Working Paper

Semiparametric Estimation of Time, Age and Cohort Effects in An Hedonic Model of House Prices

Koji KARATO, Oleksandr MOVSHUK and Chihiro SHIMIZU

Working Paper No. 256
November 17, 2010

FACULTY OF ECONOMICS
UNIVERSITY OF TOYAMA

3190 Gofuku, Toyama
930-8555 JAPAN

Semiparametric Estimation of Time, Age and Cohort Effects in An Hedonic Model of House Prices

Koji Karato*
Faculty of Economics, University of Toyama
Oleksandr Movshuk
Faculty of Economics, University of Toyama
Chihiro Shimizu
International School of Economics and Business Administration, Reitaku University

November 17, 2010

Abstract

In hedonic models of housing prices, it is impossible to estimate simultaneously the impact of selling time, age and cohort effects without introducing some restrictions on estimated effects. In this paper we address the simultaneity problem by estimating time, age and cohort effects with a semiparametric generalized additive model that allows for a nonlinearity in age and cohort effect. The model is applied to house prices in 23 Tokyo special wards between 1990 and 2008. Estimates of age effect showed lower prices for older houses, and we failed to reject the linearity restriction in this effect. On the other hand, there was a significant nonlinearity in estimates of cohort effect, which justified the application of the nonparametric regression model. We also examined the joint impact of cohort and age effect on housing prices, and found that the shape of age effect was different across cohorts of housing. Estimates of year effect indicate a declining trend in prices that was more pronounced compared with conventional hedonic models that do not include simultaneously age, time, and cohort effects on housing prices.

JEL Classification Code: C14, R21, R31
Keywords: Hedonic price index; Age effect; Cohort effect; Semiparametric model; Generalized additive model;

*Correspondence. E-mail: kkarato@eco.u-toyama.ac.jp, 3190 Gofuku, Toyama 930-8555, Japan

1 Introduction

Hedonic models of house prices commonly include three factors that are related to time: the time of sale, the age of house, and cohort or year of construction. There is an perfect collinearity between these time, age and cohort terms, because the year when the house is sold equals to the house age plus the year when the house was constructed. This identification problem results in the multicollinearlity among dependent variables when the model is estimated in regression analysis.

In this paper we address the simultaneity problem by estimating time, age and cohort effects with a semiparametric generalized additive model that allows for a nonlinearity in age and cohort effect. To break the collinearity, the typical solution has been to omit either the age or the cohort effects. However, there are good reasons for not eliminating any of the three variables, since their impact on price is likely to arise from quite different sources. The time effect measures the impact of market conditions on the general trend of housing prices, so this effect is essential for creating quality-controlled housing price indexes out of hedonic models. The age effect is measuring the physical depreciation, or the added cachet that accrues as the housing unit gets older, and in consequence the age effect is nearly always included in hedonic regressions. Finally, the cohort effect measures the impact of the year when the housing unit was constructed, and the effects could account for unmeasured style characteristics that are particularly prized in a particular area (Coulson and McMillen (2008)).

A survey of 78 hedonic studies referenced by Sirmans *et al.* (2006) found that while almost all of them included either age or vintage in the hedonic specification, and those that had multiple dates of sale included some form of time variable, but no study included both age and vintage in the specification, even in some nonlinear or dummy variable form.

An alternative solution is provided by the method of non-linearizing these variables in the functional form of the 'parametric' regression model. Unfortunately, economic theory provides little guidance concerning the functional form of dependence of house price on quality and researchers have used forms which are somewhat flexible in order to let the data 'speak'. Cropper *et al.* (1988) used a Monte Carlo study to investigate the performance of different functional forms. While Halvorsen and Pollakowski (1981) used a real world data set, Cropper *et al.* (1988) carefully specified a single type of utility function for a group of consumers and produced a market price gradient by allowing the taste parameter of this function to be randomly distributed. After considering six different models (such as translog or Box−Cox), the study found that a linear Box−Cox regression produced the most accurate estimates of marginal attribute prices. However, non-linearizing may not perfectly solve the problem and leave the high correlation among the year of the sale, house age and the year of construction. Furthermore, it is not clear which functional form should be used to specify the cohort effect.

As an alternative method, Coulson and McMillen (2008) disentangle the year, age and cohort effects on housing price using the second difference approach of McKenzie (2006). The method of McKenzie (2006) allows for simultaneous unrestricted nonparametric estimation of the year of sale, the house age, and cohort effects. As a nonparametric estimator, it removes the problem of imposing structure on a model when the structure is clearly incorrect. The only restriction that this method imposes is that that some two neighboring age effects are equal to some known constant, which in practice is

set to zero. One serious limitation of this method is that many alternative restrictions on neighboring parameters may be considered, corresponding the number of age effects in the estimated model. Fu (2008) reports evidence that the estimates can be changed dramatically with different combinations of neighboring effects that are assumed to be equal.

The remainder of this paper is organized as follows. Next section presents the basic hedonic price specification of generalized additive model that includes the time of housing sale, the housing age and the year when the housing was build. The section also describes the details of the identification problem and our approach how to solve it. Section 3 describes the estimation method of generalized additive model which is based on the algorithm of Wood (2004). Section 4 reports descriptive statistics of our data on the transactions in single-family condominiums in the special 23 wards of Tokyo (Japan), as well as variable definitions in estimated models. Section 5 report our results of estimating hedonic price models, and compare them with estimates derived from conventional models that do not include all three time-related effects on housing prices. Some concluding remarks are given in section 6.

2 Model

2.1 Time, Age and Cohort Effects

We denote the year of sale by t. If age index of a house is j at transaction year t, then the log price of i-th house is expressed by $P_{i(t,j,t-j)}$, where $t-j$ is cohort year. The k-th characteristic variable of house $X^k_{i(t,j,t-j)}$ is similarly defined. To create a pseudo-panel dataset, consider the sample average of housing units for a given year of sale t and age j. The average of log price is

$$P_{t,j,t-j} \quad = \quad \frac{1}{n_{(t,j,t-j)}} \sum_{i=1}^{n_{(t,j,t-j)}} P_{i(t,j,t-j)} \quad (t = 1, 2, \cdots, T,\ j = 0, 1, \cdots, J(t)).$$

The number of observations in transaction year t is $J(t)+1$ if there is no missing age in year. Hence the number of the sample average is $\sum_{t=1}^{T}(J(t)+1) = N$.

In our hedonic price model of house prices, the log price of a housing unit depends on three major effects: the year when the house was sold, the age of house, and the year when the house was constructed. Consider the log price $P_{t,j,\ell}$ of a house that was sold in year t (the year of construction as $\ell = t-j$). Values of ℓ will be used to differentiate between each cohort of housing. We allow a flexible shape of year, age and cohort effects, and estimated them by three sets of dummy variables for year of the sale, and for age and cohorts of housing. For instance, the number of different sale years is T, and age varies between 0 and J years. In our data $T = 19$ and $J = 50$ (the specifics of the data are discussed in section 4). Let $D^Y_t,\ t = 1, 2, \cdots, T$ be T dummy variables for the year of the sale. Similarly, we estimate age and cohort effects with the dummy variables D^A_j and D^C_ℓ, respectively.

Combining three effects on log of housing prices, we get the following model (omitting for the moment the other housing characteristics on the right-hand side):

$$P_{t,j,\ell} = b_0 + \sum_{t=1}^{T} \alpha_t D^Y_t + \sum_{j=0}^{J} \beta_j D^A_j + \sum_{\ell=1}^{L} \gamma_\ell D^C_\ell + u_{t,j,\ell} \tag{1}$$

where α_t, β_j and γ_ℓ are year, age and cohort effect, $u_{t,j,\ell}$ is the error term with zero mean and variance σ^2. In each dummy variable D_t^Y, D_j^A, and D_ℓ^C, the sum across rows is always one, which results in perfect collinearity (i.e., 'dummy variable trap') between the intercept term b_0 and each of D_t^Y, D_j^A and D_ℓ^C. Typically, the problem is solved by dropping a single dummy variable from each of D_t^Y, D_j^A and D_ℓ^C. For instance, the first dummy variable can be dropped, which restricts the corresponding regression parameter to zero, making it a convenient benchmark against which all subsequent estimates can be compared. We used this approach in our study.

2.2 Identification problem and its solutions

Identification problem occurs in equation (1) even after solving the dummy variable trap, because there is an exact linear relation among the year of the sale t, age j and the year of cohort ℓ (namely, $\ell = t - j$). Because of the perfect collinearity, a given pattern of house prices can be explained by various combinations of year, age and cohort effects. Suppose that housing prices are increasing by 2 percent a year. Due to the identification problem, it is not possible to single out a unique explanation of this general trend. One possible interpretation is a change in the year effect by 2 percent per year, with no changes in age and cohort effects. Another possible interpretation of the same pattern is by increasing age and cohort effects, when housing price increase by 2 percent per year for older houses, and the same 2 percent increase among houses in more recent cohorts (as denoted by higher values of $t - j$), while the effect remains fixed. Similar examples of alternative interpretations of age, cohort and year effects are well-known in the literature on age-cohort-year models (see, for example, Deaton and Paxson (1994) and Paxson (1996)).

The identification problem can be solved by imposing restrictions on estimated regression coefficients in (1). A recent application of year-age-cohort model to housing prices by Coulson and McMillen (2008) follows McKenzie (2006)'s second difference approach, where the identification problem is avoided by assuming that some neighboring parameters in the age effect $\beta_j - \beta_{j-1}$ are equal to some known constant λ (which Coulson and McMillen (2008) actually set to zero).

The major problem is this approach is possible sensitivity of final estimates to the choice of parameters that are assumed to be equal. While Coulson and McMillen (2008) choose to restrict the first and second parameters of age effect (i.e., $\beta_2 - \beta_1 = 0$), in fact there is a large number of alternative restrictions. For instance, with 51 age effects in our model, there are 50 possible restrictions on pairs of neighboring coefficients. Similarly, we have 19 years of data in our dataset, so the number of possible restrictions on neighboring year effects is 18. Overall, the number of possible restrictions in either age or year effects grows to $50+18 = 68$. The number increases further after we consider possible restrictions on neighboring parameters of cohort effects. With $(51 + 19) - 1 = 69$ cohort effects, the total number of restrictions in either of these three effects is $(51 - 1) + (19 - 1) + (69 - 1) = 136$. Clearly, the consideration of all these possible restrictions on neighboring parameters is not an easy task.

The solution of Coulson and McMillen (2008) would be satisfactory if final estimates of age, cohort and year effects are little changed with different combinations of restricted parameters, but Fu (2008) reports evidence that the estimates can be changed dramatically with different combinations of neighboring parameters that are assumed to be

equal.

In this paper we propose a different solution of the identification problem, in which a mild structure is imposed on the age effect, while no restrictions are applied to cohort and year effects. Specifically, we assume that age effect can be represented by a smooth function of housing age, and that the age affect may have a nonlinear impact on housing prices. We will refer to this model as the smoothing age model of housing prices.

Our model is related to the smoothing cohort model of Fu (2008), where it was also developed to solve the identification problem among age, cohort, and year effects. The only difference between our approach and the approach of Fu (2008) is that we apply the nonparametric term age effect, while in Fu (2008) it is applied to cohort effect.

After introducing a smooth nonlinear age effect, the age dummy D_j^A in (1) is replaced with a single variable A_j. Similarly, the cohort dummy D_{t-j}^C is replaced with a single cohort variable C_{t-j}. The impact of age on housing price is estimated by a smooth, but possibly nonlinear, function $s(A_j)$, resulting in the following regression model:

$$P_{t,j,t-j} = \alpha_t + s(A_j) + \gamma \cdot C_{t-j} + \boldsymbol{X}'_{t,j,t-j}\boldsymbol{b} + u_{t,j,t-j}, \tag{2}$$

where α_t is the year effect in year t of the sale, $s(\cdot)$ is smoothing function, A_j is age term, γ is the cohort effect for cohort year trend term, and

$$\boldsymbol{X}_{t,j,t-j} = \left(1, X^1_{t,j,t-j}, \cdots, X^k_{t,j,t-j}, \cdots, X^K_{t,j,t-j},\right)'$$

is vector of average characteristic variables, which the k-th characteristic variable is $X^k_{t,j,t-j} = \frac{1}{n_{(t,j,t-j)}} \sum_{i=1}^{n_{(t,j,t-j)}} X^k_{i(t,j,t-j)}$ and $u_{t,j,t-j}$ is error term.

The initial specification of the smoothing age model in (2) will be referred as Model 1. Subsequently, this initial model will be modified by several alternative specifications. For example, we will introduce in Model 2 a nonlinear cohort effect $s(C_{t-j})$ (similarly to the smoothing cohort model of Fu (2008)), while in Model 3 we will consider the possibility that there is a joint effect between cohort and age effects, both of which are estimated as nonlinear functions. These alternative specifications will be explained in more details in Section 3.2.

3 Estimation

3.1 Estimation of the basic model

The smoothing age model in equation (2) is essentially a semiparametric regression model that has two parts: a nonparametric term $s(A_j)$ and a parametric part. Originally, Fu (2008) suggested to fit the smoothing model as a generalized additive model (GAM), using the backfitting algorithm of Hastie and Tibshirani (1990). However, the stability of the backfitting algorithm was questioned in recent years, particularly in datasets with high collinearity among explanatory variables (Schimek, forthcoming). Another limitation of the traditional GAM estimator is the need to choose a smoothing parameter prior to estimation. Most often, the smoothing parameter in the GAM estimator is given by the number of degrees of freedom v that are used to approximate the nonparametric term. For example, when $v = 1$, then the conventional linear regression model becomes a special case of the GAM. On the other hand, semiparametric regression models have $v > 1$, with larger values of v indicating relatively more nonlinear

effects. While Fu (2008) claimed that setting v to 10 degrees of freedom 'yields good results' (p.341), there is no guarantee that the value of smoothing parameter will be an accurate in describing the actual shape of age effects. It is more preferable to determine the degree of smoothing of $s(A_j)$ in an endogenous way that depends on examined data.

In this paper, we use the automatic selection of v, which is possible with the Modified Generalized Cross Validation (MGCV) algorithm of Wood (2004). Compared with the backfitting algorithm, the MGCV approach has superior numerical stability, especially when explanatory variables are correlated (Schimek, forthcoming). In addition, the MGCV algorithm selects an appropriate degree of smoothness using a large variety of selection methods, including the generalized cross validation (GCV) criterion of Craven and Wahba (1979), or restricted maximum likelihood (REML) methods that represent the nonparametric part as random effects (Ruppert *et al.*, 2003). In this paper, we estimate the smoothing age model (2) by the MGCV algorithm, with the smoothness of age effect determined by minimizing the GCV criterion.

In the Appendix to this paper, we provide technical details about estimation of optimal degree of smoothness in MGCV algorithm of Wood (2004). In summary, while Fu (2008) suggested to estimate the smoothing cohort model with v fixed at 10, the MGCV algorithm searches for an optimal values of smoothing parameter λ. This algorithm in practice can select any degree of smoothing, including the special case of linear age effect, when for $v = 1$ the GCV criterion is minimized.

We will refer to the semiparametric model (2) as Model 1. The model was estimated with `mgcv` library version 1.5-5 Wood (2006) which is available in R software (R Development Core Team, 2009).

3.2 Alternative Models

In addition to Model 1, we considered several alternative specifications that include a nonlinearity in cohort effect, and also a joint impact between age and cohort effects.

In Model 2, we specified cohort effect as a nonlinear smooth function, in the same way as we specified the age effect in Model 1:

$$P_{t,j,t-j} = \alpha_t + s(A_j) + s(C_{t-j}) + \boldsymbol{X}'_{t,j,t-j}\boldsymbol{b} + u_{t,j,t-j} \tag{3}$$

In equation (3), the two smooth nonparametric terms $s(A_j)$ and $s(C_{t-j})$ have additive affect on house prices, but there is no interaction between these effects. However, age effects may not have the same pattern with different vintages of housing. For example, some old vintages of houses may have a kind of 'retro' value, which will increase their price compared with houses of the same age, but build in more recent years. To account for this joint impact, we specified Model 3 with an interaction term $s(A_j, C_{t-j})$ between age and cohort effects:

$$P_{t,j,t-j} = \alpha_t + s(A_j) + s(C_{t-j}) + s(A_j, C_{t-j}) + \boldsymbol{X}'_{t,j,t-j}\boldsymbol{b} + u_{t,j,t-j} \tag{4}$$

Finally, we considered whether our initial specification of nonlinear age effect on housing prices can be simplified further by assuming a linear effect of age on housing prices. With this modification, we obtained the following Model 4:

$$P_{t,j,t-j} = \alpha_t + \beta \cdot A_j + s(C_{t-j}) + s(A_j, C_{t-j}) + \boldsymbol{X}'_{t,j,t-j}\boldsymbol{b} + u_{t,j,t-j} \tag{5}$$

Note that after specifying the linear age effect in Model 4, there is no the identification problem between age, cohort, and year effects, since the cohort effect in Model 4 is no longer linear, but specified as a nonparametric term $s(C_{t-j})$, and this prevents the exact linear dependence between these three effects.

4 Description of the Data and Variable Definition

Our data of single-family condominiums is drawn from a weekly magazine, *Shukan Jutaku Joho* (Residential Information Weekly) published by Recruit Co., Ltd., one of the largest vendors of residential lettings information in Japan. Table 1 shows the descriptive statistics of single-family condominiums data. This dataset covers the special 23 wards of Tokyo in Japan for the sales periods from 1990 to 2008, and the sample size includes 39,218 housing transactions. When the Japan's bubble economy bursted in 1989, not only stock prices but also house prices fell sharply. While for the whole sample the average price is about 37 million yen, the average was as high as 84 million yen in 1990, but then dropped to 27 million yen in 2001. The full sample of 39,218 condominiums include houses build from 1954 to 2008, except for 1955, 1956 and 1961, for which the sample contains no data. The age varied between zero (indicating houses that were during the current year), and 50 years.

Table 2 reports the distribution of age and cohorts (construction year) at the time of sale. The frequency of construction year is the highest in the 1980s, while the housing age is the highest between 10 and 19 years. There is a negative correlation between age and construction year, with the correlation coefficient -0.779 for full sample.

To implement a pseudo-panel approach to disentangle time, age, and cohort effects, we construct a matrix of mean price and the following characteristic variables

- X^1: Log of sq. meters
- X^2: Log of time distance from Central Business District
- X^3: Log of minutes on foot to a nearly station
- X^4: Log of number of houses in condominium

by housing age $(0, 1, 2, \cdots, 50)$ and the year of sale $(1990, 1991, \cdots, 2008)$. We expected positive effects on housing prices from X^1 and X^4, while negative effects were expected from X^2 andX^3.

Our pseudo-panel dataset contained $51 \times 19 = 969$ cells with 259 missing elements, so the final sample size was 710. With this pseudo-panel data, we constructed two price indices that show price changes relatively to the base year. In the first index I_1, the housing age was fixed at 8 years, while in the second index I_2, housing cohort was fixed at 1982. These indices were defined as follows:

$$\begin{aligned} I_1 &= \exp\left(P_{t,\,8,\,t-8} - P_{1990,\,8,\,1990-8}\right) \\ I_2 &= \exp\left(P_{t,\,t-1982,\,1982} - P_{1990,\,1990-1982,\,1982}\right) \\ &\quad (t = 1990, 1991, \cdots, 2008) \end{aligned}$$

where $P_{t,j,t-j}$ is log price in year of the sale t, age j and construction year $t-j$.

Estimates of these housing indices are shown in Figure 1. Both price indices are falling sharply from 1990 to 1999. Subsequently, the first index with fixed age rebounded sharply, while the second index with fixed cohort remained steady. Though age constant price index I_1 control the age effects, the index does not differentiate between cohort and time effects, simply because cohort years are changing simultaneously with current years. Similarly, the price index I_2 with fixed cohort effect involves changing age effects. However, the second index could not differentiate between age and year effects.

These results can be explained as follows. Suppose that the log of housing prices depends on time effect α_t, age effect β, and cohort effect γ: $P_{t,j,t-j} = \alpha_t + \beta A_j + \gamma C_{t-j}$, where A_j and C_{t-j} are trend terms. Consider the first case when age is fixed at j. Then the log difference between t and s is $P_{t,j,t-j} - P_{s,j,s-j} = (\alpha_t - \alpha_s) + \gamma(C_{t-j} - C_{s-j})$. Since this price change has not only time effect but also cohort effect, the price index I_1 has bias. Next, consider the second case when the cohort year is fixed at $t - j$. Log differencing the cohort-constant price results in $P_{t,j,t-j} - P_{s,j-(t-s),t-j} = (\alpha_t - \alpha_s) + \beta(A_j - A_{j-(t-s)})$. Note that if the cohort year is fixed, age effect remains in the price change equation (such as repeat sales method of Bailey *et al.* (1963)), once again producing bias in price index.

In next section, we show the results of estimating the hedonic price model with the semiparametric estimator. Time, age and cohort terms are perfectly collinear because each term is measured annually. In order to avoid the collinearity problem, we use smoothing terms.

5 Results

5.1 Estimation results of a parametric part in semiparametric hedonic price models

As a benchmark for comparing our semiparametric models, we estimated a standard hedonic linear regression model, in which the identification problem is solved by omitting the cohort effect:

$$P_{t,j,t-j} = \alpha_t + \beta A_j + \boldsymbol{X}'_{t,j,t-j}\boldsymbol{b} + u_{t,j,t-j} \tag{6}$$

Table 3 reports results of estimating the standard linear hedonic model, as well as semiparametric models, discussed in sections 3 and 3.2. The coefficients of variables from years from 1991 to 2008 show the time effects, with the base year at 1990. The smooth term of age appears in Models 1, 2 and 3, and while in Model 4 it is represented by a linear term. Similarly, the smooth term of cohort appears in Models 2, 3 and 4, and the effect has a linear specification in Model 1. Finally, the joint smooth term $s(A_j, C_{t-j})$ for age and cohort effects is used in Models 3 and 4.

Results of estimating the standard linear hedonic model without cohort effect are shown in first column (eq.(6) titled 'Linear'). Time effect, age effect and other attributes effects are statistically significant and have expected signs. Based on the parameter estimate for age, the house depreciation rate turned out to be $100 \times \{\exp(-0.017) - 1\} = 1.69$ % per year. However, note that these coefficients may be biased since a cohort effect is omitted from equation (6).

Second column (Model 1) provides estimation results of equation (2), in which the semiparametric hedonic price model has a smooth age term. Cohort effect for is sig-

nificantly positive, indicating higher prices for more recently build houses. As for the time effect, regression estimates are almost the same compared with the result from the linear regression model in the first column up to 2000, and then estimates from Model 1 show a more significant decline in prices.

Third column (Model 2) reports estimation results of equation (3), in which semi-parametric hedonic price model has smooth nonlinear terms for both age and cohort terms. This time, estimates of time effects turned out slightly higher in Model 3 compared with the linear hedonic model and the nonparametric Model 1.

Fourth column (Model 3) reports estimation results of the parametric part of Model 3, specified by equation 4. Overall, estimated time effects are smaller compared with estimates from the hedonic linear model.

Fifth column (Model 4) provides estimation results of the parametric part in semi-parametric hedonic price with smoothing cohort effects, and the joint term for age and cohort effects. Age is specified as a linear term, and turned out negative, but not significantly different from zero. Besides, the estimates of time effects were not on the whole significant.

Table 3 also reports 'deviance explained', which is a measure of fit for nonparametric models. Similarly to R-squared statistics in linear models, deviance approaches unity with small residuals.

Overall, the deviance statistic reached a high level of 0.948 for Models 3 and 4, and the measure of fit was only marginally lower for Model 2. The generalized cross validation (GCV, see Appemdix A.1) score eq.(13) is a method to choose the degree of smoothing for fitting a model. Apart from finding an appropriate level of smoothness, the relatively low GCV score for Model 3 indicates that this Model is slightly more preferable compared to other hedonic price models in table 3. Conversely, the GCV score turned out the highest for the standard linear hedonic model, where it is as high as 0.504.

5.2 Comparison of nonlinear effects of age and cohort

Figure 2 plots the age effect, estimated by Model 1 (eq.(2)). The effect is allowed to be nonlinear, and on the whole it shows a declining effect of age on prices, implying that older housing is sold at discount in Japan. The estimated number of degrees of freedom for the nonparametric term is 3.51 (as reported in table 4), indicating a moderate nonlinearity of the estimated age effect. Besides, table 4 reports that an approximate p-value for the null hypothesis that the smoothing age term is zero is small enough, implying that the effect is statistically significant at 1 percent significance level.

In a similar way, figure 3 displays the age effect that we obtained with Model 2 (eq.(3)), in which the cohort term is estimated by a nonlinear smooth function. This smoothed cohort effect is plotted in figure 5. Compared to smoothed age effect in figure 3, the smoothed cohort effect is more nonlinear, as indicated by large number of estimated degrees of freedom 12.58, which is much larger that the comparable estimate for the degree of freedom 3.06 of the age effect in Model 2, as shown in figure 3. The slope of the cohort effect is declining, indicating a positive effect on price for houses build in the 1960s. As shown in Table 4, the smoothed cohort effect was statistically different from zero.

Figures 4, 6 and 8 show estimated of smoothed age and cohort effects, and their joint

effect on housing prices, as estimated by Model 3 (eq.(4)). Profiles of age and cohort effects were quite similar to estimates from Model 2 (eq.(3)), though the standard errors turned out larger in Model 3. As for the joint effect of age and cohort effects, Figure 8 shows that it was complex. For relatively old and recent cohorts of houses, the age effect shows a decline in prices with older houses. However, for houses in the middle of our cohort range, the age effect on prices turned out basically flat. Overall, we found that the shape of age effect on housing prices was not the same for different cohorts of houses.

How this peculiar pattern of joint age and cohort effects can be interpreted? Coulson and McMillen (2008) noted that housing cohorts measure the separate impact from the period of housing construction, such as unmeasured style characteristics. Economic conditions at the construction year also affect the decision-making of sellers and buyers, which is likely to be reflected in transaction price. There is a possibility that buyers would expect high future income if the economy at the construction year is in good shape. Such expectations might increase the bid-price of houses. Moreover, sellers might build more luxurious houses. In consequence, cohort effects in our models may pick up the effects of these general economic conditions.

However, our housing data do not have in sufficient details information about style characteristics of houses, and we have no data about attitudes of sellers and buyers during construction years. Thus, as a proxy for this missing information, we used the growth rate of Japanese real GDP G_{t-j} in construction years as an additional explanatory variable to estimate the cohort effect. Figure 9 plots the annual rate of change in the Japanese real GDP. Average growth rate was 9.5% in the 1960s, 5.2% in the 1970s, 4.4% in the 1980s, 1.5% in the 1990s and 0.7% in the 2000s.

After substituting the growth of GDP for the cohort effect, we obtained the following Model 5:

$$P_{t,j,t-j} = \alpha_t + s(A_j) + \gamma_g \cdot G_{t-j} + \boldsymbol{X}'_{t,j,t-j}\boldsymbol{b} + u_{t,j,t-j}. \tag{7}$$

Results of estimating Model 5 (eq.(7)) are reported in the rightmost column in Table 3. As expected, a higher GDP growth has positive effect on housing prices, and the estimated parameter is statistically significant. Figure 5, 6, 7 show that smoothing cohort effects in the 1960's push up the price. Replacing a cohort year variable with GDP growth rate, we may approach the true character of nonlinear cohort effect. Nevertheless, the GCV score of Model 5 was 0.413, which is much higher than comparable scores for Models 3 and 4, where we did not approximate the cohort effect by the GDP growth.

5.3 Model selection

Table 5 reports results of comparing several pairs of models, one of which is restricted, while the other is unrestricted. In other words, these two models specify the null hypothesis H_0 and alternative hypothesis H_1, respectively, as shown in the heading of table 5. Hypothesis testing is based on the deviance of generalized additive models. The table reports two sets of p values. The first set is derived from the F distribution, which is not strictly applicable to the generalized additive model due to the use of non-linear terms (Wood, 2006). The second set of p values is obtained from regression bootstrap testing. Our bootstrap approach is explained in more details in Appendix A.3.

The first row reports F value and corresponding p values for the linear model eq.(6) and Model 1 eq.(2). F value was 57.2, and both p values turned out sufficiently small, so the null hypothesis of linear model for the age effect and the lack of cohort effect was rejected. The second row compares Model 1 to Model 2 eq.(3) that has smoothed age effect and cohort effects. Once again, both p values are much lower than the significance level of 0.05, indicating that Model 2 is preferable to more restricted Model 1. The third row compares Model 3 eq.(4), which has a joint smooth term for age and cohort smooth, to Model 2 eq.(3), in which age and cohort effects are included only as additive terms. The comparison produced very low p values, indicating that the null hypothesis (i.e., Model 2 in this comparison) can be rejected at 0.05 significance level. In the fourth row, we compare Model 3 eq.(4) and Model 4 eq.(5). In the latter model, age effect is expressed by a linear term, rather than a set of dummy variables for different ages. In this comparison, p values proved inconsistent, with the first p value exceeding the significance level of 0.05, and thus proving support for the model under the null hypothesis (i.e., Model 3). On the other hand, the bootstrap p value turned out less than 0.05, indicating that the null hypothesis can be rejected, and giving support to Model 4.

How these conflicting results in comparing Models 3 and 4 can be reconciled? One additional piece of evidence is provided by GCV scores, reported at the last row of Table 3. Note that out of five estimated models, the smallest GCV score was in Model 3 (0.324), while the score was only marginally larger for Model 4 (0.325). Based on this result, our final preference is for Model 3, with Model 4 only marginally less preferable.

5.4 Estimation of hedonic price indices

In figure 10 we report the quality-adjusted price indices that can be estimated from year effects in our hedonic models. The price indices were estimated by setting the price for 1990 to 1, and then estimating indices for subsequent years as

$$\{\exp(0), \exp(\hat{\alpha}_{1991}), \cdots, \exp(\hat{\alpha}_{2008})\}.$$

Line 0 denotes the price index derived from the standard linear hedonic model that omit cohort effects eq.(6), and which is specified by OLS. Lines 1, 2, 3, and 4 denote price indices that are derived, respectively, from Models 1 (eq.(2)), 2 (eq.(3)), 3 (eq.(4)), and 4 (eq.(5)) as specified by generalized additive model. Line 5 denotes price indices from the OLS estimation result of Model 5 eq.(7)

It turned out that the major difference between these alternative indices appeared whenever we included the joint effect of age and cohort of houses. Overall, our results show that conventional hedonic price indices that do not include cohort effects on housing prices, as well as the joint effect of age and cohort, may produce an upward bias in estimated price indices (as evident, for example, in price index estimates from the standard hedonic price model).

6 Conclusion

The purpose of this paper is to solve identification problem among time-related variables. Year of the sale, age of the construction and construction year are important in estimating hedonic models of house prices. If we could observe exogenous changes

of housing markets and have satisfactory measurements of quality features of housing units, then proxy variables like the year of the sale, housing age and construction year are not necessary in the hedonic approach. However, because such quality-related data are rarely available, the use of time-related proxy variables is usually unavoidable. Unfortunately, this creates another problem, because the time-related proxy variables have perfect collinearity among them. Due to the collinearity, conventional hedonic regression models did include both age and cohort variables in their specification (Sirmans *et al.*, 2006).

To disentangle the perfect collinearity between time, age and cohort variables, this paper suggested to use a semiparametric regression approach that imposes relatively mild restrictions on estimated hedonic model.

In this approach, we approximated age and cohort effects by smooth nonparametric functions. Compared with nonlinear age effect, estimates of the cohort effect were more non-linear, as shown by a larger number of degrees of freedom that are required to approximate the nonlinear function. The cohort effect also showed relatively high prices for older cohorts of houses, especially for houses that were build in the 1960s. We may interpret this pattern by a relatively good maintenance of old houses in Japan, which allowed such houses to remain in the housing market, with sellers capable to charge premium prices.

We also tried to explain the estimated profile of the cohort effect by attributing the cohort effect to changing economic conditions during the construction years, since these conditions have effect on the decision-making of sellers and buyers. Thus, we replaced our cohort-year variable with annual growth of Japan's GDP, which we assumed to be a good proxy for the true pattern of nonlinear cohort effect. However, this model did not perform better than models with the original cohort-year variable.

Our major finding is that the introduction of smooth joint function of age and cohort effects resulted in models with the best performance in terms of explained deviance and generalized cross-validation score. The estimated pattern of the joint effect showed that the shape of age effect was not the same for different cohorts of houses, indicating that the house depreciation rate in Japan may depend on specific cohorts of housing. We also found that the omission of the joint effect of age and cohort terms may produce a bias in hedonic price indices. This implies that it may not be sufficient if hedonic price models control for only the age effect on housing prices in pooling data.

A Appendix. Outline of nonparametric estimation and hypothesis testing.

A.1 Estimation of a single nonparametric term.

Consider a reduced specification of eq.(2) that includes only the nonparametric term $f(z_i)$. Once this basic case is introduced, its extension to the full semiparametric model (2) will be trivial. In the reduced specification, the dependent variable y_i is explained by a single explanatory variable z_i with a nonlinear effect on y_i:

$$y_i = f(z_i) + \epsilon_i \tag{8}$$

where $f(\cdot)$ is an arbitrary smooth function and ϵ_i is the error term with zero mean and variance σ^2.

Let $\kappa_1 < \cdots < \kappa_M$ be a sequence of breakpoints ('knots') that are distinct numbers that span the range of z_i. In the MGCV algorithm, the smooth function $f(z_i)$ is approximated by a sequence of cubic splines. In general, splines are piecewise polynomials that are joined at the 'knots'. Due to special restrictions, the cubic splines are continuous at the knots, and also have continuous first and second derivatives. Let M denote the number of knots. Then a cubic spline can be represented by truncated cubic basis functions:

$$f(z_i) = \delta_0 + \delta_1 z_i + \delta_2 z_i^2 + \delta_3 z_i^3 + \sum_{m=1}^{M} \delta_{m+3}(z_i - \kappa_m)_+^3 \tag{9}$$

where

$$(z_i - \kappa_m)_+ = \begin{cases} 0 & z_i \leq \kappa_m \\ z_i - \kappa_m & z_i > \kappa_m \end{cases}$$

In this representation, the cubic spline has a simple interpretation of a *global* cubic polynomial $\delta_0 + \delta_1 z_i + \delta_2 z_i^2 + \delta_3 z_i^3$ and M *local* polynomial deviations $\sum_{m=1}^{M} \delta_{m+3}(z_i - \kappa_m)_+^3$. In matrix form, the truncated cubic basis becomes $\boldsymbol{y} = \boldsymbol{Z\delta} + \boldsymbol{\epsilon}$, where $\boldsymbol{Z}$ is design matrix with ith row vector $\boldsymbol{Z}_i = \begin{bmatrix} 1 & z_i & z_i^2 & z_i^3 & (z_i - \kappa_1)_+^3 & \cdots & (z_i - \kappa_M)_+^3 \end{bmatrix}$, $\boldsymbol{\delta}$ is the corresponding vector of regression parameters, and ϵ is the error term. The smooth function $f(\boldsymbol{Z}, \boldsymbol{\delta})$ is linear in $M + 4$ regression parameters, and can be fitted by minimizing the sum of squared residuals: $(\boldsymbol{y} - \boldsymbol{Z\delta})'(\boldsymbol{y} - \boldsymbol{Z\delta}) = \|\boldsymbol{y} - \boldsymbol{Z\delta}\|^2$, where $\| \cdots \|$ stands for the Euclidean norm.

By increasing the number of knots M, the model becomes more flexible in approximating y. But if the number of knots is too large, the estimates $\hat{f}(z)$ may follow y too closely. In the limit, when $M = n$, the cubic spline simply interpolates y. To prevent too much wiggliness in the estimated curve, a special term that penalizes rapid changes in $\hat{f}(z)$ is added to the fitting criteria. A common penalty is $\lambda \int [f_{zz}(z)]^2 \, dx$, which has a smoothing parameter λ and an integrated squared second derivative $f_{zz}(z)$ of $f(z)$. This results in the penalized least-squares criterion as follows:

$$Q(f, \lambda) = \|\boldsymbol{y} - \boldsymbol{Z\delta}\|^2 + \lambda \int [f_{zz}(z)]^2 \, dx.$$

If $\hat{f}(z)$ is too rough, this will increase the penalty term $\int [f_{zz}(z)]^2 \, dx$. The smoothing parameter λ controls the trade-off between the model fit $\|\boldsymbol{y} - \boldsymbol{Z\delta}\|$ and *the roughness*

penalty $R = \int [f_{zz}(z)]^2 dx$. When $\lambda = 0$, the roughness penalty R has no effect on the minimization criterion $Q(f, \lambda)$, producing unpenalized estimates $\hat{f}(x)$ that just interpolate data. In contrast, when $\lambda = +\infty$, this results in the perfectly smooth line, *i.e.*, in a linear regression line with a constant slope.

The minimization of the penalized criterion $Q(f, \lambda)$ is simplified by noting that derivatives and integrals of $f(z)$ are linear transformations of parameters $d^m(z)$ in the cubic spline basis, with $f_{zz}(z) = \sum_{m=1}^{M} \delta_m d^m_{zz}(z)$ and $\int f(z)dz = \sum_{m=1}^{M} \delta_m \int d^m(z)dz$, where $d^m(z)$ denotes a particular form of basis function (such as the truncated cubic basis function in (9)). Thus, $f_{zz}(z) = \boldsymbol{d}_{zz}(z)'\boldsymbol{\delta}$, from which it follows that $[f_{zz}(z)]^2 = \boldsymbol{\delta}'\boldsymbol{d}_{zz}(z)'\boldsymbol{d}_{zz}(z)\boldsymbol{\delta} = \boldsymbol{\delta}'F(z)\boldsymbol{\delta}$. Finally,

$$R = \int [f_{zz}(z)]^2 dz = \boldsymbol{\delta}' \left(\int F(z)dz \right) \boldsymbol{\delta} = \boldsymbol{\delta}'\boldsymbol{S}\boldsymbol{\delta}.$$

Thus, the roughness penalty R can be represented as a quadratic form in the parameter vector δ and matrix $\boldsymbol{S}$ of known coefficients that are derived from the basis function $d^m(z)$.

Substituting the roughness penalty R with $\boldsymbol{\delta}'\boldsymbol{S}\boldsymbol{\delta}$, the penalized least-squares criterion becomes

$$Q(f, \lambda) = \|\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\delta}\|^2 + \lambda\boldsymbol{\delta}'\boldsymbol{S}\boldsymbol{\delta}. \tag{10}$$

Differentiating $Q(f, \lambda)$ with respect to $\boldsymbol{\delta}$ and setting the derivative to zero produces an estimate of $\boldsymbol{\delta}$:

$$\hat{\boldsymbol{\delta}} = \left(\boldsymbol{Z}'\boldsymbol{Z} + \lambda\boldsymbol{S}\right)^{-1} \boldsymbol{Z}'\boldsymbol{y}. \tag{11}$$

The estimate of δ depends on the value of unknown smoothing parameter λ. The MGCV algorithm selects an appropriate value of λ by using the concept of hat matrix from the ordinary least-squares model. In the model, the hat matrix $\boldsymbol{H}$ projects the vector of dependent variable $\boldsymbol{y}$ into the vector of predicted values $\hat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y}$, with $\boldsymbol{H} = \boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'$. Using the estimate of $\hat{\delta}$ from (11), the hat matrix of the penalized spline model can be similarly defined as $\boldsymbol{H}_S = \boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z} + \lambda\boldsymbol{S}\right)^{-1}\boldsymbol{Z}'$. Since the matrix $\boldsymbol{H}_S$ transforms the vector of $\boldsymbol{y}$ into the vector of its smoothed values, the matrix $\boldsymbol{H}_S$ is often called a smoother matrix. In the MGCV algorithm, the optimal value of λ is found by minimizing the GCV criteria $V_g(\lambda)$ that depends on the sum of squared residuals $\|\boldsymbol{y} - \boldsymbol{Z}\hat{\boldsymbol{\delta}}\|^2$ and the trace of smoother matrix $\boldsymbol{H}_S$:

$$V_g(\lambda) = \frac{n\|\boldsymbol{y} - \boldsymbol{Z}\hat{\boldsymbol{\delta}}\|^2}{\left[n - \mathrm{tr}\left(\boldsymbol{H}_S\right)\right]^2} \tag{12}$$

where n is the number of observations, and $\mathrm{tr}\left(\boldsymbol{H}_S\right)$ is the trace of $\boldsymbol{H}_S$.

Though the MGCV algorithm selects an appropriate degree of smoothness with respect to parameter λ, this parameter is not informative in evaluating the estimated degree of smoothness. It is much easier to interpret the trace of the smoother matrix $\mathrm{tr}\left(\boldsymbol{H}_S\right)$, since it is equal to the number of degrees of freedom, needed to approximate the smoothed function $f(z)$ (Ruppert *et al.*, 2003). Let $\nu = \mathrm{tr}\left(\boldsymbol{H}_S\right)$. Since the smoothing parameter λ is a part of $\boldsymbol{H}_S$, λ and ν are correlated. In particular, a small degree of smoothing is indicated by $\lambda \to 0$ and $\nu \to \infty$. Conversely, a high degree of smoothing corresponds to $\lambda \to \infty$ and $\nu \to 0$. An important special case is when $\nu \leq 1$. This range

of ν indicates a parametric effect, when a single variable is sufficient to approximate the smoothed function $f(z)$.

The GCV criterion $V_g(\lambda)$ has one problem in selecting an optimal smoothness. Monte Carlo studies by Kim and Gu (2004) and Bacchini *et al.* (2007) demonstrated that $V_g(\lambda)$ may choose too small values of λ, which results in undersmoothing. The problem can be solved by multiplying $\mathrm{tr}\,(\boldsymbol{H}_S)$ in (12) by a parameter η that increases the cost per trace of $\boldsymbol{H}_S$:

$$\bar{V}_g(\lambda) = \frac{n\|\boldsymbol{y} - \boldsymbol{Z}\hat{\boldsymbol{\delta}}\|^2}{[n - \eta \cdot \mathrm{tr}\,(\boldsymbol{H}_S)]^2} \,. \tag{13}$$

In estimating the smoothing cohort model, we followed the recommendation in Wood (2006) that a good value for η is 1.4. In practice, the modification had little effect on our estimates of age or cohort effects.

After specifying how the smooth function $f(x)$ is estimated by spline basis functions, the basic model (8) can be easily extended to the full semiparametric model eq.(2) that adds the parametric part with cohort and year effects (in subsection 2.2). For the smoothing age model, the parametric part $\boldsymbol{W}$ includes matrices of dummy variables D_t^Y, D_ℓ^C. After the extension, the truncated cubic basis (9) still has the form $\boldsymbol{y} = \tilde{\boldsymbol{Z}}\tilde{\boldsymbol{\delta}} + \epsilon$, but the basis $\tilde{\boldsymbol{Z}}$ now includes an expanded design matrix $\tilde{\boldsymbol{Z}} = [\boldsymbol{Z}, \boldsymbol{W}]$. The estimate of $\tilde{\boldsymbol{\delta}}$ is obtained from (11), where the smoothing parameter λ is found by minimizing either $V_g(\lambda)$ or $\bar{V}_g(\lambda)$.

A.2 Estimation of a joint effect of two smooth functions.

In this subsection, we describe how we estimated the joint effects of age and cohort of housing in Model 3 and 4 (eq.(4) and (5) in subsection 3.2). While the effect of single nonparametric term z_i on y_i in eq.(8) produces a smooth line that account a possible nonlinear relationship, the joint effect of two variables a_i and c_i on y_i is given by $y_i = f(a_i, c_i) + \epsilon_i$. The joint effect of a_i and c_i on y_i produces a smooth surface, in which the effect of a_i on y_i may be not only nonlinear, but also different at various levels of c_i.

In estimating the smooth effect of two covariates a_i and c_i on y_i, we used a tensor product smoother that was introduced in Wood (2006). The smoother is closely related to the univariate smoother that we described in subsection A.1. Essentially, the joint smoother of a_i and c_i is constructed from marginal bases and penalties of each of the covariates. Consider the construction of the joint basis function of $f(a, c)$. Let marginal smoothing terms for $f_a(a)$ and $f_c(c)$ be denoted by $f_a(a) = \sum_{q=1}^{M_q} \theta_q^a d^q(a)$ and $f_c(c) = \sum_{r=1}^{M_r} \theta_r^c d^r(c)$, where θ_q^a and θ_r^c are regression parameters (similar to the parameter δ in the univariate specification eq.(9)), and $d^q(a)$ and $d^r(c)$ are basis functions for a and c. To proceed from $f_a(a)$ and $f_c(c)$ to $f(a, c)$, we first assume that θ_q^a in the basis function of $f_a(a)$ is a smooth function of c, with $\theta_q^a(c) = \sum_{r=1}^{M_r} \delta_{qr} d^r(c)$. Then the joint basis for a and c becomes

$$f(a, c) = \sum_{q=1}^{M_q} \theta_q^a(c) d^q(a) = \sum_{q=1}^{M_q} \sum_{r=1}^{M_r} \delta_{qr} d^r(c) d^q(a) \tag{14}$$

In matrix form, the joint basis regression model is written by $\boldsymbol{y} = \boldsymbol{Z}(a, c)\boldsymbol{\delta} + \boldsymbol{\epsilon}$. Essentially, the joint basis function $\boldsymbol{Z}(a, c)$ is constructed as the Kronecker product of

individual marginal smoothing bases of a and c, denoted $\boldsymbol{Z}_a$ and $\boldsymbol{Z}_c$. For example, for the univariate smooth term a, the individual smoothing base was defined by $\boldsymbol{Z}$ in subsection A.1.

The roughness penalty for the joint smoother is constructed similarly to the joint smoothing basis function $\boldsymbol{Z}$, by using marginal roughness penalties for a and c. For the univariate smooth of a, such a penalty was already defined by (10). To construct the composite penalty term, let $f_{a|c}(a)$ be a joint smooth of a and c with some fixed c. Then the roughness of $f_{a|c}$ is given by $R_a(f_{a|c})$. By integrating $R_a(f_{a|c})$ across different c, we obtain $R_a(f_a) = \int R_a(f_{a|c})dc$, which measures the total roughness of $f(a, c)$ in the direction of a.

The total roughness penalty in the direction of c is obtained similarly, by fixing a at some specific points, and integrating the total roughness penalty $R_c(f_c) = \int R_c(f_{c|a})da$ across different fixed values of a. So a reasonable penalty is

$$\lambda_a \int R_a(f_{a|c})dc + \lambda_c \int R_c(f_{c|a})da.$$

On the assumption that $f_{a|c}(a) = \sum \theta_q^a(c)d^q(a)$, we colud write $R_a(f_{a|c}) = \boldsymbol{\theta}^a(c)'\boldsymbol{S}_a\boldsymbol{\theta}^a(c)$. A simple reparameterization can be used to provide an approximation to the terms in penalty: $\boldsymbol{\theta}^{a\prime} = \boldsymbol{\Gamma}\boldsymbol{\theta}^a$. Hence the penalty coefficient matrix becomes $\boldsymbol{S}'_a = \boldsymbol{\Gamma}^{-1\prime}\boldsymbol{S}_a\boldsymbol{\Gamma}^{-1}$. Then $R_a(f_a)$ and $R_c(f_c)$ are used to create composite roughness penalties $\bar{\boldsymbol{S}}_a = \boldsymbol{S}'_a \otimes \boldsymbol{I}_{M_r}$ and $\bar{\boldsymbol{S}}_c = \boldsymbol{I}_{M_q} \otimes \boldsymbol{S}'_c$, where $\boldsymbol{I}_{M_r}$ and $\boldsymbol{I}_{M_q}$ denote identity matrices, with M_q and M_r equal to the number of 'knots' in the direction of c and a, respectively.

Using the composite roughness penalties $\bar{\boldsymbol{S}}_a$ and $\bar{\boldsymbol{S}}_c$, the penalized least-squared criterion is constructed similarly to (10), by combining the least-squares term with roughness penalties in the direction of a and c, which are multiplied by the corresponding smoothing parameters λ_a and λ_c:

$$Q(f(a, c), \lambda_a, \lambda_c) = \|\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\delta}\|^2 + \lambda_a\boldsymbol{\delta}'\bar{\boldsymbol{S}}_a\boldsymbol{\delta} + \lambda_c\boldsymbol{\delta}'\bar{\boldsymbol{S}}_c\boldsymbol{\delta} \tag{15}$$

Specific details about the construction of the joint basis function $\boldsymbol{Z}(a, c)$ and the roughness penalty are provided in Wood (2006). Similarly to the univariate case, individual smoothing parameters λ_a and λ_c are selected by minimizing the GCV criterion, as defined in (13).

A.3 Hypothesis testing with bootstrap.

Since the GAM estimator does not belong to conventional linear regression models, hypothesis testing is complicated because the finite sample distribution of test statistics is not known. The problem can be solved by using a bootstrap testing procedure that resamples residuals from a GAM fit. Consider two models, called Model A and B. Let Model A satisfy the null hypothesis, and Model B satisfy the alternative hypothesis. Denote fitted values and residuals from estimating Model A as $\hat{y}^A$ and $\hat{u}^A$. Let the actual value of test statistic be $\hat{\phi}$. To estimate a p-value for the test statistic $\hat{\phi}$, we used the following bootstrap approach from MacKinnon (2007):

1. Specify the number of bootstrap replications O, and the significance level of the test.

2. For each $o = 1, \cdots, O$, resample regression residuals from $\hat{u}^A$, and denote the bootstrap sample as $\hat{u}_o^A$. Then calculate bootstrap values of y as $y_o^A = \hat{y}^A + \hat{u}_o^A$.

3. Using y_o^A and matrix of independent variables $\boldsymbol{x}$, estimate alternative model B, and calculate a bootstrap test statistic ϕ_o^* .

4. Repeat until the last bootstrap resampling of $\hat{u}^A$ that produces test statistic ϕ_O^*.

5. Estimate a bootstrap p-value for $\hat{\phi}$ by $\hat{p}^*(\hat{\phi}) = \frac{1}{O}\sum_{o=1}^{O} I\left(\phi_o^* > \hat{\phi}\right)$. Suppose that ϕ_o^* was larger than $\hat{\phi}$ at 35 times, and $O = 1000$. Then $\hat{p}^*(\hat{\phi}) = 35/1000 = 0.035$.

6. If $\hat{p}^*(\hat{\phi}) <$ significance level, reject the null hypothesis, and otherwise, accept it.

References

Bacchini, M., Biggeri, A., Lagazio, C., Lertxundi, A. and Saez, M. (2007) Parametric and semi-parametric approaches in the analysis of short-term effects of air pollution on health, *Computational Statistics and Data Analysis*, **51**, 4324–4336.

Bailey, M. J., Muth, R. F. and Nourse, H. O. (1963) A regression model for real estate price index construction, *Journal of the American Statistical Association*, **58**, 933–942.

Coulson, E. N. and McMillen, D. P. (2008) Estimating time, age, and vintage effects in housing prices, *Journal of Housing Economics*, **17**, 138–151.

Craven, P. and Wahba, G. (1979) Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross validation, *Numerische Mathematik*, **31**, 377–403.

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

Deaton, A. S. and Paxson, C. (1994) Saving, growth, and aging in Taiwan, in *Studies in the Economics of Aging* (Ed.) D. Wise, University of Chicago Press, Chicago, pp. 331–357.

Fu, W. J. (2008) A smoothing cohort model in age–period–cohort analysis with applications to homicide arrest rates and lung cancer mortality rates, *Sociological Methods and Research*, **36**, 327–361.

Halvorsen, R. and Pollakowski, H. O. (1981) Choice of functional form for hedonic price equations, *Journal of Urban Economics*, **33**, 37–49.

Hastie, T. J. and Tibshirani, R. J. (1990) *Generalized Additive Models*, Chapman and Hall–CRC, London.

Kim, Y.-J. and Gu, C. (2004) Smoothing spline gaussian regression: more scalable computation via efficient approximation, *Journal of Royal Statistical Society (Series B)*, **66**, 337–356.

MacKinnon, J. (2007) Bootstrap hypothesis testing, Working paper no. 1127, Department of Economics, Queen's University.

McKenzie, D. (2006) Disentangling age, cohort, and time effects in the additive model, *Oxford Bulletin of Economics and Statistics*, **68**, 473–495.

Paxson, C. (1996) Saving and growth: Evidence from micro data, *European Economic Review*, **40**, 255–288.

R Development Core Team (2009) *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, `http://www.R-project.org`.

Ruppert, D., Wand, M. P. and Carroll, R. J. (2003) *Semiparametric Regression*, Cambridge University Press, Cambridge.

Schimek, M. G. (forthcoming) Semiparametric penalized generalized additive models for environmental research and epidemiology, *Envinmetrics*, p. accepted for publication.

Sirmans, G. S., MacDonald, L., MacPherson, D. A. and Zietz, E. N. (2006) The value of housing characteristics: a meta analysis, *Journal of Real Estate Finance and Economics*, **33**, 215–240.

Wood, S. (2004) Stable and efficient multiple smoothing parameter estimation for Generalized Additive Models, *Journal of the American Statistical Association*, **99**, 673–686.

Wood, S. (2006) *Generalized Additive Models. An Introduction with R*, Chapman and Hall–CRC, Boca Raton, Florida.

Table 1: Descriptive statistics

Variable	Mean	Std. Dev.	Minimum	Maximum
Log of sales price	8.003	0.603	5.966	11.771
Year of sale	1999.323	5.165	1990	2008
Year bulit	1981.737	7.397	1954	2008
Age	17.585	8.044	0	50
X^1: Log of sq. meters	4.005	0.395	2.461	6.085
X^2: Log of time distance from CBD	1.879	0.676	0	3.258
X^3: Log of minutes on foot to a nearly station	2.157	0.791	0	4.159
X^4: Log of number of houses in a condominium	4.213	0.953	2.639	7.641

Note. The full sample size is 39,218 sales of single-family condominiums in the special 23 wards of Tokyo.

Table 2: Distribution of age cohorts at time of sale

Year built	Age at time of sale					
	$0-9$	$10-19$	$20-29$	$30-39$	$40-50$	Total
$1950-1959$	0	0	0	4	6	10
$1960-1969$	0	0	614	770	82	1,466
$1970-1979$	0	3,757	6,944	2,356	0	13,057
$1980-1989$	3,275	11,677	4,627	0	0	19,579
$1990-1999$	2,605	1,765	0	0	0	4,370
$2000-2008$	736	0	0	0	0	736
Total	6,616	17,199	12,185	3,130	88	39,218

Table 3: Estimation results

Variable	eq.(6) Linear	eq.(2) Model 1	eq.(3) 2 Model 2	eq.(4) Model 3	eq.(5) Model 4	eq.(7) Model 5
Const.	6.431***	0.037***	5.942***	6.194***	7.011***	6.300***
	(0.209)	(0.005)	(0.189)	(0.264)	(2.331)	(0.184)
Year1991	−0.103***	−0.102***	−0.088***	−0.111***	−0.105	−0.100***
	(0.032)	(0.029)	(0.026)	(0.032)	(0.116)	(0.029)
Year1992	−0.267***	−0.266***	−0.245***	−0.292***	−0.278	−0.260***
	(0.031)	(0.028)	(0.025)	(0.047)	(0.231)	(0.028)
Year1993	−0.344***	−0.343***	−0.320***	−0.390***	−0.367	−0.331***
	(0.030)	(0.028)	(0.025)	(0.065)	(0.351)	(0.027)
Year1994	−0.461***	−0.460***	−0.433***	−0.525***	−0.493	−0.444***
	(0.030)	(0.027)	(0.026)	(0.083)	(0.474)	(0.027)
Year1995	−0.618***	−0.617***	−0.585***	−0.701***	−0.659	−0.596***
	(0.030)	(0.027)	(0.026)	(0.102)	(0.600)	(0.027)
Year1996	−0.711***	−0.710***	−0.674***	−0.810***	−0.761	−0.685***
	(0.031)	(0.028)	(0.028)	(0.120)	(0.729)	(0.028)
Year1997	−0.732***	−0.731***	−0.687***	−0.845***	−0.791	−0.700***
	(0.031)	(0.028)	(0.029)	(0.138)	(0.860)	(0.028)
Year1998	−0.769***	−0.771***	−0.720***	−0.898***	−0.842	−0.737***
	(0.031)	(0.028)	(0.030)	(0.156)	(0.993)	(0.028)
Year1999	−0.808***	−0.812***	−0.755***	−0.951***	−0.898	−0.772***
	(0.031)	(0.029)	(0.032)	(0.173)	(1.127)	(0.028)
Year2000	−0.861***	−0.868***	−0.800***	−1.017***	−0.970	−0.823***
	(0.031)	(0.028)	(0.033)	(0.191)	(1.263)	(0.028)
Year2001	−0.872***	−0.885***	−0.814***	−1.049***	−1.014	−0.835***
	(0.031)	(0.029)	(0.035)	(0.208)	(1.400)	(0.029)
Year2002	−0.860***	−0.878***	−0.800***	−1.051***	−1.035	−0.822***
	(0.032)	(0.030)	(0.037)	(0.224)	(1.538)	(0.030)
Year2003	−0.821***	−0.845***	−0.759***	−1.027***	−1.035	−0.785***
	(0.033)	(0.030)	(0.039)	(0.241)	(1.676)	(0.030)
Year2004	−0.818***	−0.851***	−0.756***	−1.041***	−1.082	−0.785***
	(0.033)	(0.030)	(0.041)	(0.257)	(1.814)	(0.031)
Year2005	−0.800***	−0.841***	−0.739***	−1.040***	−1.122	−0.771***
	(0.033)	(0.030)	(0.042)	(0.273)	(1.953)	(0.030)
Year2006	−0.757***	−0.809***	−0.699***	−1.014***	−1.146	−0.733***
	(0.033)	(0.030)	(0.044)	(0.288)	(2.092)	(0.031)
Year2007	−0.628***	−0.685***	−0.574***	−0.901***	−1.093	−0.606***
	(0.033)	(0.030)	(0.046)	(0.304)	(2.230)	(0.031)
Year2008	−0.593***	−0.658***	−0.532***	−0.870***	−1.132	−0.572***
	(0.033)	(0.030)	(0.048)	(0.318)	(2.368)	(0.031)
Age	−0.017***	-	-	-	−0.041	
	(0.001)				(0.059)	
s(Age)	No	Yes	Yes	Yes	No	Yes
Cohort	-	0.003***	-	-	-	-
		(0.000)				
s(Cohort)	No	No	Yes	Yes	Yes	No
s(Age,Cohort)	No	No	No	Yes	Yes	No
Growth rate	-	-	-	-	-	0.006***
						(0.001)
X^1	0.762***	0.724***	0.741***	0.720***	0.720***	0.682***
	(0.050)	(0.046)	(0.048)	(0.047)	(0.047)	(0.047)
X^2	−0.178***	−0.176***	−0.142***	−0.121***	−0.124***	−0.179***
	(0.025)	(0.023)	(0.023)	(0.022)	(0.022)	(0.023)
X^3	−0.180***	−0.163***	−0.092***	−0.091***	−0.092***	−0.156***
	(0.016)	(0.015)	(0.018)	(0.017)	(0.017)	(0.015)
X^4	0.048***	0.051***	0.042***	0.040***	0.040***	0.058***
	(0.016)	(0.014)	(0.015)	(0.014)	(0.014)	(0.014)
Deviance explained	0.912	0.927	0.942	0.948	0.948	0.929
GCV score	0.504	0.420	0.354	0.324	0.325	0.413

t values in parentheses. * significant at 10%, ** significant at 5%, *** significant at 1%.

Table 4: Approximate degree of freedom of the smooth and F test

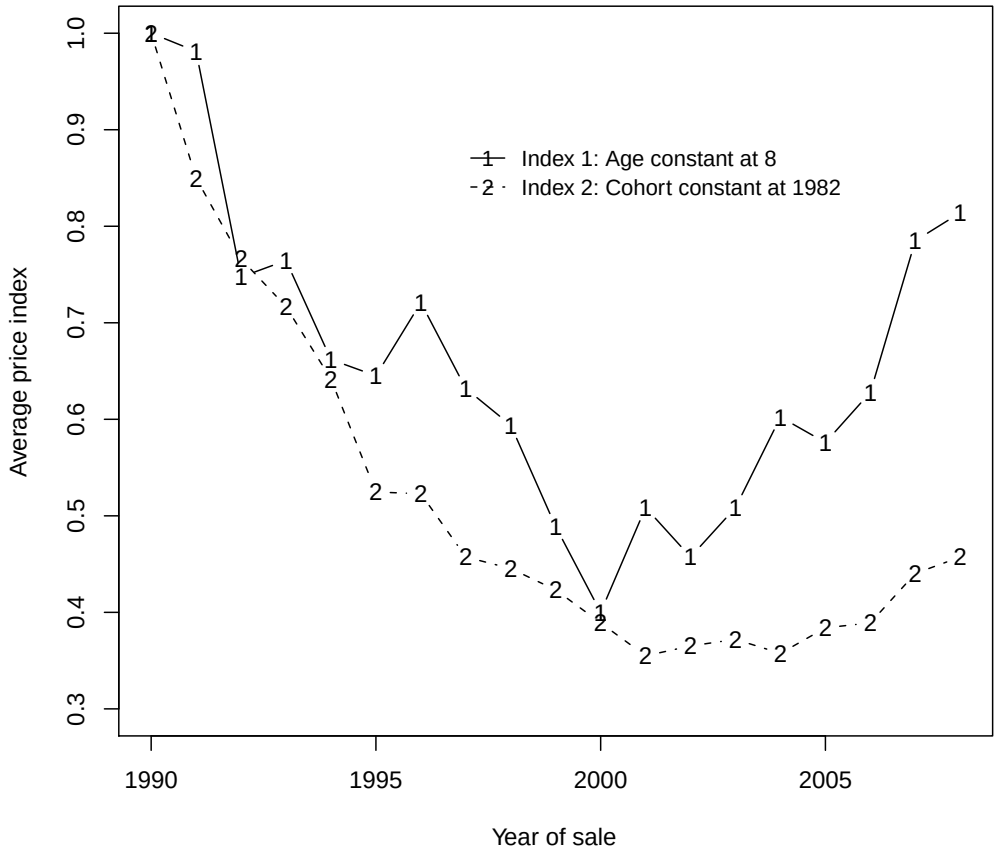
smooth term	Model 1	Model 2	Model 3	Model 4	Model 5
s(Age)	3.51^{***}	3.06^{***}	4.25^{***}	-	3.52^{***}
s(Cohort)	-	12.58^{***}	12.75^{***}	11.40^{***}	-
s(Age,Cohort)	-	-	4.50^{*}	12.70^{**}	-

note: If approximate p-value by F test (for the null hypotheses that the each smoothing term is zero) is less than .01, then ***, less than .05, then ** and less than .1, then *.

Table 5: Model selection

H_0	H_1	F-value	p-value	Bootstrap p-value
Linear Model	Model 1	57.2	0.000	0.003
Model 1	Model 2	14.0	0.000	0.002
Model 2	Model 3	13.7	0.000	0.003
Model 3	Model 4	2.0	0.110	0.003

Figure 1: Average price indices



note: Base year is 1990. Line 1: Age is fixed at 8 years. Dot line 2: Cohort is fixed at 1982.

Figure 2: Age effect (Model 1)

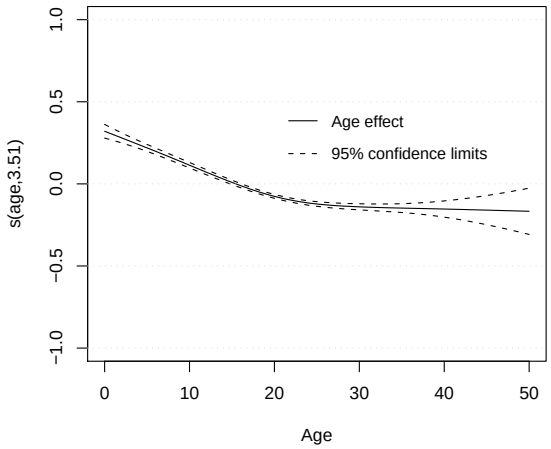


Figure 3: Age effect (Model 2)

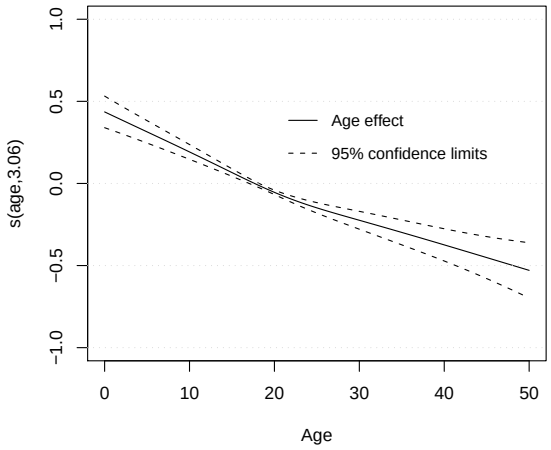


Figure 4: Age effect (Model 3)

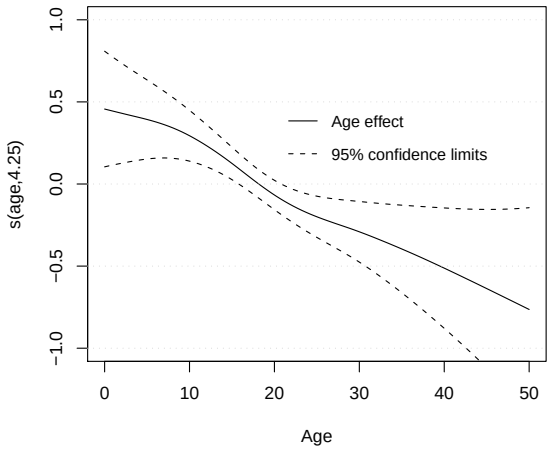


Figure 5: Cohort effect (Model 2)

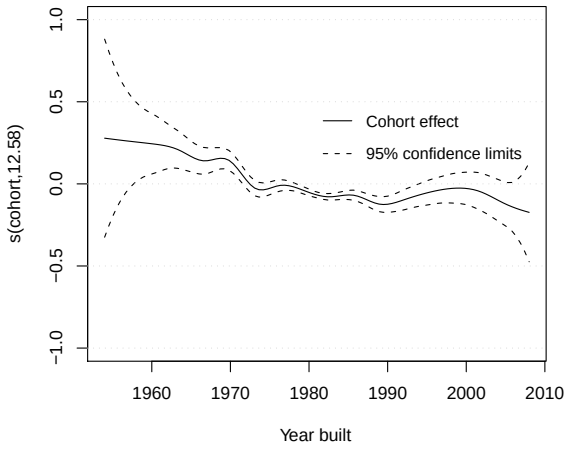


Figure 6: Cohort effect (Model 3)

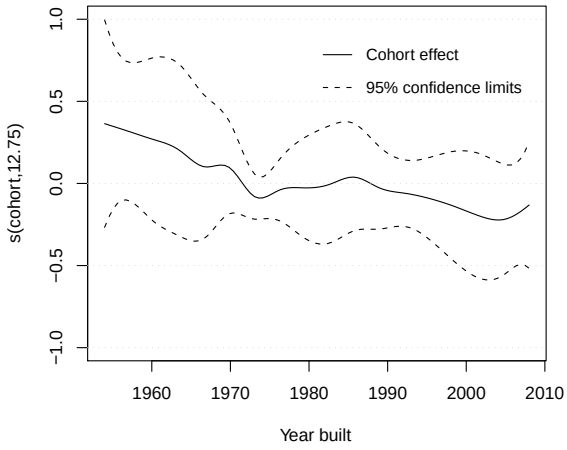


Figure 7: Cohort effect (Model 4)

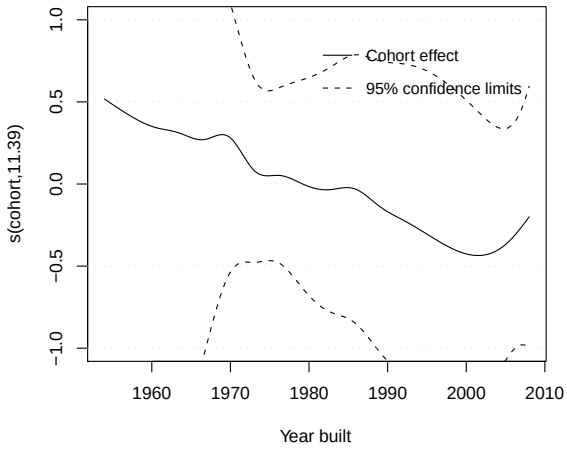


Figure 8: Joint effect of age and cohort (Model 3)

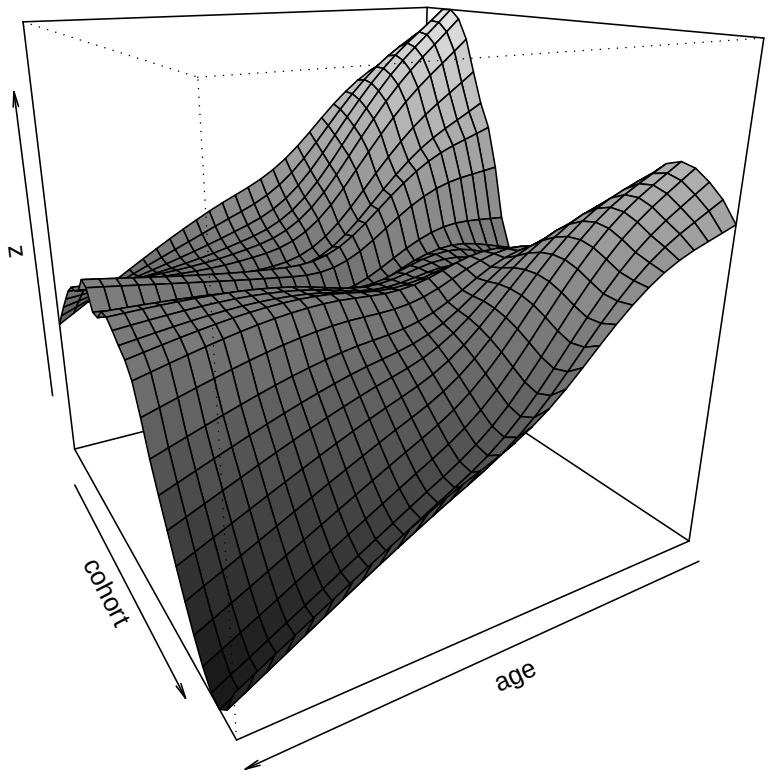
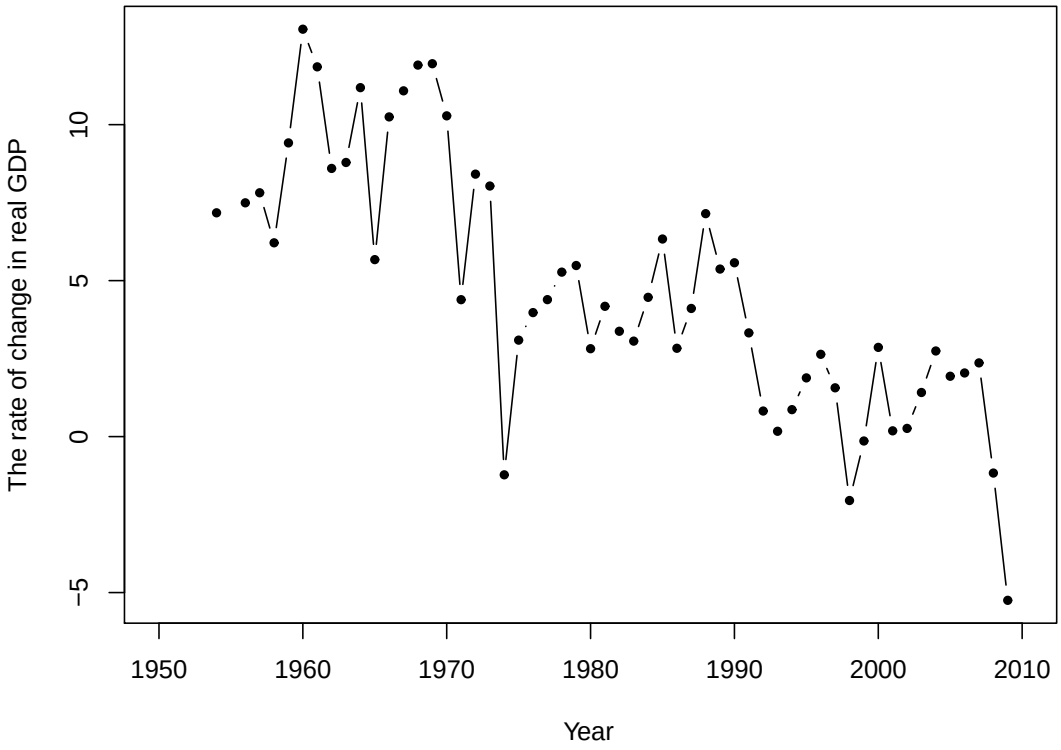


Figure 9: Annual growth rates of Japan's real GDP



Data source: Economic and Social Research Institute, Cabinet Office, Government of Japan

Figure 10: Estimates of hedonic price indices