cases. Results are based on 25 runs at each level of the quadratic coefficient Q considered.

		RPE(LS,K)	RPE(H,K)	RPE(H,LS)
Q = 0	10 Locations	2.93	2.54	0.86
	20 Locations	3.84	3.61	0.94
Q = .20	10 Locations	1.52	1.55	1.02
	20 Locations	2.09	2.18	1.04
Q = .35	10 Locations	0.78	1.14	1.45
	20 Locations	1.12	1.50	1.34

Next, the prediction results for original set of ten evenly spaced regressor locations are compared to the results for the case where normal scores are used to determine the spacing of the ten locations. When the ten regressor locations are obtained through the normal scores, the performance of the kernel method relative to least squares is much worse than it was for evenly spaced X's. (See Table IV.2.10.) This increase in RPE(LS,K) occurred at all three levels considered here for the power of the lack of fit F test. For the normal scores locations, the mean SSEP for kernel exceeds the mean SSEP for least squares even in the Q = .45 case, where the power of the lack of fit test is P = .32. The failure of the kernel method in the normal scores case causes the HATLINK predictions to suffer somewhat, relative to the least squares. When Q = 0, the Cp3 criterion is able to discern that the least squares predictions are better than those obtained by kernel, so that RPE(H,LS) remains the same for the normal scores locations as it was for the evenly spaced locations. However, the advantage held by HATLINK over least squares under model misspecification in the evenly spaced X situation is reduced in the normal scores case.

Part of the problem with the kernel method in the normal scores situation is that a global bandwidth kernel method has been used for these simulations. (See the discussion in Section II.3.A.) That is, the same bandwidth that is used to fit the close observations near the center of the interval (1,10) is also used to obtain the fit near the endpoints, where the data is more sparse. It would be better, perhaps, to use a density-adjusted variable bandwidth kernel (Section II.3.A), which employs smaller bandwidths where the data locations are more dense and larger bandwidths where the points are more spread apart.

Another part of the explanation for the poor performance of the the kernel method as indicated in Table IV.2.10, is that evaluation of prediction performance based on the measure RPE(LS,K) is not entirely fair to the kernel method. Recall that the relative prediction efficiency RPE is defined as the ratio of the mean SSEP's for the methods being compared. In computing these SSEP's, the endpoints are excluded from the summation, since their inclusion would be unfair to the kernel method. This was because of the bias at the endpoints for predictions obtained with the simple version of kernel being used in these simulations. (These issues were discussed earlier in Section III.2.) In the normal scores case, the least squares prediction at the endpoints are quite

poor when the user's model is incorrect. Therefore, the preceding analysis with the endpoint predictions deleted from the evaluation portrays the least squares method in a more favorable light than it truly deserves. To see this, consider Table IV.2.11, where the endpoint predictions have now been incorporated into the RPE computations. The values of RPE(LS,K) are now somewhat lower at all three levels of the power of the lack of fit test, indicating that the performance of kernel relative to least squares is not as bad for the normal scores case as was suggested previously in Table IV.2.10. These lower values for RPE(LS,K) in Table IV.2.11 occur despite the fact that the kernel is biased at the extreme X locations. If the simple kernel method used here was replaced with a kernel procedure that corrects for endpoint bias, such as in Rice (1984a) or Gasser and Müller (1984), the relative prediction efficiency of kernel compared to least squares would be even better than Table IV.2.11 indicates.

Table IV.2.12 shows the relative prediction efficiencies which result when twenty different X locations based on the normal scores are used. The results here are comparable to those obtained for preceding set of ten normal scores locations. That is, the measure RPE, in its original form, indicates that the kernel and HATLINK methods do not fare as well relative to least squares as they did in the evenly spaced regressor case. However, when RPE is revised to take into account predictions at the endpoints of the regressor region, the results for the kernel and HATLINK methods are improved.

Different sample size

Since the kernel method makes predictions on the basis of local information, this procedure will generally perform better for larger sample sizes. The impact of increasing the sample size is now investigated by considering prediction results for data sets of size $n = 40$. Simulations here were based on quadratic underlying functions with quadratic coefficients $Q = 0$, $Q = .15$, and $Q = .25$. There were ten evenly spaced regressor locations from $X = 1$ to $X = 10$, with four obser-

Table IV.2.10. Comparison of Relative Prediction Performances (RPE) for the Quadratic Family of Models for the Evenly Spaced (ES) and Normal Scores (NS) X Locations. Results are for 25 runs at each level of the quadratic coefficient Q considered. The Cp3 criterion was used to select the value of λ for HATLINK.

Q			RPE(LS,K)		RPE(H,K)		RPE(H,LS)	
ES	NS	Approx. Power	ES	NS	ES	NS	ES	NS
.00	.00	.05	2.93	4.01	2.54	3.46	0.86	0.86
.30	.35	.19	0.97	1.56	1.24	1.57	1.28	1.06
.40	.45	.32	0.64	1.14	1.07	1.29	1.66	1.14

Table IV.2.11. Relative Prediction Performance (RPE) for the Quadratic Family of Models with 10 Normal Scores X Locations. A comparison is made between the RPE's computed with the enpoints ("With") and RPE's computed without the endpoints ("W/O"). Results are for 25 runs at each level of the quadratic coefficient Q considered. The Cp3 criterion was used to select the value of λ for HATLINK.

		RPE(LS,K)		RPE(H,K)		RPE(H,LS)	
Q	Approx. Power	W/O	With	W/O	With	W/O	With
.00	.05	4.01	4.07	3.46	3.51	0.86	0.86
.35	.19	1.56	1.02	1.57	1.14	1.06	1.12
.45	.32	1.14	0.70	1.29	0.90	1.14	1.29

Table IV.2.12. Relative Prediction Performance (RPE) for the Quadratic Family of Models with 20 Normal Scores X Locations. A comparison is made between the RPE's computed with the enpoints ("With") and RPE's computed without the endpoints ("W/O"). Results are for 25 runs at each level of the quadratic coefficient Q considered. The Cp3 criterion was used to select the value of λ for HATLINK.

	RPE(LS,K)		RPE(H,K)		RPE(H,LS)	
Q	W/O	With	W/O	With	W/O	With
.00	3.72	3.95	2.78	3.01	0.75	0.76
.35	1.53	1.08	1.47	1.16	0.96	1.08
.45	1.12	0.91	1.23	0.75	1.09	1.77

Table IV.2.13. Comparison of Prediction Performances for the n = 20 and the n = 40 Situations for the Quadratic Family of Regressions. For both sample sizes, the X locations are evenly spaced, with two observations at each location. Results are based on 25 runs at each level of the quadratic coefficient Q for the basic quadratic family of regressions. Results for n = 20 and n = 40 are paired according to levels Q which result in approximately equal power for the usual lack of fit test. The results for HATLINK are based on the Cp3 criterion. Mean SSEP's are divided by two for the n = 40 case to allow for proper comparison of prediction for the two sample sizes.

Case			Mean SSEP			Relative Prediction Efficiency		
n	Q	Power	LS	K	H	RPE(LS,K)	RPE(H,K)	RPE(H,LS)
40	.00	.050	7.4	21.0	8.6	2.85	2.46	0.86
20	.00	.050	20.8	60.9	24.0	2.93	2.54	0.86
40	.15	.147	18.6	21.4	16.0	1.15	1.34	1.16
20	.25	.144	52.7	63.7	47.0	1.21	1.35	1.12
40	.20	.246	27.0	21.7	19.8	0.80	1.10	1.36
20	.35	.258	82.2	64.3	56.5	0.78	1.14	1.45

vations taken at each location. Again, the results are based on 25 simulations at each level of Q, each time with the same 25 sets of random errors in that were used previously.

Table IV.2.13 shows the mean SSEP's and the relative prediction efficiencies for the kernel, least squares, and HATLINK methods. Results from the original n = 20 family of regressions are included for comparison. The results for n = 40 at level Q = .15 should be compared to those for the n = 20 case at Q = .25, since the degree of model misspecification as measured by the power of the lack of fit test is about the same for these two situations. Similarly, the Q = .20 level for n = 40 is properly matched with the Q = .35 level for n = 20. In order to compare the values of the mean SSEP for the n = 20 and n = 40 cases, the values the mean SSEP's listed in Table IV.2.13 for the n = 40 case have been divided by two. This was done to achieve a fair comparison, since the SSEP measure is a sum of twice as many prediction errors when n = 40.

As expected, the mean SSEP's for all methods decreased substantially when n was increased to 40. However, the relative prediction efficiencies changed very little. A noteworthy result is that the relative efficiency of kernel compared to least squares at level Q = 0 is nearly the same for both sample sizes. Moreover, if RPE(LS,K) is viewed as a function of the power of the F test for lack of fit, then there is little difference between the n = 20 and n = 40 cases. The prediction performance of HATLINK relative to least squares and kernel regression follows the same pattern that has been observed several times previously as the degree of model misspecification ranges from zero to a moderate level.

IV.3. Sinusoidal Underlying Models in a Single Regressor

So far in this chapter, the simulation analysis has focused on one type of functional departure from one type of user's model, namely, quadratic departures from a straight line model. In order to demonstrate that the HATLINK method succeeds in general, other types of departures and user's models must be considered. Toward this end, simulation results for a second class of

underlying functions are presented in this section. Also included are some cases where the user's model is quadratic.

IV.3.A. Basic sine wave family of models.

As was done in Section IV.2, detailed consideration is first given here to the behavior of HATLINK and its associated statistics for one basic family of underlying models. Much of the simulation analysis done in Section IV.2.A for the quadratic family is repeated for the sine wave family of regressions. Additionally, the lack of fit diagnostics developed in Section III.4.D are reviewed in the present section.

The description of the basic sine wave family is as follows. Twenty observations, 2 each at 10 evenly spaced locations from X = 1 to X = 10, were generated using the underlying model,

$$Y = A \sin\left[\frac{\pi(X-1)}{4.5}\right] + 5X + \varepsilon ,$$

with $\varepsilon \sim N(0,16)$. The argument of the sine function is such that the sine function completes one full period over the interval $[1, 10]$. As before, it is assumed that the user has specified a simple linear regression model. A family of departures from the user's model is formed by allowing the amplitude A to range over the values A = 0, 2.5, 3.5, 4.0, 5.0, and 6.0. Two hundred simulations were performed at A = 0.0, one hundred at levels A = 3.5 and A = 5.0, and fifty at the remaining levels of the amplitude A. Again, the same initial seed was used to generate the sets of random errors at each of the levels of A.

The set of amplitudes considered here represent differing degrees of departure from the user's model, ranging from where the user's model is correct (A = 0), to where the misspecification is fairly substantial (A = 6.0). The power of the F test for lack of fit at each of the levels of A is shown in Table IV.3.1. The middle values of A represent situations where the model misspecification is very

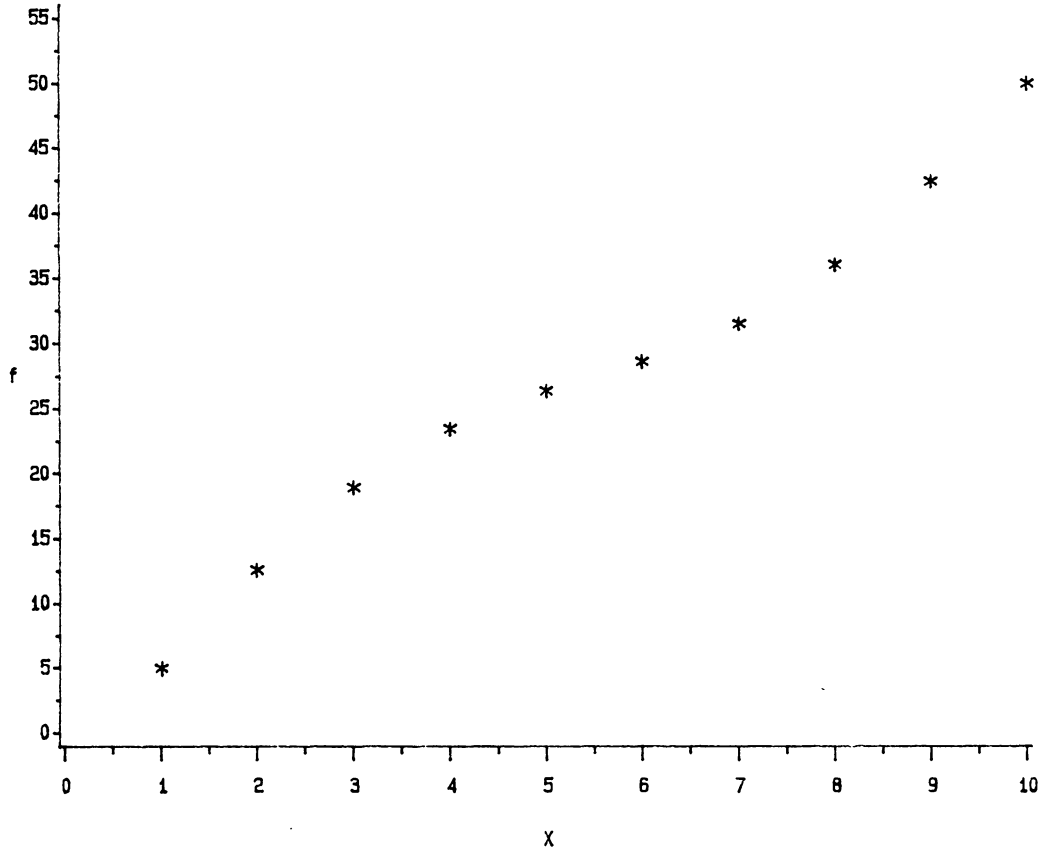


Figure IV.3.1. Plot of True Values of f(X) for the Sine Wave Model with Amplitude A=4.0.

likely to go undetected by the user. Figure IV.3.1 confirms for the A = 4.0 case that the true model departs only slightly from a straight line.

Table IV.3.1. Power (P) of the lack of fit test at α = .05 for the six levels of the amplitude (A) used in the sine wave simulations.

A	0.0	2.5	3.5	4.0	5.0	6.0
P	.05	.09	.14	.18	.26	.38

Prediction results

Table IV.3.2 summarizes the prediction performance of the least squares, kernel, and HATLINK regression methods for the basic sine wave family of models. The relative prediction efficiencies for the three methods are included in Table IV.3.2 and plotted as a function of the power of the lack of fit test in Figure IV.3.2. The values for HATLINK in this table and plot are again based on the Cp3 criterion for selecting λ . (Section III.3.B.) The same pattern that was observed previously for the relative performance of HATLINK is seen here for the sine wave models. That is, the HATLINK method provides good predictions across the range of departures from the user's model. In the case where the user has somehow specified the model perfectly (A = 0), the HATLINK predictions are almost as good, on the average, as those obtained through least squares. Once the amplitude reaches 2.5, where the power of the lack of fit test is only P = .09, the HATLINK predictions based on the PRESS* or Cp3 criterion become superior to the least squares predictions. In situations where the user's model is correct or nearly correct, the HATLINK predictions are far better than those obtained through the kernel method.

Tables IV.3.3 and IV.3.4 provide a comparison of the six methods for selecting the mixing parameter λ for HATLINK. In Table IV.3.3 it is observed that the criteria PRESS* and Cp3 tend

Table IV.3.2. Prediction Performance of the Least Squares, Kernel, and HATLINK methods for the Basic Sine Wave Family of Regressions. Results are based on 200 simulations at amplitude A = 0.0, 100 runs at A = 3.5 and A = 5.0, and 50 simulations at each of the other levels of A. Results shown for HATLINK are for the Cp3 version.

		Mean SSEP				Relative Prediction Efficiency		
A	POWER	LS	KER	H		RPE(LS,K)	RPE(H,K)	RPE(H,LS)
0.0	.05	19.6	56.4	22.9	(+K,-L)	2.87	2.46	0.86
2.5	.09	40.6	53.4	36.7	(+K)	1.32	1.46	1.11
3.5	.14	64.2	59.0	50.7	(+K,+L)	0.92	1.16	1.26
4.0	.18	73.8	55.8	51.0	(+L)	0.76	1.09	1.45
5.0	.26	109.2	64.8	66.5	(+L)	0.59	0.97	1.64
6.0	.38	142.4	62.3	66.8	(+L)	0.44	0.93	2.13

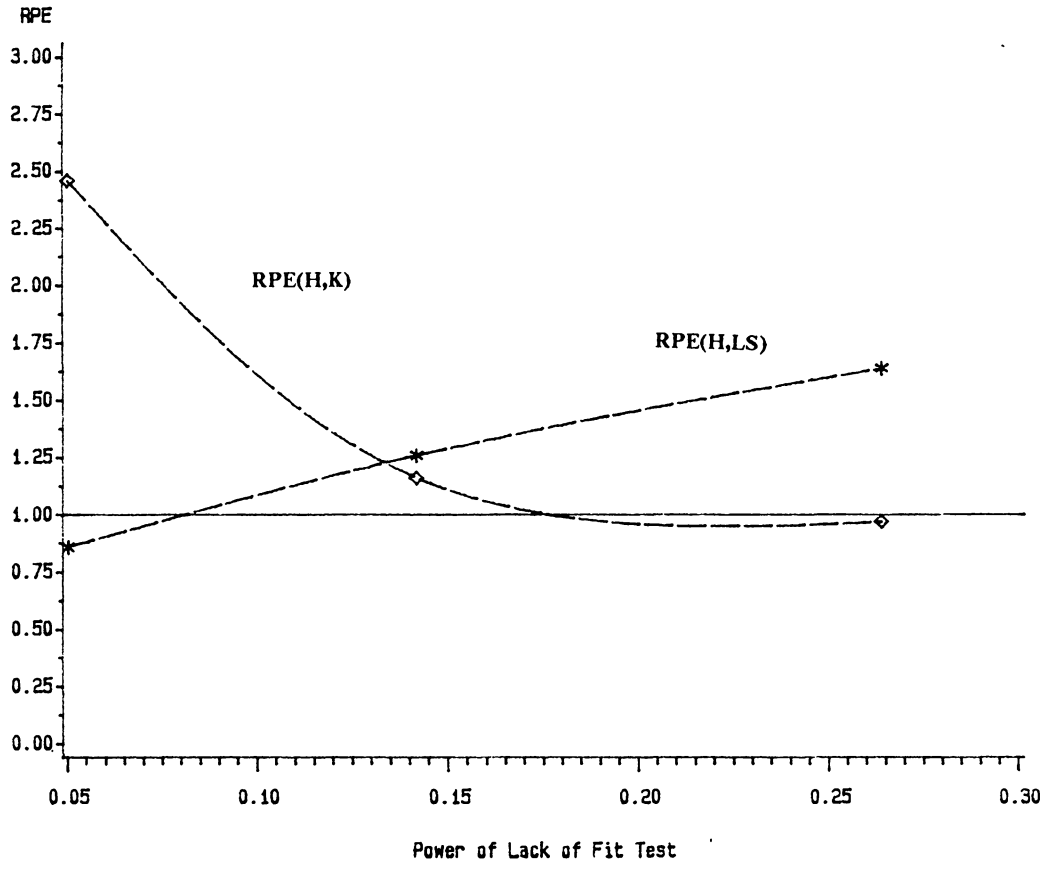


FIGURE IV.3.2. Relative Prediction Efficiencies (RPE).
Results are for the basic sine wave family of regressions.

Table IV.3.3. Mean Values of the Mixing Parameter λ for the Six Selection Criteria. Results are for the basic sine wave family of regressions.

A	POWER	PRESS	PRESS'	Cp1	Cp2	Cp3	Cp4	OPT
0.0	.05	.37	.14	.26	.65	.14	.41	.13
2.5	.09	.56	.32	.45	.56	.31	.61	.46
3.5	.14	.72	.41	.59	.75	.34	.78	.56
4.0	.17	.77	.52	.68	.77	.48	.81	.64
5.0	.26	.87	.61	.74	.85	.51	.89	.71
6.0	.38	.94	.77	.88	.92	.64	.95	.81

Table IV.3.4. Prediction Performance of the HATLINK Method for the Six Criterion for Selecting the Mixing Parameter. Results are for the basic sine wave family of regressions.

		Mean Sum of Squared Errors of Prediction									
A	POWER	LS	KER	PRESS	PRESS'	Cp1	Cp2	Cp3	Cp4	OPT	lsd
0.0	.05	19.6	56.4	36.0	24.5	30.6	40.6	22.9	38.6	18.1	3.2
2.5	.09	40.6	53.4	43.5	37.7	41.1	44.0	36.7	45.0	26.9	5.5
3.5	.14	64.2	59.0	55.1	52.5	54.1	55.9	50.7	56.3	37.9	4.3
4.0	.17	73.8	55.8	53.3	52.9	52.8	53.1	51.0	54.0	37.6	6.2
5.0	.26	109.2	64.8	64.3	67.5	65.3	64.1	66.5	64.9	49.9	5.0
6.0	.38	142.4	62.3	60.6	64.9	61.5	59.0	66.8	61.1	50.9	7.7

Table IV.3.5. Mean and Standard Deviations of the Estimates of Error Variance Obtained by the Least Squares and Kernel Methods and the Cp3 and Cp4 Versions of HATLINK. Also included are the mean degrees of freedom associated with these regression methods. Results are for 200 runs at amplitude A = 0.0, 100 runs at A = 3.5 and A = 5.0, and 50 runs at each of the other levels of A for the basic sine wave family of regressions. The true value of σ^2 is 16.0. For reference, the standard error of the mean s^2 ranges from 0.41 (for least squares at A = 0.0) to 1.15 (for least squares at A = 6.0).

	Least Squares			Kernel			HATLINK -- Cp3			HATLINK -- Cp4		
A	mean s^2	s.d. s^2	mean df	mean s^2	s.d. s^2	mean df	mean s^2	s.d. s^2	mean df	mean s^2	s.d. s^2	mean df
0.0	16.5	5.8	2	15.9	6.9	4.9	16.5	6.4	2.4	15.5	6.6	3.5
2.5	18.2	6.2	2	15.4	7.5	4.7	16.4	6.9	2.7	15.2	7.2	3.9
3.5	20.0	7.0	2	15.7	7.1	4.8	17.1	6.9	2.9	15.6	7.0	4.1
4.0	21.2	6.8	2	15.6	7.5	4.8	17.2	7.0	3.1	15.5	7.4	4.4
5.0	23.2	7.5	2	15.8	7.1	4.9	18.0	7.1	3.3	15.8	7.0	4.6
6.0	27.2	8.1	2	15.8	7.6	5.1	18.3	7.3	3.7	15.8	7.5	5.0

to produce smaller λ's than the other four methods. Consequently, these two conservative methods for selecting λ lead to superior HATLINK prediction when the model misspecification is slight, as may be seen in Table IV.3.4.

Estimation of error variance

The estimates of σ_ε^2 obtained through the least squares and kernel methods, and by the HATLINK procedure using the Cp3 and Cp4 λ selection methods were computed for each simulation at every level of the amplitude A. The means and standard deviations of these estimates are shown in Table IV.3.5, along with the corresponding mean model degrees of freedom. The results shown here for the sine wave family of models are similar to those observed previously in Table IV.2.5 for the quadratic models. When the user's specified model is correct (A = 0), all methods provide accurate estimation of σ_ε^2. The least squares estimate, however, is more precise than the others in the A = 0 case. At A = 3.5, where the user's model is just slightly off, the least squares estimate of variance is biased upward, with a mean value of 20.0. (The standard error of this mean value of s_{ols}^2 based on 100 simulations is 0.70.) As the degree of model misspecification becomes greater, the upward bias in s_{ols}^2 increases. Also, the standard deviation of this estimate rises steadily as the level of A is increased. The mean and standard deviation of kernel estimate of σ_ε^2 remain nearly constant over the range of amplitudes considered. At each level of A the kernel method tends to slightly underestimate the true value of $\sigma_\varepsilon^2 = 16$. The variance estimate associated with the Cp4 version of HATLINK shows very similar behavior to the kernel variance estimate. The variance estimate obtained through the Cp3 version of HATLINK tends to increase as A increases, but not to anywhere near the extent that this occurs for the least squares estimate. The standard deviation of the Cp3 estimates is somewhat lower than for the kernel and Cp4 variance estimates at every level of A. Summarizing the results of this table, it is apparent that the kernel estimate or either HATLINK estimate of σ_ε^2 should be favored over the least squares estimate whenever there is some doubt that the user's model has been correctly specified.

Table IV.3.6. Mean 95% Confidence Interval Widths at Location X = 6 and Percent Coverage of f(6). Intervals are formed by the least squares and kernel methods and by the Cp3 and Cp4 Versions of HATLINK. Results are for 200 simulations at amplitude A = 0.0, and 100 simulations at amplitudes A = 3.5 and A = 5.0 for the basic sine wave family of regressions. The estimated standard error of these percentage estimates ranges from 1.5 (for all methods at Q = .00) to 3.8 (for the least squares intervals at A = 5.0).

	Least Squares		Kernel		HATLINK -- Cp3		HATLINK -- Cp4	
A	mean width	% cov.	mean width	% cov.	mean width	% cov.	mean width	% cov.
0.0	3.8	94.5	7.7	97.5	4.5	94	5.6	94.5
3.5	4.2	87	7.6	96	5.5	93	6.6	93
5.0	4.6	82	7.8	95	6.3	94	7.3	94

Table IV.3.7. Mean 95% Confidence Interval Widths at Location X = 8 and Percent Coverage of f(8). Intervals are formed by the least squares and kernel methods and by the Cp3 and Cp4 Versions of HATLINK. Results are for 200 simulations at amplitude A = 0.0, and 100 simulations at amplitudes A = 3.5 and A = 5.0 for the basic sine wave family of regressions. The estimated standard error of these percentage estimates ranges from 1.5 (for all methods at Q = .00) to 4.9 (for the least squares intervals at A = 5.0).

	Least Squares		Kernel		HATLINK -- Cp3		HATLINK -- Cp4	
A	mean width	% cov.	mean width	% cov.	mean width	% cov.	mean width	% cov.
0.0	5.0	95.5	7.7	95	5.4	96.5	6.2	94.5
3.5	5.5	64	7.6	94	6.1	82	6.9	88
5.0	6.0	47	7.8	92	6.8	82	7.5	86

Confidence intervals

For each simulation run, confidence intervals for the true value of the underlying function at locations X = 6 and X = 8 were computed using the least squares and kernel methods, and for the Cp3 and Cp4 versions of HATLINK. Table IV.3.6 shows the percentage of times out of the 100 runs done at each level of A (200 simulations at level A = 0) that the confidence intervals formed by these four methods actually covered f(6). Similar results for f(8) are shown in Table IV.3.7. Location X = 6 is a point where the sine wave is near the middle of its period, so that the user's straight line least squares fit will pass close to the true value of f at this location. Location X = 8 is just beyond the point where the true function makes its maximum departure from the straight line. Therefore, predictions based on the least squares fit to a straight line model should be biased more at location X = 8 than at X = 6 when $A > 0$. Consequently, when the amplitude A of the sine wave is increased, the least squares confidence intervals should suffer more at X = 8 than at X = 6.

As was observed in Section IV.2.A for the quadratic models, Tables IV.3.6 and IV.3.7 demonstrate that percent coverage achieved by the least squares intervals drops off as the degree of model misspecification is increased. This is particularly true at X = 8 for the sine wave models, where the coverage of f(8) has decreased to 64% at A = 3.5, where the power of the lack of fit test is only P = .14. As was observed earlier for the quadratic family of models, the intervals formed using the kernel method maintain their coverage of f across the range of departures from a straight line. The Cp4 intervals perform nearly as well as those obtained by the kernel method, but the intervals formed using the Cp3 version of HATLINK provide somewhat diminished coverage of f(8).

Tables IV.3.6 and IV.3.7 display the mean widths of the confidence intervals formed at X = 6 and X = 8 using the least squares, kernel, and two HATLINK methods. As was foreseen in Section III.4.B, the intervals formed by the least squares method tend to be much narrower than the kernel intervals. The mean interval widths for two HATLINK methods are greater than for least squares

Table IV.3.8. Mean Bias of Predictions at Location X = 8 for the Least Squares and Kernel Methods, and for the Cp3 and Cp4 Versions of HATLINK. The results are based on 200 simulations at amplitude A = 0.0 and 50 simulations at A = 3.5 and A = 5.0 for the basic sine wave family. The standard error of each mean is included for each case.

	Mean Bias of Predictions (standard error)			
A	LS	KER	Cp3	Cp4
0.0	0.0 (.08)	0.1 (.13)	0.0 (.08)	0.0 (.11)
3.5	2.1 (.17)	0.2 (.26)	1.2 (.20)	0.6 (.26)
5.0	3.1 (.17)	0.4 (.27)	1.6 (.21)	0.6 (.27)

and less than for the kernel method. Apparently, the penalty required for trying to estimate any sort of underlying function with a confidence interval is a wider interval.

The loss in coverage observed for the least squares intervals is not solely attributable to the fact that these intervals are narrower, however. A major factor in this loss of coverage is the bias in the least squares estimate of f(X). Table IV.3.8 shows the mean bias in $\hat{Y}(8)$ at several levels of A for the four regression methods. (The bias in $\hat{Y}(8)$ is defined as $|\hat{Y}(8) - f(8)|$.) It is clear from this table that at least part of the reason for the improved coverage of f(8) when $A > 0$ by the kernel interval is that the kernel estimate of f(8) is more accurate.

Diagnosing lack of fit

The simplest method related to HATLINK for diagnosing when the user's model is incorrect is to consider the value of λ obtained by one of the six selection criteria. This approach was discussed in Section III.4.D, and is now investigated for the sine wave family of regressions. Recall that this series of regressions was based on 20 observations, two each at 10 evenly spaced regressor locations. First, $\alpha = .05$ and $\alpha = .10$ critical values were determined empirically from 200 runs at the null hypothesis that the user's model is correct (A = 0). Critical values are shown in Table IV.3.9 for the six criteria for selecting λ . The theoretical and empirical critical values for the F test for lack of fit are also shown in this table.

Table IV.3.10 shows the empirical powers of the tests based on the PRESS* , Cp1, and Cp3 criteria. The table shows the percentage of times out of the 50 or 100 runs at each level of A that the selected λ exceeded the corresponding empirical critical value. For purposes of comparison, the empirical and theoretical powers of the usual F test for lack of fit are also displayed. The results indicate that when the departure of the true model from the user's model are of the sine wave type, the λ method for diagnosing lack of fit is more sensitive to lack of fit than is the usual lack of fit test. This advantage is particularly true for the PRESS* criterion for selecting λ . For this method, the empirical power at $\alpha = .05$ is at least twice that achieved by the usual lack of fit test at every

Table IV.3.9. 95th and 90th percentiles of the distribution of the values of λ observed for the 200 runs under the null hypothesis A = 0.0. Results are for the PRESS*, Cp1, and Cp3 criteria for selecting λ. Also included are the 95th and 90th percentiles of the empirical distribution of the usual F statistic for testing lack of fit, plus the corresponding theoretical percentage points.

	F (theor.)	F (empir.)	PRESS*	Cp1	Cp3
95th percentile	3.07	3.272	.6246	.8445	.5544
90th percentile	2.38	2.353	.5443	.7887	.4779

Table IV.3.10. Proportion of rejections for the lack of fit test based on λ for 100 simulations at amplitudes A = 3.5 and A = 5.0, and 50 runs at each of the other levels of the amplitude A for the basic sine wave family. Critical values at α = .05 and α = .10 were determined empirically through 200 runs under the null hypothesis (A = 0.0) and are shown in Table IV.3.9. Results are presented for the PRESS*, Cp1, and Cp3 criteria for selecting λ. The theoretical and empirical power of the F test for lack of fit are also shown.

		Power of F Test		Empirical Power of λ Tests		
A	α	Theor.	Empir.	PRESS*	Cp1	Cp3
2.5	.05	.09	.06	.20	.16	.20
	.10	.17	.20	.28	.18	.26
3.5	.05	.14	.09	.30	.26	.39
	.10	.24	.27	.40	.33	.46
4.0	.05	.18	.14	.46	.38	.42
	.10	.29	.30	.52	.46	.52
5.0	.05	.26	.26	.57	.44	.55
	.10	.41	.44	.64	.56	.65
6.0	.05	.38	.36	.72	.70	.62
	.10	.54	.52	.82	.76	.65

level of A considered here. Unfortunately, the test based on λ has the drawback that the appropriate critical value must be estimated through simulation for each situation. Further, the comparatively high power achieved in this empirical study depends to some extent on the particular alternative form of the underlying model. However, the λ test has performed well for every other type of alternative model considered in these investigations. Plus, the test has the benefit that it does not require replicated observations.

IV.3.B. A second sinusoidal family of regressions

Another family of regressions will now be considered for which the user has specified the quadratic model, $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$. Departures from this model will stem from the underlying function,

$$f(X) = 2(X - 5.5)^2 + 5X + A \sin\left[\frac{\pi(X - 1)}{2.25}\right] .$$

The argument of the sine function above causes the sine to complete two full periods over the interval $[1, 10]$. As before, the errors are generated from the N(0,16) distribution, and two observations are taken at each of ten evenly spaced X locations. The true values of the underlying function at the ten regressor locations are plotted in Figure IV.3.3 for the case where the amplitude A is 3.5. A regression user would not likely detect that the quadratic model is incorrect from a scatterplot of data generated from this function plus and error term. Neither would the usual lack of fit test generally detect that the model is incorrect in this case, since the power at $\alpha = .05$ is only $P = .226$.

Table IV.3.11 and Figure IV.3.4 show the relative prediction efficiencies of the least squares and kernel methods and the Cp3 version of HATLINK. The pattern of the relative prediction efficiencies for the quadratic plus sine wave family of regressions is very similar to what was observed previously. Table IV.3.12 presents the prediction results for the least squares and kernel methods,

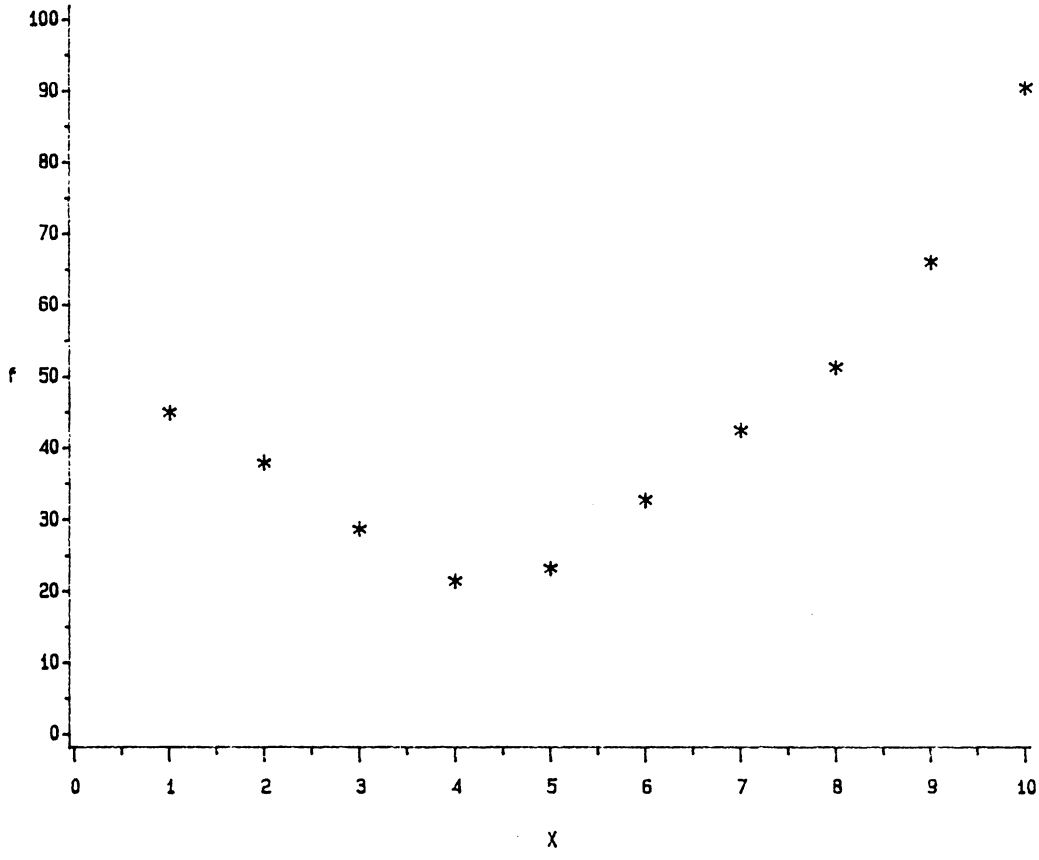


Figure IV.3.3. Plot of True Values of f(X) for the Quadratic Plus Sine Wave Model with Amplitude A=3.5.

Table IV.3.11. Prediction Performance of the Least Squares, Kernel, and HATLINK methods for the Quadratic Plus Sine Wave Family of Regressions. Results are based on 25 simulations at each level of the amplitude A considered. Results shown for HATLINK are for the Cp3 version.

		Mean SSEP				Relative Prediction Efficiency		
A	POWER	LS	KER	H		RPE(LS,K)	RPE(H,K)	RPE(H,LS)
0.0	.05	29.9	85.1	31.4	(+K)	2.87	2.71	0.95
2.0	.10	61.0	87.6	56.6	(+K)	1.32	1.55	1.08
3.0	.17	100.5	92.9	83.9	(+K,+L)	0.92	1.11	1.20
3.5	.23	126.2	95.7	98.7	(+L)	0.76	0.97	1.28
4.0	.29	155.8	98.8	112.6	(-K,+L)	0.59	0.88	1.38

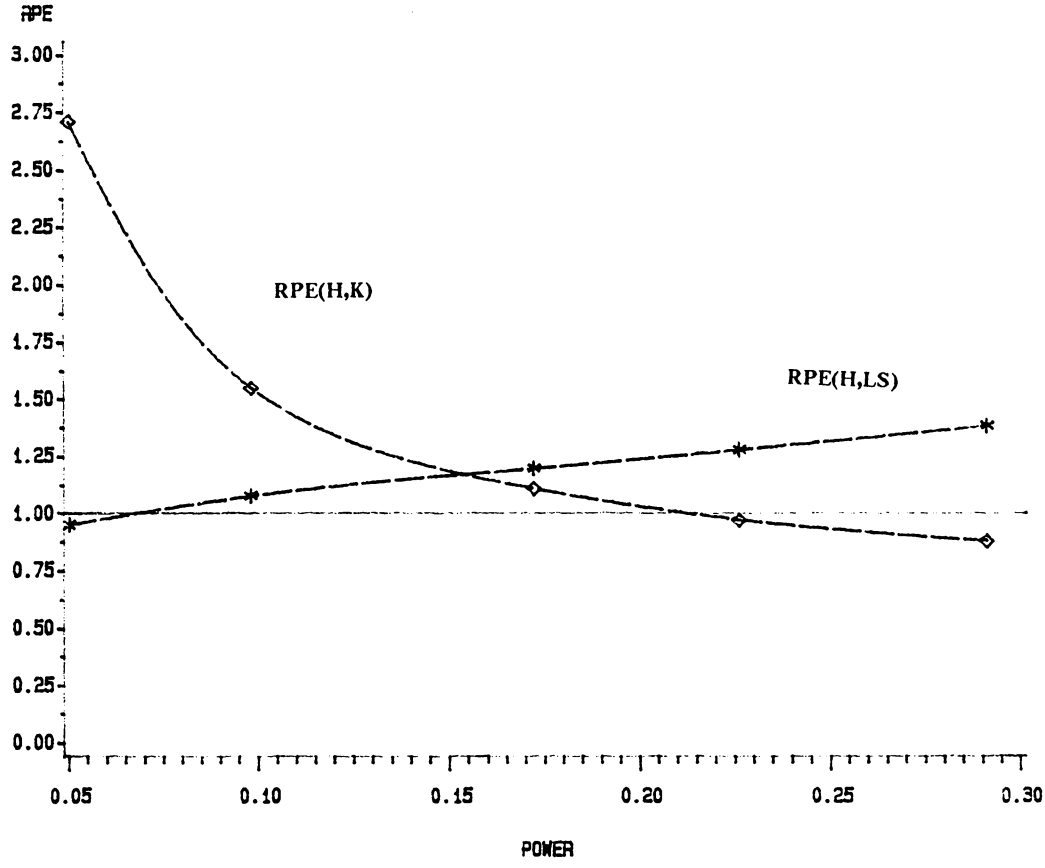


FIGURE IV.3.4. Relative Prediction Efficiencies (RPE).
Results are for the quadratic plus sine wave true mmodels.

Table IV.3.12. Prediction Performance of the HATLINK Method for the Six Criterion for Selecting the Mixing Parameter. Results are for the regressions where the underlying function is a quadratic plus a sine wave. 25 runs were used at each level of the amplitude A.

		Mean Sum of Squared Errors of Prediction									
A	POWER	LS	KER	PRESS	PRESS*	Cp1	Cp2	Cp3	Cp4	OPT	lsd
0.0	.05	29.9	85.1	38.3	31.4	36.2	72.0	31.4	44.5	25.9	10.6
2.0	.10	61.0	87.6	59.9	57.6	58.8	82.6	56.6	64.6	49.9	8.6
3.0	.17	100.5	92.9	82.1	85.9	81.7	83.5	83.9	83.4	71.4	8.3
3.5	.23	126.2	95.7	92.0	100.2	92.8	93.3	98.7	91.3	80.8	8.9
4.0	.29	155.8	98.8	99.5	113.8	100.6	108.0	112.6	97.7	88.4	10.3

and the six versions of HATLINK. As before, the "OPT" column demonstrates what would be achieved through HATLINK if λ could be selected to minimize SSEP. Again, the kernel predictions are relatively poor when the user's model is correct or nearly correct, and the least squares predictions are poor when the model is off by more than a small amouunt. In particular, four of the HATLINK versions outperform least squares at A = 2, where the lack of fit test has power of just P = .098. The Cp3 and PRESS* criteria for selecting λ result in the lowest mean SSEP's among the six criteria when the amplitude is small. When the user's model has been misspecified by a more moderate amount (A = 4.0), Cp4 and PRESS are the best criteria for making predictions through HATLINK. Perhaps predictions obtained by the Cp4 and PRESS versions of HATLINK could be improved at lower amplitudes in this case through the use of a kernel with a predetermined number of degrees of freedom, say df = 7. (See the the results for the nonstochastic kernel aprroach in Appendix A.)

Confidence intervals and estimation of error variance

Estimates of σ_ε^2 obtained through the least squares and kernel methods, and through the Cp3 and Cp4 versions of the HATLINK procedure, are shown in Table IV.3.13. The pattern observed here is very much similar to the pattern seen previously in Tables IV.2.5 and IV.3.5. Again, the least squares estimate becomes increasingly biased upward as the degree of model misspecification increases. The kernel estimate of variance remains fairly constant over the range of levels of A, while the variance estimates obtained through the two HATLINK methods increase to a lesser extent than the least squares estimate as A increases.

Table IV.3.14 presents the percentage of times the various types of 95% confidence intervals actually contain the true value of the underlying function f at X = 6. The mean widths of these intervals are also shown. The location X = 6 is a place where the sine wave term is equal to .64A, whereas the maximum departure from the user's model is 1.0A. Thus, predictions at X = 6 by the least squares method based on the user's quadratic model will not be as biased as they are at many

Table IV.3.13. Mean and Standard Deviations of the Estimates of Error Variance Obtained by the Least Squares and Kernel Methods and the Cp3 and Cp4 Versions of HATLINK. Also included are the mean degrees of freedom associated with these regression methods. Results are for 25 runs at each of the levels of amplitude A for the quadratic plus sine wave family of regressions. The true value of σ^2 is 16.0. The standard error of the mean s^2 values ranges from 1.4 (for least squares at A = 0.0) to 1.9 (for Cp3 at A = 4.0).

	Least Squares			Kernel			HATLINK -- Cp3			HATLINK -- Cp4		
A	mean s^2	s.d. s^2	mean df	mean s^2	s.d. s^2	mean df	mean s^2	s.d. s^2	mean df	mean s^2	s.d. s^2	mean df
0.0	17.4	7.2	3	18.0	9.0	6.2	17.7	8.1	3.3	16.6	8.1	4.3
2.0	19.1	7.2	3	18.3	9.1	6.4	19.3	8.6	3.5	17.6	8.6	4.8
3.0	21.4	7.6	3	18.6	9.3	6.6	21.0	9.1	3.8	18.4	9.2	5.5
3.5	23.0	7.9	3	18.8	9.4	6.8	21.9	9.4	4.0	18.7	9.4	5.9
4.0	24.7	8.2	3	19.0	9.5	6.9	22.9	9.7	4.3	19.0	9.6	6.3

Table IV.3.14. Mean 95% Confidence Interval Widths at Location X = 6 and Percent Coverage of f(6). Intervals are formed by the least squares and kernel methods and by the Cp3 and Cp4 Versions of HATLINK. Results are for 25 runs at each of the levels of amplitude A for the quadratic plus sine wave family of regressions. The estimated standard errors of the percentage estimates range from 4.4 (for all methods at A = 0.0) to 9.2 (for least squares at A = 4.0).

	Least Squares		Kernel		HATLINK -- Cp3		HATLINK -- Cp4	
A	mean width	% cov.	mean width	% cov.	mean width	% cov.	mean width	% cov.
0.0	5.2	92	8.3	100	5.4	92	6.0	92
2.0	6.1	92	9.2	96	6.4	92	7.4	100
3.0	6.4	80	10.3	96	7.5	96	8.8	96
3.5	6.7	76	10.5	96	7.9	96	9.4	96
4.0	6.9	68	10.7	100	8.3	96	10.0	100

other regressor locations. Consequently, coverage of f(6) by the least squares intervals should not suffer as much as was observed earlier for other situations as the degree of model misspecification increases. Still, Table IV.3.13 does show some loss of coverage of f(6) with the least squares intervals At A = 3.5, for example, the 95% least squares intervals contained f(6) in only 76 percent (18 out of 25) of the simulations. The kernel and two HATLINK methods for forming confidence intervals maintained their coverage of f(6) across all levels of A, but at the expense of greater average widths.

IV.4. Underlying Models in Two Regressors

Simulation results are now presented for several situations in which there are two regressor variables. First, consideration will be given in Sections IV.4.A and IV.4.B to two families of regressions where the user has specified the first order model, $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$. Then, in Section IV.4.C, a family of regressions is considered for which the user has specified a full second order model, $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2 + \beta_{11} X_1^2 + \beta_{22} X_2^2 + \varepsilon$. Finally, the problem of variable selection is addressed in Section IV.4.D.

IV.4.A. True model is first order plus a quadratic term

Simulations were carried out as follows for a family of regressions involving two X variables. In each case, it was assumed that the user had specified the first order model, $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$, while the true underlying function was of the form $f(X_1, X_2) = 5X_1 + 5X_2 + Q(X_1 - 5.5)^2$. A sample of 30 data points was used, with two observations taken at each of fifteen regressor locations. These 15 locations were used for every run, and are plotted in Figure IV.4.1. Errors were generated at random from the N(0,16) distribution.

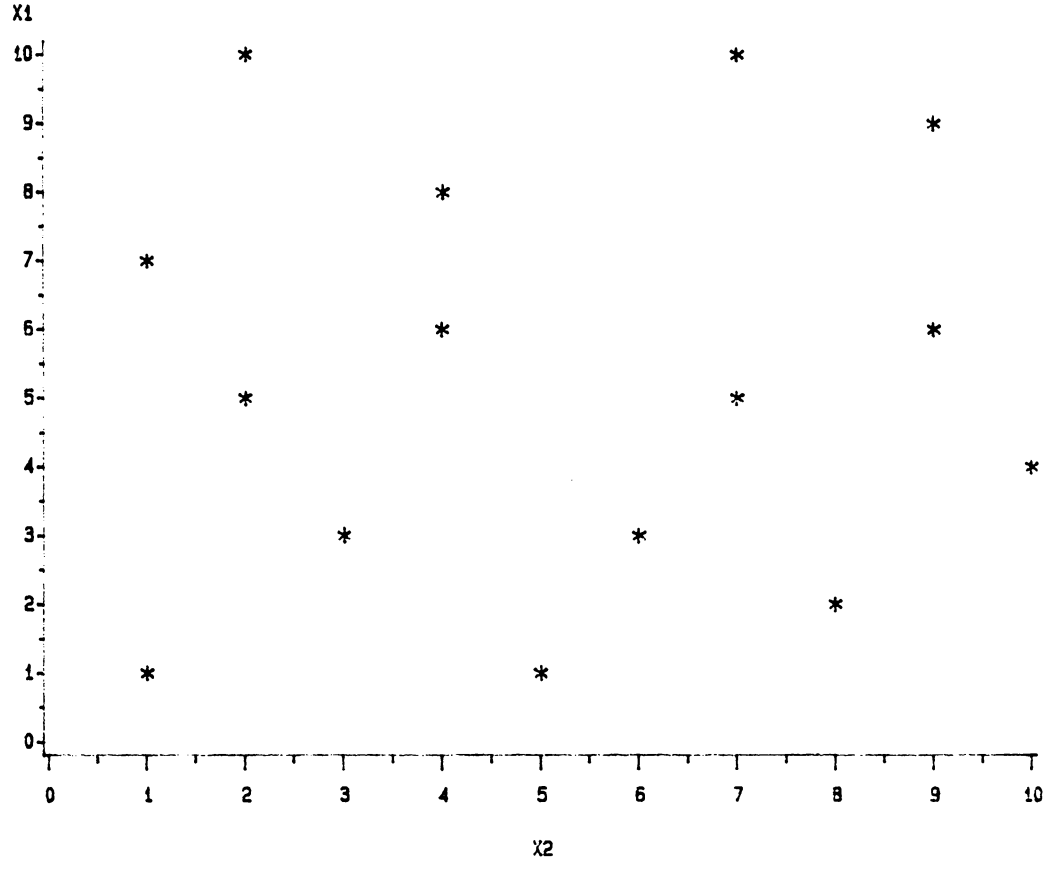


Figure IV.4.1. Plot of the 15 Data Locations for the Two Regressor Examples.

The quadratic coefficient Q was allowed to take on the values Q = 0.00, 0.15, 0.20, 0.25, and 0.30, which cover a range from no lack of fit to a moderate degree of lack of fit for the user's model. Table IV.4.1 shows the power at α = .05 of the F test for lack of fit at each of these levels of Q. One hundred simulations were carried out at levels Q = .00, .20, and .30. Again, the same 100 sets of random errors were used for these three cases, so that the results would be comparable. For levels Q = .15 and Q = .25, only 20 simulations were performed.

Table IV.4.1. Power (P) of the lack of fit test at α = .05 for the five levels of the quadratic coefficient Q used in the first order plus quadratic series of models.

Q	.00	.15	.20	.25	.30
P	.05	.10	.14	.21	.30

Table IV.4.2 shows predictions results for the set of simulations at each level of Q for this family of regressions. In the multiple regression case, the evaluator SSEP is based on all n = 30 observations. (Recall that for the single regressor situation the endpoint observations were deleted in computing SSEP in order to compensate for endpoint bias in the simple version of kernel regression being used in this investigation. In the multiple regression case, deletion of the perimeter points in the multidimensional regressor space would mean the loss of too many observations when the sample size is moderate.) Similarly, the search for λ to minimize any one of the six criteria and the search for h to minimize PRESS*(h) are now based on all the observations, as are the variance estimates formed by the kernel and HATLINK methods.

Table IV.4.2 provides evidence for the usefulness of the HATLINK procedure for making predictions in the multiple regression setting. As was observed for the single regressor case, predictions obtained through the Cp3 version of HATLINK are nearly as good as the least squares predictions when the user's model is perfectly correct, and become much better than the least squares predictions as the degree of model misspecification is increased. The table also reveals the poor performance of the kernel method in the multiple regression setting when the sample size is

Table IV.4.2. Prediction Performance of the Least Squares and Kernel Methods and for Five Versions of the HATLINK Procedure for the Two Regressor Situation. In this series of regressions the user has fit a first order model, but the true model contains an additional quadratic term in one of the variables. Results are based on 100 simulations at levels Q = .00, Q = .20, and Q = .30, and on 20 runs at levels Q = .15 and Q = .25 of the coefficient Q of the quadratic term X_1^2. The relative prediction efficiencies of the least squares, kernel, and the Cp3 version of HATLINK are included below.

		Mean Sum of Squared Errors of Prediction									
Q	Power	LS	Ker	PRESS	PRESS*	Cp1	Cp2	Cp3		OPT	lsd
.00	.05	48.6	226.1	58.6	50.7	62.8	104.0	55.6	(+K)	41.3	10.7
.15	.10	84.0	219.6	80.2	76.3	85.0	107.0	76.4	(+K)	69.5	17.6
.20	.14	122.6	222.1	107.6	107.2	113.0	127.4	104.0	(+K,+L)	86.1	8.7
.25	.21	158.0	218.2	116.5	118.5	121.0	135.2	112.2	(+K,+L)	102.6	17.3
.30	.30	215.1	221.8	145.8	156.5	150.2	157.6	143.4	(+K,+L)	119.3	8.6

Q	Power	RPE(LS,K)	RPE(H,K)	RPE(H,LS)
.00	.05	5.39	4.62	0.86
.15	.10	2.62	2.87	1.10
.20	.14	1.88	2.34	1.25
.25	.21	1.38	1.94	1.40
.30	.30	1.06	1.68	1.58

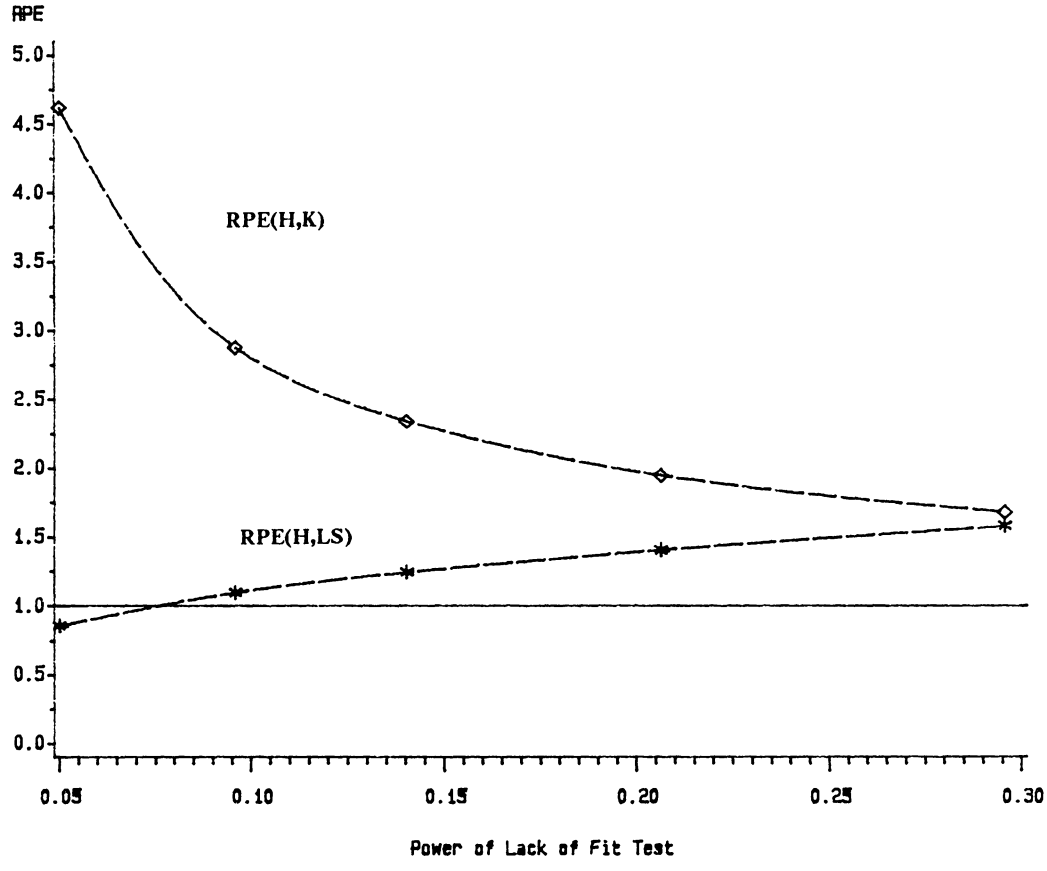


Figure IV.4.2. Relative Prediction Efficiencies (RPE).
Results are for the 1st order plus quadratic regressions.

Table IV.4.3. Mean and Standard Deviations of the Estimates of Error Variance Obtained by the Least Squares and Kernel Methods and the Cp3 and Cp1 Versions of HATLINK. Also included are the mean degrees of freedom associated with these regression methods. Results are for 100 simulations at three levels of the coefficient Q of the X_1^2 term for the first order plus quadratic series of regressions. The true value of σ^2 is 16.0. The standard errors of these mean values of s^2 range from .41 (for Cp3 at Q = .00) to .59 (for least squares at Q = .30).

	Least Squares			Kernel			HATLINK -- Cp3			HATLINK -- Cp1		
Q	mean s^2	s.d. s^2	mean df	mean s^2	s.d. s^2	mean df	mean s^2	s.d. s^2	mean df	mean s^2	s.d. s^2	mean df
.00	16.0	4.2	3	17.7	5.7	10.8	15.1	4.1	3.9	14.9	4.3	4.2
.20	18.6	5.0	3	17.6	5.5	10.6	16.0	4.2	4.9	15.6	4.3	5.4
.30	22.0	5.9	3	17.6	5.5	10.5	17.0	4.4	5.7	16.2	4.6	6.5

not too large. Considering the various criteria for selecting λ for HATLINK, the Cp3 method again appears to lead to the best predictions across the range of differing degrees of model misspecification. Figure IV.4.2 depicts the relative performance of the Cp3 version of HATLINK relative to the least squares and kernel methods.

The PRESS* version of HATLINK performed slightly better than Cp3 and the other HATLINK versions at level $Q = 0$, but did not perform quite as well for $Q \geq 0.20$. For this and other multiple regression simulations, the PRESS* criterion for selecting λ has been the most conservative. That is, it generally leans toward the least squares fit to the user's model by selecting λ'*s* that are typically lower than those selected by the other methods. Note also that results for the Cp4 criterion are not listed in Table IV.4.2. That is because the value of λ that minimizes $Cp4(\lambda)$ also minimizes $Cp1(\lambda)$ in the multiple regression setting. These two criteria differ in the single regressor case because of differences in the way they deal with the endpoint observations.

Variance estimates and confidence intervals

Table IV.4.3 shows the estimates of error variance obtained by the least squares and kernel methods and by the Cp3 and Cp1 (=Cp4) versions of HATLINK. The kernel, Cp3, and Cp1 estimates of σ_ε^2 remain fairly accurate across the range of Q values, while the least squares estimate shows an increasing amount of upward bias as Q increases. Further, the standard deviation of the least squares variance estimate is increasing in Q, whereas this is not the case for the kernel estimates. Note that the standard deviations of the kernel and HATLINK estimates of variance are now competitively small compared to the standard deviation of the least squares estimate, even at level $Q = 0$. This is partly due to the fact that these variance estimates are based on all n of the observations, whereas the endpoints had been excluded in the single regressor case due to bias in the kernel fit at these locations.

Table IV.4.4. Mean 95% Confidence Interval Widths and Percent Coverage at Locations (X_1, X_2) = (8,4), (10,2), and (6,4). Intervals are formed by the least squares and kernel methods and by the Cp3 and Cp4 Versions of HATLINK. Results are for 100 simulations at levels Q = .00, Q = .20, and Q = .30, of the coefficient Q of the quadratic term X_1^2 for the first order plus quadratic series of regressions. The estimated standard error of these percentage estimates ranges from 2.2 (for all methods at all locations when Q = .00) to 5.0 (for the least squares intervals at location (6,4) when Q = .20).

		Least Squares		Kernel		HATLINK -- Cp3		HATLINK -- Cp4	
Location	Q	mean width	% cov.	mean width	% cov.	mean width	% cov.	mean width	% cov.
(8,4)	.00	4.3	98	10.1	87	4.8	97	4.9	97
	.20	4.6	95	10.0	86	5.9	99	6.0	96
	.30	5.0	94	9.9	82	6.7	98	7.0	97
(10,2)	.00	6.8	97	10.6	94	7.1	96	7.2	96
	.20	7.1	82	11.5	96	8.4	89	8.7	90
	.30	7.7	69	12.5	95	9.7	85	10.2	85
(6,4)	.00	3.4	96	8.9	99	4.2	97	4.2	97
	.20	3.5	50	9.5	100	5.6	92	5.9	89
	.30	3.9	19	10.3	100	6.9	89	7.5	86

The percent coverage of the true value of $f(X_1, X_2)$ obtained by 95% confidence intervals using the least squares, kernel, Cp3, and Cp1 methods are shown in Table IV.4.4 for three different data locations, (X_1, X_2) = (8,4), (10,2), and (6,4). The point (6,4) has X_1 coordinate near the place (X_1 = 5.5) where the quadratic term in the true model reaches its greatest magnitude. Thus, it is seen that the least squares intervals based on the user's first order model suffer the most at location (6,4) as Q is increased. At this location the loss in coverage by the least squares intervals is quite dramatic. At the level Q = 0.20, where the lack of fit test at α = .05 has power of only P = .14, the least squares intervals contained f(6,4) in just 50% of the simulations. (The estimated standard error of this percentage estimate is 5.0.%) The kernel, Cp3, and Cp1(= Cp4) intervals provide improved coverage of f(6,4), as well as f(10,2) and f(8,4), across all levels of Q, but again these intervals tend to be somewhat wider. Mean widths of the intervals obtained at these three locations are included in Table IV.4.4.

IV.4.B. True model is first order plus interaction

A simulation analysis is presented for the situation where the user has specified a first order model, while the true model is of form $Y = 5X_1 + 5X_2 + I(X_1 - 5.5)(X_2 - 5.5) + \varepsilon$, where $\varepsilon \sim N(0,16)$. The interaction coefficient I was varied over seven values, ranging from I = 0.00 to I = 0.40. Again, 30 observations were taken, two at each of the same fifteen regressor locations that were used in the preceding section.

Table IV.4.5 shows the mean SSEP's for the various prediction methods at each level of I. The results are very similar to those obtained in Section IV.4.A for the family of quadratic departures from the first order model. The PRESS* criterion is the best version of HATLINK for low values of I, while the Cp3 version is the best when there is a moderate degree of model misspecification (I = 0.20, 0.25, and 0.30). This family of regressions also includes runs at I = .35 and I = .40, where the lack of fit of the first order model is more extreme. For these higher levels of misspecification the PRESS criterion provided the best predictions. Taken as a whole, however,

Table IV.4.5. Prediction Performance of the Least Squares and Kernel Methods and for Five Versions of the HATLINK Procedure for the Two Regressor Situation. In this series of regressions the user has fit a first order model, but the true model contains an interaction term. Results are based on 100 simulations at levels I = .00, I = .20, and I = .30, and on 20 runs at each of the other levels of the coefficient I of the term X_1X_2. The relative prediction efficiencies of the least squares, kernel, and the Cp3 version of HATLINK are included below.

		Mean Sum of Squared Errors of Prediction									
I	POWER	LS	KER	PRESS	PRESS*	Cp1	Cp2	Cp3		OPT	lsd
.00	.05	48.6	226.1	58.6	50.7	62.8	104.0	55.6	(+K)	46.0	10.7
.10	.07	62.8	228.2	67.0	60.6	70.9	94.2	64.7	(+K)	56.2	19.3
.20	.15	130.5	230.5	116.1	115.3	121.5	134.3	112.9	(+K,+L)	99.6	8.8
.25	.23	170.3	228.1	121.7	125.0	127.5	144.1	118.9	(+K,+L)	108.6	18.6
.30	.33	232.8	234.0	158.5	169.6	163.5	170.5	157.6	(+K,+L)	138.4	9.3
.35	.45	293.1	232.1	155.8	169.1	158.7	164.9	156.9	(+K,+L)	140.9	19.8
.40	.58	369.8	233.4	169.9	187.1	171.9	175.0	174.3	(+K,+L)	154.1	20.5

I	Power	RPE(LS,K)	RPE(H,K)	RPE(H,LS)
.00	.05	4.65	4.07	0.86
.10	.07	3.63	3.53	0.97
.20	.15	1.76	2.29	1.25
.25	.23	1.34	1.92	1.43
.30	.33	1.01	1.66	1.64
.35	.45	0.79	1.48	1.87
.40	.58	0.63	1.34	2.12

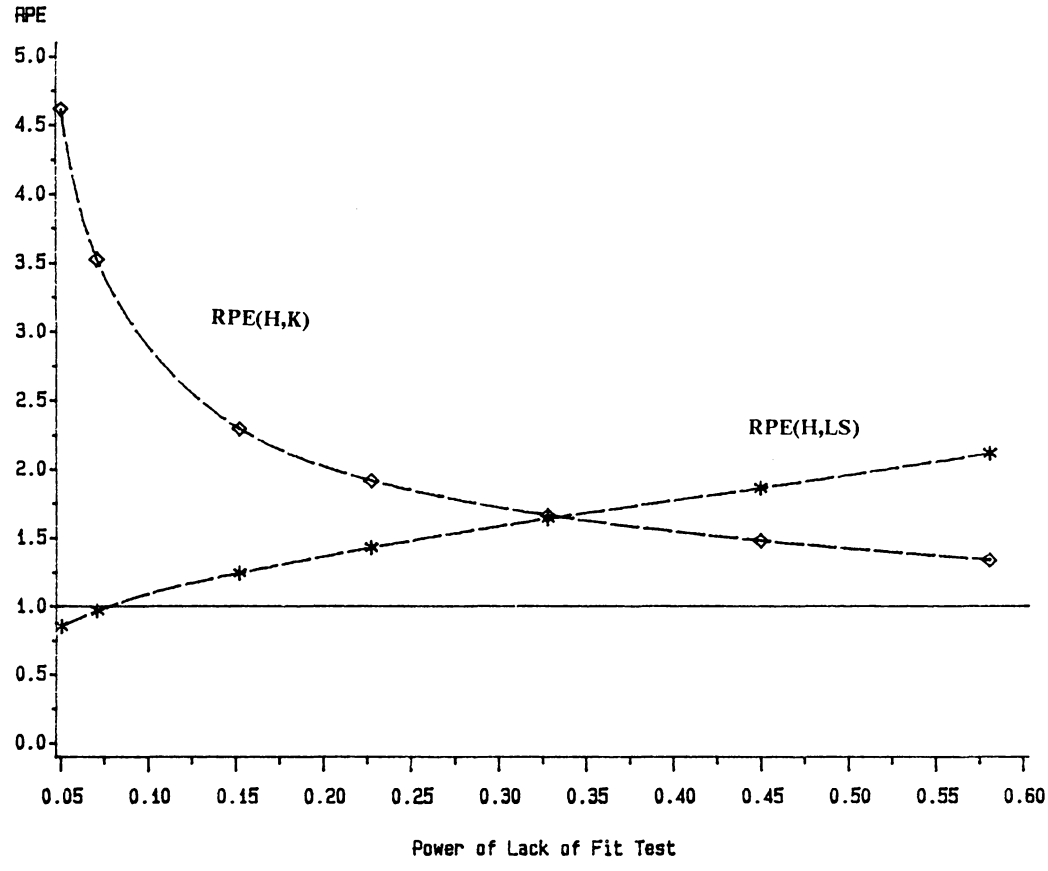


Figure IV.4.3. Relative Prediction Efficiencies (RPE).
Results are for the 1st order plus interaction regressions.

Table IV.4.6. Mean and Standard Deviations of the Estimates of Error Variance Obtained by the Least Squares and Kernel Methods and the Cp3 and Cp1 Versions of HATLINK. Also included are the mean degrees of freedom associated with these regression methods. Results are for 100 simulation at levels I = .00, I = .20, and I = .30, and 20 runs at each of the other levels of the coefficient I of the X_1X_2 term for the plane plus interaction series of regressions. The true value of σ^2 is 16.0. The standard errors of the mean s^2 values range from .41 (for Cp3 at I = .00) to 1.83 (for least squares at I = .40).

	Least Squares			Kernel			HATLINK -- Cp3			HATLINK -- Cp1		
I	mean s^2	s.d. s^2	mean df	mean s^2	s.d. s^2	mean df	mean s^2	s.d. s^2	mean df	mean s^2	s.d. s^2	mean df
.00	16.0	4.2	3	17.7	5.7	10.8	15.1	4.1	3.9	14.9	4.3	4.2
.10	18.0	5.1	3	18.0	5.9	10.8	16.4	4.5	4.4	16.1	4.6	4.9
.20	18.8	4.9	3	17.9	5.8	10.7	16.3	4.3	4.9	15.8	4.5	5.4
.30	22.5	5.8	3	18.0	5.9	10.6	17.4	4.5	5.7	16.5	4.5	6.6
.40	30.5	8.2	3	18.1	6.1	10.7	19.5	5.1	6.8	17.5	5.4	8.3

Table IV.4.7. Mean 95% Confidence Interval Widths and Percent Coverage at Locations (X_1, X_2) = (8,4), (2,8), and (10,2). Intervals are formed by the least squares and kernel methods and by the Cp3 and Cp4 Versions of HATLINK. Results are for 100 simulations at levels I = .00, I = .20, and I = .30, of the coefficient I of the interaction term X_1X_2 for the first order plus interaction series of regressions. The estimated standard error of these percentage estimates ranges from 2.2 (for the least squares and HATLINK methods at all three locations when Q = .00) to 5.0 (for the least squares intervals at location (10,2) when Q = .30).

		Least Squares		Kernel		HATLINK -- Cp3		HATLINK -- Cp4	
Location	I	mean width	% cov.	mean width	% cov.	mean width	% cov.	mean width	% cov.
(8,4)	.00	4.3	98	10.1	87	4.8	97	4.9	97
	.20	4.6	96	10.0	86	5.9	99	6.0	98
	.30	5.0	96	10.0	86	6.8	99	7.1	99
(2,8)	.00	5.4	94	10.3	99	5.8	97	5.9	96
	.20	5.8	66	10.3	98	6.7	79	6.8	79
	.30	6.4	45	10.3	96	7.5	74	7.7	81
(10,2)	.00	6.8	97	10.6	94	7.1	96	7.2	96
	.20	7.1	70	11.6	98	8.4	87	8.7	86
	.30	7.8	47	12.7	99	9.8	87	10.3	89

the results of Table IV.4.5 would seem to recommend the Cp3 version for general use. The relative performance of HATLINK compared to the least squares and kernel methods is displayed in Figure IV.4.3.

Variance estimates and confidence intervals

The means and standard deviations of the estimates of error obtained through the least squares, kernel, Cp3, Cp1, and PRESS* methods are shown in Table IV.4.6 for the first order plus interaction family of regressions. The results are much the same as have been observed for the families of regressions studied previously. Estimates of σ_ε^2 based on the two HATLINK methods appear to be the best over the range of levels of the interaction coefficient I.

Confidence interval coverages at locations $(X_1, X_2) = (8,4)$, $(2,8)$, and $(10,2)$ are shown in Table IV.4.7, as are the mean interval widths at these locations. Again it is seen that the coverage achieved by the least squares intervals is seriously affected at certain regressor locations when the user's model is incorrect. The kernel and HATLINK intervals are again wider than the least squares intervals, but maintain much better coverage of $f(X_1, X_2)$.

IV.4.C. True model is second order plus a third order interaction

A third family of regressions involving two regressors is now considered. In this case, it is assumed that the user has specified the full second order model,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2 + \beta_{11} X_1^2 + \beta_{22} X_2^2 + \varepsilon \ ,$$

while the true underlying function is of the form,

$$f(X_1, X_2) = 5X_1 + 5X_2 + 0.2(X_1 - 5.5)(X_2 - 5.5) + 0.4(X_1 - 5.5)^2 + 0.4(X_2 - 5.5)^2 + I(X_1 - 5.5)^2(X_2 - 5.5) \ .$$

Table IV.4.8. Prediction Performance of the Least Squares and Kernel Methods and for Five Versions of the HATLINK Procedure for the Two Regressor Situation. In this series of regressions the user has fit a full second order model, but the true model contains a third order interaction term. Results are based on 100 simulations at levels I = .00, I = .07, I = .10, and I = .12, and 20 runs at each of the other levels of the coefficient I of the term $X_1^2X_2$. The relative prediction efficiencies of the least squares, kernel, and the Cp3 version of HATLINK are included below.

		Mean Sum of Squared Errors of Prediction									
I	POWER	LS	KER	PRESS	PRESS*	Cp1	Cp2	Cp3		OPT	lsd
.00	.05	96.6	232.3	106.4	99.6	108.0	115.1	103.8	(+K)	91.1	8.2
.06	.11	142.2	230.9	133.6	133.1	138.0	145.9	133.3	(+K)	122.4	14.6
.07	.14	154.6	230.9	140.6	138.1	143.7	152.9	136.3	(+K,+L)	122.6	7.4
.08	.17	175.3	231.1	149.1	151.3	153.9	162.8	148.1	(+K,+L)	137.5	14.2
.10	.26	215.0	230.0	163.2	165.3	166.4	173.7	158.2	(+K,+L)	144.0	7.5
.12	.37	267.1	228.8	176.2	181.4	178.3	183.8	171.5	(+K,+L)	156.7	7.4
.14	.54	331.7	233.4	194.2	205.8	195.3	194.1	195.2	(+K,+L)	178.6	15.3

I	Power	RPE(LS,K)	RPE(H,K)	RPE(H,LS)
.00	.05	2.40	2.24	0.93
.06	.11	1.62	1.73	1.07
.07	.14	1.49	1.69	1.13
.08	.17	1.32	1.56	1.32
.10	.26	1.07	1.45	1.36
.12	.37	0.86	1.33	1.56
.14	.54	0.70	1.20	1.70

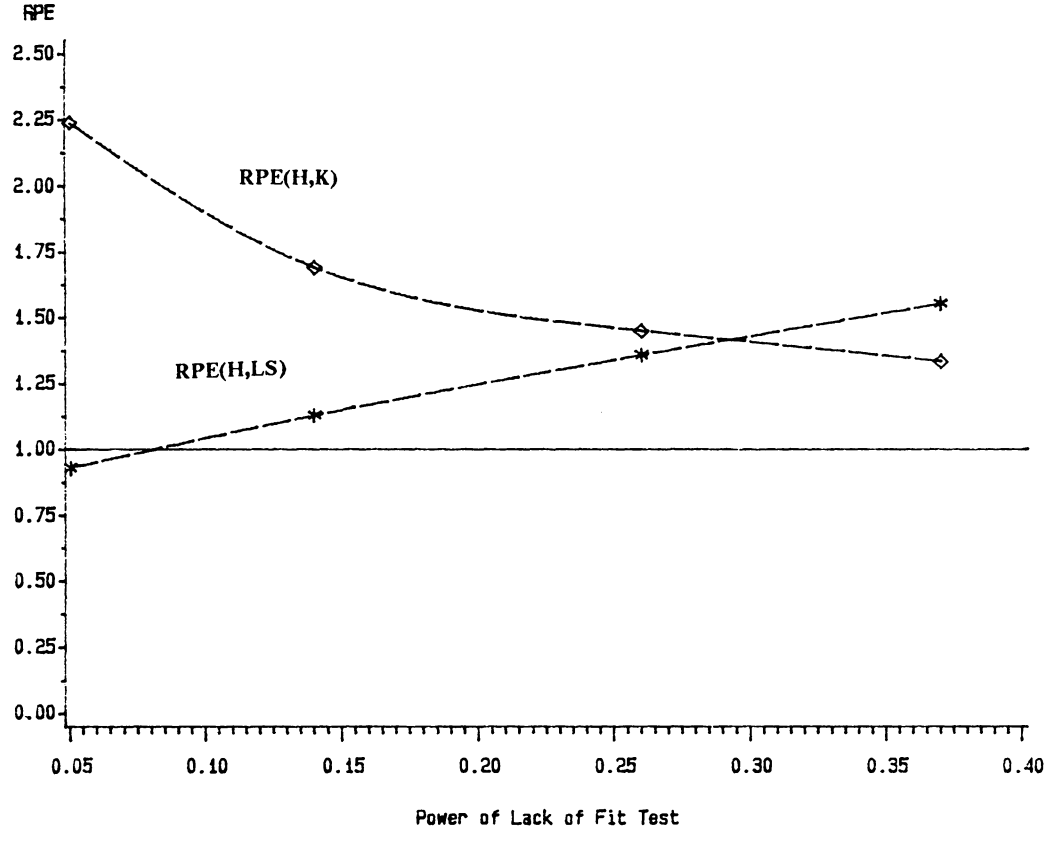


Figure IV.4.4. Relative Prediction Efficiencies (RPE). Results are for the 2nd order plus 3rd order interaction series of regressions.

The values used for the coefficient of the third order interaction term were I = 0.00, 0.06, 0.07, 0.08, 0.10, 0.12, and 0.14. The same data locations were used here as were used in the two previous 2 regressor families. The errors were once again generated from the N(0,16) distribution.

Table IV.4.8 presents the mean SSEP's for the various prediction methods for this family of regressions. The results for the cases where the interaction coefficient (I) is I = 0.00 and I = 0.12 are based on 100 runs each, with the same 100 sets of random errors for both cases. The results at the other levels of I are based on 20 runs each. Again it is observed that the HATLINK method using the Cp3 criterion for selecting λ offers relatively good prediction, regardless of the degree to which the model has been misspecified. The relative prediction performance of this version of HATLINK versus least squares and kernel regression is plotted in Figure IV.4.4.

The conservative PRESS* criterion for selecting λ again led to the lowest mean SSEP among the HATLINK versions when the user's model was correct (I = 0.00), but did not perform quite as well as the other HATLINK versions at high levels of I. On the other hand, the PRESS version of HATLINK produced nearly the same mean SSPE's as the Cp3 version at every level of the interaction coefficient I.

Variance estimates for this family are presented in Table IV.4.9, and confidence interval coverages and mean widths at locations (6,4), (2,8), and (10,2) are given in Table IV.4.10. The pattern observed in Table IV.4.9 for the least squares estimate of variance as the lack of fit of the user's model is increased is similar to what has been observed for the previous families of regressions. In this case, however, both HATLINK estimates of error variance tend to underesitimate σ^2 when the user's model is correct (I = 0.00). For example, mean estimate of σ^2 at I = 0.00 using Cp3 is 15.1, with a standard error of 0.3. In spite of this, the estimates of error variance based on Cp3 and Cp4 still appear to be more accurate than the least squares estimate across the range of levels of I. Further, the HATLINK estimates of σ^2 have the same (at I = 0.00) or lower standard deviations than the least squares estimates. In this case, the kernel estimate of error variance is biased upward somewhat at all levels of I.

The performance of the least squares confidence intervals for this family of regressions was somewhat better than for previous cases. (See Table IV.4.10.) At none of the regressor locations

considered here did the least squares intervals suffer as much damage under model misspecification as was observed earlier. However, there still was some loss of coverage by the least squares intervals at locations (2,8) and (10,2). The HATLINK intervals were able to maintain acceptable coverage with mean widths that were very close to the widths of the least squares intervals.

Diagnosing lack of fit

Tests for lack of fit using the values of λ obtained through the Cp1, Cp3, PRESS, and PRESS* criteria were carried out for the 100 runs made at $I = 0.12$. The $\alpha = .05$ and $\alpha = .10$ critical values for these tests were determined empirically through the 100 runs performed under the null hypothesis ($I = 0.00$). These critical values are shown in Table IV.4.11. Table IV.4.12 presents the empirical power of the tests based on the four λ selection methods, plus the theoretical power of the usual F test for lack of fit. The power achieved for the Cp1 method is about the same as the power of the F test, but the other three λ selection methods provided improved power, particularly the PRESS* criterion.

IV.4.D. Variable selection

In order to demonstrate the application of HATLINK to the issue of variable selection, the following situation will be considered. It is supposed that the user has decided that the variable X_1 is useful for predicting Y. However, the user is not sure whether the variable X_2 should be included in the model. A typical approach might be for the user to compare the values of R^2, R^2 *adjusted*, S^2, and PRESS for the models

(i) $Y = \beta_0 + \beta_1 X_1 + \varepsilon$, and

(ii) $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$.

Table IV.4.9. Mean and Standard Deviations of the Estimates of Error Variance Obtained by the Least Squares and Kernel Methods and the Cp3 and Cp1 Versions of HATLINK. Also included are the mean degrees of freedom associated with these regression methods. Results are for 100 simulations at each of four levels of the coefficient I of the $X_1^2X_2$ term for the full quadratic plus third order interaction series of regressions. The true value of σ^2 is 16.0. The standard errors of the mean values of s^2 range from .41 (for least squares and the two HATLINK methods at I = .00) to .60 (for least squares at I = .12).

	Least Squares			Kernel			HATLINK -- Cp3			HATLINK -- Cp1		
I	mean s^2	s.d. s^2	mean df	mean s^2	s.d. s^2	mean df	mean s^2	s.d. s^2	mean df	mean s^2	s.d. s^2	mean df
.00	16.0	4.1	6	18.1	5.7	10.3	15.1	4.1	6.8	15.0	4.1	6.4
.07	18.7	4.8	6	18.1	5.8	10.4	15.8	4.5	7.4	15.6	4.6	7.6
.10	21.3	5.5	6	18.1	5.8	10.6	16.5	4.7	7.9	16.2	4.9	8.2
.12	23.5	6.0	6	18.0	5.7	10.7	17.0	4.8	8.2	16.5	5.0	8.6

Table IV.4.10. Mean 95% Confidence Interval Widths and Percent Coverage at Locations (X_1, X_2) = (8,4), (2,8), and (10,2). Intervals are formed by the least squares and kernel methods and by the Cp3 and Cp4 Versions of HATLINK. Results are for 100 simulations at levels I = .00, I = .07, and I = .10, of the coefficient I of the interaction term $X_1^2X_2$ for the full second order plus third order interaction series of regressions. The estimated standard error of these percentage estimates ranges from 2.2 (for the least squares and HATLINK methods at all three locations when Q = .00) to 4.0 (for the least squares intervals at location (2,8) when Q = .30).

		Least Squares		Kernel		HATLINK -- Cp3		HATLINK -- Cp4	
Location	I	mean width	% cov.	mean width	% cov.	mean width	% cov.	mean width	% cov.
(8,4)	.00	5.4	97	9.9	92	6.0	98	6.0	98
	.07	5.8	94	10.0	91	6.8	97	6.9	98
	.10	6.2	94	10.0	91	7.4	98	7.5	97
(2,8)	.00	7.4	94	10.2	98	7.6	96	7.6	95
	.07	7.9	87	10.2	97	8.1	92	8.1	92
	.10	8.5	80	10.3	97	8.5	91	8.6	90
(10,2)	.00	9.7	93	10.6	90	9.9	93	9.9	93
	.07	10.5	90	11.5	98	10.9	95	10.9	94
	.10	11.2	86	12.3	97	11.7	95	11.8	95

Table IV.4.11. 95th and 90th percentiles of the distribution of the values of λ observed for the 100 runs under the null hypothesis (I = 0.0) for the full quadratic plus third order interaction series of regressions. Results are for the PRESS, PRESS*, Cp1, and Cp3 criteria for selecting λ. Also shown are the 95th and 90th percentiles of the empirical distribution of the usual F statistic for testing lack of fit, along with the corresponding theoretical percentage points.

	F (theor.)	F (empir.)	PRESS	PRESS*	Cp1	Cp3
95th percentile	2.59	2.480	.542	.316	.604	.481
90th percentile	2.09	2.188	.444	.267	.498	.399

Table IV.4.12. Proportion of rejections for the lack of fit test based on λ for the full second order plus third order interaction series of regression. Results are based on 100 simulations at each of three values for the coefficient I of the $X_1^2X_2$ term. Critical values at $\alpha = .05$ and $\alpha = .10$ were determined empirically through 100 runs under the null hypothesis (I = 0.0) and are shown in Table IV.4.11. Results are presented for the PRESS, PRESS*, Cp1, and Cp3 criteria for selecting λ. The theoretical and empirical powers of the usual F test for lack of fit are also shown.

		Power of F Test		Empirical Power of λ Tests			
I	α	Theor.	Empir.	PRESS	PRESS*	Cp1	Cp3
.07	.05	.14	.16	.19	.29	.17	.20
	.10	.24	.20	.29	.38	.26	.35
.10	.05	.26	.22	.32	.51	.25	.37
	.10	.40	.31	.58	.60	.50	.56
.12	.05	.37	.38	.53	.68	.38	.52
	.10	.53	.47	.73	.78	.61	.75

Another possibility is to perform the usual F test to compare the above two models using the "reduced" versus "full" model approach.

To simulate this situation and demonstrate the model-robustness aspect of HATLINK in the context of variable selection, two cases will be considered.

Case #1

In this case it will be assumed that the true underlying model is $Y = X_1 + \varepsilon$, $\varepsilon \sim N(0,16)$. The parametric models (i) and (ii) above will both be fit to the data by the least squares method. Each of these least squares fits will serve as the basis for a HATLINK fit to the data. The HATLINK fit based on parametric model (i) will have a kernel portion that fits the model $Y = f_1(X_1) + \varepsilon$, and the HATLINK fit corresponding to (ii) will have a kernel portion that fits $Y = f_2(X_1, X_2) + \varepsilon$.

Case #2

In the second case, the true underlying model will be $Y = X_1 + 0.40X_2^2 + \varepsilon$, $\varepsilon \sim N(0,16)$. This second case is selected as an example of where X_2 belongs in the model for predicting Y, but must be transformed in some fashion before it becomes a useful addition to the model. As in Case #1, fits will be obtained through least squares for both of the models (i) and (ii) above. The corresponding HATLINK regressions are again formed using kernel fits for the models $Y = f_1(X_1) + \varepsilon$ and $Y = f_2(X_1, X_2) + \varepsilon$ for (i) and (ii), respectively.

Table IV.4.13 shows the results for 20 runs made for each of these two cases. The various diagnostics and tests based on least squares are attempting to determine whether model (ii) is an improvement over model (i). The diagnostics based on HATLINK indicate whether there is an improvement in fitting $Y = f_2(X_1, X_2) + \varepsilon$ compared to fitting $Y = f_1(X_1) + \varepsilon$. That is, the HATLINK diagnostics provide information as to whether the inclusion of variable X_2 in any way

Table IV.4.13. Analysis of the Behavior of Several Criteria to Determine Whether the Variable X_2 Should be Included With X_1 in a model for predicting Y. Each criterion is a means for comparing regression results when only X_1 is in the model to corresponding results when X_2 is added to the model. The values in the table represent the number of times out of the 20 simulations conducted that a given criterion has indicated that variable X_2 should be included in the model. In Case 1, the true underlying model is linear in X_1 only. In Case 2, the true underlying model is linear in X_1 plus a quadratic term in X_2 . The PRESS* version of HATLINK is used here.

	Number of runs where X_2 appears to enhance the model			
	Case 1		Case 2	
Criterion	LS	H	LS	H
R^2 increased by over .05.	3	4	3	19
$R^2adj.$ increased by over .05.	3	3	3	17
PRESS decreased.	3	3	3	17
S^2 decreased.	5	6	4	20
df increased by 2.0 or more.	n.a.	1	n.a.	15
df increased by 1.5 or more.	n.a.	4	n.a.	17
Seq. F test for X_2 signif. at $\alpha = .10$	1	n.a.	2	n.a.

("n.a." denotes not applicable for that procedure.)

enhances the prediction of Y. The diagnostics and tests associated with the least squares approach are limited to detecting enhancement associated with an additive linear term in X_2. The results for HATLINK in this table are based on the PRESS* criterion for selecting λ. Virtually identical numbers were obtained for the Cp3 version of HATLINK. The values of R^2 and $R^2 adjusted$ for HATLINK are computed using the formulas presented in Section III.4.C. Observing whether the degrees of freedom for the HATLINK fit increase substantially when X_2 is added to the model is an additional piece of diagnostic information as to whether X_2 should be included in the model. Therefore, the number of times an increase of over 1.5 df occurred and the number of increases of over 2.0 df are included with the other diagnostic quantities in the table. In Table IV.4.13 it is seen that the diagnostics R^2, R^2 *adjusted*, S^2, and PRESS perform in a similar manner to their least squares counterparts in Case #1. But in Case #2, where the true model includes variable X_2 as a quadratic term, the HATLINK diagnostics are generally able to indicate the need for including X_2 in in the model.

The F^* statistic as defined in Equation III.4.12 can also be used in conjunction with HATLINK to determine if X_2 should be included in the model for predicting Y in the preceding example. Recall that this statistic takes a "full" versus "reduced" model approach to determine the usefulness of an additional variable or set of variables. As was discussed in Section III.4.E, a two-stage approach will be taken in applying the F^* test. First, it will be concluded that variable X_2 does not offer a significant improvement to the model whenever the model degrees of freedom for HATLINK increase by less than one when X_2 is added to the model. If the model degrees of freedom does increase by one or more, then the F^* statistic will be compared to the appropriate critical value found through the usual F table. (This value is found through interpolation according to the equivalent model and error degrees of freedom for the HATLINK fit.) Table IV.4.14 shows the results of using the F^* test for the two cases examined previously. It is seen that F^* offered similar effectiveness to the other statistics listed in Table IV.4.13 for diagnosing whether the variable X_2 should be included in the model.

Table IV.4.14. Use of the F^* Statistic in Two Cases to Determine Whether the Variable X_2 Should be included in the Model. Results are based on 20 simulations.

	Number Significant (Pct.)	
Significance Level	Case 1	Case 2
.05	2 (10%)	14 (70%)
.10	2 (10%)	17 (85%)

Table IV.4.15. Application of the F^o Test for Overall Model Utility for the Stepwise Approach to Variable Selection. The test statistic was formed using the PRESS* version of HATLINK. Results are based on 20 simulations.

	Number Significant (Pct.)	
Significance Level	Case 1	Case 2
.05	1 (5%)	20 (100%)
.10	2 (10%)	20 (100%)

Table IV.4.16. Application of the F^* Test to Diagnose Lack of Fit in the Addition of Variable X_2 to the Model as a First Degree Term. The test statistic was formed using the PRESS* version of HATLINK. Results are for the stepwise approach to variable selection, and are based on 20 simulatons.

	Number Significant (Pct.)	
Significance Level	Case 1	Case 2
.05	0 (0%)	18 (90%)
.10	2 (10%)	18 (90%)

Stepwise approach to variable selection

Another way to determine whether variable X_2 should be included with X_1 in the model for predicting Y is to take the stepwise approach mentioned near the end of Section III.4.E. The first step in this method is to fit a model involving only X_1 . Then the residuals from this regression are regressed against X_2 . The F^o statistic for overall model utility developed for HATLINK in Equation III.4.6 can be employed to determine whether X_2 provides additional information for predicting Y. Specifically, F^o tests the usefulness of the model $e = f(X_2) + \varepsilon$, where "e" denotes a residual from the least squares simple linear regression of Y on X_1 . The statistics F^* and F^{**} (Equations III.4.7 and III.4.9) also apply here. In this situation they indicate whether there is misspecification in the way that the user has prescribed for variable X_2 to enter the model. For example, the statistic F^* takes a full versus reduced model approach to test for possible lack of fit in the "reduced" model, $e = \beta_0 + \beta_2 X_2 + \varepsilon$. In this case, the "full" model is given by $e = f_2(X_2) + \varepsilon$.

To illustrate the use of this stepwise approach, the method is applied to the two cases in the preceding example. First the model $Y = \beta_0 + \beta_1 X_1 + \varepsilon$ was used and the residuals were obtained for the least squares fit. Then these residuals were regressed against X_2 through the user's model $e = \beta_0^* + \beta_2^* X_2 + \varepsilon$. The kernel fit for these residuals was based on X_2 only, so no interaction was considered here. (This is the "additive models" approach discussed near the end of Section III.4.E.) The stepwise method was carried out for the two cases examined previously in this section. For the case where the true underlying function was linear in X_1 only, the kernel fit to the residuals had an average of 1.72 degrees of freedom for the 20 runs. In the second case, where the true function contains the term $0.4X_2^2$, the mean kernel degrees of freedom was 3.90. The mean model degrees of freedom for the PRESS* version of HATLINK were 1.45 and 3.87 in cases one and two, respectively. Results of the F^o test for overall model utility using the PRESS* version of HATLINK are presented in Table IV.4.15. Additionally, the F^* test (Equation III.4.7) was used to determine whether X_2 should be added to the model in some form other than as a first power term. The results of this test for the two cases are shown in Table IV.4.16.

Tables IV.4.15 and IV.4.16 indicate that both pseudo F tests are capable of accomplishing the tasks they were designed for. In Case 1, where the true underlying model does not include variable X_2 , the two tests generally reflect this. When the variable X_2 belongs in the model as a second power term (Case 2), the HATLINK F^o test for overall model utility indicates that X_2 does belong in the model. Further, the F^* test reveals that variable X_2 should not just be added to the existing model as a first power term.

Note that both of these pseudo-F tests were actually carried out as two-stage diagnostic procedures. That is, if the model degrees of freedom for the PRESS* version of HATLINK was so small for a given run that the corresponding numerator df for one of the F tests was below 1.0, then that F statistic was declared nonsignificant for that run. This occurred 18 times in Case 1 for each of the F statistics, and twice in Case 2 for the F^* statistic.

Chapter V

V. APPLICATIONS OF THE HATLINK METHOD

In order to demonstrate the manner in which the HATLINK procedure can be used in practice, two real regression data sets are considered. The first data set involves one resgressor variable, while the second example involves two regressors.

V.1. Example #1: Single Regressor Variable

Table V.1.1 presents data from page 185 of Montgomery and Peck (1982). In this example, the response variable Y is the tensile strength (in psi) of paper, and the regressor variable X is the percentage of hardwood in the batch of pulp from which the paper was produced. After first centering the X's, the authors have used least squares to fit the quadratic model, $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$. Figure V.1.1 shows the least squares quadratic fit to these 19 data

Table V.1.1. Data for Example #1. Y = tensile strength (Psi), and X = percentage of hardwood.

X	Y
1	6.3
1.5	11.1
2	20.0
3	24.0
4	26.1
4.5	30.0
5	33.8
5.5	34.0
6	38.1
6.5	39.9
7	42.0
8	46.1
9	53.1
10	52.0
11	52.5
12	48.0
13	42.8
14	27.8
15	21.9

Table V.1.2. Summary of Fits Obtained by the Least Squares, Kernel and HATLINK Methods for Example #1. The least squares and HATLINK fits were based on a **quadratic** user's model. The Cp3 criterion was used to obtain the HATLINK fit.

Method	SSE	df	s^2
Least Squares	312.6	3.0	19.5
HATLINK	152.1	5.8	13.6
Kernel	41.8	10.6	4.9

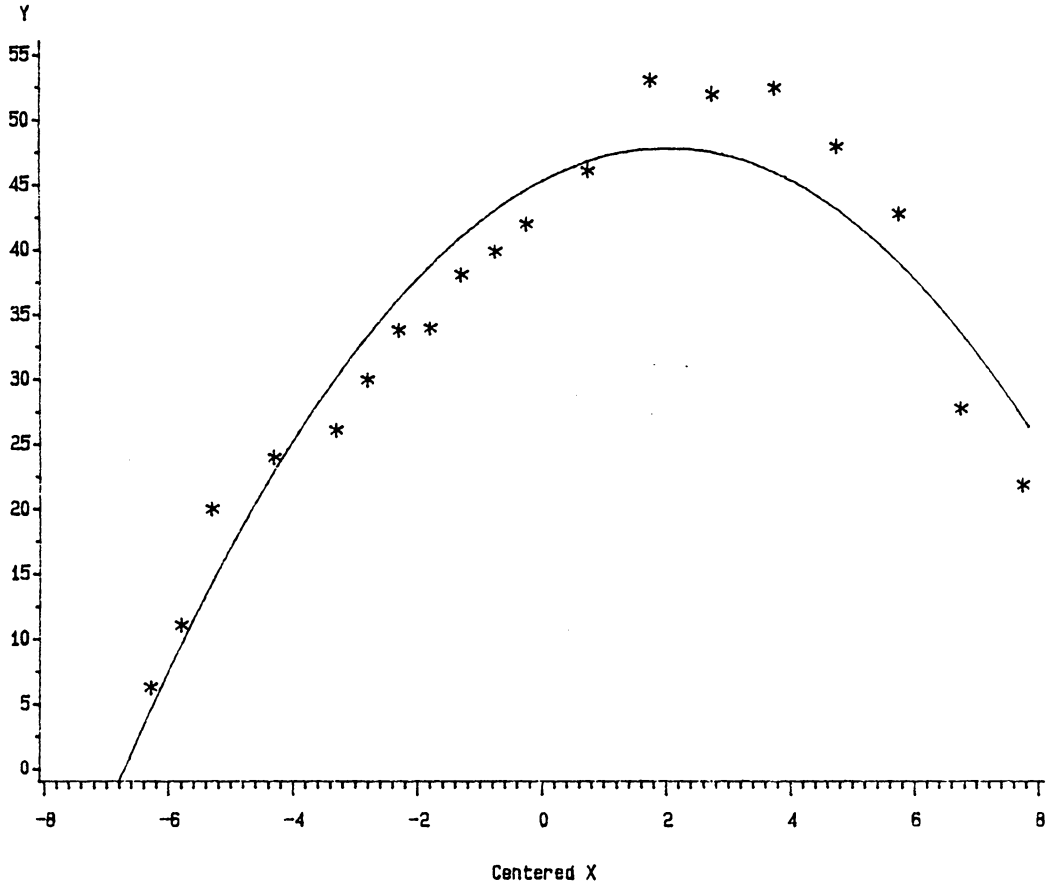


Figure V.1.1. Quadratic model fit by L.S. for Example #1.

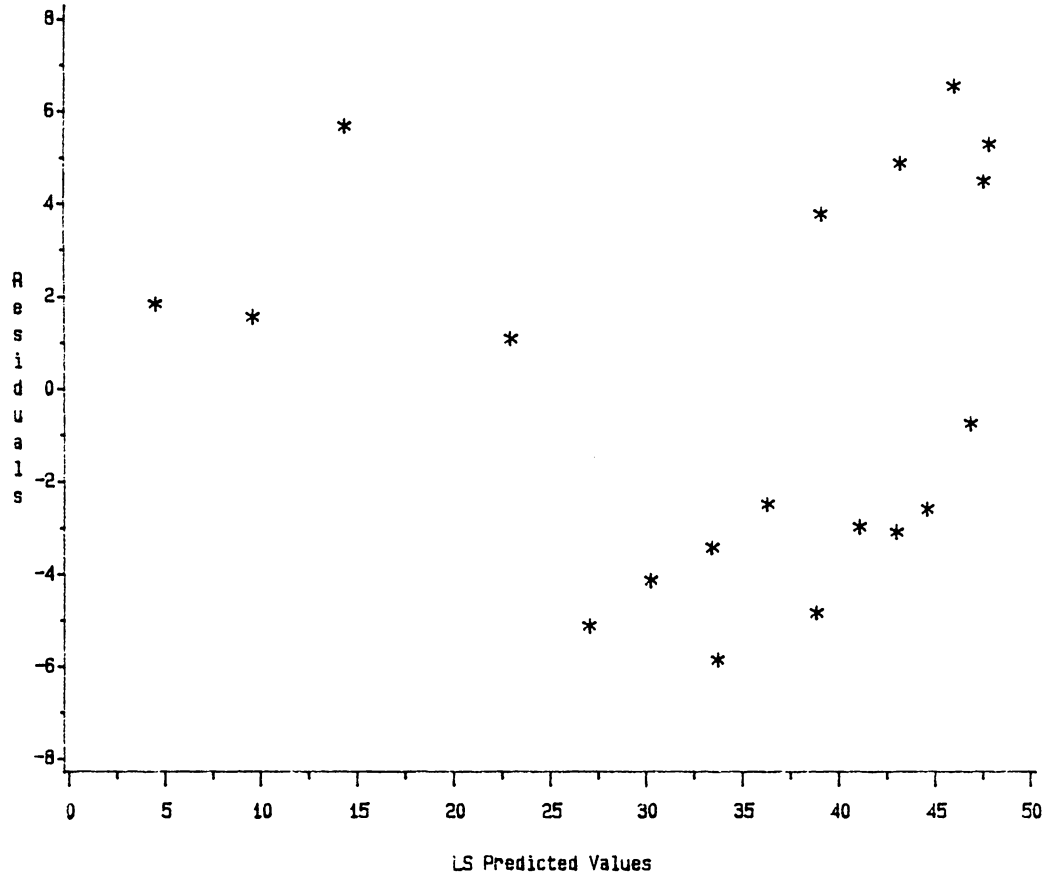


Figure V.1.2. Residuals vs. Yhats for the LS quadratic fit.

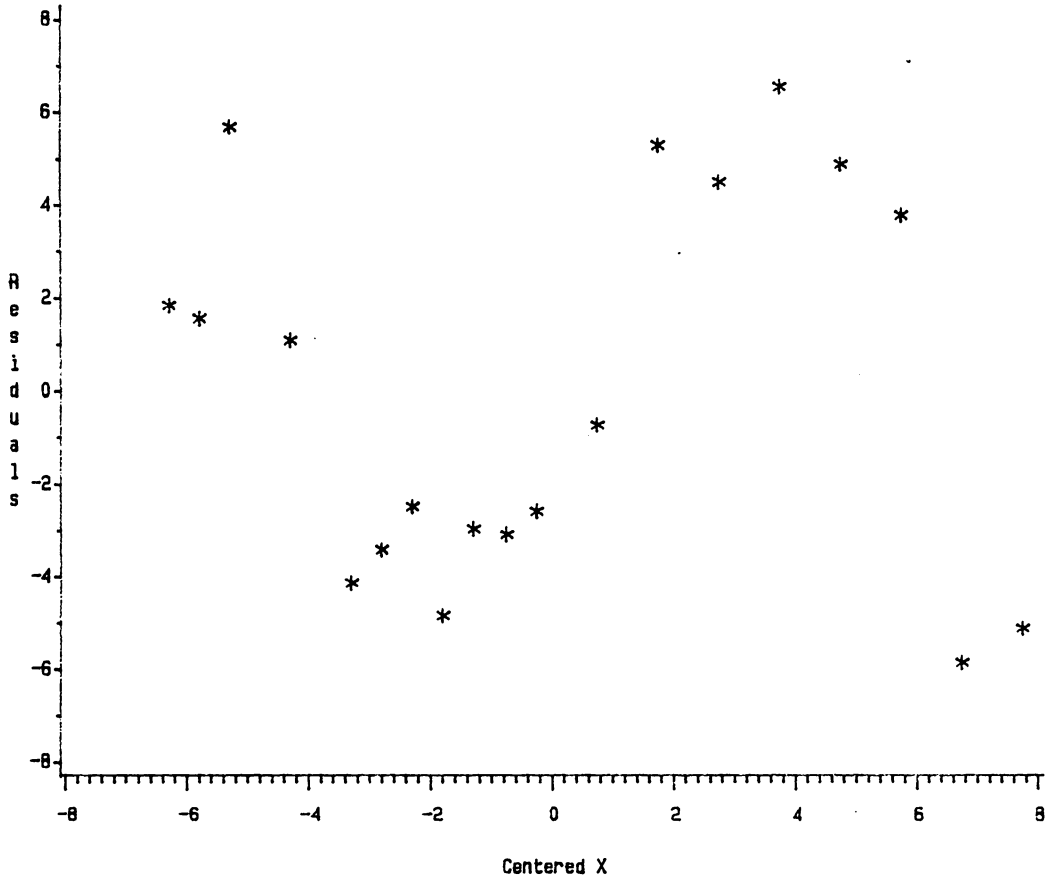


Figure V.1.3. Residuals from LS fit vs. centered X values.

points. Plots of the residuals e_i versus $\hat{Y}_i$ and $X_i - \overline{X}$ are shown in Figures V.1.2 and V.1.3. The authors comment that, "These plots do not reveal any serious model inadequacy." That conclusion is somewhat debatable, as will be seen later, but the point is that many users would be satisfied with the quadratic model in this situation.

Now consider the kernel fit for this data set, as shown in Figure V.1.4. Note how close the fitted values are to the actual data points in this figure. With the equivalent of 10.6 model degrees of freedom, it is apparent that the kernel is overfitting to the specific observations in this example. The HATLINK fit based on the quadratic user's model is graphed in Figure V.1.5. Corresponding to the selection of $\lambda = .37$ by the Cp3 criterion, there are 5.8 model degrees of freedom associated with the HATLINK fit. Figure V.1.5 indicates that the HATLINK method is not overfitting the data here. However, the sum of squared errors (SSE) for the HATLINK fit is less than half the value of SSE for the least squares quadratic fit. The values of SSE for the kernel, least squares, and HATLINK methods are given in Table V.1.2.

Additionally, the F^* test (Section III.4.D) based on HATLINK can be applied here to diagnose possible lack of fit for the quadratic model. (Recall that this test can be applied even in the absence of multiple observations at any individual regressor locations.) In this case, the F^* test is significant at $\alpha = .10$, suggesting that the quadratic model is inadequate. (The value of the test statistic, $F^* = 2.89$, was compared with an interpolated critcial value of $F_{2.84,\,13.16,\,.90} = 2.60$.)

Given the result of the F^* test, the user may elect to next fit a different model by the least squares method. Despite the conlusion of Montgomery and Peck quoted earlier, the residual plots, Figures V.1.2 and V.1.3, at least hint that the addition of a cubic term may improve the user's model. Therefore, the user may consider fitting the cubic model,

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon \ .$$

It turns out that the usual F test for testing H_0: $\beta_3 = 0$ indicates very strongly that the cubic term improves the model. (The value of the test statistic, F = 31.8, exceeds the $\alpha = .01$ critical value, $F_{1,15,.99} = 8.68$.) Figure V.1.6 shows the least squares fit for the cubic model.

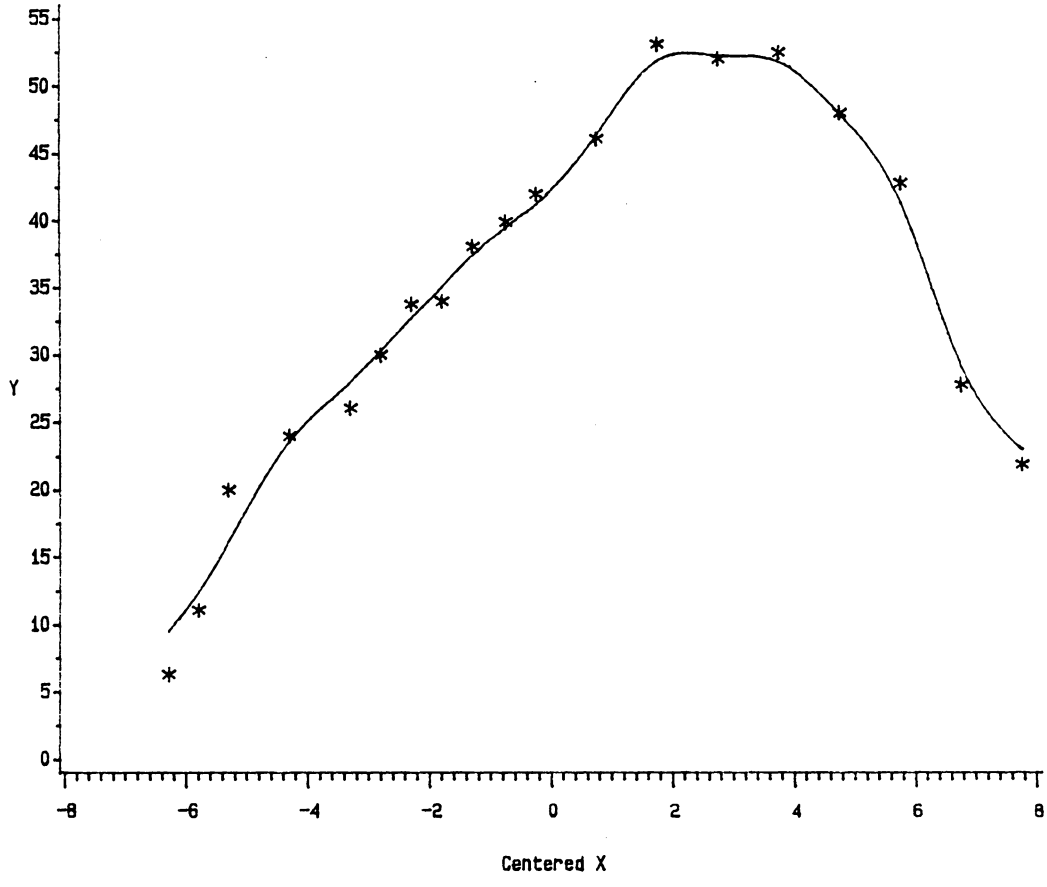


Figure V.1.4. Kernel fit to the data in Example #1.

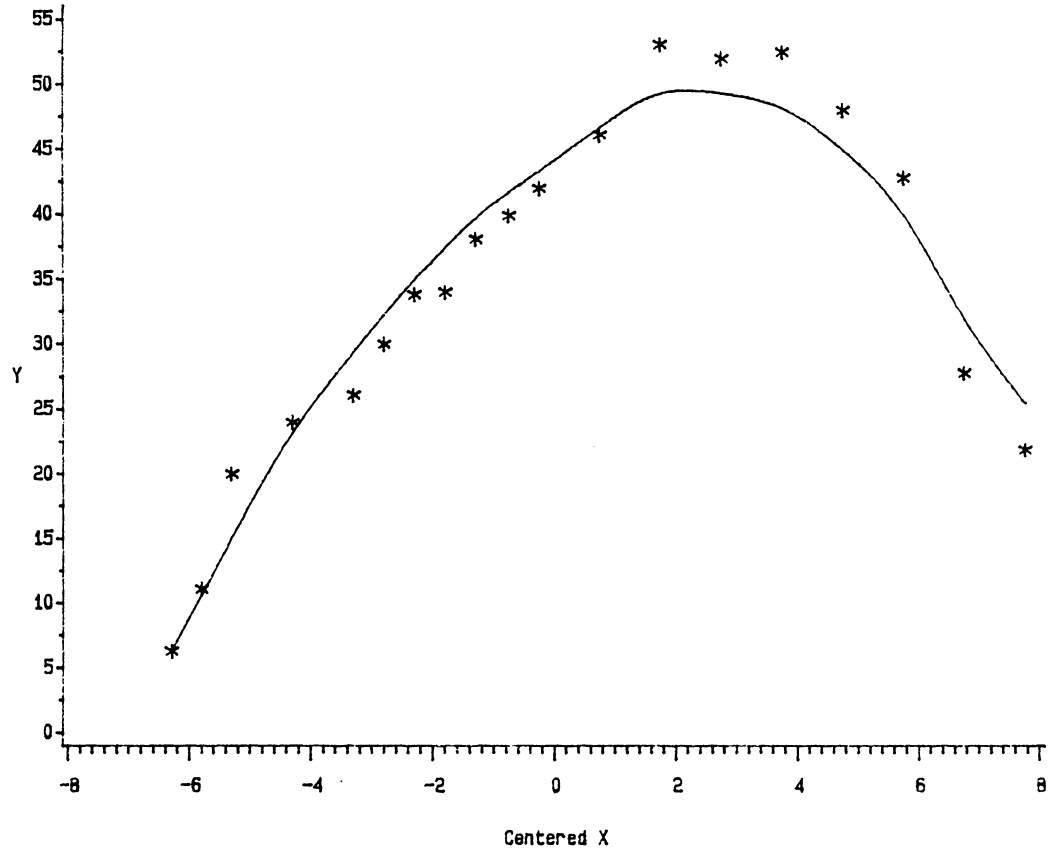


Figure V.1.5. HATLINK fit to the data in Example #1
for a quadratic model.

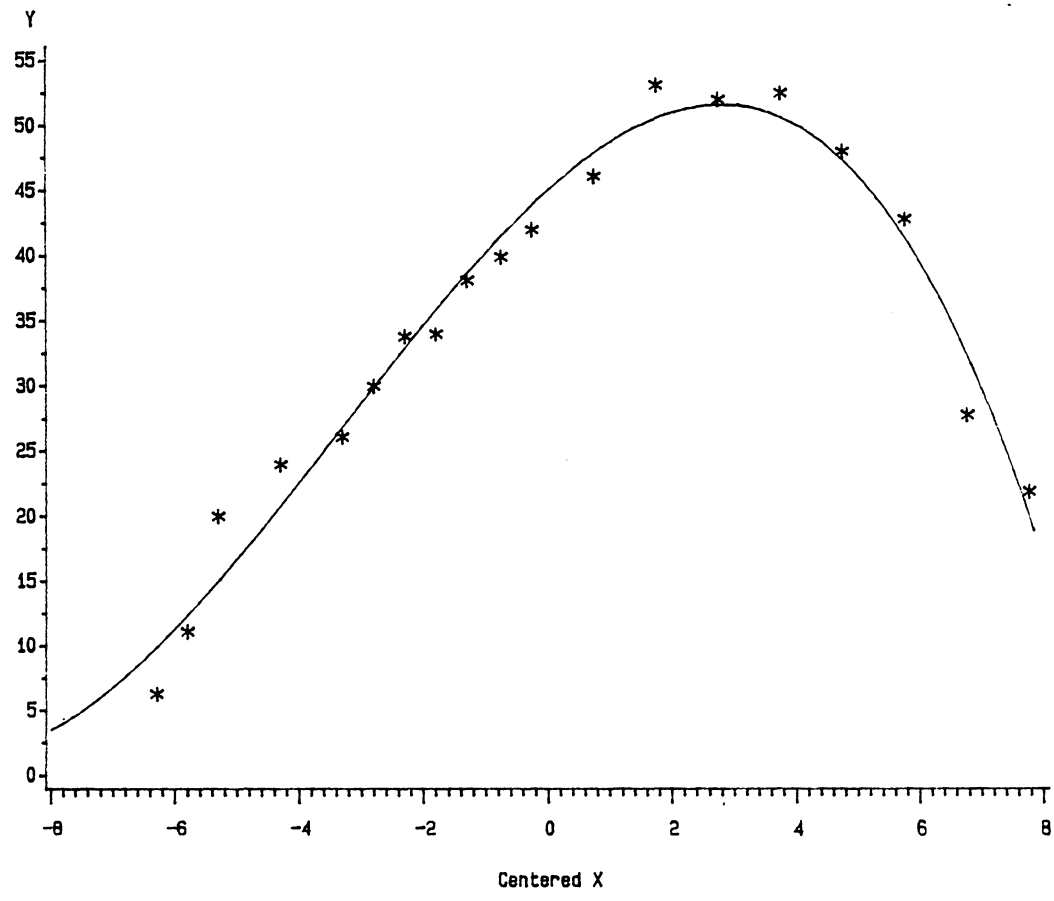


Figure V.1.6. Cubic model fit by L.S. for Example #1.

The HATLINK fit based on a cubic user's model is graphed in Figure V.1.7. Again, the Cp3 method was used to select λ . In this case, a value of $\lambda = .16$ was selected, which corresponds to 5.05 model degrees of freedom for HATLINK. Table V.1.3 shows the values of SSE for the HATLINK and least squares methods based on a cubic user's model. The HATLINK fit for this model is clearly superior to the HATLINK fit obtained previously for the quadratic user's model. The value of SSE for HATLINK dropped from 152.1 to 84.6, while the model degrees of freedom *decreased* by 0.8. The fact that the PRESS statistic for HATLINK (Table V.1.3) is somewhat lower than for the least squares and kernel methods suggests that future predictions based on the HATLINK fit may be superior.

For completeness, Figure V.1.8 is presented, which shows the HATLINK fit with 95% confidence bands (Section III.4.B) for estimating the true underlying function.

In summary, it has been observed that the HATLINK method has in several ways enhanced the regression analysis of these data. First, it has helped through the F^* test, by suggesting that the quadratic model is inadequate here. Then, when the user has fit a cubic model, the HATLINK predictions based on this model are an apparent improvement upon those obtained through least squares.

Table V.1.3. Summary of Fits Obtained by the Least Squares, Kernel and HATLINK Methods for Example #1. The least squares and HATLINK fits were based on a **cubic** user's model. The Cp3 criterion was used to obtain the HATLINK fit.

Method	SSE	df	s^2	PRESS
Least Squares	100.2	4.0	6.7	205.1
HATLINK	84.6	5.0	6.1	185.9
Kernel	41.8	10.6	4.9	189.1

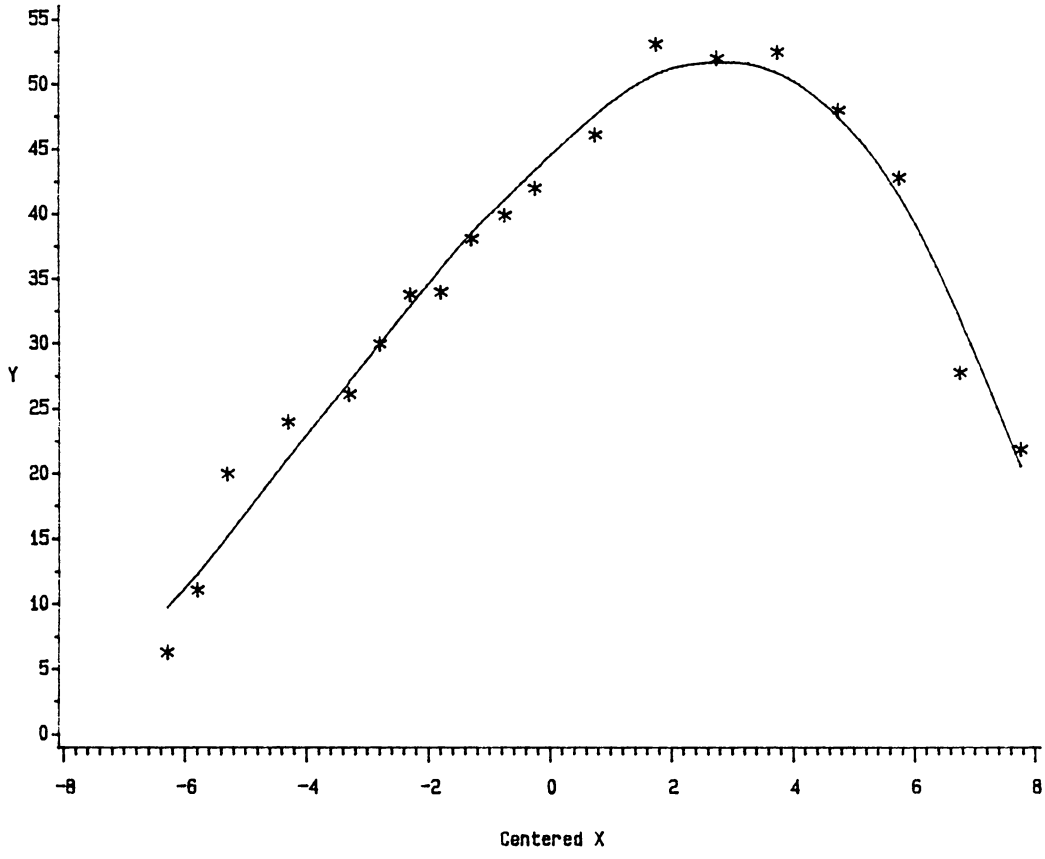


Figure V.1.7. HATLINK fit to the data in Example #1.
The fit is based on a cubic model.

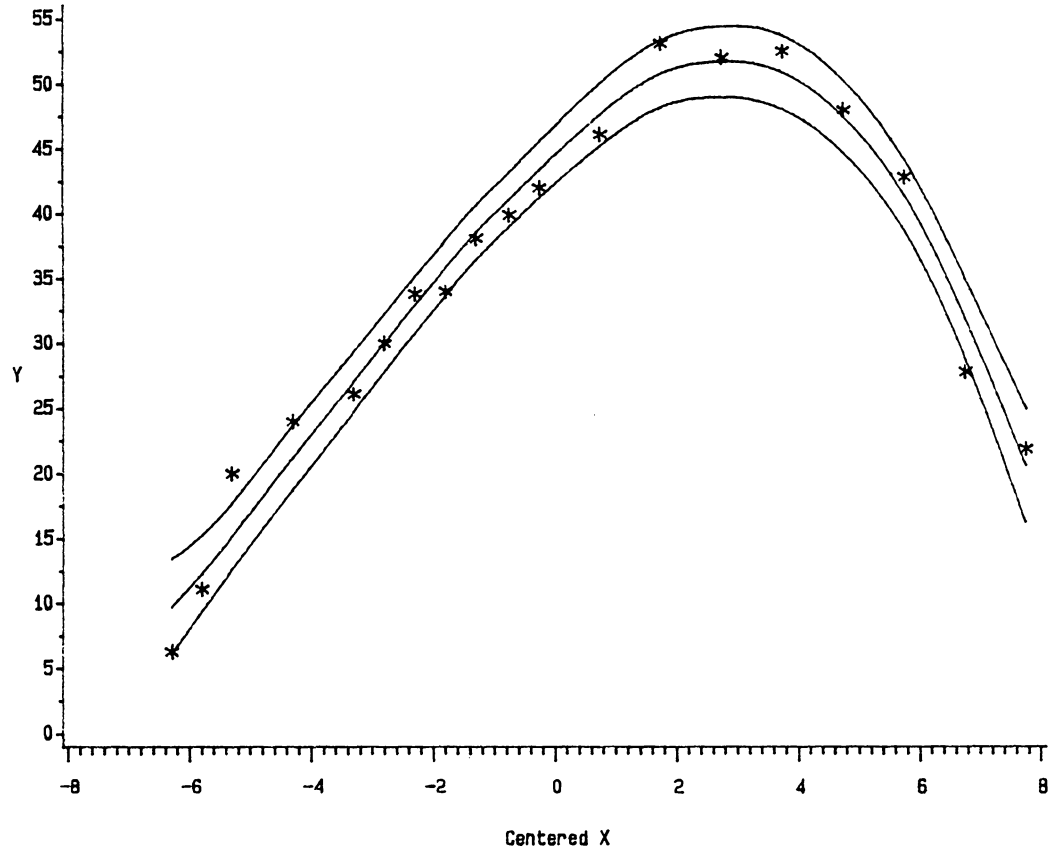


Figure V.1.8. HATLINK fit with 95% confidence bands
for a cubic model.

V.2. Example #2: Two Regressor Variables

The second data set is found on page 520 of Draper and Smith (1981), and is presented here in Table V.2.1. The data pertain to a certain chemical reaction, where Y is the fraction of original material remaining, X_1 is the reaction time in minutes, and X_2 is the temperature in degrees Kelvin. There are n = 38 observations taken at 18 distinct regressor locations. These (X_1, X_2) locations are plotted in Figure V.2.1. Note that there are only five levels of variable X_2 . This is an extremely small number of distinct X_2 locations for successful application of the kernel method.

Supposing that the user knows no theory which would suggest a particular form for the underlying model, a typical approach would be to initially fit the first order model,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon \ ,$$

using least squares. For this fit, a value of $R^2 = .86$ was obtained, so that a user may be satisfied with the first order model. Further, since the scatterplots of Y versus X_1 and Y versus X_2 (Figures V.2.2 and V.2.3) do not clearly suggest that Y must be related to other powers of these variables, a user may be content with the first order fit and stop at this point.

Now consider the kernel fit for this data set. Using the PRESS*(h) bandwidth selection criterion resulted in a kernel fit with the equivalent of 15.3 model degrees of freedom. This is an extremely large value considering that there are only 18 distinct regressor locations. Thus, in this example, the kernel is nearly fitting the mean value of the response at each of the 18 locations.

The HATLINK fit based on the first order user's model seems to offer an improvement over both the least squares and kernel fits. Using the Cp3 criterion resulted in a value of $\lambda = .65$ and a HATLINK fit with the equivalent of 10.9 model degrees of freedom. Table V.2.2 shows the sum of squared errors (SSE) for the least squares, kernel, and HATLINK fits, along with the PRESS statistic for each method. It is seen that both SSE and PRESS are substantially higher for least squares than for the other two methods. This suggests that the first order user's model may be inadequate. Also, the F^* test for lack of fit (Section III.4.D) was significant at the $\alpha = .01$ level,

Table V.2.1. Data for Example #2. Y = fraction of original material remaining, X_1 = reaction time in minutes, X_2 = temperature in degrees Kelvin.

X_1	X_2	Y
120	600	.900
60	600	.949
60	612	.886
120	612	.785
120	612	.791
60	612	.890
60	620	.787
30	620	.877
15	620	.938
60	620	.782
45.1	620	.827
90	620	.696
150	620	.582
60	620	.795
60	620	.800
60	620	.790
30	620	.883
90	620	.712
150	620	.576
60	620	.802
60	620	.802
60	620	.804
60	620	.794
60	620	.804
60	620	.799
30	631	.764
45.1	631	.688
40	631	.717
30	631	.802
45	631	.695
15	639	.808
30	639	.655
90	639	.309
25	639	.689
60.1	639	.437
60	639	.425
30	639	.638
30	639	.659

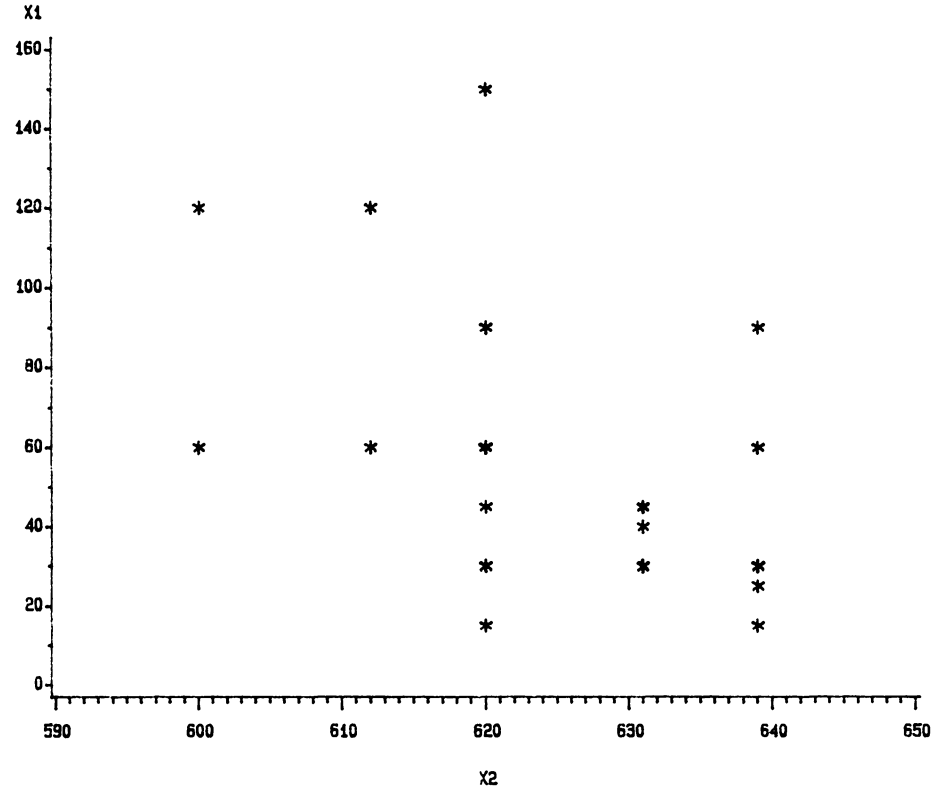


Figure V.2.1. Plot of the data locations for Example #2.

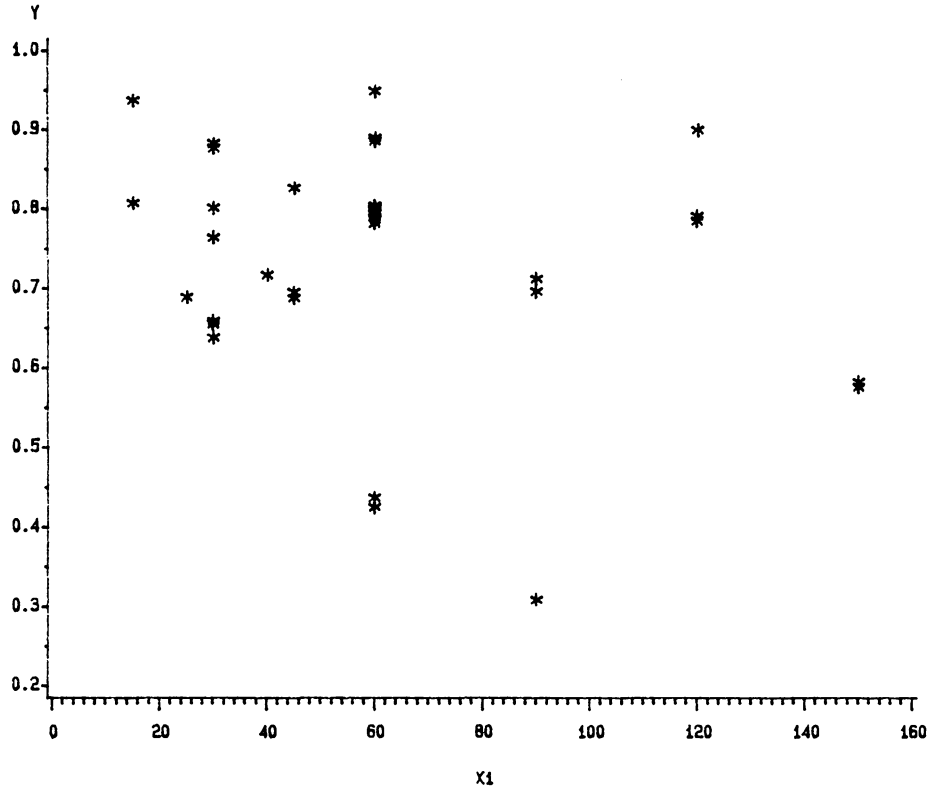


Figure V.2.2. Plot of Y versus X1 for Example #2.

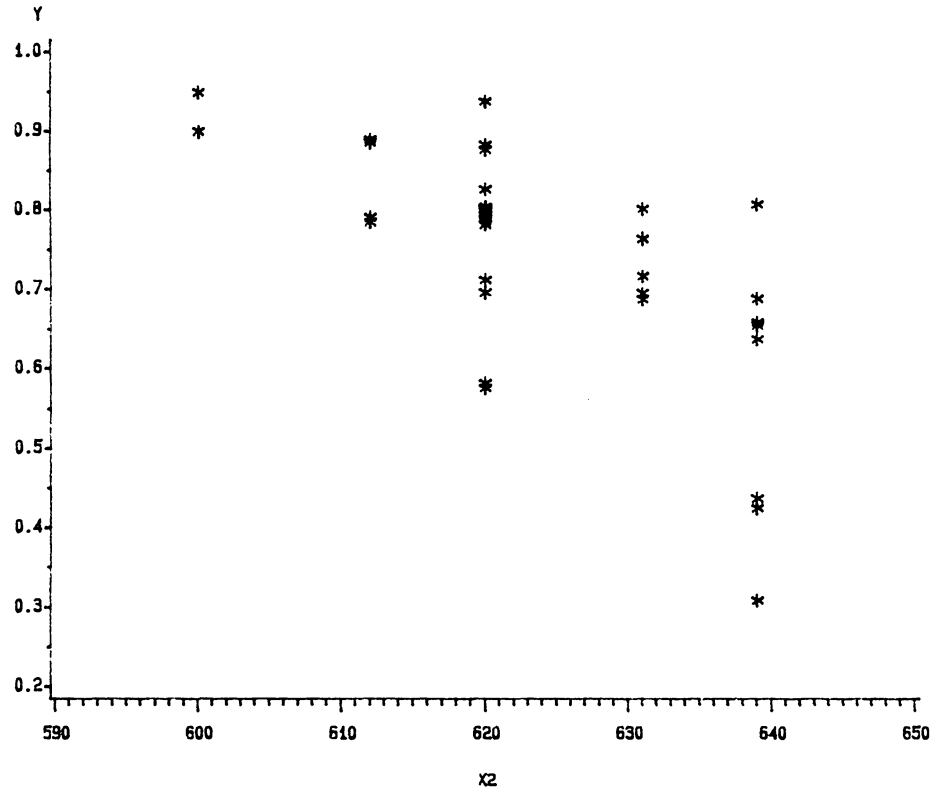


Figure V.2.3. Plot of Y versus X2 for Example #2.

thus supporting the diagnosis that the first order model is not adequate. (The statistic, $F^* = 3.41$, was compared to the interpolated critical value, $F^*_{7.9,\,27.1,\,.99} = 3.41$.)

Since the F^* test has shown that the first order model is inadequate, a second user's model is considered. This time, a full second order model,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 x_2 + \beta_{11} X_1^2 + \beta_{22} X_2^2 + \beta_{12} X_1 X_2 + \varepsilon \ ,$$

was used. Table V.2.3 shows that the sum of squared errors for the least squares fit is substantially lower for the second order model. Moreover, the usual F test for testing H_0: $\beta_{11} = \beta_{22} = \beta_{12} = 0$ indicates that adding the three second order terms results in a significant improvement in the user's model. ($F = 80.0$ was compared to $F_{3,32,.99} = 4.47$.)

The HATLINK fit using the second order user's model also is an improvement over the previous HATLINK fit with the first order user's model. As seen in Tables V.2.2 and V.2.3, the sum of squared errors dropped to less than one third of its value when the user's model was first order. For the second order model, the HATLINK fit has 11.2 model degrees of freedom, corresponding to $\lambda = .56$ chosen by the Cp3 criterion. Table V.2.3 also shows that the PRESS statistic associated with the HATLINK fit is somewhat lower than for the least squares fit, and substantially lower than the PRESS statistic for the kernel fit. This is additional evidence that the kernel has overfit to the data, and suggests that future predictions would be best obtained through the HATLINK regression.

The F^* test for lack of fit based on HATLINK was again performed, this time for the second order user's model. Once more, the test was significant, indicating that the second order model is also inadequate. ($F^* = 3.31$ was compared to $F_{5.2,\,26.8,\,.95} = 2.57$.) In this situation, the F^* statistic has apparently provided a correct diagnosis, since there is a theoretical basis for the nonlinear model,

$$Y = e^{-\theta_1 X_1 e^{\left[-\theta_2\left(\frac{1}{X_2} - \frac{1}{620}\right)\right]}} + \varepsilon \ , \qquad (V.2.1)$$

for these data (Draper and Smith, 1981, page 519). The nonlinear regression fit obtained as in Draper and Smith is graphed in Figure V.2.4. The fit obtained through HATLINK is very close

Table V.2.2. Summary of Fits Obtained by the Least Squares, Kernel, and HATLINK Methods for Example #2. The Cp3 criterion was used to obtain the HATLINK fit. Results for least squares and HATLINK are based on a first order user's model.

Method	SSE	Model df	PRESS
Least Squares	.0988	3.0	.133
HATLINK	.0175	10.9	.059
Kernel	.0040	15.3	.059

Table V.2.3. Summary of Fits Obtained by the Least Squares, Kernel, and HATLINK Methods for Example #2. The Cp3 criterion was used to obtain the HATLINK fit. Results for least squares are based on a full second order user's model.

Method	SSE	Model df	s^2	PRESS
Least Squares	.0116	6.0	.00036	.0276
HATLINK	.0054	11.2	.00020	.0237
Kernel	.0040	15.3	.00018	.0588

Table V.2.4. Sum of Squared Deviations (SSD) Between the Fitted Values at the Data Locations Obtained by Nonlinear Regression and the Fitted Values Obtained by the Least Squares, Kernel, and HATLINK methods. Results for least squares and HATLINK are based on a full second order user's model.

Method	SSD
Least Squares	.0089
HATLINK	.0049
Kernel	.0052

to that obtained through nonlinear regression for this theoretical model. The sum of squared deviations ("SSD") between the HATLINK fitted values and those for the nonlinear regression is quite small, just .0049. A substantial portion (.0021) of this sum is due to one particlular data point taken at the extreme location, $(X_1, X_2) = (15, 639)$. The sum of squared deviations from the nonliear regression fitted values was lower for the HATLINK fit than for the least squares and kernel methods. These results are shown in Table V.2.4.

The sum of squared errors for the nonlinear regression was SSE = .0043, which is close to the SSE's shown in Table V.2.3 for the kernel and HATLINK fits. Predicted values at the 18 data locations are shown in Table V.2.5 for the least squares, kernel, HATLINK, and nonlinear regression methods. At several of these locations, particularly at (X_1, X_2) = (15,620), (30,620), (45.1,620), and (25,639), the HATLINK predicted values are much closer than the kernel predicted values to the fitted values obtained through nonlinear regression. At certain other data locations, (120,600), (60,600), (15,620), and (60,639), the HATLINK method gives closer predictions to the nonlinear regression fit than the least squares method. At locations (120,600) and (60,639), 95% confidence intervals for the true underlying function formed for least squares do not contain the corresponding fitted value obtained through nonlinear regression. Similarly, 95% confidence intervals constructed as in Equation II.5.4 for the kernel method do not contain the nonlineasr regression fitted values at locations (45,620) and (90,639). On the other hand, 95% confidence intervals based on HATLINK (Section III.4.B) do contain the nonlinear regression fitted values at these locations. Table V.2.6 shows 95% confidence intervals for the three methods at the 18 regressor locations. Note that intervals formed by all three methods fail to contain the nonlinear regression predicted values at locations (15,639) and (25,639).

In summary, this example illustrates the way in which HATLINK can improve upon the usual approach to regression when the user has no theory to guide the formulation of a model. Supposing that the user is unaware that there is a theoretical basis for the nonlinear model (Equation V.2.1), then fitting the first order model would be a natural first step in the regression analysis. As mentioned earlier, it is possible that a regression user would be satisfied and stop with

Table V.2.5. Predicted Values Obtained at the 18 Data Locations in Example #2 by the Least Squares (LS), Kernel (Ker), HATLINK (H), and Nonlinear (Non) Regression Methods. The least squares fit was based on a full second order user's model, and HATLINK was formed using $\lambda = .56$, as obtained by the Cp3 criterion.

Location		Obs.	Predicted Values			
X_1	X_2	Y	LS	Ker	H	Non
120	600	.900	.9369	.9000	.9164	.9024
60	600	.949	.9278	.9490	.9396	.9500
60	612	.886	.8725	.8880	.8811	.8814
120	612	.785	.7792	.7880	.7841	.7769
60	620	.787	.7921	.7964	.7945	.7980
30	620	.877	.9005	.8801	.8892	.8933
15	620	.938	.9617	.9336	.9460	.9452
45	620	.827	.8436	.8203	.8306	.8440
90	620	.696	.7021	.7040	.7031	.7129
150	620	.582	.5775	.5790	.5783	.5689
30	631	.764	.7801	.7729	.7761	.7830
45	631	.688	.6996	.7005	.7001	.6922
40	631	.717	.7263	.7158	.7204	.7216
15	639	.808	.7529	.7725	.7638	.8095
30	639	.655	.6512	.6596	.6559	.6554
90	639	.309	.2906	.3090	.3009	.2815
25	639	.689	.6846	.6735	.6784	.7032
60	639	.437	.4611	.4310	.4443	.4289

Table V.2.6. 95% Confidence Limits for the Least Squares, Kernel, and HATLINK Methods for Example #2. The least squares and HATLINK fits are based on a cubic user's model. The Cp3 criterion was used to obtain λ for HATLINK. For comparison, the fitted values obtained by nonlinear regression ("Non") are included.

Location			LS		Ker		HATLINK	
X_1	X_2	Non	lower	upper	lower	upper	lower	upper
120	600	.9024	.9071	.9668	.8726	.9274	.8900	.9427
60	600	.9500	.8963	.9594	.9216	.9764	.9128	.9664
60	612	.8814	.8613	.8836	.8686	.9073	.8648	.8974
120	612	.7769	.7634	.7950	.7686	.8074	.7668	.8014
60	620	.7980	.7829	.8012	.7882	.8046	.7865	.8024
30	620	.8933	.8865	.9146	.8612	.8991	.8726	.9058
15	620	.9452	.9388	.9845	.9073	.9599	.9222	.9698
45	620	.8440	.8343	.8529	.7984	.8422	.8126	.8486
90	620	.7129	.6903	.7139	.6846	.7234	.6867	.7196
150	620	.5689	.5520	.6029	.5596	.5984	.5584	.5982
30	631	.7830	.7696	.7906	.7549	.7908	.7609	.7913
45	631	.6922	.6909	.7082	.6841	.7169	.6864	.7138
40	631	.7216	.7176	.7350	.6997	.7319	.7069	.7339
30	631	.7830	.7696	.7906	.7549	.7908	.7609	.7913
15	639	.8095	.7338	.7720	.7491	.7959	.7430	.7847
30	639	.6554	.6373	.6651	.6455	.6738	.6427	.6691
90	639	.2815	.2635	.3178	.2816	.3364	.2753	.3264
25	639	.7032	.6694	.6997	.6586	.6885	.6644	.6925
60	639	.4289	.4443	.4778	.4116	.4504	.4268	.4618

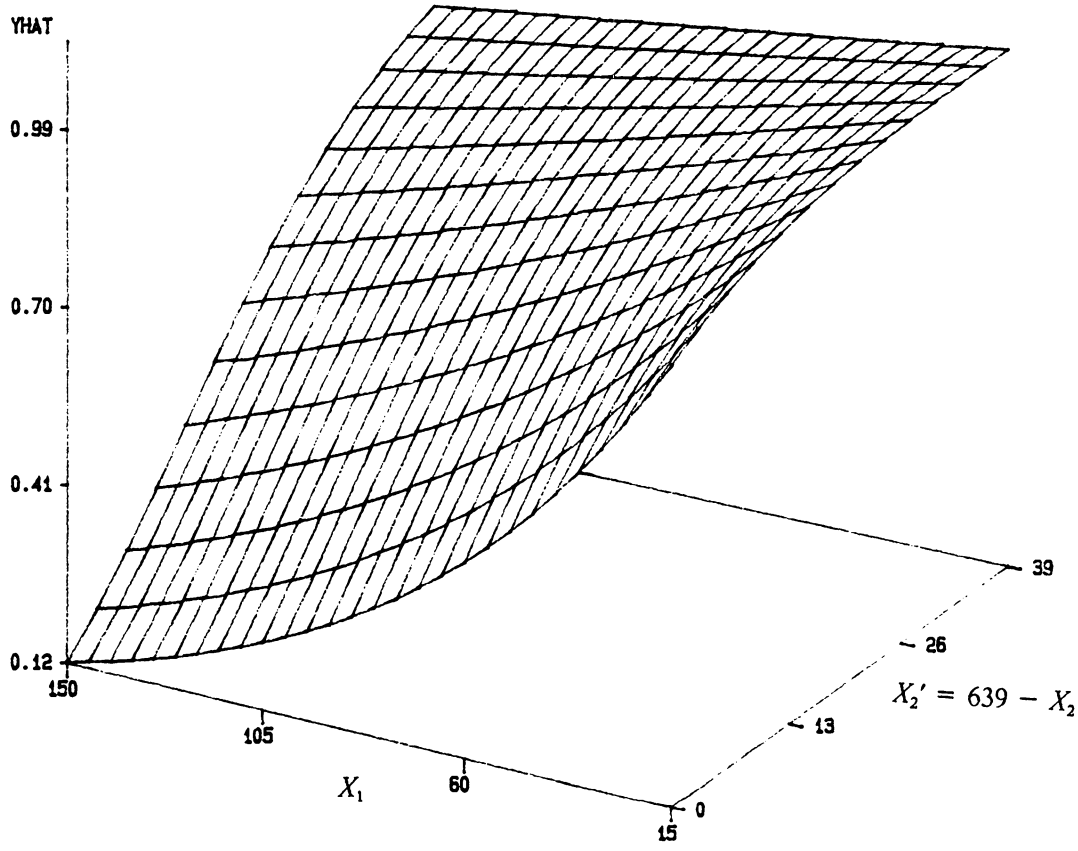


FIGURE V.2.4. PLOT OF THE NONLINEAR REGRESSION FIT FOR EXAMPLE #2.

the first order model. If so, then the HATLINK fit based on the first order user's model would have provided a much improved fit compared to least squares.

However, the F^* test based on the HATLINK fit with the first order user's model should lead the user to try a more complicated model. If, then, the full second order model is employed, the least squares fit is much improved. However, the HATLINK fit based on this model provides even further improvement. In fact, using the HATLINK method in conjunction with a second order user's model gives a fit that is very close to the fit obtained through nonlinear regression for the model in Equation V.2.1. This demonstrates that in cases where the user is unaware of any theory which would dictate a particular model, the HATLINK regression method performs well. A further benefit is that the HATLINK method directly provides estimates of error variance and confidence intervals.

Chapter VI

VI. SUMMARY AND AREAS FOR FURTHER RESEARCH

VI.1. Overview of the Performance of the HATLINK Regression Method

There are many situations which arise in practice where the regression user does not know for certain the correct parametric form of the true underlying model, but can at least prescribe a reasonable approximation to this model. The present investigation has focused on such situations by concentrating on examples where the usual lack of fit test for the user's specified model has low power. In such cases, neither the least squares approach nor the approach taken by the kernel method is entirely appropriate. The least squares fit and subsequent inferences are inexorably tied to the particular model form prescribed by the user. On the other hand, the kernel fit is obtained on the basis of local information only, without any consideration given to the approximate model

that the user has been able to specify. It is for such sittuations that the HATLINK has been developed as a compromise between the least squares and kernel methods.

The simulation results presented in Chapter IV confirm that the HATLINK method provides improved prediction in situations where the user is uncertain of the true underlying model, but can prescribe an approximate model. For each of the families of regressions examined through simulations, a similar pattern of prediction performance was observed for the least squares, kernel, and HATLINK methods. When the user's postulated model was a close approximation to the true underlying function, then the kernel method, using only local information, provided relatively poor prediction. When the user's model was misspecified to a moderate degree, then the least squares predictions were generally poor. Predictions obtained through HATLINK, however, were competitive in all cases considered. Specifically, a strong performance by the HATLINK method was observed for several different types of underlying functions and several different forms for the user's model, for different error variances, for different sample sizes, and for different sets of regressor locations. These positive results should recommend HATLINK as a "model-robust" method for obtaining predictions whenever the true underlying model is not known with certainty.

In addition to providing improved predictions, a number of the usual tasks of regression analysis can be accomplished through HATLINK. The methods for accomplishing these tasks were developed in Chapter III and their behavior studied through simulations in Chapter IV. In particular, the estimate of error variance obtained through HATLINK (Section III.4.A) performed quite well in the empirical studies. (See Tables IV.2.5, IV.3.5, IV.3.12, IV.4.3, IV.4.6, and IV.4.9.) In cases where the user's prescribed model was correct, the HATLINK estimate of σ^2 rivaled the least squares estimate in terms of both accuracy and precision. In the more common situation where the user's model is off somewhat, the HATLINK estimate of variance was more accurate than the least squares estimate.

Confidence intervals for the mean value of the true model at a given regressor location were also investigated through simulations in Chapter IV. (See Tables IV.2.6, IV.2.7, IV.3.6, IV.3.13, IV.4.4, IV.4.7, and IV.7.10.) A number of cases were observed where the usual 95% confidence intervals based on least squares suffered a substantial loss of coverage. In these instances, 95%

confidence intervals formed through HATLINK (Section III.4.B) offered much improved coverage, though sometimes falling below the desired 95% level. The advantage of improved coverage through the HATLINK intervals is somewhat offset by the fact that these intervals tend to be somewhat wider than the corresponding least squares intervals.

Another task of regression analysis that is addressed through HATLINK is testing the user's model for lack of fit. Both the empirical test based on λ (Section III.4.D) and the approximate F test using the F^* statistic (Section III.4.D) have performed quite well in diagnosing lack of fit. (See Tables IV.3.9, IV.3.10, IV.4.11, IV.4.12, and IV.4.16.) These tests have the advantage that they can be applied when there are no replicates at any of the regressor locations. Even when there are replicate observations, the simulation studies show that these tests can be be considerably more powerful than the usual F test for lack of fit. A further benefit of the HATLINK lack of fit tests was observed in Example #1 and Example #2 of Chapter V. In each of those examples it was shown how the F^* test can be useful in leading the researcher toward a more suitable model. Since such diagnostic information is provided through HATLINK, even a researcher who insists upon a final parametric model with estimated regression coefficients can be assisted through the use of HATLINK.

The issue of variable selection can also be approached through HATLINK. (See Section III.4.E.) As an illustration of how this can be done, one particular situation was considered through an empirical study in Section IV.4.D. There it was observed how HATLINK versions of quantities such as R^2 , s^2 , and PRESS, and pseudo-F tests based on HATLINK can be used to determine, in a model-robust fashion, whether a variable should be included in a regression model. (See Tables IV.4.13 - IV.4.16.)

Selection of λ

In order for HATLINK to achieve the correct compromise between the least squares and kernel fits for a given set of data, it is necessary to select the proper value of the mixing parameter λ . Several data-adaptive criteria for choosing λ were developed in Section III.3.B. The performance of these criteria was given considerable attention in the empirical studies reviewed in Chapter IV. Based on these empirical results, several recommendations can be made. For situations where the user is confident that the specified model is a good approximation to the true underlying model, the $Cp3(\lambda)$ and $PRESS^*$ criteria are recommended, since these methods tend to select conservatively small λ's . If the user is aware that the specified model is at best a crude approximation to the true model, then either the $PRESS(\lambda)$, $Cp1(\lambda)$, or $Cp4(\lambda)$ criterion should be used to select λ . For general purposes, the $Cp3(\lambda)$ is recommended on the basis of its good performance for each of the families of regressions considered in the empirical study.

VI.2. Insights into the Least Squares Method

An interesting byproduct of the empirical ivestigations has been the obervation of the performance of the least squares predictions, variance estimate, and confidence intervals under a modest degree of model misspecification. While it is not surprising that the least squares inferences should suffer somewhat when the user's model is incorrect, the extent to which they suffer and the fact that the damage often occurs when when the user's model is only slightly off, is perhaps surprising. (See, for example, Tables IV.2.5, and IV.2.6.)

VI.3. Potential Improvements

It is possible that some improvement in the HATLINK method could be achieved by using some other criterion for selecting the mixing parameter λ . However, the greatest potential for improving HATLINK lies in the kernel component. In the present study, a very basic kernel procedure has been incorporated in order to keep computational expenses to a reasonable level. However, the HATLINK procedure could be improved through the use of one of the variations on the kernel method listed in Chapter II. In particular, the use of a local bandwidth kernel (Section II.3.A) should improve both the kernel and HATLINK predictions.

VI.4. Areas for Further Research

The chief area for future work in conjunction with HATLINK is the multiple regression setting. The present investigation has made use of a very simple kernel approach based on Euclidean distances among the scaled regressors. Such an approach would be inadequate for cases where the response varies rapidly in one regressor, but slowly in another. For such cases, the use of a single bandwidth parameter would be inadequate. Rather, it would be desirable to use a narrower bandwidth to accommodate one variable and a wider bandwidth for the other variable. An approach of this type is taken in the kernel method used by Staniswalis, McCrady, Carter, Campbell, and Carchman (1987), but it is not clear that their approach is the best that is possible. Moreover, much more work needs to be done in terms of developing the multiple regression version of the kernel method, and in studying the behavior of competing approaches through simulations. Any improvement achieved for the kernel method should, in turn, lead to an improvement in HATLINK.

Also worthy of future investigation is the problem of variable selection through HATLINK. More empirical studies are needed in order to firmly establish the HATLINK approach to variable selection. Also, it would be beneficial to develop a computer routine to deal with the variable selection issue through HATLINK when there is a set of several candidate regressor variables.

Another issue regarding HATLINK that needs to be considered is that of experimental design. In situations where the researcher can select the regressor locations, some guidelines are needed as to a strategy for selecting these locations. For the single regressor case, Müller (1984) presents asymptotically optimal designs for a version of the kernel estimator under certain conditions. However, much more work regarding the design issue remains to be done for the kernel method, as well as for HATLINK.

For a given set of data it may be desirable to estimate the location of the optimal (maximum or minimum) value of the mean response function over a given region. Staniswalis, McCrady, Carter, Campbell, and Carchman (1987), have done this using kernel regression, and their approach could be extended to HATLINK. Using HATLINK in this regard should be an improvement, particularly for small to moderate sized data sets.

One final area of potential future investigation is that of diagnosing influential observations and outliers. Mehods for detecting outliers and influential points for linear regression are presented in Belsley, Kuh, and Welsch (1980), Cook and Weisberg (1982), and Myers (1986). Some of these regression diagnostics have been adapted for spline regression by Eubank (1984, 1985). A number of related diagnostics for outliers and influential observations were proposed by Einsporn (1986), but have not been investigated, thus remaining in the category of "partly baked" ideas (Good, 1985). Further work is needed to refine the forms of these diagnostic measures and to establish their usefulness in practice.

Chapter VII

VII. REFERENCES

Allen, D. A. (1974) "The Relationship Between Variable Selection and Data Augmentation and a Method for Prediction," *Technometrics* , 16, 125-127.

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

Benedetti, J. K. (1975) "Kernel Estimation of Regression Functions," *Procedings of the Computer Science and Statistics Eighth Annual Symposium on the Interface* , 405-408.

Butler, G. A. (1975) "Heuristic Regression for Large Commercial Problems," *Procedings of the Computer Science and Statistics Eighth Annual Symposium on the Interface* , 398-404.

Cheng, K. F., and Lin, P. E. (1981) "Nonparametric Estimation Function," *Z. Wahr. und Verw. Gebiete* , 57, 223-233.

Cleveland, W. S. (1979) "Robust Locally Weighted Regression and Smoothing Scatterplots," *J. Amer. Statistical Assoc.* , 74, 828-836.

Cleveland, W. S., and Devlin, S. J. (1986) "Locally Weighted Regression: An Approach to Regression Analysins by Local Fitting," AT&T Bell Laboratories Statistical Research Report.

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

Cover, T. M. (1968) "Estimation by the Nearest Neighbor Rule," *I.E.E.E. Transactions on Information Theory* , 14, 50-55.

Cover, T. M., and Hart, P. E. (1967) "Nearest Neighbor Pattern Classification," *I.E.E.E. Transactions on Information Theory* , 13, 21-27.

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

Denby, L. (1987) "Introduction to Regression Fitting Strategy," tutorial presented at the 1987 Winter Conference of the American Statistical Association, Jan. 7, 1987.

Draper, N. R., and Smith H. (1981) *Applied Regression Analysis* , second edition, Wiley, New York.

Einsporn, R. L. (1986) "A Link Between Least Squares and Nonparametric Curve Estimation," dissertation proposal, Virginia Polytechnic Institute and State University.

Eubank, R. L. (1984) "The Hat Matrix for Smoothing Splines," *Stat. and Prob. Letters* , 2, 9-14.

Eubank, R. L. (1985) "Diagnostics for Smoothing Splines," *J. Royal Stat. Soc. B* 47, 332-341.

Gasser, T., and Müller, H. G. (1979) "Kernel Estimation of Regression Functions" in *Smoothing Techniques for Curve Estimation* (T. Gasser and M. Rosenblatt, eds.), 23-68 Springer-Verlag, Heidelberg.

Gasser, T., and Müller, H. G. (1984) "Estimating Regression Functions and Their Derivatives by the Kernel Method," *Scand. J. of Statistics* , 11, 197-211.

Geisser, S. (1975) "The Predictive Sample Reuse Method with Applications," *J. Amer. Stat. Assoc.,* 70, 320-328.

Georgiev, A. A., and Greblicki, W. (1986) "Nonparametric Function Recovering from Noisy Data," *J. Stat. Planning and Inference* , 13, 1-14.

Good, I. J. (1985) "Partly Baked Ideas: A Souffle of Ideas on the Edge of Collapse," *CHEMTECH,* 200-202.

Hastie, T., and Tibshirani, R. (1986) "Generalized Additive Models" (with discussion), *Statistical Science* , 1, 297-318.

Hastie, T., and Tibshirani, R. (1987) "Generalized Additive Models: Some Applications," *J. Amer. Statistical Assoc.* , 82, 371-386.

Hoaglin, D. C., and Welsch, R. (1978) "The Hat Matrix in Regression and ANOVA," *Amer. Statistician* , 32, 17-22. 12, 55-67.

Krutchkoff, R. G. (1987) *KANOVA* , diskette, R. G. Krutchkoff, Virginia Polytechnic Institute and State University.

Li, K. C. (1984) "Consistency for Cross-validated Nearest Neighbor Estimates in Nonparametric Regression," *Annals of Statistics* , 12, 230-240.

Mallows, C. L. (1973) "Some Comments on Cp," *Technometrics* , 15, 661-675.

Montgomery, D. C., and Peck, E. A. (1982) *Introduction to Linear Regression Analysis* , Wiley, New York.

Müller, H. G., (1984) "Optimal Designs for Nonparametric Kernel Regression," *Statistics and Probability Letters* , 2, 285-290.

Müller, H. G., (1987) "Weighted Local Regression and Kernel Methods for Nonparametric Curve Fitting," *J. Amer. Statistical Assoc.* , 82, 371-386.

Müller, H. G., and Stadtmüller, U. (1985) In discussion of Silverman, *J. Royal Stat. Soc. B* , 47, 39.

Müller, H. G., and Stadtmüller, U. (1987) "Variable Bandwidth Kernel Estimators of Regression Curves," *Annals of Statistics* , 15, 182-201.

Marriot, F. H. C. (1985) In discussion of Silverman, *J. Royal Stat. Soc. B* , 47, 24.

Myers, R. H. (1986) *Classical and Modern Regression with Applications* , Duxbury, Boston.

Nadaraya, E. A. (1964) "On Estimating Regression," *Theory of Prob. and its Applications* , 9, 141-142.

Priestley, M. B., and Chao, M. T. (1972) "Nonparametric Function Fitting," *J. Royal Stat. Soc. B.* , 34, 384-392.

Reinsch, C. (1967) "Smoothing by Spline Functions," *Numerische Mathematik* , 10, 177-183.

Rice, J. (1984a) "Boundary Modification for Kernel Regression," *Communications in Statistics, Ser. A. -- Theory and Methods,* 13, 893-900.

Rice, J. (1984b) "Bandwidth Choice for Nonparametric Regression," *Annals of Statistics* , 12, 1215-1230.

SAS Institute Inc. (1982) *SAS User's Guide: Statistics,* 1982 edition, SAS Institute, Cary, North Carolina.

Silverman, B. W. (1984) "A Fast and Efficient Cross-validation Method for Smoothing Parameter Choice in Spline Regression," *J. Amer. Stat. Assoc.* , 79, 584-589.

Silverman, B. W. (1985) "Some Aspects of the Spline Smoothing Approach to Nonparametric Regression Curve Fitting (with discussion)," *J. Royal Stat. Assoc. B* , 47, 1-53.

Staniswalis, J.G., McCrady, C. W., Carter, W. H., Jr., Campbell, E. D., and Carchman, R. A. (1987) "The Use of Kernel Estimators in the Estimation and Exploration of a Response Surface," (submitted to *Technometrics* .)

Stone, M. (1974) "Cross-validatory Choice and Assessment of Statistical Predictions" (with Discussion), *J. Royal Stat. Soc. B.* , 36, 111-147.

Stone, C. J. (1977) "Consistent Nonparametric Regression"(with Discussion), *Annals of Statistics,* 5, 595-645.

Tibshirani, R., and Hastie, T. (1987) "Local Likelihood Estimation," *J. Amer. Statistical Assoc.* , 82, 559-567.

Watson, G. S. (1964) "Smooth Regression Analysis," *Sankhya Ser. A* , 26, 359-372.

Wong, W. H. (1983) "On the Consistency of Cross-validation in Kernel Nonparametric Regression," *Annals of Statistics* , 11, 1136-1141.

Appendix A

A. Empirical Results for Alternative Methods of Bandwidth Selection

A.1. Nonstochastic Kernel Approach

As was mentioned in Section III.3.A, it is sometimes advantageous to allow the user of kernel regression to select the bandwidth. Rather than arriving at the bandwidth automatically through minimizing PRESS, for example, the user may want to choose the bandwidth that corresponds to a certain value for the kernel degrees of freedom. By selecting a bandwidth that corresponds to a sensible number of degrees of freedom, the problem of possible overfitting (or underfitting) by the kernel may be avoided to some extent. In preliminary empirical investigations it was found that the fits obtained through the kernel method were generally better when the model degrees of freedom was set at about 5 to 7 for the single regressor case. This section presents results of a series of simulations for the sine wave family where the bandwidth was predetermined so that the kernel fit would have six degrees of freedom.

Table A.1.1. Prediction Performance of the Least Squares, Kernel and HATLINK methods for the Sine Wave Family of Regressions. Here the kernel bandwidth is pretermined so that the kernel will have 6 df. Values reported are based on 25 simulations at each level of the amplitude A. Results shown for HATLINK are for the Cp3 version.

		Mean SSEP				Relative Prediction Efficiency		
A	POWER	LS	KER	H		RPE(LS,K)	RPE(H,K)	RPE(H,LS)
0.0	.05	20.8	53.9	21.8	(+K)	2.60	0.95	2.47
1.0	.06	23.4	53.6	23.7	(+K)	2.29	0.99	2.26
2.0	.08	33.0	53.4	31.6	(+K)	1.62	1.04	1.69
3.0	.11	49.6	53.5	44.6	(+K,)	1.08	1.15	1.23
3.5	.14	60.5	53.6	49.5	(+L)	0.89	1.22	1.08
4.0	.18	73.1	53.7	55.8	(+L)	0.74	1.31	0.96
5.0	.26	103.6	54.2	67.0	(+L,-K)	0.52	1.55	0.81
6.0	.38	141.0	54.9	75.4	(+L,-K)	0.39	1.87	0.73

Table A.1.2. Prediction Performance of the HATLINK Method for the Six Criterion for Selecting the Mixing Parameter for the Sine Wave Family of Regressions. The kernel bandwidth was predetermined so that the kernel fit would have six degrees of freedom. Results are based on 25 runs at each level of the amplitude A.

		Mean Sum of Squared Errors of Prediction									
A	POWER	LS	KER	PRESS	PRESS*	Cp1	Cp2	Cp3	Cp4	OPT	lsd
0.0	.05	20.8	53.9	26.9	22.9	21.6	53.9	21.8	24.8	20.3	6.5
1.0	.06	23.4	53.6	29.4	24.4	23.4	53.4	23.7	26.6	21.4	7.1
2.0	.08	33.0	53.4	36.6	32.0	31.5	53.3	31.6	33.7	26.2	7.8
3.0	.11	49.6	53.5	44.6	43.8	44.6	54.9	43.3	41.6	31.7	8.8
3.5	.14	60.5	53.6	48.9	50.0	51.2	53.6	49.5	45.9	34.4	9.4
4.0	.18	73.1	53.7	53.4	56.3	57.9	51.9	55.8	50.6	37.0	10.1
5.0	.26	103.6	54.2	57.4	66.6	69.1	58.8	67.0	57.3	41.7	12.1
6.0	.38	141.0	54.9	60.0	72.1	77.9	68.1	75.4	61.1	45.9	14.3

Tables A.1.1 and A.1.2 provide a summary of the prediction performances of the kernel and least squares methods, and for the six versions of HATLINK. The values displayed in these tables are based on 25 runs each, with the same 25 sets of random errors used in each case. The set-up (data locations, error variance, underlying function) is the same as for the basic sine wave family presented in Section IV.3.A. Comparing Table A.1.1 to Table IV.3.2, it is seen that the mean SSEP's for the kernel method with df fixed at 6 are somewhat lower than the corresponding values for the earlier runs. It is perhaps surprising that at A = 0 the mean SSEP for kernel is slightly lower with the df fixed at 6 than for the stochastic version of kernel, which has the opportunity of using fewer degrees of freedom to fit the underlying straight line in that case. However, at A = 0 kernel method with 6 df is still badly outperformed by the least squares and HATLINK methods.

Table A.1.2 shows that the six criteria for selecting λ in HATLINK behave differently when the bandwidth is predetermined. (See Table IV.3.4.) In this case, the Cp1 version yielded the best prediction at amplitudes A = 0.0, 1.0, and 2.0, while the Cp4 method was the best at A = 3.0, 3.5, 4.0, and 5.0. This is not surprising, since the Cp4 criterion was designed to work well in conjunction with kernels having high degrees of freedom. (See Appendix B.2.) Moreover, the good performance of Cp4 across the entire range of amplitudes suggests that this version of HATLINK be recommended for use whenever a predetermined bandwidth corresponding to a fairly high value for the kernel degrees of freedom is employed.

Diagnosing lack of fit with a nonstochastic kernel

When a kernel fit with a certain preselected number of degrees of freedom is used for the simulations, the statistics such as $F^{\bullet}$ and $F^{\bullet\bullet}$ become more useful for diagnosing lack of fit. (This approach was discussed previously in Section III.4.D.) For this type of kernel, these pseudo-F statistics formed using λ = 1 will be based on the same number of degrees of freedom for every simulation. To illustrate the sensitivity of the $F^{\bullet\bullet}$ to lack of fit of the user's model, results of an empirical power study are presented in Table A.1.3. Critical values for $F^{\bullet\bullet}$ and the usual F test for lack of fit were based on the 25 runs at the null hypothesis (A = 0). Recall that the subsequent sets of 25 runs at the other levels of A were based on the same 25 sets of random errors. While the scale

of this study is small, the results do confirm that F^{**} , using $\lambda = 1$, is somewhat more sensitive to lack of fit of the sine wave variety than is the usual F test.

Table A.1.3. Percent Rejections at Empirical $\alpha = .05$ Critical Values for Usual Lack of Fit F Statistic and the F^{**} Statistic Based on a 6 df Kernel Fit. Results are for the sine wave family of regressions with 25 runs made for each value of the amplitude A. Also included are the theoretical values of the power of the usual F test.

	Amplitude							
Statistic	0.0	1.0	2.0	3.0	3.5	4.0	5.0	6.0
F^{**}	4	4	4	8	16	28	52	68
F	4	0	0	4	12	16	24	36
(Theor.)	5	6	7	11	14	18	26	38

A.2. Bounded Kernel Degrees of Freedom Approach

A simulation analysis for the sine wave family was performed where the kernel degrees of freedom was required to be at least two but no more than six. As was done previously, the function PRESS*(*h*) (equation III.1.6) was minimized over h. If, however, the minimizing value of h resulted in a kernel with df below 2 (or above 6), then the bandwidth corresponding to 2 df (or 6 df) was used for the kernel fit. The bounds 2 and 6 were used in an attempt to avoid either underfitting or extreme overfitting with the kernel. (See Section III.3.A for a discussion of the bounded kernel approach.)

Prediction results for this bounded kernel approach are shown in Table A.2.1. Comparing the mean SSEP's in this table to those in Table A.1.1 for the predetermined 6 df kernel method, it is seen that the prediction performance for the two kernels is about the same at every level of A considered. Predictions obtained by HATLINK based on the bounded degrees of freedom approach are worse than for the 6 df kernel in most cases. (See Table A.1.2.) Notable exceptions are for the Cp2 method, which does better at lower amplitudes, and for all HATLINK versions at the higher amplitudes A = 5.0 and A = 6.0. However, given a choice between the bounded kernel, the predetermined kernel, and the original stochastic kernel, Table A.2.1 provides little reason for recommending the bounded kernel method for use in conjunction with HATLINK, unless the user is certain that the prescribed model is at least moderately inaccurate.

Table A.2.2 summarizes the performance kernel and various versions of HATLINK for the stochastic, fixed, and bounded kernel approaches. Specifically, this table compares the performance of each prediction method for the kernel approach that appears to work the best for the particular method. For example, results for the Cp4 version of HATLINK are presented for the fixed df kernel approach, since the performance of Cp4 was generally better when the kernel df was fixed at 6, than it was when the kernel bandwidth was chosen stochastically with or without bounds on the model degrees of freedom. From this table it is observed, for the sine wave family of underlying

Table A.2.1. Prediction Performance of the Least Squares and Kernel Methods and for the Six Versions of the HATLINK Procedure. Results for are for the bounded kernel df approach, where the kernel was fit under the constraint that the degrees of freedom were between 2 and 6. 25 runs were performed at each of four levels of the amplitude A for the sine wave family of regressions.

		Mean Sum of Squared Errors of Prediction									
A	POWER	LS	KER	PRESS	PRESS'	Cp1	Cp2	Cp3	Cp4	OPT	lsd
0.0	.05	20.8	52.9	30.6	24.7	24.2	42.0	24.1	28.6	19.9	7.2
2.0	.08	33.0	52.8	42.2	36.6	36.7	40.9	36.3	39.9	26.2	7.6
3.5	.14	60.5	53.3	51.7	54.0	52.6	51.1	52.4	50.4	35.7	8.6
5.0	.26	103.6	54.5	54.9	61.9	61.1	72.3	61.3	54.4	43.6	10.9

Table A.2.2. Summary of the Best Prediction Performances Obtained for the Sine Wave Family of Regressions by the Kernel Method and by Six Versions of the HATLINK Procedure. Results indiicated by (1) are those obtained for the original version of the kernel where the bandwidth is not constrained. Results noted (2) are based on the predtermined bandwidth kernel with 6 df, and those marked (3) are for kernel fits with df restricted to fall between 2 and 6. For the kernel method and each version of HATLINK, results are presented for the specific type of kernel, (1), (2), or (3), which led to the best overall predictions for that method.

			Mean Sum of Squared Errors of Prediction						
A	POWER	LS	KER (3)	PRESS (2)	PRESS* (1)	Cp1 (3)	Cp2 (1)	Cp3 (1)	Cp4 (2)
0.0	.05	20.8	52.9	26.9	24.5	24.2	40.6	22.9	24.8
3.5	.14	60.5	53.3	48.9	48.2	52.6	50.1	46.5	45.9
5.0	.26	103.6	54.5	57.4	59.1	61.1	56.6	58.6	57.2

models, that the Cp3 method using a stochastic kernel and the Cp4 method based on the fixed df kernel provide the best overall prediction.

Appendix B

B. Some Mathematical Considerations

B.1. Bias Term in Equation III.3.5

It is shown that the bias term in equation III.3.5 can be expressed as in equation III.3.7:

$$\frac{1}{\sigma^2} \sum_{i=1}^{n} [\text{ Bias } \hat{Y}(X_i)]^2 = \frac{1}{\sigma^2} \{ E(SSE) - \sigma^2 tr[(I - H)'(I - H)] \} . \quad \text{(III.3.7)}$$

The derivation will begin by considering the expectation of SSE, the sum of squared errors associated with the fitted values $\hat{Y}_i$.

$$E(SSE) = E \sum_{i=1}^{n} (Y(X_i) - \hat{Y}(X_i))^2$$

$$= E[(\underline{Y} - H\underline{Y})'(\underline{Y} - H\underline{Y})]$$

$$= E[\underline{Y}'(I - H)'(I - H)\underline{Y}]$$

$$= \sigma^2 tr[(I - H)'(I - H)] + E(\underline{Y})'(I - H)'(I - H)E(\underline{Y})$$

$$= \sigma^2 tr[(I - H)'(I - H)] + [E(\underline{Y}) - E(H\underline{Y})]'[E(\underline{Y}) - E(H\underline{Y})]$$

$$= \sigma^2 tr[(I - H)'(I - H)] + \sum_{i=1}^{n} [E(Y(X_i)) - E(\hat{Y}(X_i))]^2$$

$$= \sigma^2 tr[(I - H)'(I - H)] + \sum_{i=1}^{n} [\text{Bias}(\hat{Y}(X_i))]^2 \ .$$

Solving for the Bias term and dividing through by σ^2 then yields equation III.3.7.

B.2. Limiting Behavior of Cp4

To consider the behavior of the $Cp4(\lambda)$ criterion when the kernel has a small bandwidth h, consider the limiting case where $h \to 0$. In this case, the kernel prediction at any given X location is the mean of the observations at that location. That is, the hat matrix H_{ker} for kernel is the same as the matrix H_m used for the standard F test for lack of fit. (See equation III.4.8.) If the data points are ordered so that replicate observations are grouped together, then this matrix is block diagonal, with blocks corresponding to the groups of replicates. Moreover, the matrix H_m is symmetric, idempotent, and has the property that $H_m H_{ols} = H_{ols}$.

For this limiting situation, it is now shown that the value of λ found through minimizing $Cp4(\lambda)$ is related to the usual F statistic for detecting lack of fit. First, the quantity $Cp4(\lambda)$ (equation III.3.16) can be rewritten as follows.

$$\begin{aligned} Cp4(\lambda) &= 2(tr[H(\lambda)] - p) + \frac{(SSE(\lambda) - SSE_{ols})(n - tr[H_{\text{ker}}])}{SSE_{\text{ker}}} \\ &= 2\,tr[H(\lambda)] - 2p + \frac{SSE(\lambda)(n - tr[H_{\text{ker}}])}{SSE_{\text{ker}}} - \frac{SSE_{ols}(n - tr[H_{\text{ker}}])}{SSE_{\text{ker}}} . \end{aligned}$$

Next, through differentiating with respect to λ , the value of λ which minimizes $Cp4(\lambda)$ is determined.

$$\frac{\partial\, Cp4(\lambda)}{\partial \lambda} = 2\frac{\partial}{\partial \lambda}(tr[H(\lambda)]) + \frac{(n - tr[H_{\text{ker}}])}{SSE_{\text{ker}}} \frac{\partial}{\partial \lambda}[SSE(\lambda)] . \qquad \text{(B.2.1)}$$

The derivative of the term $tr[H_{\text{ker}} - H_{ols}]$ in the above expression is

$$\begin{aligned} \frac{\partial}{\partial \lambda}\{tr[H(\lambda)]\} &= \frac{\partial}{\partial \lambda}\{\lambda tr[H_{\text{ker}} - H_{ols}] + tr[H_{ols}]\} \\ &= tr[H_{\text{ker}} - H_{ols}] . \end{aligned} \qquad \text{(B.2.2)}$$

Before differentiating $SSE(\lambda)$, note that $SSE(\lambda)$ can be rewritten as

$$\begin{aligned} SSE(\lambda) &= \underline{Y}'[I - H(\lambda)]'[I - H(\lambda)]\underline{Y} \\ &= \underline{Y}'[I - \lambda(H_{\text{ker}} - H_{ols}) + H_{ols}]'[I - \lambda(H_{\text{ker}} - H_{ols}) + H_{ols}]\underline{Y} \\ &= \underline{Y}'[I + \lambda^2(H_{\text{ker}} - H_{ols})'(H_{\text{ker}} - H_{ols}) + 3H_{ols} - \lambda(H_{\text{ker}} - H_{ols})' \\ &\quad - \lambda(H_{\text{ker}} - H_{ols}) - \lambda(H_{\text{ker}} - H_{ols})'H_{ols} - \lambda H_{ols}(H_{\text{ker}} - H_{ols})]\underline{Y} \\ &= \underline{Y}'[I + \lambda^2(H_{\text{ker}} - H_{ols}) + 3H_{ols} - 2\lambda(H_{\text{ker}} - H_{ols})]\underline{Y} \ . \end{aligned}$$

The last equality above follows from the fact that H_{ker} is symmetric, idempotent, and $H_{\text{ker}}H_{ols} = H_{ols}$ for this limiting case. Then,

$$\frac{\partial}{\partial\lambda}\{SSE(\lambda)\} = 2\underline{Y}'[\lambda(H_{\text{ker}} - H_{ols}) - (H_{\text{ker}} - H_{ols})]\underline{Y} \ . \qquad \text{(B.2.3)}$$

Substituting the derivatives obtained in equations B.2.2 and B.2.3 into equation B.2.1, the derivative of Cp4 can be expressed as

$$\frac{\partial\, Cp4(\lambda)}{\partial\lambda} = 2\,tr[H_{\text{ker}} - H_{ols}] + 2\frac{(n - tr[H_{\text{ker}}])}{SSE_{\text{ker}}}\underline{Y}'[(\lambda - 1)(H_{\text{ker}} - H_{ols})]\underline{Y} \ .$$

This derivative is equal to zero when

$$\begin{aligned} \lambda &= 1 - \frac{tr[H_{\text{ker}} - H_{ols}]SSE_{\text{ker}}}{(n - tr[H_{\text{ker}}])\underline{Y}'(H_{\text{ker}} - H_{ols})\underline{Y}} \\ &= 1 - \frac{\underline{Y}'(I - H_{\text{ker}})\underline{Y} \,/\, (n - tr[H_{\text{ker}}])}{\underline{Y}'(H_{\text{ker}} - H_{ols})\underline{Y} \,/\, tr[H_{\text{ker}} - H_{ols}]} \\ &= 1 - \frac{1}{F_{LOF}} \ , \end{aligned}$$

where F_{LOF} denotes the test statistic for the usual regression test for lack of fit (equation III.4.8). Thus, the value of λ selected through the $Cp4(\lambda)$ criterion is, in this limiting case, an increasing

function of the usual F statistic for detecting lack of fit. For situations where the kernel bandwidth is small, the Cp4 criterion therefore should perform in an appropriate manner in selecting a suitable value for λ .

Appendix C

C. Computing Considerations

For a given set of data, predicted values at the regressor locations can be obtained for HATLINK very quickly through SAS PROC MATRIX (SAS Institute, 1982). The following steps are required to obtain the HATLINK predictions using the PRESS*(h) criterion (equation III.1.6) for the kernel bandwidth and the $Cp3(\lambda)$ criterion (equation III.3.15) to select the mixing parameter λ.

(1) Input the data.

(2) Define the user's parametric model.

(3) Obtain the least squares hat matrix for that model.

(4) Obtain the kernel hat matrix.

  (a) Create a matrix of distances among the regressor locations.

  (b) Obtain $H_{ker}(h)$, the kernel hat matrix, as a function of the distance matrix and bandwidth h.

  (c) Search for h so that $H_{ker}(h)$ provides the minimum value of PRESS*(h).

(5) Obtain the HATLINK hat matrix.

(a) Obtain $H(\lambda)$ for $\lambda = 0, .5,$ and 1.0 . These determine the actual quadratic function $H(\lambda)$.

(b) Determine the minimum value of $Cp3(\lambda)$ based on the quadratic function $H(\lambda)$.

(6) Output the predicted values, $\hat{\underline{Y}} = H(\lambda)\underline{Y}$.

The following excerpts from the PROC MATRIX program show, in more detail, the steps required to obtain the kernel and HATLINK hat matrices for the single regressor case.

OBTAINING THE KERNEL HAT MATRIX

```
D = (J(N,1,1)*X')-(X*J(1,N,1));
  * D is an N by N matrix of distances among the regressor locations;
  * X is the vector of N regressor locations and J is a vector of ones;
DSQ = D##2;        * Matrix of squared distances among the X points;
DSQSUM = DSQ(+,+); * Sum of all entries in DSQ
                     (used as a scaling constant);
```

```
**** Set Initial Values in the Search for Optimal ALPHA ****;
    ** ALPHA is a function of the bandwidth h ** ;
```

```
INIT = 5/15/25;
AL = .05#INIT#N##3;
```

```
**** Loop To Obtain Kernel Predictions and Values for
        the PRESS Statistic for the 3 Initial Bandwidths ****;

DO A=1 TO 3;
ALPHA=AL(A,);
DINT=(-1)#ALPHA#(DSQ#/(DSQSUM#J(N,N,1)));
      * Obtain exponent in Equation III.1.5 by scaling the distance
                  matrix by the bandwidth;
DO I=1 TO N; DO K=1 TO N;
  EXPD(I,K)=EXP(DINT(I,K));   * Compute the numerator of the normal
density type kernel function (Equation III.1.5) at each data location;
  END; END;
ESUM=EXPD(,+);        * Row sum of numerator values;
HK=EXPD#/(ESUM*J(1,N,1)); * Kernel hat matrix obtained by scaling the
                  numerator values so that row sums are one;
YHATK=HK*Y;     * Vector of kernel predictions at the data locations;
YHKNOEND=YHATK(2:9,)//YHATK(10:19,); *Kernel predictions w/o endpts;


******** Compute the PRESS Statistic for the Kernel Fit *****;
* Compute the minus i version of the kernel hat matrix HK;
HKMI=HK;       * Set minus i hat matrix equal to kernel hat matrix;
DO I=1 TO N;
  HKMI(I,I)=0; * Set diagonal entries of minus i hat matrix to zero;
  END;
HSUM=HKMI(,+); * Obtain revised row sums of minus i hat matrix;
HKMI=HKMI#/(HSUM*J(1,N,1)); * Mimus i hat matrix obtained by
                  rescaling entries to make row sums equal 1;
YHATKMI=HKMI*Y;       * Vector of minus i predictions;
YHKMINOE=YHATKMI(2:9,)//YHATKMI(10:19,); * Set aside endpoints;
PRSS(A,)=(YNOEND-(YHKMINOE))'*(YNOEND-(YHKMINOE))#/(N-TRACE(HK));
      *PRESS statistic adjusted for error df excluding endpoints;
```

```
        * See Equation III.1.6;
      *YNOEND is the vector of observations Y excluding the endpts;
END;
```

```
***** Second Loop Searches for the Bandwidth to Minimize PRESS *****;
```

```
DO ITER=1 TO 8;    * Eight iterations are used;
 MINI=MIN(PRSS);          * Find minimum value of PRESS for the
                          3 runs in the initial loop above
**** Determine what value of ALPHA (bandwidth) to try next ***;
* Then reset the initial vector of 3 ALPHA's accordingly;
 IF MINI=PRSS(1,) THEN DO;
   AMIN=1;
   SET1=AL(1,);
   SET2=2#AL(1,)-AL(2,);
   SET3=AL(1,)#/2;
   VEC=SET3//SET2;
   AL(1,)=MAX(VEC);
   AL(3,)=AL(2,);
   AL(2,)=SET1;
   ADD=AL(1,);
   END;
 IF MINI=PRSS(2,) THEN DO;
   AMIN=2;
   MAXI=MAX(PRSS);
    IF MAXI=PRSS(1,) THEN DO;
      AMAX=1;
      AL(1,)=AL(2,);
      AL(2,)=(AL(2,)+AL(3,))#/2;
      END;
    IF MAXI=PRSS(3,) THEN DO;
      AMAX=3;
```

```
      AL(3,)=AL(2,);
      AL(2,)=(AL(2,)+AL(1,))#/2;
      END;
   ADD=AL(2,);
   END;
IF MINI=PRSS(3,) THEN DO;
   AMIN=3;
   SET1=AL(3,);
   AL(3,)=2#AL(3,)-AL(2,);
   AL(1,)=AL(2,);
   AL(2,)=SET1;
   ADD=AL(3,);
   END;

*** Obtain the kernel fit for the new value of the bandwidth ***;
ALPHA=AL(AMIN,);
DINT=(-1)#ALPHA#(DSQ#/(DSQSUM#J(N,N,1)));
DO I=1 TO N; DO K=1 TO N;
   EXPD(I,K)=EXP(DINT(I,K));
   END; END;
ESUM=EXPD(,+);
HK=EXPD#/(ESUM*J(1,N,1));
YHATK=HK*Y;

* Compute PRESS for this new bandwidth;
HKMI=HK;
DO I=1 TO N;
   HKMI(I,I)=0;
   END;
HSUM=HKMI(,+);
HKMI=HKMI#/(HSUM*J(1,N,1));
YHATKMI=HKMI*Y;
```

```
YHKNOEND = YHATK(2:9,)//YHATK(10:19,);  *Kernel predictions w/o endpts;
YHKMINOE = YHATKMI(2:9,)//YHATKMI(10:19,);  * Minus i pred w/o endpts;
MINPRESS = (YNOEND-(YHKMINOE))'*(YNOEND-(YHKMINOE))#/(N-TRACE(HK));
      * Value of PRESS for the current bandwidth for this iteration;

*** Reset positions of ALPHA and PRSS vectors depending on the value
          of PRESS for the current value of the bandwidth  ***;
  IF AMIN = 1 THEN DO;
    PRSS(3,) = PRSS(2,);
    PRSS(2,) = PRSS(1,);
    PRSS(1,) = MINPRESS;
    END;
  IF AMIN = 3 THEN DO;
    PRSS(1,) = PRSS(2,);
    PRSS(2,) = PRSS(3,);
    PRSS(3,) = MINPRESS;
    END;
  IF AMIN = 2 THEN DO;
    IF AMAX = 1 THEN
      PRSS(1,) = PRSS(2,);
    IF AMAX = 3 THEN
      PRSS(3,) = PRSS(2,);
    PRSS(2,) = MINPRESS;
    END;   END;    * Search for optimal bandwidth ends;
```

```
YHATK = HK*Y;  * Kernel predictions for the bandwidth  h  that
                          minimizes PRESS;
```

OBTAINING LAMBDA TO MINIMIZE Cp3

```
LVEC=0/0.5/1.0;     * Start with lambda's of 0, .5, and 1.0;
      ***** For these 3 lambda's compute the value of Cp3 ***;
DO K=1 TO 3;
 G=LVEC(K,1);          * G plays the role of lambda here;
 HADJ=(G#HK)+((1-G)#H); * HATLINK hat matrix with mixing parameter G;
 YHATA=HADJ*Y;          * HATLINK predicted values;
 YHANOE=YHATA(2:9,)//YHATA(10:19,);    * Endpoints set aside;
 TRHA=TRACE(HADJ);
 VARHATC3=(YNOEND-(YHANOE))'*(YNOEND-(YHANOE))#/(N-TRHA-2);
            * Variance estimate based on Cp3;
 CP3(K,)=TRHA+((VARHATC3-(VHATNOE))#(N-TRHA-2)#/(VHATNOE));
           * Value of CP3 computed;
 END;

**** Find minimum of Cp3 -- a quadratic function in lambda determined
        by the three values of Cp3 computed above ***;
LMAT=J(3,1,1)||LVEC||LVEC##2;
INVLL=INV(LMAT'*LMAT);
BETAHAT=INVLL*LMAT'*CP3;
L3MIN=-(BETAHAT(2,)#/(2#BETAHAT(3,)));
                    * Value of lambda that minimizes Cp3;
IF L3MIN<0 THEN L3MIN=0;    * If min falls outside 0 to 1 then reset;
IF L3MIN>1 THEN L3MIN=1;
HADJ=(L3MIN#HD)+((1-L3MIN)#H); * Hat matrix for HATLINK
                      with lambda chosen to minimize Cp3;
YHATA=HADJ*Y;   * HATLINK predictions at the vector of data locations;
```

The vita has been removed from the scanned document