SEMIPARAMETRIC ESTIMATION OF A HEDONIC PRICE FUNCTION

PAUL M. ANGLIN AND RAMAZAN GENÇAY

Department of Economics, University of Windsor, Windsor, Ontario, N9B 3P4, Canada

E-mail: gencay@uwindsor.ca

SUMMARY

Previous work on the preferred specification of hedonic price models usually recommended a Box–Cox model. In this paper we note that any parametric model involves implicit restrictions and they can be reduced by using a semiparametric model. We estimate a benchmark parametric model which passes several common specification tests, before showing that a semiparametric model outperforms it significantly. In addition to estimating the model, we compare the predictions of the models by deriving the distribution of the predicted log(price) and then calculating the associated prediction intervals. Our data show that the semiparametric model provides more accurate mean predictions than the benchmark parametric model.

1. INTRODUCTION

The basic theory of hedonic prices has been well known for many years and was less formally understood for many-years before that (see, for example, Griliches, 1961; Rosen, 1974; Follain and Jiminez, 1985). For this reason, it has been widely used to study many problems where economic variables need to be adjusted for obvious qualitative differences, especially in the housing market. Unfortunately, economic theory provides little guidance concerning the functional form of the dependence of price on quality [1] and researchers have used forms which are somewhat flexible in order to let the data 'speak'. This paper uses the method of semiparametric estimation to study the conditional mean of log(price) of a residential housing unit.

One of the first applications of the Box–Cox model within the context of residential housing prices is presented by Goodman (1978). Goodman finds that a linear functional form is overly restrictive and generally rejected in favour of the Box–Cox transformation. Halvorsen and Pollakowski (1981) also discuss the proper specification of the hedonic price function and recommend the use of a Box–Cox transformation. In their comment on Halvorsen and Pollakowski, Cassel and Mendelson (1985) note four problems with a Box–Cox transformation, only two of which are relevant here. First, they argue that the coefficients of a non-linear transformation are 'cumbersome' to use properly. Rasmussen and Zuehlke (1990) counter this argument by saying that coefficients are not necessarily harder to use and interpret if a linear

---

[1] Coulson (1989), Colwell (1993) and Arguea and Hsiao (1993) present some hypotheses based on arbitrage opportunities.

model is used and the interactions are explicitly recognized. We will counter Cassel and Mendelson's argument in a different way by arguing that a simple parametric form may provide a simple interpretation but, to be useful, it should be accurate. Their second relevant point relates to the criterion of selecting a model. They argue that the study of hedonic prices is interesting because of the information revealed about the marginal prices of individual regressors. However, they do not offer a better criterion for selecting a model. Our analysis identifies the preferred model using the specification testing because that is most consistent with the model that is presumed to generate the data.

More recently, Cropper, Deck and McConnell (1988) use a Monte Carlo study to investigate the performance of different functional forms. Where Halvorsen and Pollakowski and Cassel and Mendelson use a 'real world' data set, Cropper *et al.* carefully specify a single type of utility function for a group of consumers and produce a market price gradient by allowing the taste parameters of this function to be randomly distributed. Six different models (such as translog or Box–Cox) are considered. They find that a linear Box–Cox regression produces the most accurate estimates of marginal attribute prices, implicitly arguing that Cassel and Mendelson's argument about the proper selection criterion is correct.[2] However, they qualify this conclusion by pointing out that if some variables are unobserved or if a proxy variable is used, then a simple linear function may outperform the others using the same criterion.

One way to make the specification of the regression function robust is to use a model which does not impose an *a priori* parametric specification such as a non-parametric model. Several techniques can be used to estimate such models including 'cubic splines', 'nearest neighbours', 'series approximators' and 'kernel estimates'. A useful and common alternative to pure non-parametric regression is the semiparametric regression (Robinson, 1988) which incorporates some parametric information into a non-parametric regression.

Stock (1989, 1991) uses this alternative of a semiparametric regression to estimate the effect of removing hazardous waste on house prices. Using a Monte Carlo simulation in his 1989 paper, he does not *contrast* functional forms but does show that the semiparametric estimate of a response coefficient associated with waste clean-up, his variable of interest, is biased toward zero. His simulations indicate that this bias is most apparent near the extreme points of observed data and when the number of regressors is large. The second paper uses data on 324 houses in the Boston area that are near waste-disposal sites to find the effect of a clean-up. This paper compares models and spends much time discussing the effects of 'bandwidth' and 'kernel function', which are defined and discussed at length below in our review of how non-parametric models can be estimated. Stock shows that the choice of kernel function can significantly affect the estimates, with a mild recommendation for a Gaussian kernel, but that the estimates are relatively insensitive to the choice of bandwidth. However, he does not provide a complete comparison of any parametric model versus any semiparametric model which is the focus of our article.

The paper which may be closest in spirit to our own is the recent paper by Pace (1993). In it, Pace reviews the literature on alternative estimators and presents two examples with some comparison. Our paper attempts to provide a conclusive comparison including more careful consideration of the benchmark parametric model to show that a simple alternative semiparametric model can outperform the benchmark parametric model. The message of our paper is more serious because the benchmark parametric model that we reject passes many of the specification tests that researchers use to reassure themselves that the parametric specification is reasonable.

---

[2] They do not provide data on the frequency that the two criteria provide conflicting advice.

Since the economic theory of hedonic prices is well known and not in question, we assume that readers are familiar with the relationship between price and quality that can be supported by a market. The next section discusses the econometric theory of non-parametric and semiparametric estimation. Section 3 describes the data set that we use and Section 4 presents our results. The results of this paper suggest that even the best parametric model imposes restrictions that substantially reduce the explanatory power of a regression equation. In Sections 4.3 and 4.4 we study the statistical significance of different models by comparing their predictions.

2. ESTIMATION TECHNIQUES

2.1. Semiparametric Regression

Consider a model

$$p_i = M(x_i) + \varepsilon_i \qquad E(\varepsilon_i \,|\, x_i) = 0 \qquad (i = 1, 2, \ldots, n)$$

where p_i is the ith observation of the regressand (i.e. log(price) for the ith house) to be explained by variations in the regressor(s) x_i. x_i is of dimension $1 \times k$. Estimating and testing such a model often requires imposing many strong assumptions on $M(\cdot)$ to produce results. Parametric estimation techniques, which are familiar to most economists, would assume that $M(\cdot)$ is a linear function of x_i or a linear function of some power of x_i,

$$M(x_i) = x_i \gamma$$

where γ is a $k \times 1$ parameter vector. Non-parametric models use a more general functional form which, while more robust, tend to be less precise (see Silverman, 1986; Ullah, 1988; Härdle, 1991). This problem is especially severe when there are many regressors because the rate of convergence of the non-parametric response coefficients slows down as the number of regressors increases.

An intermediate strategy is to employ a semiparametric form, which incorporates some parametric information into the non-parametric form:

$$p_i = z_i \beta + q(x_i) + \varepsilon_i \qquad E(\varepsilon_i \,|\, z_i, x_i) = 0 \qquad (i = 1, 2, \ldots, n) \tag{1}$$

where p_i is the ith observation on a dependent variable, z_i is a $1 \times p$ vector, x_i is a $1 \times k$ vector and β is a $p \times 1$ vector. ε_i has mean zero and a conditional variance var$(\varepsilon_i \,|\, z_i, x_i)$. The functional form $q(\cdot)$ is unknown to the researcher. Robinson (1988) shows that this model can be rewritten as

$$p_i - E(p_i \,|\, x_i) = (z_i - E(z_i \,|\, x_i))\beta + \varepsilon_i \tag{2}$$

suggesting that β can be estimated in a two-step procedure: First, the unknown conditional means, $E(p_i \,|\, x_i)$ and $E(z_i \,|\, x_i)$, are estimated using a non-parametric estimation technique. Second, the estimates are substituted in place of the unknown functions in equation (2) and ordinary least squares is used to estimate β. Robinson (1988) shows that the resulting estimates are asymptotically equivalent to those where the true mean functions are known and used in the estimation. We apply these steps where $E(p_i \,|\, x_i)$ and $E(z_i \,|\, x_i)$ are estimated using a non-parametric kernel estimator.

Delgado and Mora (1993) discuss the procedures for estimating regression models where regressors are discrete and apply this discussion to semiparametric inference problems. They show that the parameter estimates of the linear part of the regression are $\sqrt{n}$-consistent, just as when regressors are continuous (i.e. the case considered in Robinson, 1988).

2.2. Non-parametric Kernel Estimation

If the sample size is n, then the usual kernel estimator $M(x) = E(p \mid x)$ is written as

$$\hat{M}(x) = \sum_{i=1}^{n} p_i \frac{K_i(i)}{\hat{f}(x)} \tag{3}$$

where

$$\hat{f}(x) = \sum_{i=1}^{n} K_i(x) \tag{4}$$

$$K_i(x) = (na_n^k \det(\Sigma_n)^{1/2})^{-1} K(a_n^{-1}\Sigma_n^{-1/2}(x - x_i)) \tag{5}$$

$\hat{f}(x)$ is the multivariate density estimate x_i evaluated at x while a_n is the bandwidth parameter. Σ_n is the sample covariance matrix of the regressors. In the present paper we use the Gaussian kernel function $K(x) = (2\pi)^{-k/2} \exp(-x^T x/2)$. Existing literature also shows that the choice of the window width can be important: too large a value of a_n induces bias and too small a value induces imprecise estimates. Delgado and Robinson (1992), Härdle (1990) and Ullah (1988) detail the conditions that the kernel function and the bandwidth parameters a_n must satisfy in order to obtain the desired asymptotic properties of the regression function estimator. Recent research examines the properties of the various automatic bandwidth selection procedures. Some of the available choices are the 'cross-validation', 'penalizing functions' and the 'plug-in' methods (e.g. Devroye and Penrod, 1984; Gasser *et al.*, 1984; Li, 1984; Muller, 1984; Rice, 1984; Stone, 1984; Marron, 1985; Härdle, Hall and Marron, 1988). A seemingly natural method of choosing the bandwidth is to minimize the sum of squared residuals of the regression equation:

$$MSE = n^{-1} \sum_{i=1}^{n} (p_i - z_i\beta - q(x_i))^2 \tag{6}$$

Because p_i is used when estimating $q(x_i)$, the mean square error (MSE) in equation (6) can be reduced arbitrarily by decreasing the bandwidth until effectively all weight in $q(x_i)$ is placed on p_i. We choose the bandwidth parameter by cross-validation which avoids this problem by choosing a bandwidth which minimizes the sum of squared residuals from the cross-validated regression. In this approach, the bandwidth minimizes

$$MSE_{cv} = n^{-1} \sum_{i=1}^{n} (p_i - z_i\beta - q(x_i))^2 \tag{7}$$

where the estimator of $q(x_j)$ is

$$\hat{q}(x_i) = \sum_{i \neq 1}^{n} (p_i - z_i\hat{\beta})K_j(x)/\hat{f}(x) \tag{8}$$

Using MSE_{cv} avoids the difficulty that arises when using MSE because p_i is not used in estimating $q(x_i)$.

3. DATA DESCRIPTION

Like people in many cities, many of the quarter-million people who live in Windsor and area live in houses that are bought and sold. Our data was provided by the Windsor and Essex County

Table I. Data summary

Variable	Mean	St. dev.	Min	Max.
Sale price, P	68122	26703	25 000	190 000
LOT	5150	2169	16 50	16 200
$BDMS$	2·965	0·737	1	6
REC	0·178	0·383	0	1
STY	1·808	0·868	1	4
$FFIN$	0·350	0·477	0	1
GHW	0·046	0·209	0	1
CA	0·317	0·466	0	1
GAR	0·692	0·861	0	3
DRV	0·859	0·348	0	1
REG	0·234	0·424	0	1
FB	1·286	0·502	1	4

Real Estate Board and describes residential houses sold during July, August and September of 1987 through the local Multiple Listing Service. The 546 records contain information describing the key features of each house. Table I provides a summary of the data. The variables are defined as follows.

- $DRV = 1$ if the house has a driveway
- $REC = 1$ if the house has a recreational room
- $FFIN = 1$ if the house has a full and finished basement
- $GHW = 1$ if the house uses gas for hot water heating
- $CA = 1$ if there is a central air conditioning
- GAR shows the number of garage places
- $REG = 1$ if the house in located in a preferred neighbourhood of the city: Riverside or South Windsor
- LOT is a continuous variable showing the lot size of the property in square feet
- $BDMS$ is the number of bedrooms
- FB is the number of full bathrooms (i.e. including, at least, a toilet, sink, and bathtub)
- STY represents the number of stories, excluding the basement.

In general, Windsor is a typical mid-size Canadian city.

4. EMPIRICAL RESULTS

4.1. Benchmark Parametric Model

The specification of the benchmark model is

$$\log(P_i) = \beta_0 + \beta_1 DRV_i + \beta_2 REC_i + \beta_3 FFIN_i + \beta_4 GHW_i + \beta_5 CA_i + \beta_6 GAR_i$$
$$+ \beta_7 REG_i + \gamma_1 \log(LOT_i) + \gamma_2 \log(BDMS_i) + \gamma_3 \log(FB_i)$$
$$+ \gamma_4 \log(STY_i) + u_i \tag{9}$$

and includes ten variables on the characteristics of a house plus a neighbourhood variable. The mean of u_i, conditional on the explanatory variables, is zero. The characteristic variables are driveway (DRV), recreation room (REC), finished basement ($FFIN$), gas heating (GHW), central air (CA), garage (GAR), neighbourhood dummy variable (REG), lot size in square feet

(*LOT*), number of bedrooms (*BDMS*), number of full bathrooms and the number of stories (*STY*). We choose not to use data on the number of rooms, even though such data is available, because it is highly collinear with the number of bedrooms, number of full bathrooms, the presence of a finished basement, or a recreation room which are included separately.

This specification is a special case of a more general Box–Cox functional form. We tested whether a more general form would better explain this data using a double-length regression test (Davidson and Mackinnon, 1985) and the test resulted in a calculated DLR *t*-test statistic equal to 1.462. Thus we feel justified in using a loglinear specification for the benchmark model. Table II shows the estimated coefficients from this model.

To test for a potential misspecification of the regression function, the RESET test of Ramsey (1969) is calculated by adding the squared and cubed fitted values of the regression equation (9) as additional regressors to equation (9).[3] If squared fitted values are added as the only extra regressor, the calculated *t*-test (with heteroscedasticity robust standard errors) on its coefficient is 1·031. If squared and cubed fitted values are included as additional regressors, then an *F*-test on excluding these variables gives a test statistic equal to 1·020. Given that the 1% tabulated *t*-statistic is 2·576 and $F_{0 \cdot 01}(2,532) = 4.61$, this test does not provide evidence of misspecification.

The RESET test is commonly used but it is not the only possible test. We supplied our data to one of the referees who used them for a different specification test based on Wooldridge (1992). By regressing $\log(P)$ on the variables used in the benchmark model, expressed in levels instead of logs, plus squared and cubed terms for *LOT*, *BDMS*, *FB* and *STY*, the referee obtained fitted values $\log(\hat{P})$. When these fitted values are added to equation (9), the resulting *t*-statistic was 3·47. Thus the referee's suggested test rejects our benchmark model. We discuss the implications of these two tests at the end of the next section.

The OLS model with the discrete variables, *BDMS*, *FB* and *STY* in levels, is presented in Table III. The performance of this model is not substantially different from the benchmark

Table II. Ordinary least squares estimation of the parametric benchmark model dependent variable: $\log(P)$

	Coefficients	*t*-statistics
Constant	7·921	36·137
DRV	0·110	3·863
REC	0·060	2·283
FFIN	0·096	4·417
GHW	0·173	3·929
CA	0·171	8·031
GAR	0·049	4·238
REG	0·130	5·693
$\log(LOT)$	0·313	11·623
$\log(BDMS)$	0·089	2·031
$\log(FB)$	0·264	8·450
$\log(STY)$	0·165	6·627
SSR	23·838	
R^2	0·684	
$\bar{R}^2$	0·677	
Number of observations	546	
Log likelihood function	80·119	

---

[3] In the RESET test equation the regressand $\log(P)$ can be replaced by the residual vector of equation (9) which would result in the same sum of squared residuals as in regression equation with regressand $\log(P)$.

model except that the interpretation of coefficients corresponding to *FB*, *BDMS* and *STY*. To further check the specification of the benchmark model, we considered adding higher-order terms and interactions, or using dummy variables for the different categories of the discrete variables. If the fit to these augmented models improves and gets closer to the fit of the semiparametric model, then we would conclude that these additional terms should account for much of the superior performance of the semiparametric model. An augmented model which include higher-order terms is presented in Table IV. The parameter estimates on the cross-terms are statistically insignificant and the R^2's are about the same as that of the benchmark parametric model. *F*-tests for the exclusion restrictions indicate that the added variables are statistically insignificant at the 1% levels. Thus the non-linearities which characterize the housing price behaviour cannot be accounted by simple interactions between regressors.

A full set of dummies are introduced for the *BDMS*, *FB* and *STY*. *BDMS*, *FB* and *STY* take 6, 4 and 4 distinct values (i.e. *BDMS*1 = 1 if and only if *BDMS* = 1; *BDMS*2 = 1 if and only if *BDMS* = 2; ...). This totals 64 interactions. Full interaction with log(*LOT*) produces another 64, and with the square of log(*LOT*), 64 more which total 192 interactions. The coefficients of the appropriate OLS model with this many interactions cannot be estimated due to singularity problems. There are not too many observations for *BDMS* = 1, 5 and 6 and *FB* = 3 and 4. Accordingly, we re-estimated the OLS model by excluding the interactions corresponding to *BDMS* = 1, 5 and 6 and *FB* = 3 and 4. This overparametrized model yields an R^2 of 0·734 and an adjusted R^2 of 0·682, both of which improve on the benchmark model's R^2's.

4.2. Semiparametric Estimation

An alternative to the benchmark parametric model is the semiparametric model (1). Robinson (1988) shows that if β is estimated using

$$p_i - \hat{E}(p_i \mid x_i) = (z_i - \hat{E}(z_i \mid x_i))\beta + \text{error}_i \tag{10}$$

Table III. Ordinary least squares estimation of the parametric model with discrete variables in levels: dependent variable: log(*P*)

	Coefficients	*t*-statistics
Constant	7·745	35·801
DRV	0·110	3·904
REC	0·058	2·225
FFIN	0·105	4·817
GHW	0·179	4·079
CA	0·166	7·799
GAR	0·048	4·179
REG	0·132	5·816
log(*LOT*)	0·303	11·356
BDMS	0·344	2·410
FB	0·166	8·154
STY	0·092	7·268
SSR	23·638	
R^2	0·687	
$\bar{R}^2$	0·680	
Number of observations	546	
Log likelihood function	82·412	

Table IV. Ordinary least squares estimation of the augmented parametric model:
dependent variable: $\log(P)$

	Coefficients	t-statistics
Constant	7·634	7·845
DRV	0·108	3·770
REC	0·062	2·327
FFIN	0·102	4·534
GHW	0·169	3·799
CA	0·171	7·955
GAR	0·050	4·276
REG	0·127	5·422
$\log(LOT)$	0·347	2·970
$\log(BDMS)$	0·454	0·457
$\log(FB)$	−0·719	−1·051
$\log(STY)$	0·383	0·695
$\log(LOT) \times \log(BDMS)$	−0·042	−0·357
$\log(LOT) \times \log(FB)$	0·102	1·265
$\log(LOT) \times \log(STY)$	−0·022	−0·339
$\log(BDMS) \times \log(FB)$	0·061	0·415
$\log(BDMS) \times \log(STY)$	−0·040	−0·352
$\log(FB) \times \log(STY)$	0·059	0·814
SSR	23·698	
R^2	0·686	
$\bar{R}^2$	0·676	
Number of observations	546	
Log likelihood function	81·721	
F-test of exclusion restrictions for the cross-terms	0·520	
$F_{0 \cdot 01}$ (6,528)	2·81	

then it is $\sqrt{n}$-consistent. The specification of the semiparametric model is[4]

$$\log(P_i) = \beta_1 DRV_i + \beta_2 REC_i + \beta_3 FFIN_i + \beta_4 GHW_i + \beta_5 CA_i + \beta_6 GAR_i$$
$$+ \beta_7 REG_i + q[LOT_i, BDMS_i, FB_i, STY_i] + \varepsilon_i \qquad (11)$$

where the mean of ε_i, conditional on the explanatory variables, is zero.   In the specification of the semiparametric regression, the dummy variables enter into the linear part and the discrete and continuous variables enter into the unknown function $q(\cdot)$. The dummy variables would only cause scale effects if they were included in $q(\cdot)$ but would not affect the curvature of this function. Therefore, the dummy variables are modelled in an additive fashion whereas all discrete variables and the continuous variables are modelled in the unknown function $q(\cdot)$. Table V presents estimates of β from equation (11). We choose the bandwidth parameter by cross-validation as described in Section 2.2. The bandwidth parameter is minimized at 0·446 which we use in the estimation of the semiparametric regression model in equation (11).

The semiparametric model explains substantially more variation in the dependent variable than the benchmark parametric model,[5] $R^2 = 0 \cdot 923$ versus 0·684. If the effective degrees of freedom of the semiparametric model is large, it is more appropriate to compare the adjusted

---

[4] As mentioned in the Introduction, one advantage of using $\log(P)$ as the dependent variable is that doing so provides a common basis of comparison between the different models.
[5] $\hat{p} = \hat{E}(p \mid x) + (z - \hat{E}(z \mid x)) \in \beta$. R^2 of the semiparametric model is calculated by $R^2 = \hat{p}^T \hat{p} / p^T p$.

Table V. Semiparametric regression: dependent variable: $\log(P)$

	Coefficients	t-statistics
DRV	0·147	3·081
REC	0·078	2·830
FFIN	0·097	4·245
GHW	0·191	4·245
CA	0·158	7·153
GAR	0·064	4·788
REG	0·124	4·898
SSR	5·809	
R^2	0·923	
Number of observations	546	

R^2's of both models. This effective degrees of freedom, however, is not known to us. Therefore, it is more appropriate to judge the performance of these two competing models by looking into their out-of-sample predictions and with tests involve the comparison of these two models.

There are a number of tests that compare parametric and semiparametric models (see Robinson, 1988; Whang and Andrews, 1993; Delgado and Stengos, 1994, among others). Here we will employ the tests proposed by the first two papers, since the Delgado and Stengos (1994) paper deals with the non-nested models.

As in Hausman (1978), one may base a test on a vector of contrasts, that is, the vector of differences between two vectors of estimates, one of which will be consistent under weaker conditions than the other. We calculated the Hausman-type specification test given in Robinson (1988):

$$H_n = n(\tilde{\beta} - \hat{\beta})^{\mathrm{T}} \Phi^{-1} (\tilde{\beta} - \hat{\beta})$$

where $\tilde{\beta}$ and $\hat{\beta}$ are the estimates for β from equations (11) and (9), respectively and Φ is the variance–covariance matrix[6] of $(\tilde{\beta} - \hat{\beta})$. The null hypothesis for H_n is that the underlying model is linear, so that the ordinary least squares is the best linear unbiased estimator. The alternative hypothesis is that the model is semiparametric, linear with a non-linear component. In fact, H_n tests the linearity of the $q(z)$ and rejection of the null hypothesis implies that the ordinary least squares estimates would be inconsistent. H_n is distributed $\chi^2(7)$ and the calculated test statistic is 19·216. Given that $\chi^2_{0·01} = 18·48$ and $\chi^2_{0·05} = 14·07$, the null hypothesis of linearity is rejected.

Recently, Whang and Andrews (1993) proposed a test of a linear model against a semiparametric model. The specification of this test statistic is

$$G_n = n\bar{l}(\hat{\beta}, \hat{q})^{\mathrm{T}} \Psi^{-1} \bar{l}(\hat{\beta}, \hat{q})$$

where $\bar{l}(\hat{\beta}, \hat{q})$ is a $p \times 1$ vector where

$$l_i(\hat{\beta}, \hat{q}) = [p_i - \hat{E}(p_i \mid x_i) - (z_i - \hat{E}(z_i \mid x_i))\hat{\beta}] \times [(z_i - \hat{E}(z_i \mid x_i))]^{\mathrm{T}}$$

$\hat{\beta}$ is the estimator of β from the benchmark parametric model and Ψ is a consistent variance–covariance estimator. The calculated test statistic is 18·950. Given that $\chi^2_{0·01}(7) = 18·48$ and $\chi^2_{0·05}(7) = 14·07$, the null hypothesis of loglinear specification is rejected.

---

[6] Recently, Fan, Li and Stengos (1995) show that the Robinson (1988) semiparametric model can be estimated under conditional heteroscedasticity of errors.

An implication of these specification tests is that the RESET test, which was applied to the parametric model above and is supposed to detect the higher-order interactions between regressors, may not have enough power to detect the effects which semiparametric techniques exploit. Other specification tests,[7] such as the one used by one of the referees to this paper, are able to raise doubts about the specification of our benchmark model but they are not as widely used. As might be expected and as the previous section demonstrated, adding extra regressors or some higher-order terms can improve the performance of a parametric model but, as this section shows, a semiparametric model provides better in- and out-of-sample performance in comparison to the benchmark parametric model.

4.3. Prediction and Prediction Intervals

As demonstrated above, it is relatively easy to compute a point estimate for a prediction. However, it is less common to compute a prediction interval for semiparametric models even though it provides important information. Here we discuss two ways to compute a prediction interval.

4.3.1. Asymptotic prediction intervals

A semiparametric model can be written as

$$p_i - z_i\beta = q(x_i) + \varepsilon_i \tag{12}$$

Therefore, for any x, $q(x)$ is estimated by a kernel regression of $(p_i - z_i\epsilon\beta)$ on x:

$$\hat{q}(x) = \sum_{i=1}^{n} K_i(x)(p_i - z_i\beta)/f(x) \tag{13}$$

where $\epsilon\beta$ is the estimate of β obtained from equation (11). We are interested in predicting $E(p^* \mid x^*, z^*)$ where $p^* = z^*\beta + q(x^*) + \varepsilon^*$:

$$\hat{p}^* = z^*\epsilon\beta + \hat{q}(x^*) \tag{14}$$

A prediction interval for $\hat{p}^*$ is based on the limiting distribution of $\hat{p}^* - \bar{p}^*$ where $\bar{p}^*$ is the mean value of p^*. By applying Theorems 2.2.2 and 2.2.3 in Bierens (1987), the limiting distribution is

$$\sqrt{na^k}(\hat{p}^* - \bar{p}^*) \to N\left(\lambda \frac{b(x^*)}{f(x^*)}, \sigma_{p^*}^2\right) \tag{15}$$

$\sigma_{p^*}^2$ is estimated by

$$\hat{\sigma}_{p^*}^2 = \frac{\hat{\sigma}_{\varepsilon,x^*}^2}{f(x^*)} \int K^2(x^*)\, dx^* \tag{16}$$

$$\hat{\sigma}_{\varepsilon,x^*}^2 = \left(\sum_{i=1}^{n} (p_i - z_i\beta - \hat{q}(x_i)^2 K_i(x)\right) \bigg/ f(x^*) \tag{17}$$

---

[7] Lee, White, and Granger (1993) compare the RESET test to other tests for neglected nonlinearity and show that the RESET test does not perform well with some specifications.

and $a \to 0$, $na^k \to \infty$ and $a^2\sqrt{na^k} \to \lambda$ with $0 \leqslant \lambda \leqslant \infty$. Since the estimate of the limiting distribution is biased, it requires a correction. We use undersmoothing to correct for this bias, since it is one of the methods suggested by Bierens (1987).

The 99% prediction interval for the semiparametric model is

$$\log(\hat{P})^* \pm 2.576 \left(n \prod_{i=1}^{k} a_i \right)^{-0.5} \hat{\sigma}_{p^*} \tag{18}$$

The 99% prediction interval for the parametric model is

$$\log(\hat{P})^* \pm t_{0.005} s \sqrt{x^*(X^TX)^{-1}x^{*T}} \tag{19}$$

where X is an $n \times (p+k)$ matrix of regressors of equation (9), x^* is a $1 \times (p+k)$ row vector which contains the description of the reference house and s is the standard error of the parametric regression in equation (9).

4.3.2. Prediction intervals with wild bootstrap

An alternate and well-known method of computing prediction intervals is to use bootstrap methods. For our purposes, the *wild* bootstrap method is the most appropriate in terms of accounting for the bias, discussed above, in non-parametric and semiparametric regression. We follow the approach in Härdle and Marron (1991).

Their essential idea is to resample the estimated residuals to construct an estimator whose distribution will approximate the distribution of the original estimator. To retain the conditional distributional characteristics of the estimate, Härdle and Marron (1991) do not resample the entire set of residuals, but rather use the idea of *wild* bootstrap as in Härdle and Mammen (1989), where each bootstrap residual is drawn from a two-point distribution which has zero mean and variance equal to the square of the residual and the third moment equal to the cube of the residual. In particular define a random variable ε_i^* having a two-point distribution $\hat{G}_i$, where $\hat{G}_i = \gamma\delta_a + (1 - \gamma)\delta_b$ is defined through the parameter γ, and where δ_a and δ_b denote point measures at a and b, respectively. Härdle and Marron (1991) show that $a = \epsilon\varepsilon_i(1 - \sqrt{5})/2$, $b = \epsilon\varepsilon_i(1 + \sqrt{5})/2$ and $\gamma = (5 + \sqrt{5})/10$. This ensures that $E\varepsilon^* = 0$, $E\varepsilon^{*2} = \epsilon\varepsilon_i^2$ and $E\varepsilon^{*3} = \epsilon\varepsilon_i^3$. In a certain sense, the resampling distribution $\hat{G}_i$ can be thought of as attempting to reconstruct the distribution of each residual through the use of one single observation. Therefore, it is called *wild* bootstrap. After resampling, new observations are

$$p_i^* = z_i\beta + \hat{q}(x_i) + \varepsilon_i^* \tag{20}$$

The observations $\{p_i^*, x_i, z_i\}$; $(i = 1, 2, ..., n)$ can be used to estimate β and $q(\cdot)$ and prediction intervals can be constructed for $(\hat{p}_i - p_i)$ because its distribution is approximately equal to that of $(\hat{p}_i^* - \hat{p}_i)$. Our results are based on 1000 replications.

4.4. Housing Price Predictions

4.4.1. Prediction with a reference house

Often the purpose of estimation is to find a prediction given particular values of the regressors. In the case of a housing market, we would be interested in predicting the log(price) of a house described by $z_i = Z$, $x_i = X$ such that $z_i\epsilon\beta + \hat{q}(x_i)$ provides a point estimate for this purpose. Thus for a given house, ordinary statistical arguments can be applied. The natural place

to start is with a 'reference house' such as

- *DRV* = 1
- *BDMS* = 2
- *STY* = 1
- *FB* = 1
- *REC* = 1
- *FFIN* = 1
- *GHW* = 1
- *CA* = 1
- *GAR* = 1
- *REG* = 1
- *LOT* = 4000

Figures 1 to 4 indicate pointwise predictions of different models and 99% prediction intervals for the semiparametric model. In general, prediction intervals appropriate for the parametric model have been omitted to reduce the clutter. Figure 1 shows how the predicted $\log(P)$ changes as the number of bedrooms changes in our reference house, *ceteris paribus*. The solid line between the prediction intervals is the semiparametric prediction while the solid line below it represents the prediction of the parametric model. In Figure 1, the prediction interval is calculated using the asymptotic standard errors whereas in Figure 2, the prediction intervals are constructed with wild bootstrap technique. Clearly, the standard errors calculated with the wild bootstrap are wider than the asymptotic ones.

Figures 1 and 2 show that both models predict that $\log(P)$ is a concave function of the number of bedrooms. For the same type of houses as in Figures 1 and 2, Figure 3 shows that the
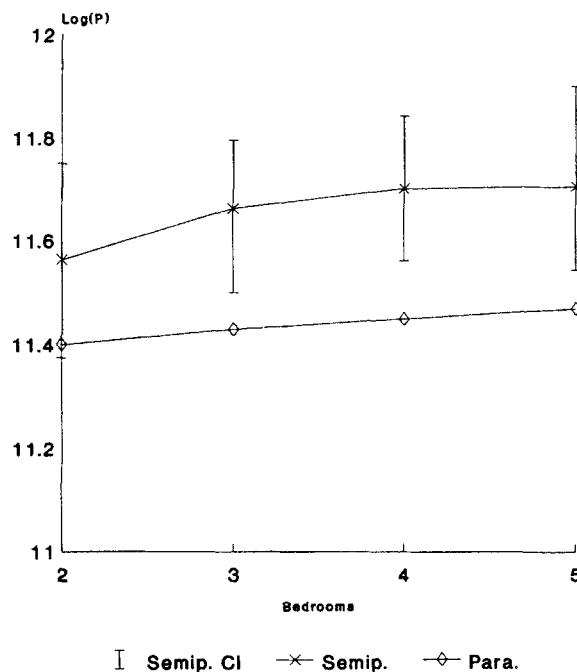
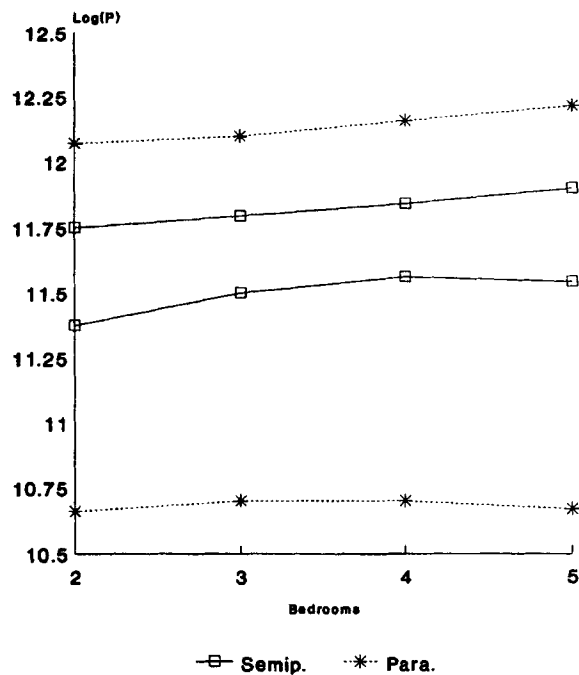Figure 1. Predicted log price
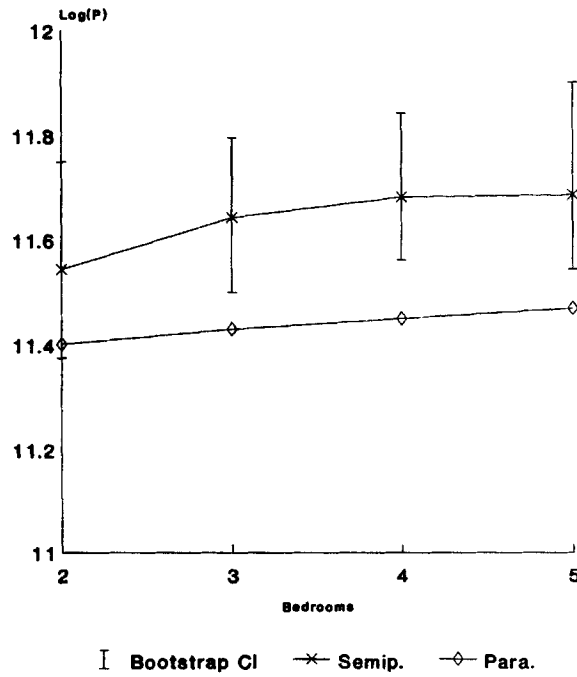
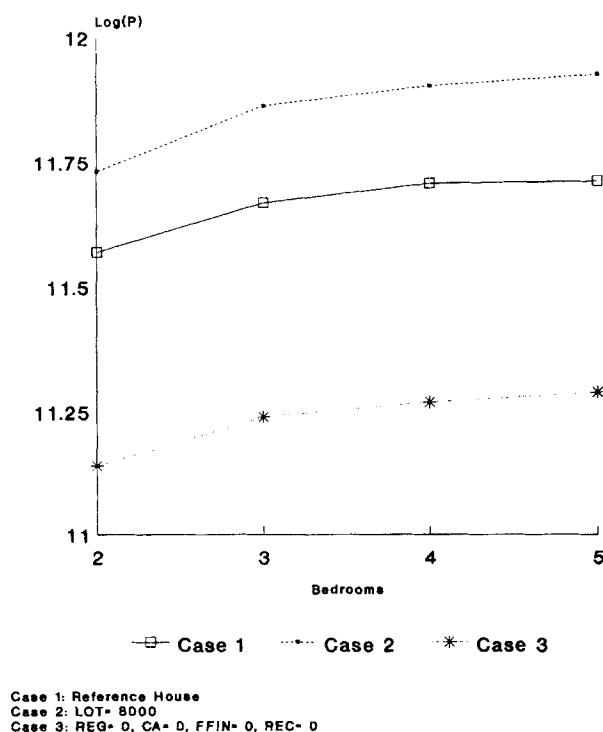Figure 2. Predicted log price

Figure 3. Confidence intervals

Figure 4. Alternative houses

prediction interval for the parametric model is larger than that of the semiparametric model. The reason for this is that the variance of the residuals from the semiparametric model is relatively smaller in comparison to the residual variance from the OLS model. Figure 4 shows the effect of two changes in the reference house on the level of the market value of the reference house and its rate of change with respect to the number of bedrooms. The middle line with boxes in Figure 4 repeats that shown in Figure 1; the upper line (Case 2) represents the market value of a house that is identical to the first house except with a 8000 square foot lot; the lower line (Case 3) shows the market value of a house identical to the first house but without central air conditioning, without a recreation room, without a fully finished basement and not in a preferred neighbourhood (i.e. the values of some of the regressors in the parametric portion of the model have been changed).

4.4.2. Out-of-sample predictions

As a final method of comparing the performances of the two estimators, we calculate the *out-of-sample* mean square prediction error (MSPE). The remaining observations are used for in-sample fit. Table VI reports the MSPE for ordinary least squares using the benchmark specification (denoted OLS) and the MSPE of the semiparametric model (denoted S). The bandwidth parameter for the out-of-sample horizon require undersmoothing to minimize the bias as explained in Section 4.3.1. Here, we adopt up to 50% undersmoothing in comparison to the in-sample bandwidth parameter adopted in Section 4.2. The out-of- sample predictions beyond 50% undersmoothing does not improve the out-of- sample predictions. We measure the out-of-sample forecasts using two different horizons: 10 and 20 observations. In both horizons, the semiparametric regression provides smaller MSPE than the parametric benchmark model.

With 10 houses in the forecast sample the MSPE of the parametric benchmark model is 1·35 times that of the semiparametric regression and this ratio is 1·53 with 20 houses in the forecast sample. Thus the superior performance of the semiparametric estimator should be due to better explanatory power rather than overfitting.

5. CONCLUSIONS

We have shown how semiparametric estimation techniques can be used to predict the (log) sale price of a house. In comparison tests, a parametric model was rejected when compared to a semiparametric model that allows arbitrary interaction between many of the regressors. While rejection of the parametric model on statistical criteria should not be unexpected, the fact that it was rejected after passing several common specification tests suggests that these specification tests on the parametric model may not be as powerful as desired. Our graphical analysis also illustrated one way that a semiparametric regression model might be applied and visually showed its superior performance. The figures revealed that the prediction intervals of a semiparametric model are tighter than those of a parametric model.

REFERENCES

Arguea, N. and C. Hsiao (1993), 'Econometric issues of estimating hedonic price functions with an application to the U.S. market for automobiles', *Journal of Econometrics*, **56**, 243–267.

Bierens, H. J. (1987), 'Kernel estimation of regression functions', in T. F. Bewley (ed.), *Advances in Econometrics: Fifth World Congress* Cambridge University Press, Cambridge.

Cassel, E. and R. Mendelson (1985), 'The choice of functional forms for hedonic price equations: comment', *Journal of Urban Economics*, **18**, 135–142.

Colwell, P. (1993), 'Semiparametric estimates of the marginal price of floorspace: comment', *Journal of Real Estate Finance and Economics*, **7**, 73–77.

Coulson, N. E. (1989), 'The empirical content of the linearity-as-repackaging hypothesis', *Journal of Urban Economics*, **25**, 295–309.

Cropper, M., L. Deck and K. McConnell (1988), 'On the choice of functional form for hedonic price functions', *Review of Economic and Statistics*, **70**, 668–675.

Davidson, R. and J. G. MacKinnon (1985), 'Testing linear and loglinear regressions against Box–Cox alternatives', *Canadian Journal of Economics*, **18**, 499–517.

Delgado, M. A. and J. Mora (1993), 'Nonparametric and semiparametric estimation with discrete regressors', Universidad Carlos III de Madrid, manuscript.

Delgado, M. A. and P. Robinson, (1992), 'Nonparametric and semiparametric methods for economic research', *Journal of Economic Surveys*, **6**, 201–249.

Delgado, M. A. and Stengos, T. (1994), 'Semiparametric specification testing of nonnested econometric models', *Review of Economic Studies*, **61**, 291–303.

Devroye, L. and C. S. Penrod (1984), 'The consistency of automatic kernel density estimates', *Annals of Statistics*, **12**, 1231–1249.

Fan, Y., Q. Li and T Stengos (1995), 'Root-N-consistent semiparametric regression with conditionally heteroskedastic disturbances', *Journal of Quantitative Economics*, forthcoming.

Follain, J. and E Jiminez (1985), 'Estimating the demand for housing characteristics: a survey and critique', *Regional Science and Urban Economics*, **15**, 77–107.

Gasser, T., H. G. Muller, W. Kohler, L. Molinari and A. Prader (1984), 'Nonparametric regression analysis of growth curves', *Annals of Statistics*, **12**, 210–229.

Goodman, A. C. (1978), 'Hedonic price, price indices and housing markets', *Journal of Urban Economics*, **5**, 471–484.

Griliches, Z. (1961), 'Hedonic price indexes for automobiles: an econometric analysis of quality change', in *Price Statistics of the Federal Government*, General Series No. 73, 137- 196, National Bureau of Economic Research, New York.

Halvorsen, R. and H. Pollakowski (1981), 'Choice of functional form for hedonic price equations', *Journal of Urban Economics*, **10**, 37–47.

Härdle, W. (1990), *Applied Nonparametric Regression*, Econometric Society Monographs No: 19, Cambridge University Press, Cambridge.

Härdle, W., Hall, P. and J. S. Marron (1988), 'How far are automatically chosen regression parameters from their optimum.' *Journal of the American Statistical Association*, **83**, 86–101.

Härdle, W. and J. S. Marron (1991), 'Bootstrap simultaneous error bars for nonparametric regression', *Annals of Statistics*, **19**, 778–796.

Härdle, W. and E. Mammen (1989), 'Comparing nonparametric versus parametric regression fits', Discussion paper A–177, University of Bonn.

Hausman, J. A. (1978), 'Specification tests in econometrics', *Econometrica*, **46**, 1251–1271.

Lee, T.-H., H. White and C. Granger (1993), 'Testing for neglected nonlinearity in time-series models', *Journal of Econometrics*, **56**, 269-290.

Li, K. C. (1984), 'Consistency for cross-validated nearest neighbour estimates in nonparametric regression', *Annals of Statistics*, **12**, 230–240.

Marron, J. S. (1985), 'An asymptotically efficient solution to the bandwidth problem of kernel density estimation', *Annals of Statistics*, **13**, 1019–1023.

Muller, H. G. (1984), 'Smooth optimum kernel estimates of densities, regression curves, H modes', *Annals of Statistics*, **12**, 766–774.

Pace, R. K. (1993), 'Nonparametric methods with applications to hedonic models' *Journal of Real Estate Finance and Economics*, **7**, 185–204.

Ramsey, J. B. (1969), 'Tests for specification errors in classical linear least squares regression analysis', *Journal of the Royal Statistical Society, Series B*, **31**, 350–371.

Rasmussen, D. and T. Zuehlke (1990), 'On the choice of functional form for hedonic price functions', *Applied Economics*, **22**, 431–438.

Rice, J. (1984), 'Bandwidth choice for nonparametric regression', *Annals of Statistics*, **12**, 1215–1230.

Robinson, P. M. (1988), 'Root-N-consistent semiparametric regression', *Econometrica*, **56**, 931–954.

Rosen, S. (1974), 'Hedonic prices and implicit markets: product differentiation in perfect competition', *Journal of Political Economy*, **82**, 53–76.

Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.

Stock, J. (1989), 'Nonparametric policy analysis', *Journal of the American Statistical Association*, **84**, 567–575.

Stock, J. (1991), 'Nonparametric policy analysis: an application to estimating hazardous waste cleanup benefits', in W. Barnett, J. Powell and G. Tauchen (eds), *Nonparametric and Semiparametric Methods in Econometrics and Statistics: Proceedings of the Fifth International Symposium in Economic Theory and Econometrics*, Cambridge University Press, New York.

Stone, C. J. (1984), 'An asymptotically optimal window selection rule for kernel density estimates', *Annals of Statistics*, **12**, 1285–1298.

Ullah, A. (1988), 'Non-parametric estimation of econometric functionals', *Canadian Journal of Economics*, **21**, 625–658.

Whang, Y.-J. and D. W. K. Andrews (1993), 'Tests of specification for parametric and semiparametric models', *Journal of Econometrics*, **57**, 277–318.

Wooldridge, J. M. (1992), 'A test for functional form against nonparametric alternatives', *Econometric Theory*, **8**, 452–475.