Regensburger
DISKUSSIONSBEITRÄGE
zur Wirtschaftswissenschaft

University of Regensburg Working Papers in Business,
Economics and Management Information Systems

On Nonparametric Estimation of a Hedonic Price Function

Harry Haupt[*], Joachim Schnurbus[**] and Rolf Tschernig [***]

December 5, 2008

Nr. 429

JEL Classification: C1, C14

Keywords: Nonparametric modeling, specification testing, forecast evaluation

[*] Harry Haupt holds the chair of Econometrics at the Department of Business Administration and Economics and is member of the Centre for Statistics at the University of Bielefeld, PO Box 100 131, 33501 Bielefeld, Germany. Phone: +49-521-106-4873, E-mail: hhaupt[at]wiwi.uni-bielefeld.de.

[**] Joachim Schnurbus is research and teaching assistant at the Department of Economics and Econometrics at the University of Regensburg, 93040 Regensburg, Germany.
Phone: +49-941-943-2739, E-mail: joachim.schnurbus[at]wiwi.uni-regensburg.de

[***] Rolf Tschernig holds the chair of Econometrics at the Department of Economics and Econometrics at the University of Regensburg, 93040 Regensburg, Germany.
Phone: +49-941-943-2737, E-mail: rolf.tschernig[at]wiwi.uni-regensburg.de

On Nonparametric Estimation of a Hedonic Price Function

Harry Haupt[*], Joachim Schnurbus[†], and Rolf Tschernig[‡]

Abstract

Recently, using data on Canadian housing, Parmeter, Henderson, and Kumbhakar (2007) found that a nonparametric approach for estimating hedonic prices is superior to formerly suggested parametric and semiparametric specifications. We carefully analyze this data set by applying a nonparametric specification test and simulation based forecast comparisons. For the case at issue our results suggest that a previously proposed parametric specification cannot be rejected.

Keywords: Nonparametric modeling, specification testing, forecast evaluation.

JEL classification: C1, C14

1 Introduction

Anglin and Gencay (1996), hereafter AG, proposed a semiparametric framework for the estimation of hedonic prices using data on the Canadian housing market. Based on their data and results, Parmeter, Henderson, and Kumbhakar (2007), hereafter PHK, apply a completely nonparametric kernel regression method suggested by Li and Racine (2004, 2007) and Racine and Li (2004), where both continuous and discrete regressors are smoothed within a multivariate regression framework. In terms of in-sample fit they find the nonparametric approach superior to the specifications considered by AG.

PHK recognize that two points make the kernel regression methods for mixed data a natural candidate for the data set used by AG. First, hedonic theory suggests that regressions concerning the heterogenous good housing should reflect the intrinsic nonlinearity

[*] Corresponding author. Bielefeld University, Centre for Statistics, Department of Business Administration and Economics, PO Box 100131, 33501 Bielefeld, Germany. Phone: +49 521 106 4873, fax: +49 521 106 6425. hhaupt@wiwi.uni-bielefeld.de.

[†] University of Regensburg, Germany. joachim.schnurbus@wiwi.uni-regensburg.de

[‡] University of Regensburg, Germany. rolf.tschernig@wiwi.uni-regensburg.de

in the relationship between house prices and housing characteristics, though nothing is known a priori about a specific functional form. Second, data on housing characteristics typically consist of a few continuous and rather many ordered and unordered categorical variables.

In this paper we carefully analyze the data used by AG and PHK with respect to general specification testing, variable selection, and forecast performance. For testing the null of correct specification of a parametric benchmark suggested by AG (see Table III. in AG), we apply the specification test of Hsiao, Li, and Racine (2007) that explicitly allows for mixed continuous and discrete regressors. We are not able to reject the null hypothesis of a correct specification at any reasonable significance level. In order to check whether the fit-improvement of the nonparametric specification is largely due to overfitting, we additionally conduct a simulation based forecast comparison between the parametric benchmark specification and the nonparametric approach of PHK. Again, our results are in favor of the parametric specification. Finally, we provide an explanation for these findings and complement them with some remarks on the relevance of the kernel choice for discrete regressors for estimation efficiency and elimination of irrelevant regressors.

2 Nonparametric Analysis

AG and PHK analyze a data set for the Canadian housing market, where the conditional mean of the natural log of price is explained by eleven characteristics of the houses. The data set consists of 546 observations, one continuous regressor, the lot size (in logs), and ten discrete regressors, six of them binary. AG apply different parametric specifications, as well as the semiparametric estimator of Robinson (1988). PHK reject the null hypothesis of correct semiparametric specification using the test of Delgado and González Manteiga (2001). Hence they apply the nonparametric generalized kernel estimation method presented in detail in Li and Racine (2007). We use the specification test of Hsiao et al. (2007), which is also based on the nonparametric generalized kernel estimation and is more efficient than traditional[1] nonparametric tests for a setting of mixed data regressors. We prefer the Hsiao et al. (2007) test to the test of Delgado and González Manteiga (2001) since the latter is based on the same bandwidth value for all nonparametric regressors and so is unlikely to work reliably when discrete ordered and continuous covariates are treated jointly like continuous variables.

[1] In traditional tests, the sample has to be split up for the category-combinations of the discrete regressors.

We begin our analysis by testing the null hypothesis of correct specification of the linear parametric benchmark model against the alternative of parametric misspecification. We use the same configuration of the test as in the original paper of Hsiao et al. (2007). They apply a local-constant estimator with a Gaussian kernel for the continuous regressor (in level-form) and kernels proposed by Li and Racine[2] as weighting functions for the discrete (binary and non-binary) regressors. The bandwidths are selected by least-squares cross-validation. Note that all of our calculations are done in R (version 2.7.1) and the nonparametric estimations and tests are conducted with the np-package (version 0.20-1) of Hayfield and Racine (2008). For the standardized version of the test statistic we get an asymptotic p-value of 0.51, implying that the parametric specification cannot be rejected. Hsiao et al. (2007) point out, that "the asymptotic normal approximation does not work well in finite-sample settings", suggesting the use of bootstrap methods. We therefore conduct the test using three different bootstrap methods (iid, wild, wild with Rademacher variables) and obtain quite similar p-values between 0.22 and 0.24. Hence the parametric specification cannot be rejected at any reasonable significance level.

As it may be possible that in the current setting even the bootstrap-versions of the Hsiao test may not have enough power, we also apply simulation techniques in order to compare the predictive ability of the parametric benchmark specification to that of the nonparametric specification of PHK. In each of the 399 replications, we first split the sample randomly into two subsamples: 90% of the observations are used for estimating each specification. Based on these estimates, the mean squared errors (MSE) for the predicted (log-)prices are calculated for the remaining 10% of the houses. Next we calculate the relative MSE as the fraction of the MSE for the parametric specification and that of the nonparametric specification. Figure 1 shows the simulated density for this relative mean squared error. Values above one indicate the nonparametric specification to predict better, values below one indicate better predictions of the parametric specification. Evidently, in most of the 399 replications, the relative mean squared error takes a value less than 1. In fact, in about 92% of the replications the predictions of the parametric specification exhibit a lower mean squared error. For the mean absolute error and mean absolute percentage error, the shares are of the same magnitude (meaning that the parametric specification is preferable for prediction in at least nine out of ten cases)[3].

[2] These kernels are optionally available in np. Details can be found in Racine and Li (2004).

[3] By using alternative sample-split proportions, such as 80-20 or 70-30, the results are even more in favor of the parametric specification.

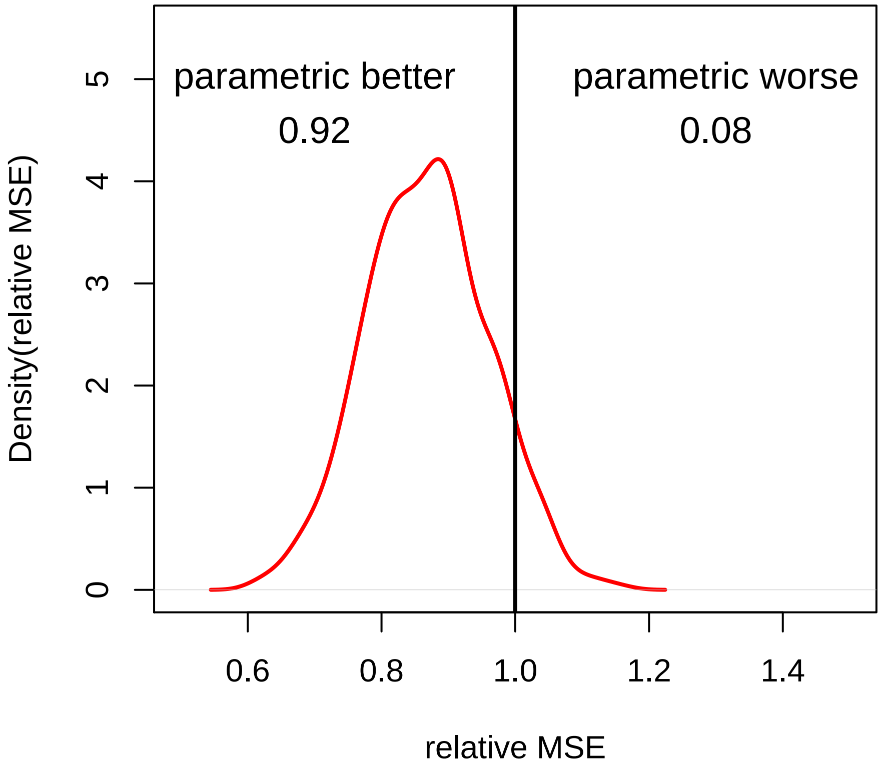

Figure 1: Density of relative mean squared errors of prediction.

Why do our prediction results differ so much from the findings of PHK? Three issues have to be considered here: first, the relevance of nonparametrically estimating the impact of the continuous regressor, second, the optimal degree of smoothing of discrete variables and, finally, the possibility of smoothing out irrelevant variables. To begin with, reconsider the estimation setup of PHK. They use a local linear estimator with a second order Gaussian kernel for the continuous regressor combined with the weighting functions of Wang and van Ryzin (1981) and Aitchison and Aitken (1976) for the discrete regressors. The bandwidths are selected with a modified version of the Akaike Information Criterion suggested by Hurvich, Simonoff, and Tsai (1998). We were able to reproduce their estimation results and obtain exactly the same gradient values as in Table I. of PHK. Since the selected smoothing parameters are crucial in nonparametric estimation, we display the estimated bandwidths in Table 1 because they are not presented in PHK. In addition, the type of each regressor, the applied kernel as well as the

Regressor	Regressortype	Weighting Function	Est. Bandwidth	Max. Bandwidth
`ln(lot)`	continuous	Gaussian (2. order)	888595.2	∞
`ca`	binary	Aitchison and Aitken	0.1409	0.5
`drv`	binary	Aitchison and Aitken	0.1206	0.5
`ffin`	binary	Aitchison and Aitken	0.1296	0.5
`ghw`	binary	Aitchison and Aitken	0.0302	0.5
`rec`	binary	Aitchison and Aitken	≈ 0.5	0.5
`reg`	binary	Aitchison and Aitken	0.2253	0.5
`bdms`	ordered	Wang and van Ryzin	0.6153	< 1
`fb`	ordered	Wang and van Ryzin	0.2667	< 1
`gar`	ordered	Wang and van Ryzin	≈ 1	< 1
`sty`	ordered	Wang and van Ryzin	0.5063	< 1

Table 1: Estimated bandwidths of local linear kernel regression.

maximum of the admissible bandwidth range are displayed.

Now inspect the first row in Table 1. It contains the estimated bandwidth for the single continuous regressor, `ln(lot)`, the logarithm of the lot size. The extremely large bandwidth relative to the observed values of the regressor strongly suggests that no smoothing is needed for this variable. Taking logs is already sufficient for capturing the nonlinearity in lot size. It is therefore not very surprising to find the parametric model to be superior to the nonparametric alternative since it is well known that within the local linear estimation framework a very large bandwidth suggests that the impact of the corresponding regressor is best represented in a linear way. Thus, the direct influence of `ln(lot)` does not require nonparametric modeling.

Before inspecting the discrete regressors with respect to their entries in the remaining rows of Table 1, we remark that, following Li and Racine (2007, p. 145), a desirable property of weighting functions for discrete variables is that irrelevant regressors are smoothed out. This requires that in case of an irrelevant discrete regressor there exists a smoothing parameter that implies identical weights across all categories of the regressor considered.

For the case of unordered binary regressors this requirement is fulfilled by the choice of the Aitchison and Aitken (1976) weighting function: if the bandwidth lies at its upper bound 0.5, the corresponding regressor is irrelevant. Inspection of Table 1 reveals that the variable `rec` (recreational room exists/does not exist) is irrelevant for the estimated house price since the estimated bandwidth is in the close vicinity of 0.5.

The remaining discrete regressors are treated in the model as ordered with the weighting function of Wang and van Ryzin (1981). In the following we show that the choice of the Wang and van Ryzin (1981) kernel comes with two drawbacks: the kernel weights are limited to a certain range such that the degree of smoothing is limited and irrelevant regressors cannot be smoothed out. In any case, such restrictions may lead to inefficiencies in finite samples due to a higher mean squared error. It will turn out that for the case at issue the limited smoothing capabilities matter.

These limitations become apparent by inspecting the Wang and van Ryzin (1981) weighting function

$$l(\delta,\lambda) = \begin{cases} 1-\lambda & \text{for} \quad \delta = 0, \\ & \\ \frac{1}{2}(1-\lambda)\lambda^{|\delta|} & \text{for} \quad \delta \in \mathbb{Z}, \end{cases}$$

where λ is the selected bandwidth for the discrete regressor and δ represents the distance between the category of interest and any category used for estimation. Note that a bandwidth λ of exactly 1 implies zero kernel weights for any δ and is thus no admissible bandwidth. In contrast, a zero bandwidth implies that only those observations are used that are in the category of interest, implying no smoothing at all. Thus, for smoothing to occur, the bandwidth has to take values in $0 < \lambda < 1$. Now consider the ratio of the kernel weights for the category of interest, $\delta = 0$, and any other category $\delta \neq 0$. As an example, you may be interested in the price for a house that has two garages (`gar`$= 2$), so all observations with two garages exhibit $\delta = 0$ and are therefore weighted by $1-\lambda$, while observations with one or three garages have $|\delta| = 1$ and are assigned weights $\frac{1}{2}(1-\lambda)\lambda$, respectively.

In general, the ratio of weights is given by

$$\frac{1-\lambda}{\frac{1}{2}(1-\lambda)\lambda^{|\delta|}} = \frac{2}{\lambda^{|\delta|}}, \quad 0 < \lambda < 1, \tag{1}$$

where the equal sign only holds within the indicated range of bandwidths. Clearly, this ratio decreases monotonically with an increasing bandwidth and approaches its infimum 2 as the bandwidth λ tends to 1.

Thus, this ratio cannot take any value smaller than 2. This clearly limits the amount of smoothing since almost equal weights are not possible. In particular, equal weights corresponding to a ratio of 1 are ruled out and therefore smoothing out of irrelevant variables cannot occur. Inspecting the remaining four rows of Table 1 shows that these limitations become relevant for the ordered regressor `gar`. The corresponding bandwidth is estimated to be extremely close to 1 with a precise value of 0.999999999995943. This indicates that more smoothing might be needed for a further reduction of the MSE.

Such inefficiencies can be avoided by applying kernels that explicitly allow for smoothing out irrelevant ordered regressors such as the kernel of Li and Racine. We therefore use the latter in our simulations concerning the predictive ability although the results turn out to be robust with respect to the kernel choice. Furthermore, excluding irrelevant variables such as `rec` does not change the results. For the ordered regressor `gar` the corresponding ratio of weights is found to be between 1 and 2, indicating the relevance of this variable on the one hand and the need for strong smoothing on the other hand.[4]

3 Conclusion

Multivariate nonparametric estimation methods are an exciting edge of research in applied econometrics. Especially through the collected works of Li, Racine, Hall and others, who achieved that nonparametric specifications can be used in multiple regression estimation and testing settings with discrete regressors without noticeable efficiency losses, these methods are of increasing appeal.

We re-estimated the results of PHK and AG and extended their analysis based on a careful reconsideration of their specification. It turns out that the choice of the specification test matters. An adequate test should explicitly allow for the mixed data structure. In addition, it was pointed out that kernels for discrete data differ with respect to their smoothing capabilities and efficiency properties. For the housing data set under consideration, all our results indicate that a complete nonparametric specification does not seem to be necessary.

We conclude that advanced nonparametric methods are an indispensable tool for modeling, either for considerably strengthening the support of simple parametric specifications or for estimation purposes when a robust parametric specification cannot be determined.

References

[1] Aitchison J. and C.G.G. Aitken. 1976. Multivariate Binary Discrimination by the Kernel Method. *Biometrika* **63**: 413-420. DOI: 10.1093/biomet/63.3.413

[2] Anglin P.M. and R. Gencay. 1996. Semiparametric Estimation of a Hedonic Price Function. *Journal of Applied Econometrics* **11**: 633-648.

[4] For the Li and Racine kernel the ratio of weights is $\lambda^{-|\delta|}$. Given the bandwidth estimate for `gar` of 0.6875 and choosing $|\delta| = 1$, one obtains a ratio of 1.45.

[3] Delgado M.A. and W. González Manteiga. 2001. Significance Testing in Nonparametric Regression Based on the Bootstrap. *The Annals of Statistics* **29**: 1469-1507. DOI: 10.1214/aos/1013203462

[4] Hayfield T. and J.S. Racine. 2008. np: Nonparametric kernel smoothing methods for mixed datatypes. R package version 0.20-1.

[5] Hsiao C., Q. Li and J.S. Racine. 2007. A consistent model specification test with mixed discrete and continuous data. *Journal of Econometrics* **140**: 802-826. DOI: 10.1016/j.jeconom.2006.07.015

[6] Hurvich C.M., J.S. Simonoff and C.L. Tsai. 1998. Smoothing Parameter Selection in Nonparametric Regression using an Improved Akaike Information Criterion. *Journal of the Royal Statistical Society* Series B **60**: 271-293.

[7] Li, Q. and J.S. Racine. 2004. Cross-validated Local Linear Nonparametric Regression. *Statistica Sinica* **14**: 485-512.

[8] Li Q. and J.S. Racine. 2007. *Nonparametric Econometrics: Theory and Practice.* Princeton University Press.

[9] Parmeter C.F., D.J. Henderson and S.C. Kumbhakar. 2007. Nonparametric Estimation of a Hedonic Price Function. *Journal of Applied Econometrics* **22**: 695-699. DOI: 10.1002/jae.929

[10] Racine J. and Q. Li. 2004. Nonparametric Estimation of Regression Functions with both Categorical and Continuous Data. *Journal of Econometrics* **119**: 99-130. DOI: 10.1016/S0304-4076(03)00157-X

[11] Robinson P.M. 1988. Root-N-consistent semiparametric regression. *Econometrica* **56**: 931-954.

[12] Rosen S. 1974. Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. *Journal of Political Economy* **82**: 34-55. DOI: 10.1086/260169

[13] Wang, M.C. and J. van Ryzin. 1981. A class of smooth estimators for discrete distributions. *Biometrika* **68**: 301-309. DOI: 10.1093/biomet/68.1.301