Semiparametric Measurement of Environmental Effects

by

Diego Rodriguez and Thomas M. Stoker

MIT-CEEPR 93-008WP June 1993

SEMIPARAMETRIC MEASUREMENT OF ENVIRONMENTAL EFFECTS

Diego Rodriguez and Thomas M. Stoker*

January 1993

revised June 1993

* Department of Economics, MIT and Sloan School of Management, MIT, respectively. This research was funded by a grant from the MIT Center for Energy and Environmental Policy Research. The authors wish to thank R. Schmalensee, W. Newey and audiences at seminars at Brigham Young University, Princeton, CentER at Tilburg University, the CEEPR, and the NBER Summer Institute for helpful comments.

Abstract

This paper gives the results of a semiparametric analysis of pollution effects on housing prices using the Boston Housing Data. The exposition introduces the basic ideas of modeling pollution impacts with hedonic price methods, discusses the standard log-linear model, and then introduces nonparametric estimation and semiparametric index models. We focus on the intuitive content and substantive results of the semiparametric analysis. We find that the impact of pollution is smaller than that previously estimated, and varies dramatically depending on the status level of the community. We give various interpretations of the findings, and contrast our methods with those used in previous analysis of the Boston Housing Data.

SEMIPARAMETRIC MEASUREMENT OF ENVIRONMENTAL EFFECTS

1. Introduction

Policy measures designed to affect the environment involve complex interactions between consumption and production technology and economic behavior. Policy evaluation requires benefit-cost comparisons that are somewhat elusive, because of difficulties in measuring overall benefit or cost components. An economic approach to evaluating environmental conditions makes use of implicit valuations revealed by individual actions, or explicit valuations when variations in condition levels are adequately priced in a market. Such an approach is most suitable when the variation in environmental conditions is observed directly, such as a comparison of housing prices before and after the discovery of a nearby toxic waste dump, or of the cost structure of a plant before and after the use of an environmentally cleaner technology.[1]

The purpose of this paper is to present the results of a semiparametric analysis of air pollution effects on housing prices. The semiparametric analysis involves flexible econometric methods, that are designed to give as realistic a depiction as possible of pollution effects. The focus of our analysis is the Boston Housing Data of Harrison and Rubinfeld (1978a,b). This is a well known data set used in a seminal study of the economic evaluation of auto emission abatement policies.

The exposition begins with a general discussion of the economic approach to evaluating environmental effects, including the hedonic approach to explaining differences in housing prices. We then discuss issues in the specification of the equations used to measure pollution effects, introducing our semiparametric method.[2] This sets the stage for the empirical depiction of pollution effects from the Boston Housing Data.

Our results indicate that traditional methods missed substantial nonlinearity in the pollution effect. While a systematic pollution effect can be found for lower status communities, no such effect occurs for higher status communities. The overall impact of changes in pollution levels is considerably smaller than that measured in earlier studies. We give graphical depictions of these features, as well as some interpretation of the findings.

2. Issues in the Economic Analysis of Pollution Effects

To fix ideas, consider evaluating air pollution differences through their impact on housing prices. Suppose that a given house is priced at $p(a)$ when the air pollution level is a. Suppose further, that the pollution level is increased by Δa to $a + \Delta a$. Since the house is now less desirable to live in, its price falls to $p(a + \Delta a)$. The economic value (to homeowners) of the pollution change Δa is the change in the housing price

$$p(a + \Delta a) - p(a) = p_a \Delta a ,$$

where p_a is the per unit value (or cost, since it is negative) of the change in the pollution level

$$p_a = \frac{p(a + \Delta a) - p(a)}{\Delta a} .$$

A measure of the "pollution price" p_a therefore gives the economic value of the environmental change Δa in air pollution level. To the extent that the change in the level of pollution occurs for many houses in an area, the total impact is the total change in all prices. This impact can be summarized through the values of the pollution price p_a measured for each house in the affected area.[3]

While simple in concept, the empirical measurement of pollution prices involves at least two well known blocks of issues. The first concerns the

comparability of "before" and "after" pollution situations as discussed above. In particular, to attribute the entire change in observed house price to pollution neglects other market features that impinge on demand and supply, an omission which becomes more problematic as the time interval between "before" and "after" increases. For instance, suppose the group of home buyers increases in size, or that tastes change to increase the overall attractiveness of the location of the house under consideration. In that case the observed price difference has two components, one based on the pollution difference and the other based on the increase in demand. Alternatively, the supply of housing could have increased in a fashion to depress overall house prices, so that the observed price decrease is again comprised of the pollution effect and the supply effect. With supply influences taken into account, many outcomes are possible, ranging from the competitive situation where observed house prices reflect construction costs plus capitalized value of the land (which would reflect the pollution difference), to situations of over or under adjustment due to time lags in construction. At any rate, many outcomes are possible, and to properly isolate the pollution effect involves explicit modeling of the supply response. Unfortunately, no well accepted econometric methods have been developed for this, and so it is standard to measure pollution effects under the assumption of no supply response in the stock of housing, and we do so here.[4]

The second block of issues, which are more germane to our study, concern when "before" and "after" assessments are made by comparing different houses at the same point in time. In particular, suppose that one could view two houses as entirely identical (including locational aspects), except that one exists in an area with pollution level a and the other in an area with pollution level $a + \Delta a$. In this context, the price difference between these houses could be attributed to the only thing that differed between them,

namely the pollution difference Δa, and the pollution price p_a could be measured as above from the price difference.

This type of comparison is the most widely used method of assessing pollution differences, and is the only available method when the observed data consists of prices for different houses in a single time period. The overall approach of regarding a house as a bundle of attributes, with the pollution level as one, is known as the hedonic price approach.[5] The attributes that rationalize housing prices include number of rooms and other characteristics of the structure, as well as aspects of location, such as lot size and proximity to schools, etc. If we denote all relevant (non-pollution) attributes of a given house as h, then the housing price is determined as $p(a,h)$, and the pollution price is determined as

$$p_a = \frac{p(a + \Delta a,h) - p(a,h)}{\Delta a} \quad .$$

This formulation reflects how the actual "house" is held constant through h, with the pollution price reflecting the change in value associated with the pollution change Δa.

The main ingredient to a "hedonic" study of pollution effects is a statistical characterization of the house price function $p = p(a,h)$. Differences in the distribution of housing prices under alternative pollution scenarios can be calculated from this function, as well as the distribution of pollution prices across different kinds of houses. For small changes in pollution levels, pollution prices are given as the partial derivative

$$p_a = \frac{\partial p(a,h)}{\partial a} \quad ,$$

where again, other housing attributes h are held constant.

3. Specification of the Hedonic House Price Equation

We have spelled out the basic issues above to retain focus on what is being measured in empirical studies of pollution effects. The results depend intrinsically on the hedonic equation representing housing prices, which is characterized statistically. In this section, we discuss standard issues in specifying the hedonic equation, in order to introduce nonparametric and semiparametric methods. In the following section, we give the results of analyzing the Boston Housing Data.

The logic of the hedonic approach - house prices determined by housing characteristics - is hardly controversial, provided that all relevant (valued) housing characteristics could be included in the analysis. In practice, however, the observed characteristics are but a fraction of what one could sensibly regard as an exhaustive list. As such, unobserved characteristics are modeled as stochastic elements, and the treatment of such attributes is an integral part of a housing price model. Including the pollution level, we denote the observed attributes by $x = (a, x_0)$ and a stochastic disturbance that represents unobserved attributes by ϵ, so that the full list of housing characteristics is $(a,h) = (a, x_0, \epsilon) = (x, \epsilon)$. The overall object of the statistical analysis is the estimation of the connection between housing price p and observed attributes x, as well as the hedonic price of pollution p_a.

The simplest form of hedonic price equation $p(a,h) = p(x,\epsilon)$ is the standard linear model in price levels:

$$p = x^T\beta + \epsilon$$

$$= \beta_a a + x_0^T\beta_0 + \epsilon$$

where the unobserved characteristics have mean $\alpha = E(\epsilon|x)$. Here, the coefficients (β_a, β_0) are directly interpreted as hedonic prices, with the pollution price above given as $p_a = \beta_a$, which is constant regardless of the level of pollution or other housing characteristics. The relative importance to housing price of any two housing characteristics is given by the ratio of their respective coefficients. This model is dictated if arbitrage exists under competition, when houses can be easily repackaged (or their characteristics unbundled, with an effective market for each characteristic). In particular, if one house has half the attribute values of another, then it is priced at half as much. To the extent that a house representing any "bundle" of housing attributes can be purchased, competition will result in each attribute being uniquely priced, and housing prices will necessarily follow the linear model above.

While giving easily interpreted results, it is well known that the linear model does not perform well statistically (in terms of goodness-of-fit) in hedonic price applications, including studies of housing prices. Moreover, the kind of unbundling that dictates a linear model is especially unreasonable for housing (for any small number of characteristics), even approximately. Because different locations of houses are associated with persistent differences in prices, it is natural to question whether a house could be "unbundled" from the location on which it sits. On statistical grounds, the linear model takes all unobserved house price differences to be additive, which implies that unobserved differences in comparable house prices should be the same as one moves from one locale to another, which does not even hold up under casual observation. The practical consequence of this has been to make the characterization of hedonic price equations a purely statistical problem, with the proper form to be decided on the basis of goodness of fit to the data.

The most commonly used hedonic price equation is the log-linear model, that relates log-price variations to characteristics as

$$y = \ln p = x^T\beta + \epsilon \quad ,$$

where again, $E(\epsilon|x) = \alpha$. Here, the coefficients are interpreted as the proportional changes in prices associated by changes in characteristics, holding specific location features constant. The proportional impact on house price of a characteristic x_j relative to x_k is summarized compactly as β_j/β_k. This model dictates that when the location value ϵ is changed, the same proportional configuration of housing prices exists along the lines of the observed characteristics.

It should be noted that because arbitrage is not relied on in the specification of the log-linear model, the characteristic vector "x" can include transformations of basic attribute values: for example, the model does not rely on a unique "price per room," so that polynomial terms in "number of rooms" could be included as part of the statistical investigation. This flexibility makes the log-linear model a considerably richer framework for empirical analysis than the linear model based on competition.

Despite these advantages, the log-linear model has also been recently questioned as to its statistical adequacy in various hedonic price contexts, for instance it is rejected for several data bases in Berndt, Showalter and Wooldridge (1990).[6] The most common parametric method of assessing the log-linear model is to estimate a model based on a more general transformation of prices than the logarithm, and check if the logarithm is suggested by the results. In particular, one estimates

$$y^{(\lambda)} = x^T\beta + \epsilon$$

where $y^{(\lambda)}$ is the "Box-Cox" transformation of prices

$$y^{(\lambda)} = \frac{p^{\lambda} - 1}{\lambda} \quad , \lambda \neq 0,$$

$$= \ln p \quad , \lambda = 0,$$

and then does a statistical test of whether the value of λ is different from zero. While this estimation involves some delicate econometric issues, the basic point is that a value of λ different from zero rejects the log-linear model, giving $y^{(\lambda)}$ as the transformation of prices that is suggested by the data. In this context, the interpretation of the values of the coefficients β becomes somewhat obscure, as they translate changes in x to changes in $y^{(\lambda)}$; i.e. the only situations of easy interpretation of β are for $\lambda = 0$ (proportional changes) or $\lambda = 1$ (level changes).

The linear, log-linear and transformation models discussed above constitute the primary parametric models of hedonic prices, such as housing prices, where "parametric" refers to the fact that the model is determined by the few parameter values, namely β, α and λ. Because of focusing attention on a few parameters, each of these approaches can potentially miss important variations in the price data, and therefore give mismeasurements of hedonic prices for attributes, or in our case, for pollution levels. For example, it is easy to verify that the "Box-Cox" transformation model always implies that prices are monotonically increasing in $x^T\beta$ (or characteristics with positive coefficients). Moreover, the connection between price and $x^T\beta$ is convex if $\lambda < 1$, implying that all characteristics x have larger impacts (hedonic prices) on expensive houses than less expensive houses. When $\lambda > 1$, the opposite is true.

To the extent that the empirical investigator is lucky, in that the impositions of the parametric model used adequately hold in the data, there

will not be any systematic mismeasurement of the price equation or mismeasurement of the hedonic prices. The only way to be certain of this, however, is to use methods that are more flexible still, by permitting quite general patterns between house prices and characteristics. Nonparametric and semiparametric methods are designed to permit this kind of "imposition free" determination of the pricing equation.

The nonparametric approach to determining the hedonic price equation is based on an unstructured model connecting prices to characteristics; in log form, one could write this "model" as

$$y = \ln p(x,\epsilon) \quad .$$

Statistical analysis can be based on characterizing the mean of log-price for different characteristic values, or the "regression" function

$$E(y|x) = E[p(x,\epsilon)|x] \equiv m(x) \quad .$$

Nonparametric methods measure the function m(x) directly, without restricting it to take on a linear or log-linear form as above. The main requirement is that the function m(x) is suitably smooth, so that small changes in characteristics x are associated with small (continuous, possibly differentiable) changes the mean log-price m(x).

Regression functions can be estimated with any method that permits arbitrarily fine approximation with a large amount of data. For instance, m(x) could be measured with series approximation, such as polynomials or Fourier series, or with local averages, such as so-called "nearest neighbor" or kernel regression estimators.[7] In the next section we present kernel estimators, so here we explain the local average method of measuring the log-price equation. This discussion reveals the advantages and the drawbacks to fully general nonparametric regression, as well as the role of

semiparametric methods.

Local averages measure the mean log-price m(x) by averaging log-price values of houses that have characteristics close to the value x. A useful way to consider these statistical tools is to consider the familiar method of "market analysis" used by realtors in appraising homes. When putting one's home up for sale, the realtor will find "comparable" homes in the locale, look up their selling prices, and estimate the "market value" of the home using an average selling price of the comparables. Local averages (such as kernel estimators) implement this idea with data - namely "comparables" are defined by similar x values, and observations with virtually identical characteristic values are given higher weight in the estimation than observations with similar, but not as close, characteristic values.

With a bit more formality, suppose that one has observations for N homes, with log-price and observed attributes for the i^{th} house denoted as (y_i, x_i), for $i = 1,\ldots,N$. A local average estimator of the mean of log-price y given the values of attributes x takes the form

$$\hat{m}(x) = N^{-1} \sum w_i(x)\, y_i$$

where $w_i(x)$ denotes the local weight applied to y_i, which is larger for observations with x_i close to the evaluation point x, and smaller (possibly zero) for observations with x_i far from x. The results presented below are based on kernel regression estimators, where the local weight $w_i(x)$ is specified as

$$w_i(x) = \frac{\mathcal{K}[(x-x_i)/h]}{\sum_j \mathcal{K}[(x-x_j)/h]}$$

where $\mathcal{K}(.)$ is a prespecified density function giving the shape of the local weights, and h is the bandwidth parameter that gauges the "proximity" of x_i to

x.[8] Figure 0 illustrates local average estimation, depicting it as an enhanced method of curve or surface fitting; or as a flexible method of indicating the structure of the basic data. This amounts to "market analysis" with a systematic method of determining "comparables," given by the formulation of local weights.

The advantages of local averages are obvious - estimating house values using similar houses is a quite natural method of evaluation. The statistical drawbacks are also fairly clear. In particular, if only a few approximately comparable houses can be found, the resulting estimate of house price will be quite imprecise. This kind of problem is exacerbated when there are many characteristics to take into account in the analysis, wherein it becomes increasingly difficult to find comparable matches to all observations. In a different light, this issue says that a fully general nonparametric approach will yield precise log-price mean values only when there is a great deal of data, so that a fair number of comparables can be found for any given observation. This issue is referred to in technical parlance as the "curse of dimensionality" for nonparametric estimation. For our application, there are nine characteristics and 506 census tract observations. In this context one faces the additional problem of how to display the results of nonparametric estimation, because m(x) is a function of nine arguments.

Semiparametric methods combine features of parametric and nonparametric methods, to retain simple interpretability of the results, and to avoid arbitrary mismeasurement by an incorrect parametric formulation of the hedonic price model. In particular, the hedonic price equation has some parametric structure, utilizing parameters to summarize key features of attribute-price connections, but also permits other parts of the hedonic relationship to be measured flexibly, with nonparametric estimators. With reference to the log-linear model, one semiparametric generalization is to assume that the mean

log price y is determined by an "index" $x^T\beta$ of characteristics, but that the connection may be nonlinear

$$E(y|x) = G_0(x^T\beta)$$

or a so-called "single index" model. Here, the coefficients β can be estimated directly, and then the (univariate) function G_0 estimated with a kernel regression estimator.[9] With regard to our motivation above, the problem of finding "comparables" is substantially reduced (via "dimension reduction"), because comparability here means similarity in the value of the index $x^T\beta$, and the univariate function $G_0(.)$ can generally be estimated with greater precision than a relation with more than one argument. This specification gives nonparametric treatment to a similar kind of transformation as the "Box-Cox" model above.

While a substantial generalization over the log-linear model, the single index model is still a restrictive specification relative to a fully general regression model, such as the one estimated by $\hat{m}(x)$ above. If the single index model is statistically rejected against a general regression, one needs to consider more general semiparametric specifications, and many can be devised. For instance, partial index models employ an "index" specification for certain characteristics, but a fully general structure for other variables. For instance, suppose that the impact of pollution level was not adequately represented by the "index" formulation. A partial index model could be constructed as

$$E(y|x) = G_1(a, x_0{}^T\beta_0) \quad ,$$

which permits a much more flexible pollution structure. In this model, the coefficients β_0 are estimated,[10] and the two-dimensional function G_1 is estimated nonparametrically. "Comparables" are determined via close values of

a and $x_0^T\beta_0$, and the impact of pollution a on mean log house price is not necessarily connected to the impact of any of the other characteristics.

In a related paper, Rodriguez and Stoker (1992), we devise a testing procedure for assessing what degree of functional structure is called for in data sets on log prices and characteristics, along the lines above. For the Boston Housing Data, we found that the log-linear and single index models gave statistically equivalent depictions of the data, but both were rejected against a general kernel regression estimator. We concluded that a certain partial index model gave a statistically adequate description, in that it was not rejected against nonparametric regression. We now turn to a discussion of the specifics of the data and model, and a description of the basic results on log housing prices and the hedonic price of pollution. The reader is referred to the above reference for the details of our testing procedure, and our aim in this paper is to illustrate how one can interpret our final model, in terms of house price - attribute relationships as well as the associate hedonic price of pollution.

4. Environmental Effects in the Boston Housing Data

4.1 Some Preliminaries

The Boston Housing Data consists of 506 observations on the average value of housing in census tracts in the Boston area in 1970. This data was first used in the analysis of pollution abatement policy by Harrison and Rubinfeld (1978a, hereafter HR, and 1978b), and the structure of their hedonic house price equation has received considerable attention since then. For instance, Belsley, Kuh and Welsch (1980, hereafter BKW) analyze the basic log-linear equation for robustness, including identifying various influential observations and other outliers. The variables in the data set are listed in Table 1, where we have focused our analysis on the nine predictors found

significant in the Harrison-Rubinfeld and Belsley-Kuh-Welsch work.[11] Likewise, we have retained the transformations of the basic observed variables as used by these authors.[12] The ordinary least squares (OLS) estimates of the coefficients of the log linear model are given in Table 2.

The coefficient estimates used for index models are also given in Table 2 under the heading of "average derivatives." These coefficients are measures of the average of the effects (slopes) of the attributes on log-prices, treating the regression as arbitrarily general. If $m(x) = E(y|x)$ denotes the true statistical relationship, then the average derivatives are $\delta = E(\partial m/\partial x)$, and these parameters can be measured more precisely than the attribute specific effects $\partial m(x)/\partial x$ for any given value x of attributes. Stoker (1992) gives a lengthy treatment of average derivatives, their estimation, and how they provide consistent estimates of the coefficient parameters of single and partial index models.[13] If the log-linear model were, in fact, statistically adequate, the average derivatives would measure the same values as the OLS coefficients.

The average derivative estimates have the same qualitative pattern as the OLS coefficients, save for the race effect, with a negligible average derivative estimate.[14] The average derivatives indicate weaker effects than OLS for pollution, distance to employment and access to radial highways, pupil-teacher ratio and lower status, and stronger effects for crime rate, number of rooms and taxes. Despite these difference in magnitudes, it is not possible to reject the hypothesis that the OLS coefficients and the average derivatives are measuring the same values, as indicated by the Wald statistic at the bottom of Table 2. In other words, from just looking at coefficient estimates, there are no grounds for rejecting the basic log-linear model.

However, Rodriguez and Stoker (1992) note that the log-linear model is indeed rejected against a general regression model, so that apart from the

coefficients, there are systematic departures from the log-linear model in the data. The same is true of the strict single index model (with average derivative estimates as coefficients), so that using a single index to summarize the effects of all the attributes likewise misses some nonlinear structure. Our testing procedure concluded with a partial index model that omitted the pollution variable, and the lower status variable, from the index summarizing the remaining attributes. In particular, our procedure failed to reject such a partial index model against general regression, so that the nonlinearity not accounted for by the log-linear model arises from the treatment of pollution and lower status effects. Since the pollution impact is a primary concern, we focus on the hedonic price of pollution after describing the basic model below.

The partial index model that passes our criteria gives the mean log housing price as

$$(*) \qquad E(y|x) = \hat{G}_2(\text{NOXSQ}, \text{LSTAT}, \text{INDEX})$$

where the index variable is constructed from the remaining seven predictors as

$$\begin{aligned} \text{INDEX} = {} & (-.0256\ \text{CRIM}) + (.0106\ \text{RMSQ}) + (-.0746\ \text{DIS}) + (.0669\ \text{RAD}) \\ & + (-.0009\ \text{TAX}) + (-.0175\ \text{PTRATIO}) + (-.0526\ \text{B}) \end{aligned}$$

The INDEX variable can be regarded as a linear representation of log house prices, omitting the influence of air quality (NOXSQ) and the (lower) status position of the communities. The function $\hat{G}_2()$ gives a nonparametric description of the impact of pollution and lower status, with remaining housing attributes controlled for via the index representation. We now give a graphical description of the pollution impacts.

4.2 Graphical Analysis of the Boston Housing Data

The estimated log-price function $\hat{G}_2(.)$ is a function of three variables, so we cannot depict it entirely on one diagram. Since our focus is on the pollution effect, we first give three-dimensional diagrams of mean log-price $\hat{G}_2(.)$ over values of pollution and lower status, for various values of the index of other housing attributes. We then give diagrams of $\hat{G}_2(.)$ over values of pollution and the index, for various values of lower status. For interpretability, we detransform the pollution-squared and log-lower status variables as used in estimation, plotting $\hat{G}_2(.)$ over values of pollution and lower status.[15] As discussed above, $\hat{G}_2(.)$ represents the mean of log house prices for given values of its arguments: twists, bumps and other kinds of curvature indicate nonlinearity in the hedonic house price relationship.[16]

Figure 1a graphs the relation between log house prices and pollution and lower status with INDEX set to its mean value. For lower status communities, we see that the pollution effect is negative as expected, or that lower house prices are associated with higher pollution levels. However, for higher status (low LSTAT) communities, the opposite is true, namely that higher pollution levels are associated with higher house prices. This counter intuitive finding indicates immediately how the log-linear model fails to adequately account for the empirical features of higher status communities.

Before interpreting this feature, it could be that there are just a few high price - high pollution - high status communities observed at the mean index level; in other words, there may be a few outliers that arise from a unfortunate choice of plotting at the average value for INDEX. To consider this, Figure 1b gives an analogous plot with the index value set to a higher value (mean plus 1.5 standard deviation) and Figure 1c gives an analogous plot with the index variable set to a lower value (mean minus 1.5 standard deviation). These figures illustrate the same phenomena, namely a negative

pollution effect for lower status communities and a somewhat positive effect for higher status communities. As such, the misspecification of pollution effects for higher status communities is a robust feature regardless of the overall level of house prices (omitting pollution and status effects).

There are some subtle differences in Figures 1a-c worth noting. Figure 1c depicts a relatively smaller upturn in prices for pollution values in high status communities than Figures 1a and 1b. This is quite natural, as the level of house prices can be regarded as an alternative measure of a "status" phenomena, so that the counter intuitive pollution effect with high status would be less for low price communities. These subtle differences give further indications that the generality of model (*) is important for an adequate data description.

An alternative depiction of these results is given from the implied hedonic price of pollution for the model (*) versus the log-linear model. For a given value of the observed attributes x, the hedonic price of pollution is obtained by differentiating (*), giving

$$\hat{P}_{NOX} = \hat{G}_2(x) \frac{\partial \hat{G}_2(x)}{\partial NOXSQ} \hat{\delta}_{NSQ} \, NOXSQ$$

where $\hat{\delta}_{NSQ}$ is the estimated average derivative for NOXSQ from Table 2. For the log-linear model, the implied hedonic price of pollution is

$$\tilde{P}_{NOX} = \hat{L}(x) \hat{\beta}_{NSQ} \, NOXSQ$$

where $\hat{L}(x)$ is the fitted value from the log-linear model, and $\hat{\beta}_{NSQ}$ is the OLS coefficient of NOXSQ.

Figure 2 depicts the hedonic price as implied by model (*) (for INDEX set to its mean value). It shows a smoothly increasing cost (negative price) of higher pollution levels in lower status communities. As the status of the

community is increased, the prices become positive, as we would expect from the relations in Figures 1a-c. For further illustration of this, Figures 3a-c illustrate the pollution price as a function of pollution level, and include the pollution price that is implied by the log-linear model. Here INDEX is set to its mean value, and Figures 3a, 3b, 3c depict the pollution price for high, middle and low status communities respectively.[17] This figures illustrate more clearly the sensitivity of the pollution effect to the status of the community, as well as the substantial differences between semiparametric and log linear estimates of the hedonic price of pollution.[18] Reliance on a log linear model amounts to asserting serious impositions on the pollution price that are in conflict with the observed data patterns. The hedonic price of pollution has a considerably more complicated structure than that given by the standard log linear model.

The same basic lesson arises when we consider the interrelationship between log house price, pollution level, the index of other housing attributes. Figure 4 depicts this structure at the mean value of lower status. While somewhat accentuated by the orientation of the graph, the increase in housing prices at quite high pollution levels is evident, especially for high values of the index.

A more vivid depiction of the intrinsic nonlinearity of the basic hedonic relationship is found by considering the hedonic price of pollution as compared to the index, at different levels of community status. Figure 5a, 5b and 5c depict the hedonic price at high, middle and low status levels as indicated by the value of LSTAT. Most evident is how the lower status communities give a strongly negative impact of increases in pollution on housing prices, especially for high pollution ranges. This effect is mollified for middle status communities, and apparently reversed for high status communities. While yet just another method of isolating the

substantial nonlinearity in the log house price - pollution relationship, the differences between the estimated prices over status ranges are quite striking. Finally, for comparison, Figure 6 presents the hedonic prices implied by the log-linear model at the mean status level. Noting differences in the plotting range of hedonic prices, one can see how the log linear model predicts smooth pollution impacts in a fairly narrow range as compared to those seen in Figures 5a, 5b and 5c.

4.2 Interpretation

One of the more powerful features of the semiparametric methods we have used is the ability to clearly depict the structure of the data, giving a visual indication as to where a standard model can fail. In this regard, it is worth noting some findings from previous analysis of this data which are consistent with our findings above.

First, the original Harrison and Rubinfeld study found that their estimates of pollution effects were not robust to perturbations of the basic model, and their estimates of aggregate benefits of air quality improvements could be decreased by 60% by considering alternative specifications (HR, p.78). From Table 2, the average derivative estimate for NOXSQ is considerably below the OLS estimate, and considerably less precise, reflecting this feature. The substantial nonlinearities evidenced from our graphical depiction of model (*) would lead to very sensitive estimates based on varying specifications of a log-linear model. In essence, different log-linear specifications amount to different weightings of the observed pollution effect over different ranges of the data. Therefore, the substantial differences in the hedonic price of pollution noted above would lead to wide variation in summary estimates such as OLS coefficients from differing specifications.

The analysis of Belsley, Kuh and Welsch carried out careful

diagnostic analysis on the residuals of the basic log-linear model, and found strong evidence of outliers, or particularly influential observations. In particular, they note a correlation between the size of the residuals and the census tract to which they belong, concluding that misspecification has occurred with respect to a factor that is correlated with the geographic district assignments of the data. More specifically, they note

> *The influential data tend to be quite heavily concentrated in a few neighborhoods and these are, for the most part, the central city of Boston, which leads us to believe that the housing price equation is not as well specified as it might be* (BKW, p. 243)

These authors do not pursue model variations that isolate these influential observations.[19]

Our results find systematic departures from the model in terms of nonlinearity of the pollution - status relationship. While we have not included "Boston" as an explanatory factor, we have noted how the pollution - status effect is basically robust to the levels of the other variables in the data. Since there are high status census tracts other than those in Boston, we consider what further interpretation can be given to model inadequacy beyond the fact that "Boston may be special."

The covariation of pollution and housing prices could arise because of genuine disutility associated with breathing poorer air, but it could also arise from various locational features that are associated with air quality. For example, the air quality in a community could be poor because of the presence of high density traffic areas. Another example concerns proximity to factories or other pollution-emitting sources. Moreover, the possibility that pollution is a proxy for other locational effects is consistent with our

semiparametric findings. The fact is that lower status communities are likely to be those in which high density traffic areas or industrial sources are to be found. Consequently, the lower status communities would display a systematic negative pollution effect as we have found. Alternatively, if the higher status communities were relatively free of traffic and other problems, the measured pollution level may not be systematically associated with lower prices, also as we have found.

The possibility of pollution acting as a proxy for other kinds of locational effects also places a proviso on the applications of hedonic results to the assessment of pollution abatement policies. Suppose, for instance, that laws were passed that severely limited automobile emissions. Over a period of time, housing prices in previously polluted areas would adjust upward to the extent that the improved air quality improved living conditions. This impact would likely be most relevant for communities where the pollution effect were highly associated with traffic and congested downtown areas. But such an effect would not be so pronounced for areas where pollution arose from an industrial site, or other area that was not affected by the law to limit automobile emissions.

Moreover, even in areas where the major polluters are automobiles, the emissions law would not mollify other factors associated with traffic density that impact on housing prices. For instance, the noise level of close busy roadways, or issues of child safety near busy roadways would go unchanged, and their implicit valuation in observed housing prices would remain. As in many contexts where there is the potential for important omitted variables, the measured pollution effect may overstate the true benefits of environmental policies.

5. Conclusion

The purpose of this paper is to introduce many of the ideas of semiparametric modeling in the context of measuring the implicit value of changes in air quality. We have illustrated these ideas with an application to the Boston Housing data, where we have graphically depicted the inadequacy of a standard log-linear model of housing prices. While our results are reflective of findings of previous analysis of this data, we have isolated more systematic features of where the pollution effect is stable and where it is not. In particular, a quite different pollution-house price structure exists for low status housing as for high status housing areas.

Our application of flexible nonparametric methods involves more complicated issues and computational methods than standard log-linear modeling. Moreover, since our application involved nine predictor variables, there was the possibility that no simplified semiparametric model would give a statistically adequate fit. In such a case, we would have been hard pressed to give a very conclusive description of the true empirical relationship, because of the difficulties of describing a nine dimensional function.

Our discovery of a statistically adequate partial index model permitted graphical analysis and interpretation, but even our basic model involved estimation of a function of three arguments (NOXSQ, LSTAT and INDEX). The analysis of our final model was considerably more complicated than familiar analysis of log-linear regression estimates. With this in mind, one might think that it is in some ways better to use a log-linear model, relying on the estimated OLS coefficient of a pollution variable as an "average effect" applicable across the entire data sample.

We would argue strongly against such an approach, which amounts to ignoring model specification issues on the grounds of expediency. It is clear that "back of the envelope" policy evaluation is aided greatly by the use of a

single summary effect, but there is no systematic reason to believe that OLS estimation of a log-linear model will give anything like an adequate summary effect. For instance, in Table 2, the OLS coefficient of NOXSQ was almost twice as large as our measure of the average derivative, or average effect of NOXSQ (measured using flexible methods). Moreover, the apparent precision of the OLS coefficient masks the fact that the empirical effect of pollution varies quite widely across the data, with a systematic negative effect for low status communities combined with a rather unstable measured effect for high status communities. If one is to use a summary statistic of the basic effect, it is important to utilize statistics whose connections to the empirical effects are well understood. One possibility, among many, is the average derivative, or average effect across the sample. Unless the data relationship is demonstrably consistent with a log-linear model, there is no clear way to trace the connection between OLS coefficient estimates and the underlying empirical effects.

As such, a major conclusion of our empirical analysis is that reliance on a log-linear framework, even when augmented with the standard tools of inference for linear models, can miss systematic nonlinear structure in the basic data. The only way to allow for a more complete range of possibilities is to adopt a framework that is sufficiently flexible to account for such a complete range, and permits methods of interpretation of the basic findings. Our use of kernel estimators permitted examination of the nonlinear house price - pollution relation on diagrams, which make clear what features are not captured by the log-linear model. While our analysis is definitely more complicated than what is available from familiar, more standard methods, the result is a much richer understanding of the house price - pollution relationship in the Boston Housing data.

Notes

[1] Environmental impacts that have not been observed can also be valued with an economic approach, to the extent that the relevant futures markets accurately price differing environmental situations. Such applications are necessarily somewhat problematic, as many aspects of future situations need to be given explicit attention. A good discussion of the issues of measuring the effects of future global warming is given in the 1991 Economic Report of the President.

[2] As described later, our semiparametric model is the final result of a detailed study of model specification given in Rodriguez and Stoker (1992). This paper introduces the model, and then focuses on the intuitive content and substantive implications of the results. Related literature includes Stock (1989), who uses another kind of semiparametric approach to measure the average costs associated with toxic waste sites, and Palmquist (1988), who discusses the use of nonparametric price estimates in environmental analysis.

[3] One could take the evaluation a further step, relating the change in house value to "constant utility" income differences (as was carried out in the original study of Harrison and Rubinfeld (1978a,b). Here, we just focus on the house price equations, in part because of difficulties in obtaining the original income data required for this latter step.

[4] For interesting discussion of how the price mechanism matches consumer tastes and product attributes, see Tinbergen (1956), and for a useful discussion of competition in supply in differentiated product markets, see Rosen (1974).

[5] Palmquist (1989) surveys the use of hedonic price methods in studies of environmental effects.

[6] The recent surveys by Case, Pollakowski and Wachter (1991) and Smith and Huang (1991) each stress how the functional form of hedonic equations remains a major issue in environmental and other types of studies.

[7] Nonparametric regression methods are surveyed by Härdle (1991).

[8] Our empirical analysis used product kernels $K(u) = \Pi k(u_j)$, with $k(u_j) = 15/16\ (1 - u_j^2)^2$, and set bandwidths via generalized cross-validation (c.f. Rodriguez and Stoker 1992 and Stoker 1992 for details on the estimation).

[9] It is possible to estimate β and $G_0(.)$ simultaneously, but that involves delicate computational issues. We normalize β to be the "average derivative" $E[\partial m/\partial x]$ and estimate it directly using an instrumental variables estimator (with nonparametric instruments). The function $G_0(.)$ is then estimated by nonparametric regression of y on $x^T\hat{\beta}$, the estimated index. This approach and related references are detailed in Stoker (1992). While the use of an average derivative estimator for β may be less efficient than in simultaneous estimation (c.f. Newey and Stoker (1993)), it greatly facilitates single- and partial-index model estimation, as well as avoids many of the computational issues.

[10] With reference to the last note, β_0 is estimated by the subvector of the average derivatives associated with x_0.

[11] The data as listed in BKW (p. 245-261), and we want to acknowledge some useful conversations with D. Rubinfeld on the data set.

[12] The variable LSTAT is very important in the analysis that follows. It is the log of "proportion of residents of lower status," which is defined as the percentage of adults who are laborers and who do not have a high school education.

[13] We have used IV estimators of the average derivatives, that use score estimates as instruments for linear coefficients of y on x.

[14] The curious transformation used by HR and BKW for the race percentage means that a positive coefficient indicates a negative impact of the presence of minorities; $B = (B_k - .63)^2 = B_k^2 - 1.26\ B_k + (.63)^2$, so that for small proportions of black residents B_k, B will vary as $-1.26\ B_k$. Consequently, the OLS coefficient indicates a substantial negative race effect, with the average derivative indicating no effect of race.

[15] It is worthwhile noting that since $\hat{G}_2(.)$ is a nonparametric estimate with NOXSQ and LSTAT as free arguments, the (invertible) transformations used in their construction do not affect the estimate of mean log house price. In other words, $\hat{G}_2(.)$ is sufficiently flexible to "undo" the transformations if that were dictated by the data.

[16] For comparison, it should be noted that the log-linear model would be depicted as smooth planar surfaces on the diagrams that follow. It would not result in flat planes per se, as our detransformation would add some minimal curvature due to removing the "square" from NOXSQ and the "log" from LSTAT. This comparison is also not exact, because of the difference between INDEX and the weighted sum of the other attributes (OLS coefficients) of the log-linear model.

[17] "Middle Status" is defined by setting LSTAT equal to its mean value, "High Status" by setting LSTAT to its mean value minus 1.5 its standard deviation, and "Low Status" by setting LSTAT to its mean value plus 1.5 its standard deviation.

[18] Stoker (1993b) argues how kernel regression estimators can under estimate derivatives. We have not studied this phenomena here, but it would not affect the qualitative features of these diagrams, but could imply an exacerbation of the basic differences noted. This issue of derivative bias is not relevant for our method of estimating the average derivatives (Stoker 1993a), and so has not caused the difference between the average derivative of NOXSQ and its OLS regression coefficient.

[19] It is worth remarking how the BKW "bounded influence" methods represent a different conceptual approach to the misspecification of the log house price equation. In particular, influential observations such as outliers are likely to be associated with departures from the basic log-linear model. BKW downweight such observations, to obtain a more "representative" set of coefficient estimates. Our approach is to describe the data configuration as it exists with a semiparametric model, instead of considering reweighted versions of a linear equation.

REFERENCES

Belsley, D.A., E. Kuh and R.E. Welsch (1980), *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, New York, Wiley.

Berndt, E.R., M.H. Showalter and J.M. Wooldridge (1990), "A Theoretical and Empirical Investigation of the Box-Cox Model and a Nonlinear, Least Suares Alternative," MIT Sloan School of Management Working Paper Number WP3187-90.

Brown, J.N. and H. Rosen, (1982), "On the Estimation of Structural Hedonic Price Models", *Econometrica*, 50, 3, 765-768.

Case, B., Pollakowski, H.O. and S.M. Wachter (1991), "On Choosing Among House Price Index Methodologies," draft, August.

Cropper, M.L., L.D. Deck and K.E. McConell, (1988), "On the Choice of Functional Form for Hedonic Price Functions", *The Review of Economics and Statistics*.

Economic Report of the President (1991), February, United States Printing Office, Washington, D.C.

Graves, P., J.C. Murdoch, M.A. Thayer and D. Waldman, (1988), "The Robustness of Hedonic Price Estimation: Urban Air Quality", *Land Economics*, 64, 3, 220, 233.

Härdle, W. (1990) *Applied Nonparametric Regression*, Cambridge University Press.

Harrison, D. and D.L. Rubinfeld, (1978a), "Hedonic Housing Prices and the Demand for Clean Air," *Journal of Environmental Economics and Management*, 5, 81-102.

Harrison, D. and D.L. Rubinfeld, (1978b), "The Distribution of Benefits from Improvements in Urban Air Quality, *Journal of Environmental Economics and Management*, 5, 313-332.

Newey, W.K. and T.M. Stoker (1993), "Efficiency of Weighted Average Derivatives," MIT Department of Economics Working Paper, forthcoming in *Econometrica*.

Palmquist, R., (1988), "Welfare Measurement for Environmental Improvements Using the Hedonic Model: the Case of Nonparametric Marginal Prices", *Journal of Environmental Economics and Management*, 15, 297-312.

Palmquist, R. B. (1989), "Hedonic Methods," in *Measuring the Demand for Environmental Improvements*, J. B. Braden and C. D. Kolstad, ed., Institute for Environmental Studies, University of Illinois at Urbana.

Rodriguez, D. and T.M. Stoker (1992), "A Regression Test of Semiparametric Index Model Specification," MIT Center for Energy and Environmental Policy Working Paper 92-006.

Rosen, S. (1974) "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition," *Journal of Political Economy*, 82, 34-55.

Smith, V.K. and J.C. Huang (1991), "Can Hedonic Models Value Air Quality? A Meta-Analysis," draft, November.

Stock, J.H. (1989), "Nonparametric Policy Analysis," *Journal of the American Statistical Association*, 84, 567-575.

Stoker, T.M. (1992), *Lectures on Semiparametric Econometrics*, CORE Foundation, Louvain-la-Neuve, Belgium.

Stoker, T.M. (1993a), "Smoothing Bias in Density Derivative Estimation," Sloan School of Management Working Paper, forthcoming *Journal of the American Statistical Association*.

Stoker, T.M. (1993b), "Smoothing Bias in the Measurement of Marginal Effects," Sloan School of Management Working Paper, January.

Tinbergen, J. (1956) "On the Theory of Income Distribution," *Weltwirtschaftliche Archiv*, 77, 155-173.

TABLE 1: VARIABLE SPECIFICATION IN THE BOSTON HOUSING DATA

$y = \ln p$	LMV	log of home value
x_1	NOXSQ	nitrogen oxide concentration
x_2	CRIM	crime rate
x_3	RMSQ	number of rooms squared
x_4	DIS	distance to employment centers
x_5	RAD	accessibility to radial highways
x_6	TAX	tax rate
x_7	PTRATIO	pupil teacher ratio
x_8	B	$(B_k - .63)^2$, where B_k is proportion of black residents in neighborhood
x_9	LSTAT	log of proportion of residents of lower status

TABLE 2: COEFFICIENT ESTIMATES FOR THE HOUSING PRICE EQUATION

Dependent Variable: y = LMV (ln p)

		Average Derivatives $\hat{\delta}$	OLS $\hat{\beta}$
x_1	NOXSQ	-.0034	-.0060
		(.0035)	(.0011)
x_2	CRIM	-.0256	-.0120
		(.0056)	(.0012)
x_3	RMSQ	.0106	.0068
		(.0025)	(.0012)
x_4	DIS	-.0746	-.1995
		(.0504)	(.0265)
x_5	RAD	.0669	.0977
		(.0468)	(.0183)
x_6	TAX	-.0009	-.00045
		(.0003)	(.00011)
x_7	PTRATIO	-.0175	-.0320
		(.0152)	(.0047)
x_8	B	-.0526	.3770
		(7.514)	(.1033)
x_9	LSTAT	-.2583	-.3650
		(.0370)	(.0225)

(Standard Errors in Parentheses)

WALD TEST OF $\delta = \beta$: W = 13.44, Prob($\chi^2(9) > 13.44$) = .143

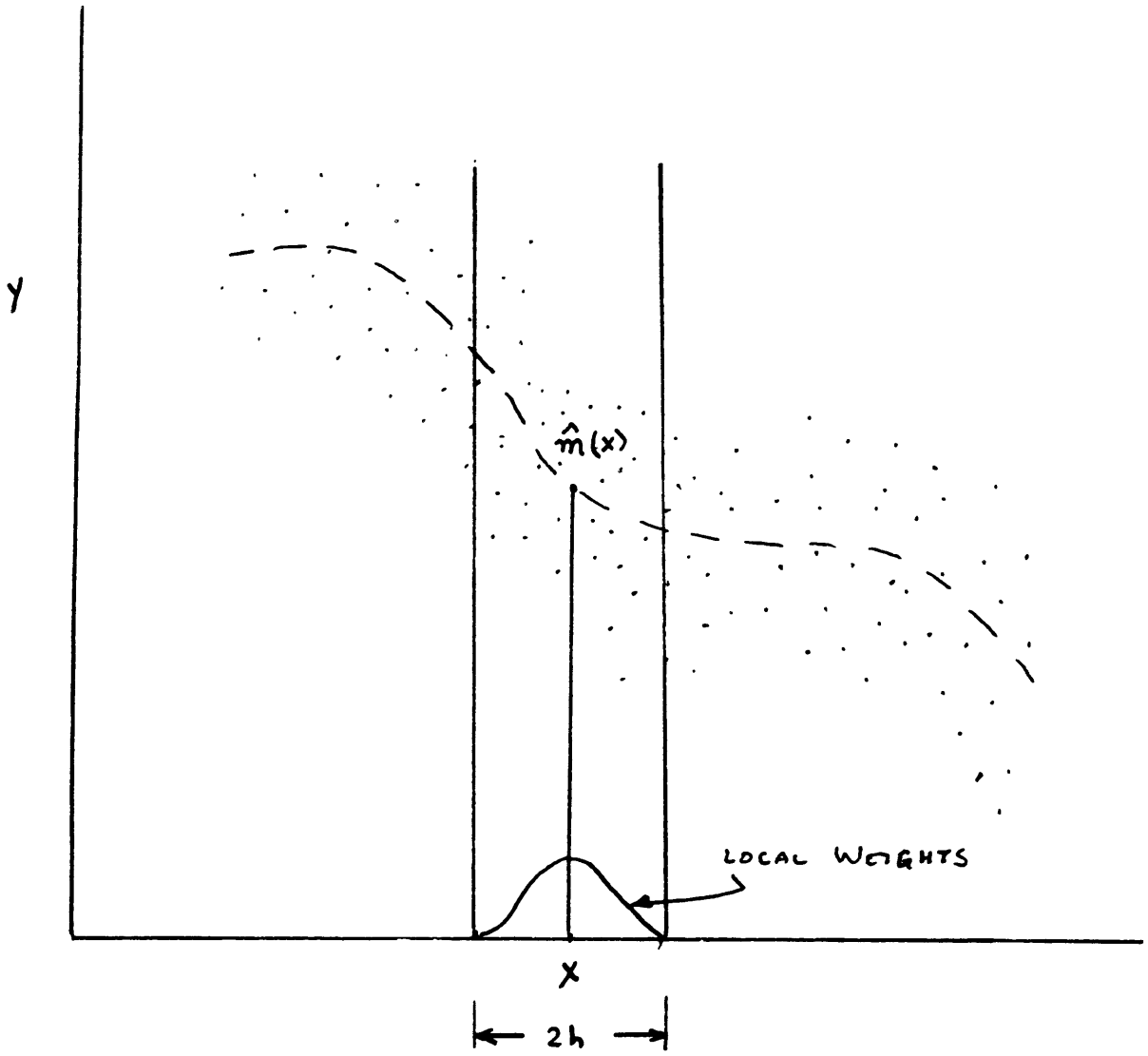

FIGURE 0: REGRESSION ESTIMATION VIA LOCAL AVERAGES

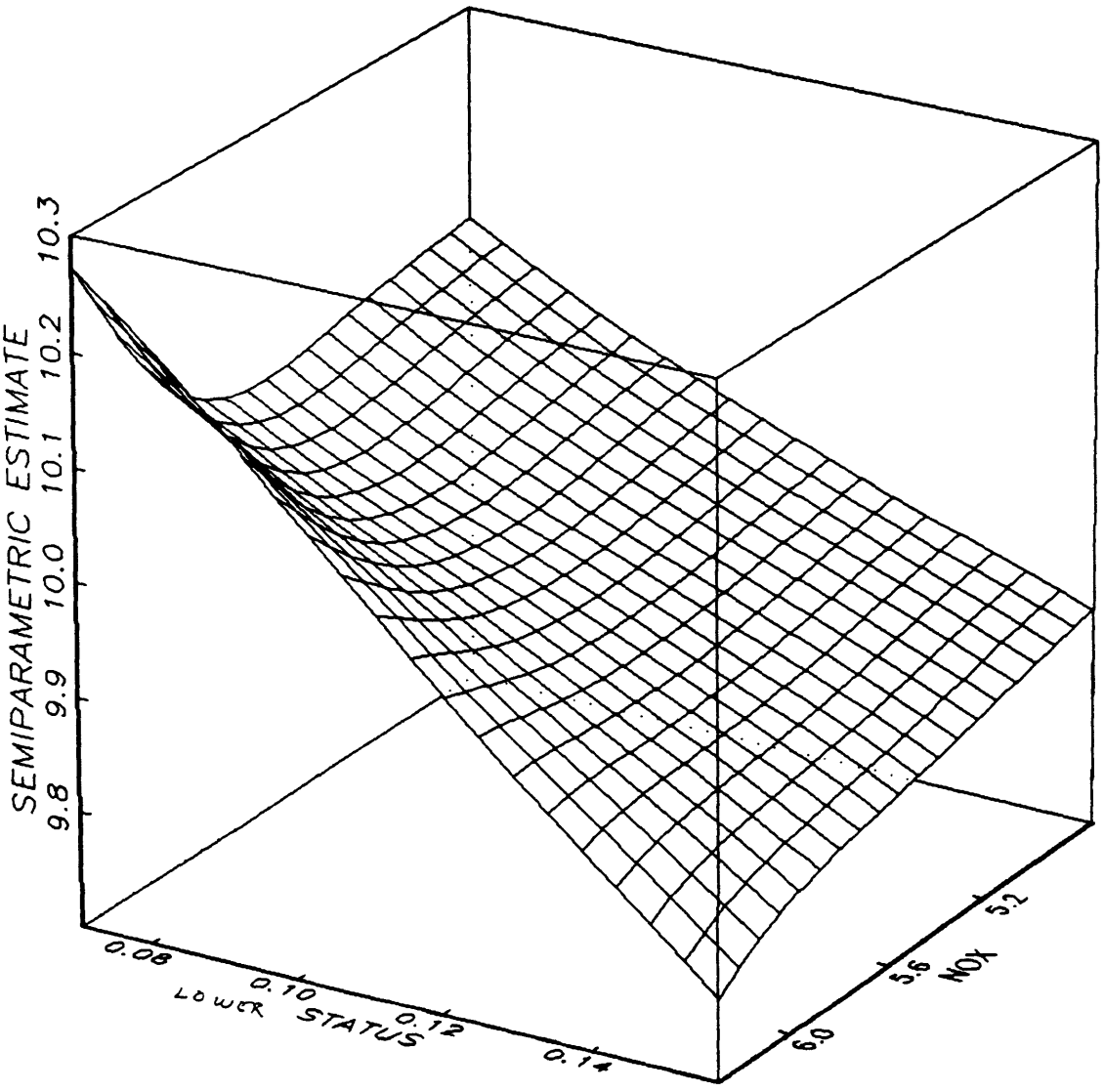

FIGURE 1a: LOG HOUSE PRICE AS RELATED TO POLLUTION AND LOWER STATUS

(Middle INDEX Value)

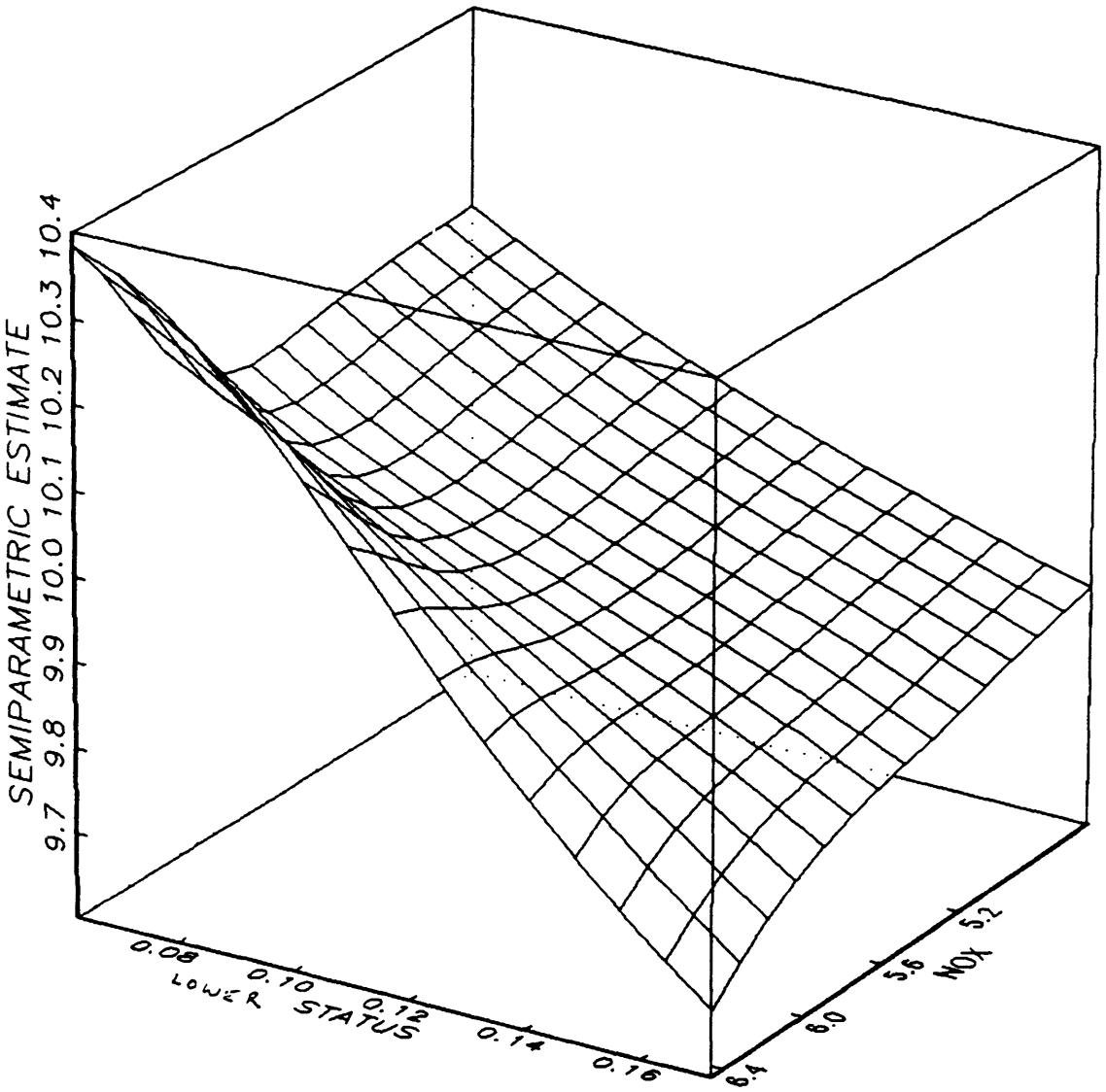

FIGURE 1b: LOG HOUSE PRICE AS RELATED TO POLLUTION AND LOWER STATUS

(High INDEX Value)

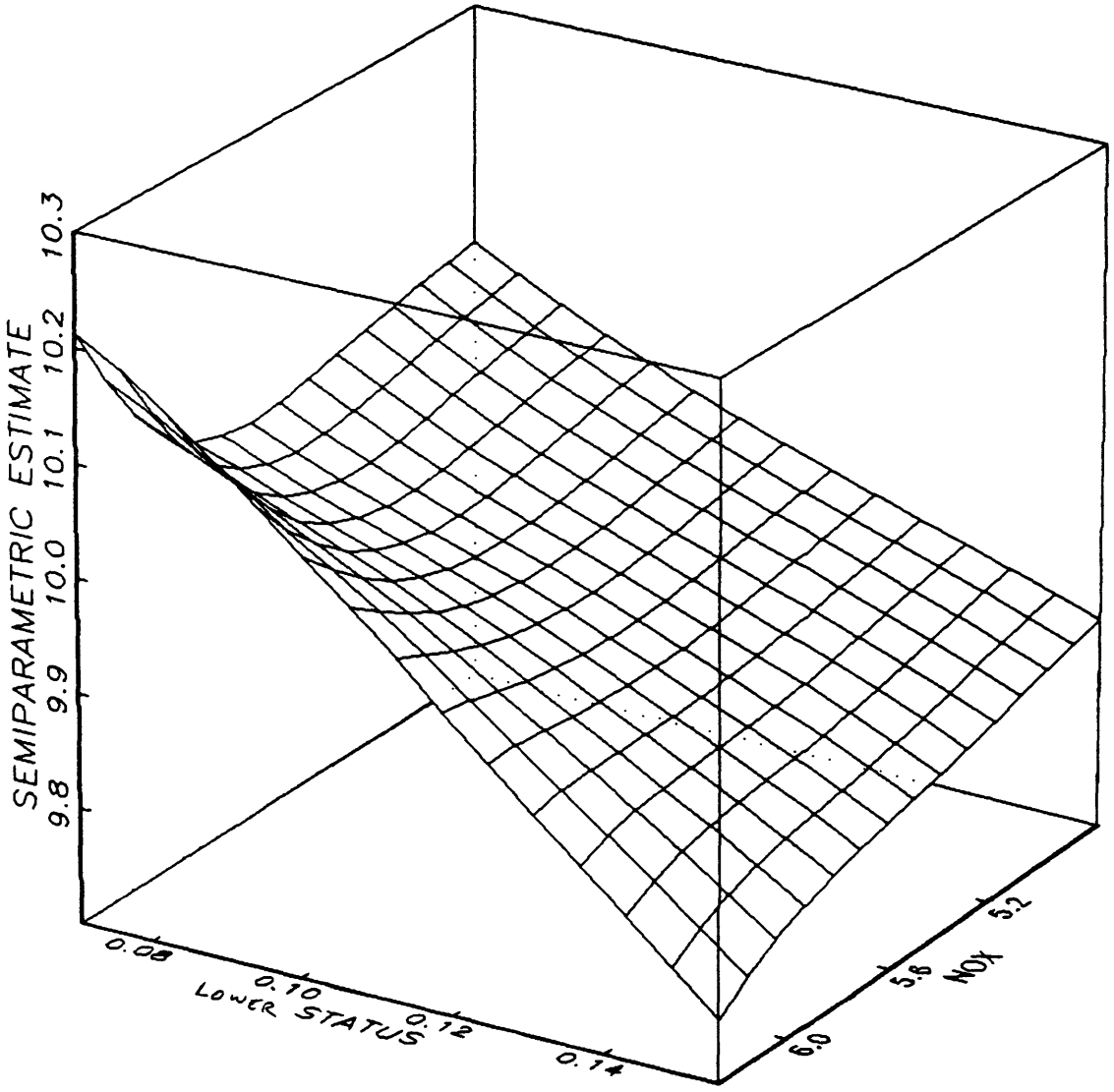

FIGURE 1c: LOG HOUSE PRICE AS RELATED TO POLLUTION AND LOWER STATUS

(Low INDEX Value)

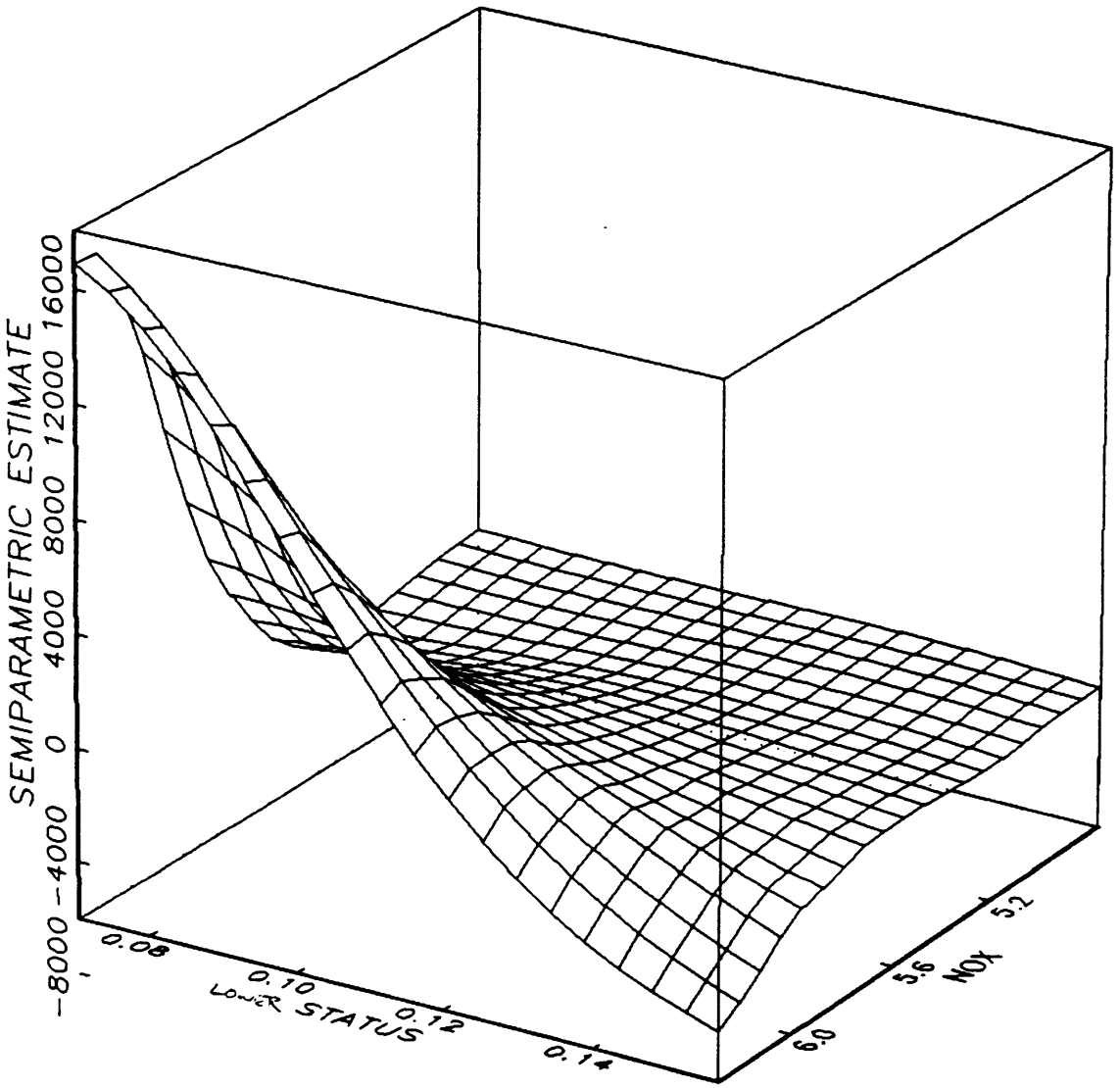

FIGURE 2: HEDONIC PRICE OF POLLUTION

(Middle INDEX Value)

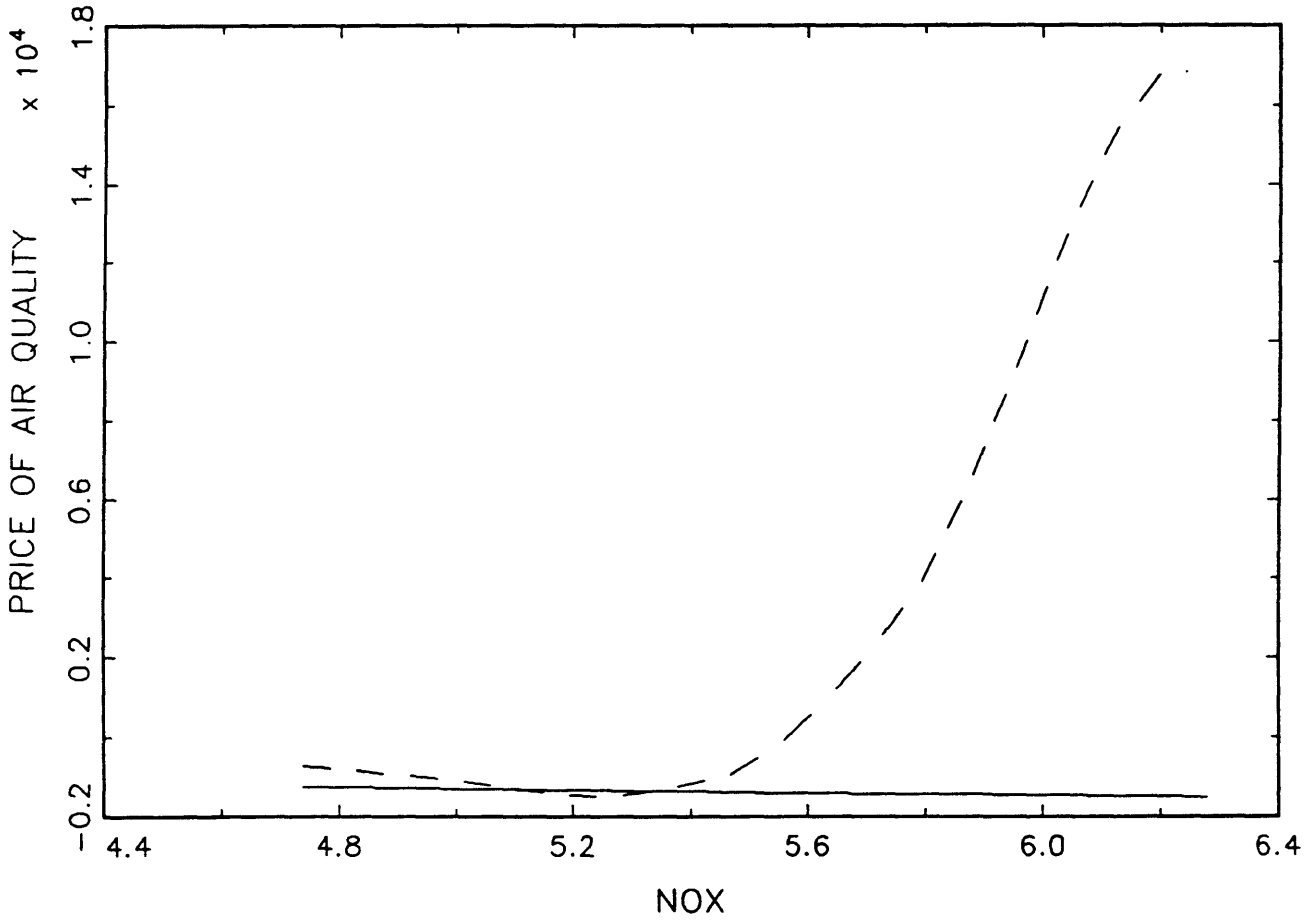

FIGURE 3a: HEDONIC PRICE OF POLLUTION: HIGH STATUS COMMUNITIES

(Middle INDEX Value; --- Semiparametric Model; ——— Log Linear Model)

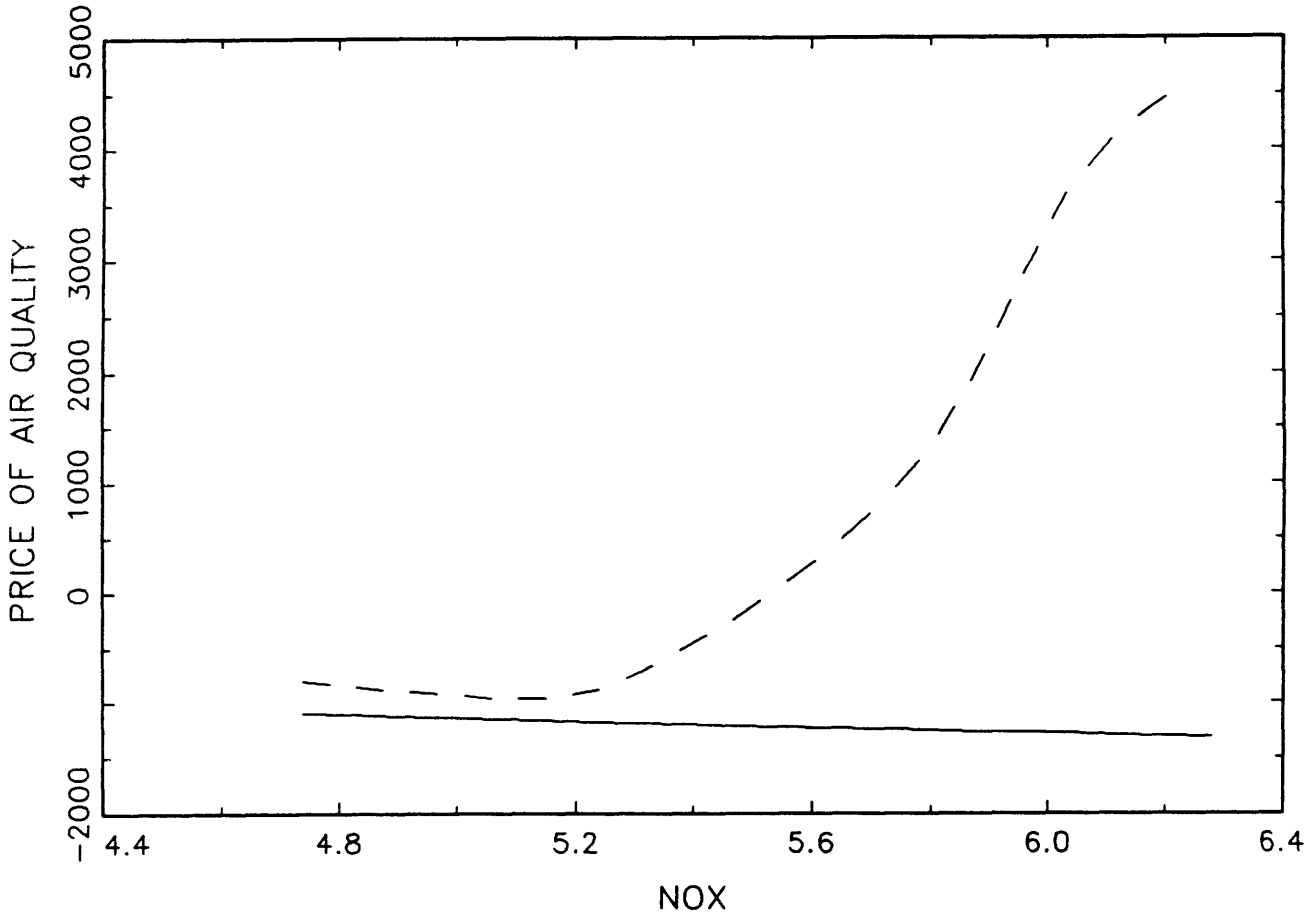

FIGURE 3b: HEDONIC PRICE OF POLLUTION: MIDDLE STATUS COMMUNITIES

(Middle INDEX Value; --- Semiparametric Model; ——— Log Linear Model)

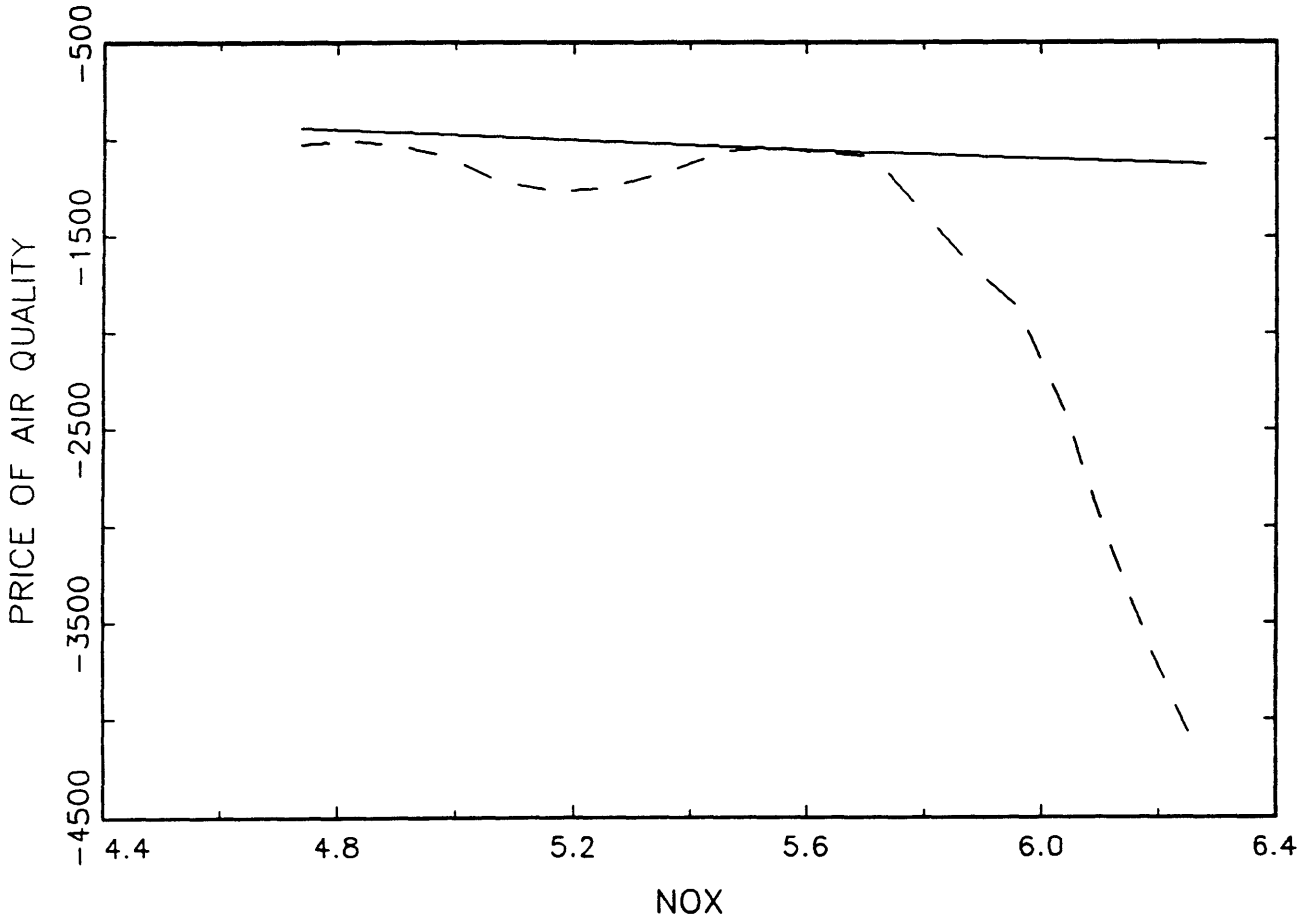

FIGURE 3c: HEDONIC PRICE OF POLLUTION: LOW STATUS COMMUNITIES

(Middle INDEX Value; --- Semiparametric Model; ——— Log Linear Model)

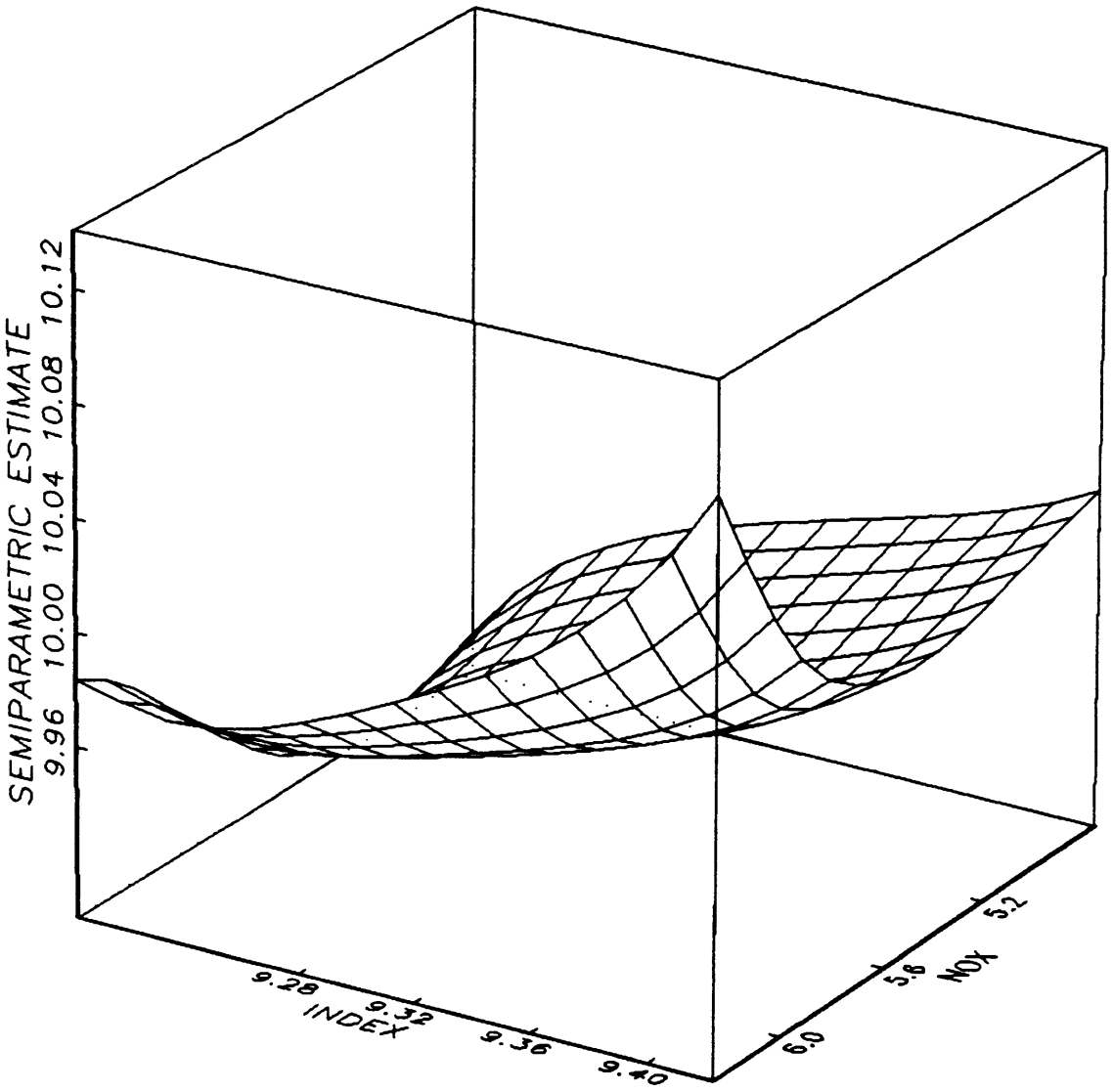

FIGURE 4: LOG HOUSE PRICE AS RELATED TO POLLUTION AND INDEX OF OTHER HOUSING ATTRIBUTES

(Middle LSTAT Value)

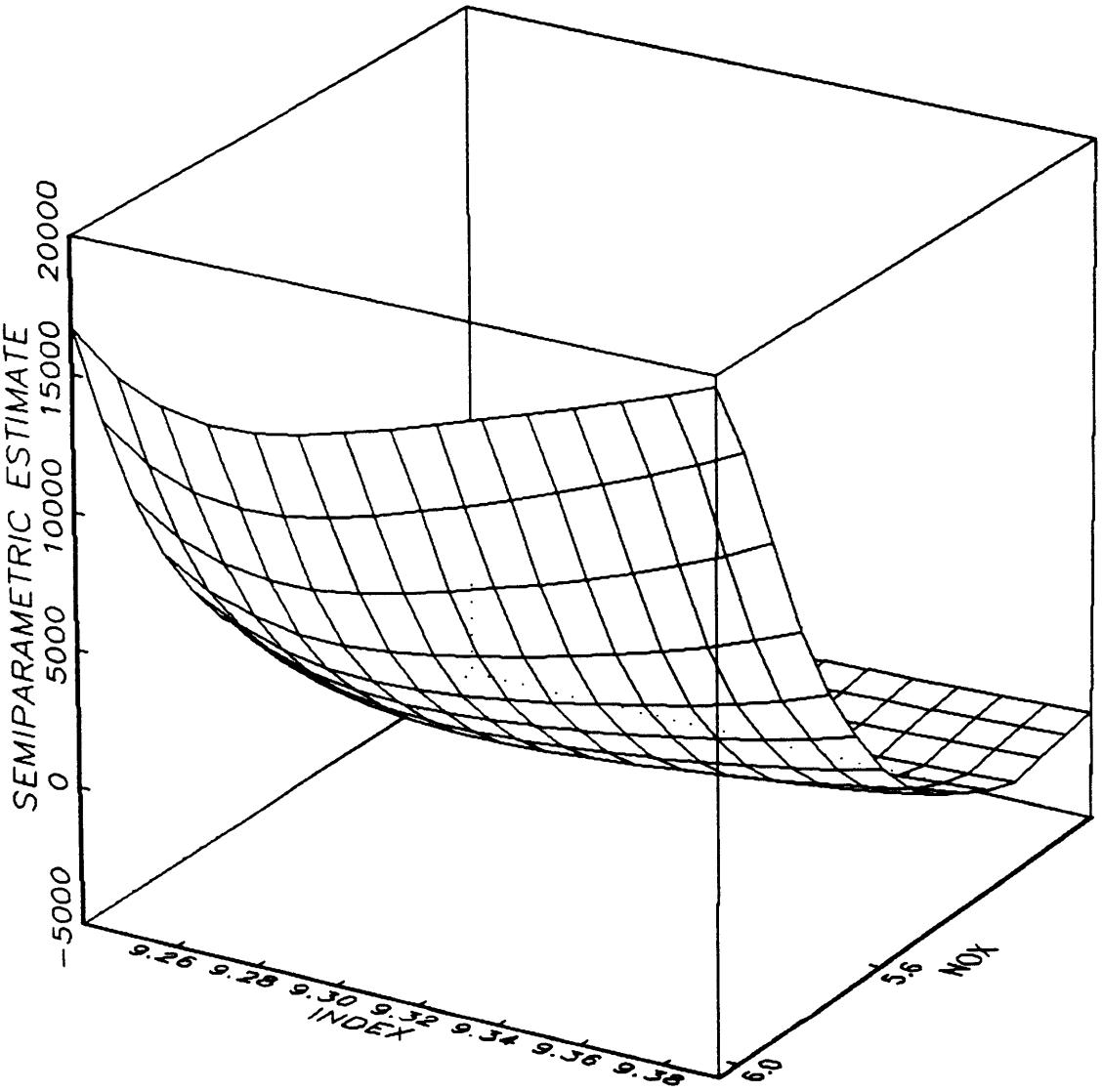

FIGURE 5a: HEDONIC PRICE OF POLLUTION: HIGH STATUS COMMUNITIES

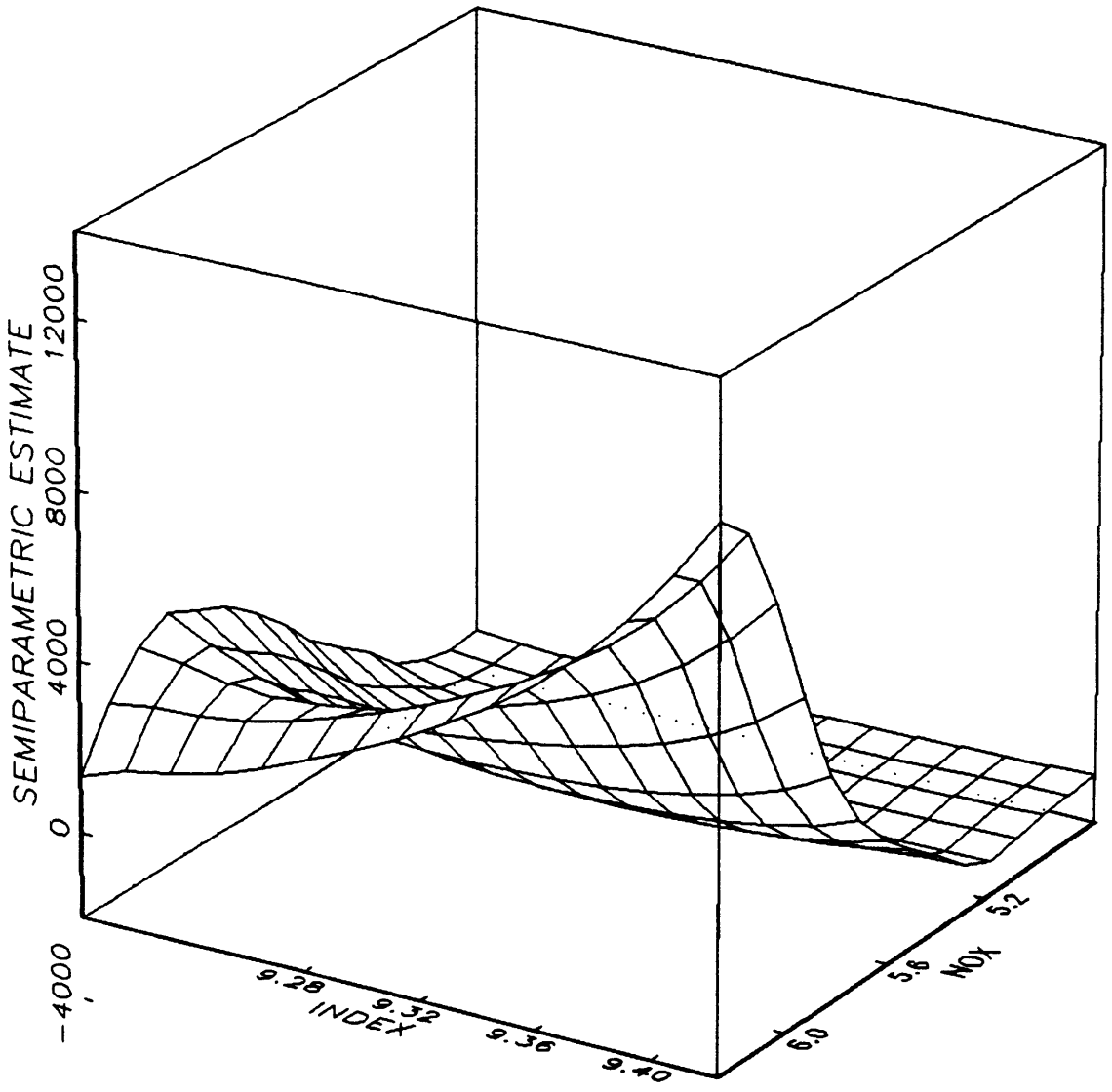

FIGURE 5b: HEDONIC PRICE OF POLLUTION: MIDDLE STATUS COMMUNITIES

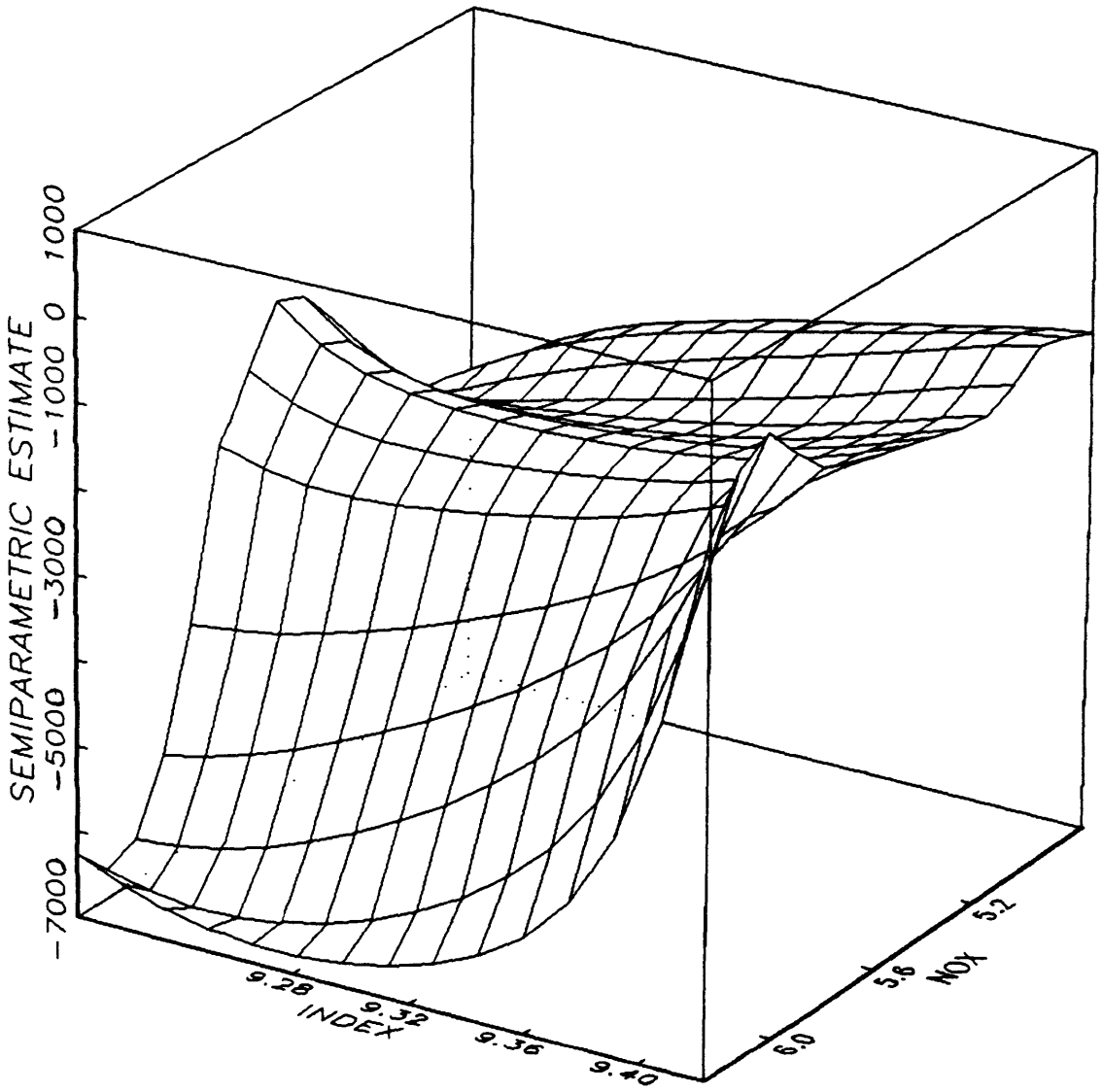

FIGURE 5c: HEDONIC PRICE OF POLLUTION: LOW STATUS COMMUNITIES

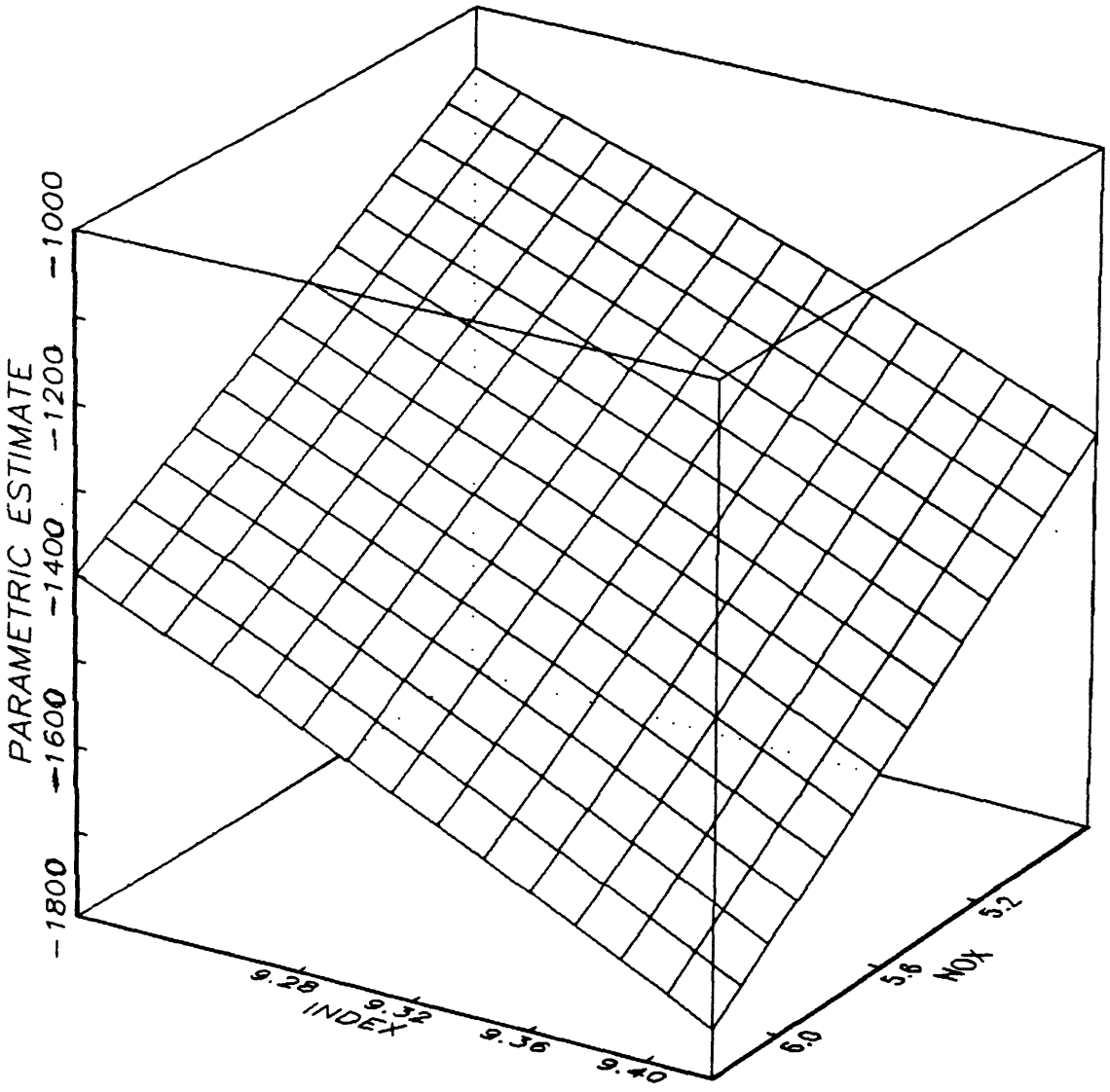

FIGURE 6: HEDONIC PRICE OF POLLUTION: LOG LINEAR MODEL