

A Theory-Based Approach to Hedonic Price Regressions with Time-Varying Unobserved Product Attributes: The Price of Pollution*

Patrick Bajari

University of Minnesota and NBER

Jane Cooley

University of Wisconsin

Kyoo il Kim

University of Minnesota

Christopher Timmins

Duke University and NBER

November 9, 2010

Abstract

We propose a new strategy for a pervasive problem in the hedonics literature—recovering hedonic prices in the presence of time-varying correlated unobservables. Our approach relies on an assumption about homebuyer rationality, under which prior sales prices can be used to control for time-varying unobservable attributes of the house or neighborhood. Using housing transactions data from California’s Bay Area between 1990 and 2006, we apply our estimator to recover marginal willingness to pay for reductions in three of the EPA’s “criteria” air pollutants. Our findings suggest that ignoring bias from time-varying correlated unobservables considerably understates the benefits of a pollution reduction policy.

*This paper has benefited from the comments of seminar participants at Brown University and the University of Toronto.

1 Introduction

In a hedonic regression, the economist attempts to consistently estimate the relationship between prices and product attributes in a differentiated product market. The regression coefficients are commonly referred to as implicit (or hedonic) prices, which can be interpreted as the effect on the market price of increasing a particular product attribute while holding the other attributes fixed. Given utility-maximizing behavior, the consumer's marginal willingness to pay for a small change in a particular attribute can be inferred directly from an estimate of its implicit price; moreover, these implicit prices can be used to recover marginal willingness to pay functions for use in valuing larger changes in attributes (Rosen, 1974).

Hedonic regressions suffer from a number of well-known problems. Foremost among them, the economist is unlikely to directly observe all product characteristics that are relevant to consumers, and these omitted variables may lead to biased estimates of the implicit prices of the observed attributes. For example, in a house-price hedonic regression, the economist may observe the house's square-footage, lot size, and even the average education level in the neighborhood. However, many product attributes such as curb appeal, the quality of the landscaping and the crime rate may be unobserved by the econometrician. If these omitted attributes are correlated with the observed attributes, ordinary least squares estimates of the implicit prices will be biased.

When correlated unobservables are time-invariant and panel data are available, the unobservables can be accounted for with fixed effects. When correlated unobservables vary over time or when panel data are not available, previous research has relied instead on instrumental variables, regression discontinuity, or other forms of quasi-experimental variation to avoid this bias. Chay and Greenstone (2005), Greenstone and Gallagher (2008), and Black (1999) have proposed quasi-experimental approaches to this problem, exploiting either a discontinuity in the application of a regulation or a structural break due to a boundary. If the regulation or boundary is exogenous and generates large movements in housing attributes, these methods may be attractive for estimating implicit prices for at least two reasons. First, they allow the econometrician to remove the bias from omitted variables that may confound estimates of implicit prices. Second, the identifying assumptions are transparent and the estimators are simple to implement (often using well-known statistical packages).

Why would anyone choose not to adopt one of these straightforward approaches? We argue that, in many important hedonic applications, this sort of identification cannot be achieved. First, a source of quasi-randomness that generates exogenous variation in product characteristics may not be available in a particular application, or may be available, but only under very strong assumptions. Such is the case, for example, in Chay and Greenstone (2005), which exploits quasi-random variation in EPA air quality attainment status to recover the effect of air pollution on housing prices. To use this strategy, they must impose the assumption that the United States is comprised of a single, unified housing market.

Second, even if a natural experiment is found, implicit prices may not be precisely estimated because the instruments implied by that experiment are weak. Third, a regulatory discontinuity or structural break caused by a boundary may only identify policy impacts over a narrow range, rather than over the full range of interest to policy-makers. It would therefore be useful to have an alternative set of assumptions with which to identify implicit prices. At a minimum, this alternative approach would provide a way to check the robustness of the results from a quasi-experiment; in other situations, it would provide a viable estimation strategy when quasi-random variation in the product attribute of interest cannot be assumed.

In this paper, we propose such a method. It is based on three simple identifying assumptions. The first assumption is that home price in a local market at a point in time can be written as a function of a home's attributes. Importantly, we assume that this includes attributes that are observed by home buyers but not by the economist (i.e., attributes that are omitted from the regression specification). This assumption is maintained in theoretical models underlying hedonic regressions including Rosen (1974), Epple (1987), Ekeland, Heckman and Nesheim (2004), Heckman, Matzkin and Nesheim (2003) and Bajari and Benkard (2005).

In most applications, it seems reasonable to assume that buyers have superior information about home attributes compared to the economist. For example, it is difficult, if not impossible, for the economist to directly measure the "curb appeal" of a home. However, anyone who has purchased a home knows this is an important consideration for many buyers. Our first assumption implies that curb appeal and other attributes like it are priced by the market even if the econometrician fails to measure them. As a consequence, the residual from a hedonic regression contains information that the researcher can use to price home attributes that she does not directly observe.

Our second identifying assumption is a parameterization of the process that determines the dynamic evolution of the value of the omitted attribute. This assumption turns out to not be terribly restrictive. In particular, the parameterization can be made increasingly flexible with more repeat-transactions observations of the same house.

Our third and final identifying assumption is that homebuyers are rational with respect to their predictions about how the omitted housing attribute evolves over time. Put differently, homebuyers do not make systematic errors in predicting its evolution. The practical implication of this assumption is that the stochastic innovation in the evolution of the omitted housing attribute is uncorrelated with their current information set. Along with the first two assumptions, this allows us to construct estimating equations that yield consistent estimates of implicit prices, even in the presence of time-varying correlated unobservable attributes.

The intuition behind our estimator is straightforward. Suppose that we observe a home that is sold in 1998 and again in 2003. Our first assumption allows us to use the 1998 sales price to impute a market value for the omitted housing attributes in 1998. If the market price was abnormally positive (negative) after controlling for the covariates in the econometrician's data set, we would infer that the home had a large positive (negative) value for characteristics that were not observed by the economist. Our second assumption allows us to say how the value of omitted housing attributes evolves over time in expectation. From this process, we can recover the stochastic innovation in the value of the omitted attribute. Our third assumption provides us with an orthogonality condition based on this stochastic innovation that is similar to conditions used in well-known GMM estimators in financial econometrics.

Our estimator is also similar in spirit to quasi-differencing or differencing approaches to dealing with correlated time-varying unobservables (e.g., see Arellano and Bond (1991) for general panel models and Blundell and Bond (1998) and Akerberg, Benkard, Berry, and Pakes (2007) for production functions and demand models). One key difference from this literature is that our model permits time-varying coefficients (implicit prices).

We also show that our approach can be extended to nonparametric models using a control function approach (see Appendix A) by casting our problem in the framework of Ai and Chen (2003). In contrast, approaches that exploit quasi-randomness may frequently require a parsimonious functional form because instrumental variables do not have adequate variation

to identify models with many parameters.

We admit that our identifying assumptions are an approximation of the way housing markets function in reality. For example, our first assumption does not hold perfectly because home prices are often determined by negotiation and therefore cannot be explained exactly by the home’s characteristics. We argue, however, that there are not many opportunities for a free lunch in a housing market with many buyers and sellers. Finding “steals”, where the asking price significantly understates the value of a home’s attributes, is the exception rather than the rule. Only rarely can a buyer find twice the home for half the price.

Our third assumption is also an approximation of real world housing markets. It may fail to hold if certain types of houses earn above-market returns, even after controlling for observable and unobservable attributes (as measured through prior prices). In that case, home buyers might be able to predict earning excess returns given information available today. Unlike many other identifying assumptions used in this literature, however, this is a possibility for which we can provide supporting evidence – our rationality assumption is a necessary (although by no means sufficient) condition for housing market efficiency as described by Case and Shiller (1989). While their test for housing market efficiency is therefore an overly stringent requirement for our homebuyer rationality assumption to hold, we can use it to provide evidence in support of our assumption.

As an application of our approach, we consider the value individuals place on a marginal improvement in air quality, as revealed by their home buying decisions. In particular, we analyze three of the EPA’s “criteria pollutants” (i.e., pollutants used by the EPA in setting emissions regulations) – particulate matter (PM10), sulfur dioxide (SO2), and ground-level ozone (O3) – all of which are known to have adverse health consequences and impose aesthetic costs. Importantly, we expect there to be many more salient determinants of individual housing choice that our data do not describe. There is, therefore, good reason to be concerned about omitted variables bias. If changes in pollution are correlated with changes in these omitted variables, a fixed-effect approach still gives biased estimates of the implicit prices.

Using data describing housing transactions in California’s Bay Area between 1990 and 2006, we show evidence in support of the hypothesis that the market is efficient, providing support for our rationality assumption. Using our estimator, we recover implicit prices for the three criteria air pollutants described above. In contrast to simple cross-sectional or

fixed-effect estimators, marginal willingnesses to pay for a reduction in all three pollutants (considered individually or together) are all statistically significant, have the expected sign, and are on the high-end of the range of estimates found elsewhere in the literature. Considered together, PM10, SO2, and O3 exhibit house price elasticities of -0.07, -0.16, and -0.60, respectively. We contrast these results with those from a simple fixed-effects model and find that controlling for time-varying unobservables appears to be extremely important for all three pollutants.

We conclude the introduction by emphasizing that our goal in this paper is to provide a method for the unbiased estimation of the hedonic price function and the associated implicit prices described by the hedonic gradient. This has been the primary focus of the applied hedonic literature to this point. As noted by Rosen (1974), however, these implicit prices can be used as inputs in the recovery of consumers' marginal willingness to pay functions. This exercise introduces a host of additional identification issues that have limited the application of this procedure – see, for example, Brown and Rosen (1982), Mendelsohn (1985), Bartik (1987) and Epple (1987). Recent papers have sought to overcome these identification issues (Bajari and Benkard, 2005, Ekeland, Heckman, and Nesheim, 2004, Bishop and Timmins, 2010a, 2010b). In this paper, we focus simply on recovering implicit prices, as is the case in the quasi-experimental hedonic literature.

This paper proceeds as follows. Section 2 describes our estimator of implicit prices in a simple parametric model. We generalize that model in the Appendix. Section 3 describes the data that we use for our application. Section 4 presents results from our model, and compares them to results from traditional cross-sectional and fixed-effects specifications. Section 5 concludes.

2 Model: Estimating Implicit Prices

In this section, we consider the traditional hedonic framework – a model of demand in a differentiated products market in which a consumer maximizes utility. The primary application we have in mind is housing, however, many of the methods we propose could carry over to other differentiated product markets where our assumptions are maintained. We treat the consumer as being forward looking in her decision-making, but unconstrained by adjustment costs; in a model of home buying without adjustment costs, forward-looking

agents maximize utility with respect to current house attributes (so that the model mimics the standard static hedonic framework).

Houses, indexed by $j = 1, \dots, J$, can be completely described by a finite vector of attributes. Let \bar{x}_j denote a 1 by K vector of attributes such as the number of square feet, the lot size, or the year built, all of which are commonly observed by the econometrician and the consumer. In addition, ξ_j denotes a scalar that captures an omitted attribute of the house that is observed by the consumers, but not by the economist. For instance, while data sets on housing are quite detailed, they typically do not report features such as the curb appeal of a home or its state of repair, both of which may be important to buyers. For notational and expositional simplicity, we require these omitted attributes to be captured in a single product attribute, ξ_j , though many of our results allow for a more general error term with vector-valued omitted attributes. To summarize, from the perspective of consumers $i = 1, \dots, I$, product j can be completely summarized by the 1 by $(K + 1)$ vector (\bar{x}_j, ξ_j) .

Equilibrium prices can be written as $p_j = \mathbf{p}(\bar{x}_j, \xi_j)$. We will refer to \mathbf{p} as the hedonic price function. This is a map between the product characteristics (\bar{x}_j, ξ_j) and the price of good j (p_j). The hedonic price function \mathbf{p} is determined in equilibrium by the interactions of buyers and sellers. Bajari and Benkard (2005) show that consumer rationality plus mild restrictions on consumer preferences imply that \mathbf{p} is a function, not a correspondence. As discussed in the introduction, the existence of the function \mathbf{p} is our first key assumption derived from economic theory.

In empirical applications, the economist is generally concerned with estimating $\mathbf{p}(\bar{x}_j, \xi_j)$ using data on the observed prices, p_j and characteristics, \bar{x}_j . Hedonic price regressions are commonly conducted assuming that $E[\xi_j | \bar{x}_j] = 0$, that is, the omitted product attributes are mean independent of the observed attributes. This assumption is frequently criticized in the literature, going back to Small (1975). Returning to our earlier example, suppose that ξ_j reflects the curb appeal of a home. The above moment condition would imply that the expected value of curb appeal is the same for small homes in low income neighborhoods as it is for million dollar homes in exclusive neighborhoods. However, in practice we expect higher values of desirable omitted attributes to be positively correlated with higher values of desirable observed attributes. Thus, failure to correct for this omitted variable would bias upward estimates of implicit prices of desirable attributes.

The only proposed solutions to this problem rely on quasi-random sources of variation

such as breaks in geography (Black, 1999) or discontinuities in the application of regulations (Chay and Greenstone, 2004; Greenstone and Gallagher, 2007). While these are important contributions to the empirical literature, they may face limitations like those discussed in the introduction.

We propose an alternative approach to estimating implicit prices. Begin by considering cases in which there are data on repeat sales so that the price of home j is observed in several time periods among $t = 1, 2, \dots, T$. Note that the price does not need to be observed in all time periods. Our empirical strategy will require as few as two sales for each house. To simplify notation, consider the case where there is a single observed, time-varying characteristic, $x_{j,t}$, which enters linearly into the hedonic price function. All of our results apply in the more general case where this is a vector of characteristics and where the characteristics enter nonparametrically; this situation is described in the Appendix (we consider observed attributes that do not vary over time in Section 2.2). Suppose the system of hedonic pricing equations is:

$$\begin{aligned} \ln(p_{j,1}) &= \alpha_1 + \beta_1 x_{j,1} + \xi_{j,1} \\ &\vdots \\ \ln(p_{j,T}) &= \alpha_T + \beta_T x_{j,T} + \xi_{j,T}. \end{aligned} \tag{1}$$

Since we can observe prices of homes only when they actually transact, we have an unbalanced panel where some of $\ln(p_{jt})$'s are never observed in (1).

In what follows, we assume that agents in the market are uncertain about the evolution of $\xi_{j,t}$. This uncertainty could come from one of two sources. The first is that the omitted characteristics themselves change over time periods. For example, a noisy neighbor may move in next door to home j or an infestation may make it necessary to cut down all the large trees in home j 's neighborhood. The second is that the implicit price of even time-invariant omitted attributes could change over time. Both of these situations would look the same from the point of view of the hedonic model.

In our model, we assume that the omitted product attribute evolves according to a first-order Markov process,¹

$$\xi_{j,t'} = \gamma(t, t')\xi_{j,t} + \eta(j, t, t'). \tag{2}$$

¹We assume that the unconditional mean of the omitted attribute equals to zero, $E[\xi_{j,t}] = 0$ without loss of generality because if not, the intercept in the hedonic pricing equation (1) can subsume the nonzero mean.

This is our second key assumption. Here $\gamma(t, t')\xi_{j,t}$ is the expected value of the omitted attribute at time t' conditional on its value at time t , and $\eta(j, t, t')$ is the stochastic innovation in the omitted attribute. Later in the description of the model we illustrate that this could be, for example, extended to a second-order Markov process if the researcher has access to three repeat sales observations for each house. In our application, we limit our attention to houses that sell twice.

Our third assumption requires that

$$E[\eta(j, t, t')|I_t] = 0. \quad (3)$$

where I_t denotes the information available to the buyer at time t . In words, given all the information available at time t , homebuyers predict that $\xi_{j,t'}$ will equal $\gamma(t, t')\xi_{j,t}$ in expectation. Note that this condition is required if individuals are to be unable to use information in I_t to predict excess appreciation rates for particular houses, which is a necessary condition for full informational efficiency of the housing market as described by Case and Shiller (1989). While we do not require full informational efficiency of the housing market for our estimator, we can use the Case and Shiller (1989) test of informational efficiency as an overly stringent test of our third assumption. We do this in Section 4.1.

2.1 Lagged Prices and Consistent Estimation

We rewrite our hedonic price function for period t' using information from the previous sale of house j (i.e., in period t) to eliminate $\xi_{j,t'}$. In particular, rewriting $\xi_{j,t'}$ as a function of $\xi_{j,t}$ using (2) and substituting $\ln(p_{j,t}) - \alpha_t - \beta_t x_{j,t}$ for $\xi_{j,t}$, we get,

$$\begin{aligned} \ln(p_{j,t'}) &= \alpha_{t'} + \beta_{t'} x_{j,t'} + \xi_{j,t'} & (4) \\ &= \alpha_{t'} + \beta_{t'} x_{j,t'} + \gamma(t, t') [\ln(p_{j,t}) - \alpha_t - \beta_t x_{j,t}] + \eta(j, t, t') \\ &= (\alpha_{t'} - \gamma(t, t')\alpha_t) + \gamma(t, t') \ln(p_{j,t}) \\ &\quad - \gamma(t, t')\beta_t x_{j,t} + \beta_{t'} x_{j,t'} + \eta(j, t, t'). \end{aligned}$$

We note that $x_{j,t'}$ could be correlated with $\eta(j, t, t')$, for example, the innovation in “curb appeal” between t and t' might be correlated with an observable characteristic such as test scores in local public schools. Thus, a regression based on (4) produces inconsistent

estimates of the hedonic price function. For the parametric model of (4), we can use the 2SNLS approach to recover parameters under our maintained assumption of (3). In the first stage we replace $x_{j,t'}$ with its projected value based on $x_{j,t}$ and other observed variables $w_{j,t}$ in the information set I_t using

$$x_{j,t'} = \pi_0(t, t') + \pi_1(t, t')x_{j,t} + \pi_2(t, t')w_{j,t} + v_{j,t,t'}, \quad E[v_{j,t,t'}|I_t] = 0.$$

The first stage uses the assumption that the innovation in the observed attributes is orthogonal to time t information.

We show in the appendix that by exploiting the process that describes the evolution of $x_{j,t}$ over time and using a control function approach, we can still obtain consistent estimates of the key structural parameters even when the characteristics enter nonparametrically in the hedonic pricing equations.

Intuitively, our approach uses the information in lagged prices, $p_{j,t}$ to impute the lagged value of the omitted attribute. For example, if the price for home j is unusually high after controlling for $x_{j,t}$, we would infer that $\xi_{j,t} = \ln(p_{j,t}) - \alpha_t - \beta_t x_{j,t}$ is also large. This is where our first economic assumption – that prices reflect attributes that are observed by consumers – has “bite”. We also assume that the stochastic innovations ($v_{j,t,t'}$) in the observed attributes are orthogonal to current information. Furthermore, after controlling for I_t , the stochastic innovation in the omitted attribute, $\eta(j, t, t')$, is assumed to be mean zero. If these assumptions were not true, it would be possible to earn excess returns in the housing market.

2.2 Measurement Error

Suppose the observed price is measured with error as $\ln(p_{j,t}^*) = \ln(p_{j,t}) + m_{j,t}$ where $p_{j,t}$ denotes the true price and $m_{j,t}$ is the measurement error. Then using the same differencing strategy as in (4), we obtain

$$\begin{aligned} \ln(p_{j,t'}^*) &= \alpha_{t'} + \beta_{t'}x_{j,t'} + \xi_{j,t'} + m_{j,t'} & (5) \\ &= (\alpha_{t'} - \gamma(t, t')\alpha_t) + \gamma(t, t')\ln(p_{j,t}^*) \\ &\quad - \gamma(t, t')\beta_t x_{j,t} + \beta_{t'}x_{j,t'} + \eta(j, t, t') + m_{j,t'} - \gamma(t, t')m_{j,t}. \end{aligned}$$

Then the error contains three terms: (1) the stochastic innovation in the omitted attribute, (2) the measurement error, and (3) the lagged measurement error. The differencing approach exacerbates the measurement error problem in two ways. The first, and more serious problem, is that the 2SNLS estimator is no longer consistent because $\ln(p_{j,t}^*)$ is correlated with the error due to the term $m_{j,t}$. Second, the lagged measurement error is added to the composite error, so the variance of the error term is increased.

Assuming that measurement error is not serially correlated, the endogeneity problem of the lagged measurement error can be resolved if we have more than two observations of sales by using the moment condition $E[\eta(j, t, t') + m_{j,t'} - \gamma(t, t')m_{j,t} | I_{\tilde{t}}, x_{j,t}] = 0$ with $\tilde{t} < t$ instead. Intuitively, in the first stage we replace $x_{j,t'}$ and $\ln(p_{j,t}^*)$ with their projected values based on $x_{j,t}$ and other observed variables in the information set from the prior period $I_{\tilde{t}}$.

Further suppose the house characteristic $x_{j,t}$ is also measured with error as $x_{j,t}^* = x_{j,t} + m_{j,t}^x$ for each t where $x_{j,t}$ denotes the true characteristic and $m_{j,t}^x$ is the measurement error. Then (5) becomes

$$\begin{aligned} \ln(p_{j,t'}^*) &= \alpha_{t'} + \beta_{t'} x_{j,t'} + \xi_{j,t'} + m_{j,t'} \\ &= (\alpha_{t'} - \gamma(t, t')\alpha_t) + \gamma(t, t') \ln(p_{j,t}^*) \\ &\quad - \gamma(t, t')\beta_t x_{j,t}^* + \beta_{t'} x_{j,t'}^* + \eta(j, t, t') + m_{j,t'} - \gamma(t, t')m_{j,t} - \beta_{t'} m_{j,t'}^x + \gamma(t, t')\beta_t m_{j,t}^x \end{aligned}$$

and the problem is further exacerbated because both $x_{j,t'}^*$ and $x_{j,t}^*$ are correlated with the composite error due to the measurement errors (2SNLS is no longer consistent) and the variance of the composite error term is increased (more terms in the error). Again assuming that measurement error is not serially correlated, the endogeneity problem can be resolved if we have more than three observations of sales by using the moment condition $E[\eta(j, t, t') + m_{j,t'} - \gamma(t, t')m_{j,t} - \beta_{t'} m_{j,t'}^x + \gamma(t, t')\beta_t m_{j,t}^x | I_{\tilde{t}}] = 0$ with $\tilde{t} < t$ where $I_{\tilde{t}}$ includes information from at least two transactions prior to t .

This method of correcting for classical measurement errors requires three or more observations of sales for a given house, which only exists in a very small subset of the data in our application. If the time invariant house attributes are also measured with errors, the method of correcting for measurement errors using lagged variables will not work and we will need instruments that are correlated with those attributes but not with errors. Obviously finding such instruments will be difficult or infeasible in our application.

While ideally it would be useful to correct for measurement error(s) in applications in this way where it is feasible, we proceed assuming the price and other house characteristics are measured without error in the paper. Note that our data come from the Dataquick Information Corporation, a real estate data aggregator that assembles official information collected by local governments in support of housing transactions. While measurement error may generally be an issue in housing applications, we contend that the problem will be minimized in our application (in particular, compared with applications that use census data, where owners self-report a home value and all housing attributes).

2.3 Time-Invariant Covariates and Model Restrictions

When houses have only time-invariant attributes, some of the parameters of the model described above are not identified without further restrictions. To see this, replace $x_{j,t}$ with the time-invariant covariate z_j in (4):

$$\ln(p_{j,t'}) = (\alpha_{t'} - \gamma(t, t')\alpha_t) + \gamma(t, t') \ln(p_{j,t}) - \gamma(t, t')\beta_t z_j + \beta_{t'} z_j + \eta(j, t, t').$$

We cannot therefore identify $\beta_{t'}$ separately from β_t , i.e., a multicollinearity problem. Adding the restrictions that $\alpha_t = \alpha_0$ and $\beta_t = \beta_0$ for all t , the above equation becomes

$$\ln(p_{j,t'}) = \alpha_0 (1 - \gamma(t, t')) + \gamma(t, t') \ln(p_{j,t}) + (1 - \gamma(t, t'))\beta_0 z_j + \eta(j, t, t').$$

With these restrictions, we can identify $\gamma(t, t')$ from the coefficient on $\ln(p_{j,t})$, α_0 from the constant term, and β_0 from the coefficient on z_j . An alternative approach is to normalize $\beta_1 = 1$. Then β_t , $t > 1$ is identified recursively up to this normalization using the fact that $-\gamma(t, t')\beta_t + \beta_{t'}$ can be recovered in each period.

Imposing some structure on $\gamma(t, t')$ yields a set of over-identifying restrictions. For example, we can let $\gamma(t, t') = \gamma(t, \tilde{t})\gamma(\tilde{t}, t')$ for \tilde{t} between t and t' . Even with time-varying x 's, the above restrictions may be useful to obtain more stable and robust estimates when the variation over time is relatively small.

2.4 Simple Parametric Model with Two Transactions

It will often be the case that only two transactions per house are available in the data (as in our application). Thus, we focus on the two-transaction setting as a straightforward illustration of how our estimator can be applied in many contexts. We describe the generalization in the Appendix.

Assuming that the implicit prices are time-invariant $\beta_t = \beta$ for all t , the model simplifies to

$$\begin{aligned} \ln(p_{j,t_b}) &= \alpha_{t_b} - \gamma(t_a, t_b)\alpha_{t_a} + \gamma(t_a, t_b) \ln(p_{j,t_a}) + \gamma(t_a, t_b)\beta x_{j,t_a} \\ &\quad + \beta x_{j,t_b} + \eta(j, t_a, t_b). \end{aligned} \quad (6)$$

where t_a denotes the time period of the first sale and t_b denotes the time period of the second sale. The estimation can proceed as a simple application of the two-stage nonlinear least squares (2SNLS). We can rewrite (6) as

$$\begin{aligned} x_{j,t_b} &= \pi_0(t_a, t_b) + \pi_1(t_a, t_b)x_{j,t_a} + \pi_2(t_a, t_b)w_{j,t_a} + v_{j,t_a,t_b} \\ \ln(p_{j,t_b}) &= \alpha_{t_b} - \gamma(t_a, t_b)\alpha_{t_a} + \gamma(t_a, t_b) \ln(p_{j,t_a}) + \gamma(t_a, t_b)\beta x_{j,t_a} \\ &\quad + \beta (\pi_0(t_a, t_b) + \pi_1(t_a, t_b)x_{j,t_a} + \pi_2(t_a, t_b)w_{j,t_a}) + u_{j,t_a,t_b} \end{aligned} \quad (7)$$

where $u_{j,t_a,t_b} = \beta v_{j,t_a,t_b} + \eta(j, t_a, t_b)$. $E[u_{j,t_a,t_b}|I_{t_a}] = 0$ because v_{j,t_a,t_b} is the projection error in the first step and $E[\eta(j, t_a, t_b)|I_{t_a}] = 0$ by assumption (3).

In (7) $w_{j,t}$ denotes other observable variables in I_t , including p_{j,t_a} . Importantly, we do not need $w_{j,t}$ to include any additional information. In other words, exclusion restrictions are not needed to identify the key structural parameters ($\gamma(t_a, t_b)$, β , and $\alpha_{t_b} - \gamma(t_a, t_b)\alpha_{t_a}$), as can be seen in (7). Our key identification condition is $E[u_{j,t_a,t_b}|I_{t_a}] = 0$. This condition differs considerably from the standard approach to estimating hedonic models, which rely instead on $E[\xi_{j,t}|x_{j,t}] = 0$, i.e., that omitted attributes are conditionally mean independent of observed attributes, as discussed above.

We let $\theta(t_a, t_b) = (\alpha_{t_b} - \gamma(t_a, t_b)\alpha_{t_a}, \gamma(t_a, t_b), \beta)'$. The coefficients in the second step equation in (7) are nonlinear functions of $\theta(t_a, t_b)$ where $\pi_0(t_a, t_b)$, $\pi_1(t_a, t_b)$, and $\pi_2(t_a, t_b)$

are obtained in the first step. We obtain estimates by solving

$$\begin{aligned}\widehat{\pi}(t_a, t_b) &= \operatorname{argmin}_{\pi(t_a, t_b)} \sum_{j=1}^J \{x_{j,t_b} - \pi_0(t_a, t_b) - \pi_1(t_a, t_b)x_{j,t_a} - \pi_2(t_a, t_b)w_{j,t_a}\}^2 \\ \widehat{\theta}(t_a, t_b) &= \operatorname{argmin}_{\theta(t_a, t_b)} \sum_{j=1}^J \{\ln(p_{j,t_b}) - g(\ln(p_{j,t_a}), x_{j,t_a}, \widehat{x}_{j,t_b}; \theta(t_a, t_b))\}^2\end{aligned}$$

where $\widehat{x}_{j,t_b} = \widehat{\pi}_0(t_a, t_b) + \widehat{\pi}_1(t_a, t_b)x_{j,t_a} + \widehat{\pi}_2(t_a, t_b)w_{j,t_a}$ and

$$\begin{aligned}g(\ln(p_{j,t_a}), x_{j,t_a}, \widehat{x}_{j,t_b}; \theta(t_a, t_b)) &= \alpha_{t_b} - \gamma(t_a, t_b)\alpha_{t_a} + \gamma(t_a, t_b)\ln(p_{j,t_a}) \\ &\quad + \gamma(t_a, t_b)\beta x_{j,t_a} + \beta\widehat{x}_{j,t_b}.\end{aligned}$$

We can also impose some parametric restrictions on $\pi(t_a, t_b)$'s and $\gamma(t_a, t_b)$'s in the above.

Note that the first step estimation contributes to the asymptotic variance of the second step estimators. We obtain correct standard errors by applying Murphy and Topel (1985). Denote

$$\begin{aligned}\sqrt{J}(\widehat{\pi}(t_a, t_b) - \pi(t_a, t_b)) &\rightarrow_d N(0, V(t_a, t_b)) \\ G(\cdot; \theta(t_a, t_b)) &= \frac{\partial}{\partial \theta(t_a, t_b)} g(\cdot; \theta(t_a, t_b)) \\ \Omega_0(t_a, t_b) &= E[\eta^2(\cdot, t_a, t_b)G(\cdot; \theta(t_a, t_b))G(\cdot; \theta(t_a, t_b))'] \\ Q_0(t_a, t_b) &= E[G(\cdot; \theta(t_a, t_b))G(\cdot; \theta(t_a, t_b))'] \\ Q_1(t_a, t_b) &= E[G(\cdot; \theta(t_a, t_b))\beta(1, x_{j,t_a}, w_{j,t_a})]\end{aligned} \tag{8}$$

Then, we have

$$\sqrt{J}(\widehat{\theta}(t_a, t_b) - \theta(t_a, t_b)) \rightarrow_d N(0, \Sigma(t_a, t_b))$$

where

$$\Sigma(t_a, t_b) = Q_0(t_a, t_b)^{-1} [\Omega_0(t_a, t_b) + Q_1(t_a, t_b)V(t_a, t_b)Q_1(t_a, t_b)'] Q_0(t_a, t_b)^{-1}.$$

We obtain a consistent estimate of the heteroskedasticity robust variance matrix $\Sigma(t_a, t_b)$ using the sample counterparts of (8) below. The variance can also be clustered using a

standard method instead.

$$\begin{aligned}
\widehat{V}(t_a, t_b) &= \left(\frac{1}{J} \sum_{j=1}^J (1, x_{j,t_a}, w_{j,t_a})' (1, x_{j,t_a}, w_{j,t_a}) \right)^{-1} \left(\frac{1}{J} \sum_{j=1}^J \widehat{v}_{j,t_a,t_b}^2 (1, x_{j,t_a}, w_{j,t_a})' (1, x_{j,t_a}, w_{j,t_a}) \right) \\
&\quad \times \left(\frac{1}{J} \sum_{j=1}^J (1, x_{j,t_a}, w_{j,t_a})' (1, x_{j,t_a}, w_{j,t_a}) \right)^{-1}, \\
\widehat{\Omega}_0(t_a, t_b) &= \frac{1}{J} \sum_{j=1}^J \widehat{\eta}^2(j, t_a, t_b) G(\cdot; \widehat{\theta}(t_a, t_b)) G(\cdot; \widehat{\theta}(t_a, t_b))', \\
\widehat{\eta}(j, t_a, t_b) &= \ln(p_{j,t_b}) - g(\ln(p_{j,t_a}), x_{j,t_a}, x_{j,t_b}; \widehat{\theta}(t_a, t_b)), \\
\widehat{Q}_0(t_a, t_b) &= \frac{1}{J} \sum_{j=1}^J G(\cdot; \widehat{\theta}(t_a, t_b)) G(\cdot; \widehat{\theta}(t_a, t_b))', \\
\widehat{Q}_1(t_a, t_b) &= \frac{1}{J} \sum_{j=1}^J G(\cdot; \widehat{\theta}(t_a, t_b)) \widehat{\beta}(1, x_{j,t_a}, w_{j,t_a}), \text{ and} \\
\widehat{\Sigma}(t_a, t_b) &= \widehat{Q}_0(t_a, t_b)^{-1} \left[\widehat{\Omega}_0(t_a, t_b) + \widehat{Q}_1(t_a, t_b) \widehat{V}(t_a, t_b) \widehat{Q}_1(t_a, t_b)' \right] \widehat{Q}_0(t_a, t_b)^{-1}.
\end{aligned}$$

3 Data

We demonstrate the role of efficient housing markets in controlling for time-varying, correlated unobservables by measuring the marginal willingness to pay to avoid exposure to three of the EPA’s “criteria” air pollutants – particulate matter (PM10), sulfur dioxide (SO2), and ground-level ozone (O3).² Without extremely detailed data describing the evolution of neighborhood attributes, correlated unobservables are likely to play an important role in such an application.

We consider housing transactions from California’s Bay Area (specifically, Alameda, Contra Costa, Marin, San Francisco, San Mateo, and Santa Clara counties) over the period 1990-2006. These data were purchased from the DataQuick Corporation and contain information describing the universe of housing transactions (i.e., buyers’, sellers’ and lenders’ names, dates, loan amounts, and transaction prices) and the houses that transacted (i.e., square footage, lot size, year built, number of rooms, and how many of those rooms are bedrooms or bathrooms). Important for our purposes, the data also provide the exact street address

²The list of criteria pollutants also includes nitrogen oxides, lead, and carbon monoxide. This list forms the basis for the EPA’s primary (health) and secondary (environmental and aesthetic) emissions reduction targets. Of the six criteria pollutants, particulate matter and ground-level ozone are commonly considered to pose the greatest health threat. (<http://www.epa.gov/air/urbanair>)

of each home, with which we can impute pollution measures using data from thirty-seven monitors located throughout the Bay Area.

3.1 Housing Data

DataQuick reports a house’s attributes as they were measured at the time of the last sale entered in our data. Because houses may have been altered (either improved or suffered some severe damage), these attributes may not be applicable to all observed transactions. We therefore carry-out a number of data cuts to avoid this problem. First, we drop all houses that are reported to have experienced major improvements. Next, we consider the appreciation rate exhibited by each house over each pair of sales that we observe in the data. From this, we deduct the average appreciation rate for all houses that sold in the same pair of years. We then drop the houses in the top and bottom 10% of the resulting distribution of normalized appreciation rates. As such, we eliminate any house that appreciated or depreciated at a very high rate relative to other houses on the market at the same time. While we admit that this may, to some extent, ‘stack the deck’ in favor of our finding evidence in support of efficient housing markets, we need to do this because we only observe house attributes at the last sale, but must treat attributes as fixed over time. Attributes are less likely to be fixed for houses that exhibit large changes in price; these houses are likely to have experienced some sort of structural change.

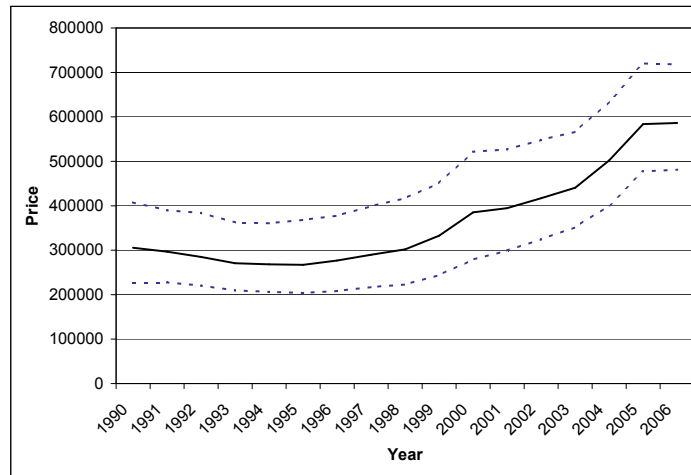
Second, we drop problematic observations – for example, all observations where “year built” is missing, or where “year built” comes after the transaction date (signaling a purchase of land on which a house was then constructed). We also drop all properties that fail to report a transaction price or a latitude and longitude (or the latitude and longitude changes), houses with outlier attributes, and all observations with housing attributes that appear to be coded with error – in particular, houses where the number of bedrooms or bathrooms is greater than five. We also drop any house more than 5,000 square feet in size, or which sits on more than a 70,000 square-foot lot. We finally drop all homes that sell more than two times in the seventeen year period we are considering. This is primarily for the sake of convenience, as it allows us to implement our estimator using the simple specification described in Section 2.4. In the end, these cuts leave us with data describing repeat transactions for 93,321 unique housing units. Table 1 summarizes the attributes of these houses.

Table 1: House Attributes (N=93,321)

	Mean	Std Dev	Minimum	Maximum
Lot size	6,884	5,429	1,000	69,809
Square feet	1701	617	500	5,000
No. of bathrooms	2.103	0.688	1	5
No. of bedrooms	3.231	0.829	1	5
No. of rooms	7.036	1.818	1	15
Year built	1966	22.66	1868	2005

Figure 1 describes the median transaction price in each year of our data. This makes clear that there were periods of (slow) depreciation and (rapid) appreciation in the Bay Area over the period we are considering.

Figure 1: Median Transaction Price by Year With 25th and 75th Percentiles



3.2 Air Quality Data

We measure individuals' average marginal willingness-to-pay (MWTP) to avoid three of the EPA's major criteria air pollutants.³ The MWTP is a key determinant of the benefits of any new air pollution regulation, such as the Clean Air Act Amendments of 1990 that allowed for trading in permits to emit sulfur dioxide. The other main source of value from a new air pollution regulation comes from avoided mortality; this is typically measured by ascribing the value of a statistical life (VSL) to each death avoided by the policy.

We first consider PM10, which denotes particles less than ten micrometers in diameter. These particles (especially those smaller than 2.5 micrometers) can travel deep into the lungs and even into the bloodstream. This can lead to a variety of health problems, including asthma, chronic bronchitis, and heart attack.⁴ Fine particles also reduce visibility, and prolonged exposure to PM10 can damage structures and stain building materials. While not necessarily as important as health effects from a welfare perspective, these aesthetic effects may have a marked impact on housing prices. We consider the average annual PM10 concentration, which is measured in micrograms per cubic meter ($\mu\text{g}/\text{m}^3$). PM10 concentration at each house is imputed with an inverse-squared-distance weighted average of the concentrations measured at each of the thirty-seven monitoring stations in the Bay Area.

Our second pollutant is sulfur dioxide (SO₂). The primary health consequences of sulfur dioxide come in the form of breathing difficulties, especially for those who suffer from asthma. Like PM10, SO₂ can create haze that impairs visibility. Acid rain (or acid fog), which is produced when SO₂ reacts with water and other chemicals in the air, will damage building materials and kill vegetation. SO₂ (and the remainder of our pollutants) is measured in parts per million (ppm), and we use the maximum one-hour observation observed over the course of the year at each monitor (imputed for each house again using an inverse-squared-distance weighted average of all monitors' observations). The maximum one-hour observation is an important figure used by the California Air Resources Board in determining whether or not

³Information on the health and aesthetic costs of each of the pollutants discussed in this section can be found at the EPA's web-site (<http://www.epa.gov/air/>).

⁴The Harvard "Six City" Study (Dockery et al., 1993) established many of these effects, which have been confirmed by numerous studies since that time. Lin et al., 2002; Norris et al., 1999; Slaughter et al., 2003; and Tolbert et al., 2000) have demonstrated detrimental effects, particularly for the young and elderly suffering from asthma. Hong et al., 2002; Tsai et al., 2003, and D'Ippoliti et al., 2003 provide evidence of increased risk of heart attack and stroke. Ghio et al. 2000 finds evidence of lung tissue inflammation, while Pope et al., 2002 finds increased risk of lung cancer. More recently, Samet et al., 2004 has found evidence of increased risk of heritable diseases from exposure to fine particulates.

an air district is in compliance with state regulations.

Third, we consider ground-level ozone (O3). Similar to smog, ozone can cause a variety of severe respiratory problems including coughing, wheezing, breathing pain, aggravated asthma, and increased susceptibility to bronchitis. Exposure to peak concentrations of ground-level ozone can have acute effects, and repeated exposure to even moderate levels can lead to permanent lung damage. In addition to its health consequences, O3 has detrimental impacts on the growth of vegetation (particularly trees and other plants in urban settings), which can have important aesthetic consequences for housing prices.

Figure 2 describes the time path of three pollution measures (along with nitrogen oxides) over the sample period. To make the numbers more easily interpretable on the same graph, we express PM10 pollution in $(\mu\text{g}/\text{m}^3) \cdot (1/1000)$.

Figure 2: Median One Hour Maximum Pollution Concentrations

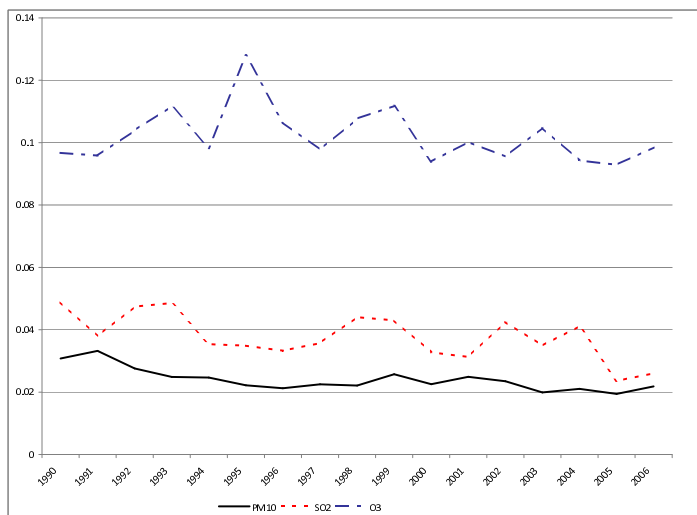


Table 2 describes the correlations across all three pollutants observed at the time of every transaction in our sample. While PM10 and SO2 are fairly highly correlated, O3 has a much lower correlation with SO2 and is even negatively correlated with PM10. This is consistent with the intuition that O3, which is formed in a complicated photochemical process, is more affected by weather patterns and is not restricted to the more polluted places. Although collinearity may be an issue for separately identifying the effects of PM10 and SO2, we are still able to estimate a model with all three pollutants appearing simultaneously. In

addition, we measure the MWTP for each pollutant considered individually.

Table 2: Correlations of Pollutants

	PM10	SO2	O3
PM10	1.000		
SO2	0.516	1.000	
O3	-0.097	0.217	1.000

A final feature of these pollutants that we do not address is the fact that the disutility from each may be a complicated nonlinear function of the concentrations of all the other pollutants. This is a result of the photochemical processes through which they interact. See Muller, Tong, and Mendelsohn (2009) for an example of research that considers these interactions.

4 Results

4.1 Evidence in Support of our Identification Assumptions

Before applying our estimation strategy, we provide empirical evidence in support of our identifying assumptions on homebuyer rationality, a limited form of the efficient housing market assumption, by approximating Case and Shiller’s (1989) test of full informational efficiency for the subset of houses that sold three times over our sample period. In particular, we check to see whether or not observable attributes in the information set at time t have any explanatory power for price changes after that time. This is a stronger test than we require. In particular, our estimator allows for predictable changes in observables and unobservables based on their current values to influence price changes. However, if we find that the predictive power of current observables is weak, statistically and/or economically, it bolsters our assumption that $E[\eta(j, t, t')|I_t] = 0$.

Begin by letting t_a , t_b , and t_c denote the times at which sales are observed for houses in our data set that transact three times, $t_a < t_b < t_c$. a , b , and c can be different for each of these houses. We regress the annualized return, $\frac{\ln(p_{j,t_c}) - \ln(p_{j,t_b})}{t_c - t_b}$ on the average return of previous sales (which is allowed to differ across counties and years), housing attributes,

pollutants, and county fixed effects.⁵ While many of the coefficients are statistically significant, their economic magnitudes in terms of marginal returns are negligible. Overall, the total explanatory power is quite small, with an R^2 of only 0.036. Table 3 summarizes the results.⁶

Table 3: Efficient Market Hypothesis(N=16656)

A. Regression of Annualized Return on House Attributes and Average Return ($R^2 = 0.036$)								
	Avg. ret.	Lotsize	Sqft	# Bath	# Bed	PM10	SO2	O3
Coeff	0.0205	0.0000	-0.0000	-0.0024	0.0077	-0.0040	1.0785	0.2496
	[3.45]	[1.28]	[-8.61]	[-1.21]	[4.48]	[-6.89]	[6.06]	[3.10]
B. Excess Returns Per Year in Dollar Amounts								
Purchase Price	Avg. ret.	Lotsize	Sqft	# Bath	# Bed	PM10	SO2	O3
	10% ↑	100 sf ↑	100 sf ↑	1 ↑	1 ↑	2 $\mu\text{g}/\text{m}^3$ ↑	5 ppb ↑	10 ppb ↑
0.4M	820	13	-916	-953	3,080	-3,241	2,157	998
1M	2,050	33	-2,290	-2,382	7,699	-8,103	5,393	2,495
2M	4,100	66	-4,580	-4,764	15,398	-16,206	10,785	4,991

t-statistics are reported in brackets and calculated from clustered robust standard errors, clustered by tract. The dollar amounts in panel B are calculated from the estimates in panel A, based on purchasing prices of .4M,1M and 2M respectively for the given change in the observed attribute.

To give a sense of the extent to which knowing housing attributes can help to generate excess returns, we calculate dollar amounts of excess returns from the estimation results in Panel A of Table 3 under the scenario that the previous average return is higher by 10 percent, the lot size is larger by 100 square feet, the home size is larger by 100 square feet, the number of bathrooms is larger by 1, the number of bedrooms is larger by 1, and the three pollutant measures PM10, SO2, and O3 are higher by 2 $\mu\text{g}/\text{m}^3$, 5 ppb, 10 ppb, respectively. We also assume that the home prices at the time of purchase are 0.4 million, 1 million, and 2 million dollars, respectively. The results are reported in Panel B of Table 3. We see that the amounts are small compared to the home prices (i.e., less than 1%). For example, when the home price is 1 million at the time of purchase, one would make excess returns of 33 dollars per year by purchasing a home with its lot size larger by 100 square feet and would earn

⁵The average return from previous sales is obtained as the fitted values from the regression of $\ln(p_{j,t_b}) - \ln(p_{j,t_a})$ on dummies indicating the year of the first and the second sales and the county in which the house is located.

⁶We also consider a more flexible estimator that includes a fourth order polynomial in observable attributes. The conclusions are similar.

7,699 dollars of additional annual returns by purchasing a home with five bedrooms instead of four bedrooms. One can interpret calculations of excess returns in other cases similarly.

While not a formal test of our identification assumption, these results suggest that information available at time t cannot be used to affect, in an economically significant way, the returns derived from a house purchase decision. This implies that information available at that time is not particularly useful in predicting $\eta(j, t, t')$; hence, supporting our assumption that $E[\eta(j, t, t')|I_t] = 0$. In general, this strategy provides a way for researchers to determine if our approach is appropriate for their particular data set and application.

4.2 The Marginal Willingness-to-Pay to Avoid Air Pollution

In our application, we allow Bay Area housing prices to be determined by different hedonic price functions in each of three separate periods: (1) 1990-1994, (2) 1995-2000, and (3) 2001-2006.⁷ These periods correspond (roughly) to periods of depreciation, appreciation, and very rapid appreciation in this housing market, as shown in Figure 1. They also correspond to periods of changing enforcement of air pollution standards, as the Bay Area moved from non-attainment to attainment and back to non-attainment with respect to federal ozone regulations.

We report results for three different econometric models. First, we estimate a simple cross-sectional model for each of the three time periods in our data set. This approach does nothing to control for omitted attributes (time-varying or time-invariant) that may be correlated with pollution:

Cross-Sectional Model

$$\ln(p_{j,1}) = \alpha_1 + x'_{j,1}\beta_1 + z'_j\phi_1 + \xi_{j,1},$$

$$\ln(p_{j,2}) = \alpha_2 + x'_{j,2}\beta_2 + z'_j\phi_2 + \xi_{j,2},$$

$$\ln(p_{j,3}) = \alpha_3 + x'_{j,3}\beta_3 + z'_j\phi_3 + \xi_{j,3},$$

where the subscripts $\{1, 2, 3\}$ correspond to each of the three time periods, $x'_t \equiv \{PM10, SO2, O3\}$ captures the pollutants and z includes the housing attributes described in Table 1 and a vec-

⁷We divide the data into three periods primarily for tractability, in order to keep the number of parameters we need to estimate down to a reasonable size.

tor of county fixed effects.

Second, we estimate a house fixed-effect model that uses the panel aspect of the data to control for time-invariant omitted attributes that could potentially be correlated with pollution. We constrain the derivative of $\ln(P)$ with respect to each pollutant to be constant over time. This constraint allows us to recover an implicit price for pollution from the fixed-effect specification. We allow the marginal effects of other housing attributes to vary over time, meaning that we can only recover the change in the implicit prices of these attributes.

House Fixed-Effect Model

$$\begin{aligned}\ln(p_{j,3}) - \ln(p_{j,2}) &= \rho_{2,3} + (x_{j,3} - x_{j,2})'\beta + z'_j\chi_{2,3} + u_{j,2,3}, \\ \ln(p_{j,3}) - \ln(p_{j,1}) &= \rho_{1,3} + (x_{j,3} - x_{j,1})'\beta + z'_j\chi_{1,3} + u_{j,1,3}, \\ \ln(p_{j,2}) - \ln(p_{j,1}) &= \rho_{1,2} + (x_{j,2} - x_{j,1})'\beta + z'_j\chi_{1,2} + u_{j,1,2}.\end{aligned}$$

where $\rho_{t,t'} = (\alpha_{t'} - \alpha_t)$ and $\chi_{t,t'} = (\phi_{t'} - \phi_t)$.

Finally, we estimate a constrained specification of the model described in equation (7). We restrict $\beta_1 = \beta_2 = \beta_3 = \beta$. Constraining the marginal effect of pollution on price to be constant over time assists with model identification and makes the results more directly comparable to those of the house fixed-effect model, where this assumption is required.⁸ $\psi_{t,t'}$ replaces the intercept in equation (4), $\psi_{t',t} = \alpha_{t'} - \gamma(t, t')\alpha_t$, and $z'_j\delta_{t,t'}$ controls flexibly for any attributes that do not vary over time.⁹ This implies the following specification:

Efficient Housing Market Model

$$\begin{aligned}\ln(p_{j,3}) &= \psi_{2,3} + \gamma(2, 3)\ln(p_{j,2}) - x'_{j,2}\gamma(2, 3)\beta + x'_{j,3}\beta + z'_j\delta_{2,3} + \eta_{j,2,3}, \\ \ln(p_{j,3}) &= \psi_{1,3} + \gamma(1, 3)\ln(p_{j,1}) - x'_{j,1}\gamma(1, 3)\beta + x'_{j,3}\beta + z'_j\delta_{1,3} + \eta_{j,1,3}, \\ \ln(p_{j,2}) &= \psi_{1,2} + \gamma(1, 2)\ln(p_{j,1}) - x'_{j,1}\gamma(1, 2)\beta + x'_{j,2}\beta + z'_j\delta_{1,2} + \eta_{j,1,2}.\end{aligned}$$

Depending upon the time periods a particular house sells, one of these three equations will apply to it. The first equation applies when $t = 2$ and $t' = 3$, the second equation applies

⁸We perform a robustness check and find that this constraint is important for identification in our context; without it, results are unstable across years and sensitive to the chosen specification.

⁹In particular, if $z'_j\phi_t$ represents the contribution of time-invariant attributes z_j to $\ln p_{j,t}$, then $z'_j\delta_{t,t'} = z'_j(\phi_{t'} - \gamma(t, t')\phi_t)$. For convenience, we label $\delta_{t,t'} = \phi_{t'} - \gamma(t, t')\phi_t$.

when $t = 1$ and $t' = 3$, and the third equation applies when $t = 1$ and $t' = 2$.

Given the linearity of the hedonic pricing equation, we implement a 2SNLS approach to deal with the endogeneity of $x_{t'}$, as described in Section 2.1. In particular, we first estimate the following regression equations using all the exogenous and predetermined variables as instruments:

$$\begin{aligned} x_{j,3} &= \Pi_0 Year_j + \Pi_1(3,2)x_{j,2} + \Pi_2(3,2)\ln(p_{j,2}) + \Pi_3(3,2)z_j + \Pi_4(3,2)County_j + v_{j,2,3}, \\ x_{j,3} &= \Pi_0 Year_j + \Pi_1(3,1)x_{j,1} + \Pi_2(3,1)\ln(p_{j,1}) + \Pi_3(3,1)z_j + \Pi_4(3,1)County_j + v_{j,1,3}, \\ x_{j,2} &= \Pi_0 Year_j + \Pi_1(2,1)x_{j,1} + \Pi_2(2,1)\ln(p_{j,1}) + \Pi_3(2,1)z_j + \Pi_4(2,1)County_j + v_{j,1,2}. \end{aligned}$$

where *County* denotes a vector of county dummies and *Year* denotes a vector of year dummies indicating t' . The first equation applies to houses that sell in periods $t = 2$ and $t' = 3$, the second equation applies to houses that sell in periods $t = 1$ and $t' = 3$, and the third equation applies to houses that sell in periods $t = 1$ and $t' = 2$. We then use the two sets of equations described above and estimate using 2SNLS as described in Section 2.4.¹⁰

Table 4 reports the results of a cross-sectional specification that considers all three pollutants simultaneously (panel A), along with specifications that consider each pollutant individually (panel B). For many pollutant-year combinations, MWTP exhibits the counterintuitive (i.e., positive) sign. Only the implicit price of O3 consistently has the expected negative sign. Moreover, for every pollutant, results are unstable across years. These results suggest the presence of (possibly time-varying) omitted attributes that are correlated with the pollutants we are studying.

A straightforward way to control for time-invariant unobservables, which may be confounding the cross-sectional estimates, is to estimate the fixed-effect model as described above. Table 5 describes these results, first including all pollutants and then for each pollutant in a separate regression. MWTP estimates for SO2 and O3 are stable across the two specifications (including all pollutants or not), have the expected sign, and are small but not unreasonable in magnitude. For instance, the results including all pollutants suggest that at the median housing price of \$417,800, a homebuyer would be willing to pay \$54 to avoid a 1ppb reduction in SO2 and \$90 for a similar reduction in O3. MWTP for PM10, however, has a counterintuitive sign, suggesting the presence of some sort of omitted attribute that

¹⁰Note that other observable attributes of the house (in z_j) are time-invariant because we only observe them at the time of last purchase, so we do this correction only for the pollutants.

Table 4: Implicit Price of Pollution: Cross-Sectional Estimates

	Period 1			Period 2			Period 3		
	Coeff	Elast. [§]	WTP [‡]	Coeff	Elast. [§]	WTP [‡]	Coeff	Elast. [§]	WTP [‡]
A. Regressions Controlling for All Pollutants									
PM10 ($\mu\text{g}/\text{m}^3$)	0.0101 [0.0008]	0.2390 [0.0183]	295.11 [22.58]	0.0200 [0.0028]	0.4726 [0.0662]	583.59 [81.72]	-0.0371 [0.0015]	-0.8798 [0.0363]	-1086.48 [44.86]
SO2 (ppm)	1.0304 [0.4985]	0.0378 [0.0183]	30.14 [14.58]	0.9270 [0.9124]	0.0340 [0.0335]	27.11 [26.68]	-4.4084 [0.3264]	-0.1617 [0.0120]	-128.93 [9.54]
O3 (ppm)	-1.3176 [0.4152]	-0.1362 [0.0429]	-38.54 [12.14]	-2.6934 [0.2786]	-0.2784 [0.0288]	-78.77 [8.15]	-2.6262 [0.3320]	-0.2715 [0.0343]	-76.81 [9.71]
B. Separate Regression for Each Pollutant									
PM10 ($\mu\text{g}/\text{m}^3$)	0.0112 [0.0006]	0.2656 [0.0147]	327.99 [18.11]	0.0222 [0.0022]	0.5268 [0.0524]	650.53 [64.65]	-0.0417 [0.0017]	-0.9868 [0.0403]	-1218.63 [49.82]
SO2 (ppm)	0.5543 [0.4596]	0.0203 [0.0169]	16.21 [13.44]	2.5803 [0.6764]	0.0947 [0.0248]	75.46 [19.78]	-5.5051 [0.3077]	-0.2020 [0.0113]	-161.00 [9.00]
O3 (ppm)	-1.8572 [0.3547]	-0.1920 [0.0367]	-54.31 [10.37]	-2.7917 [0.2752]	-0.2886 [0.0284]	-81.65 [8.05]	-3.4350 [0.4095]	-0.3551 [0.0423]	-100.46 [11.98]
N	45,041			70,075			71,526		

Standard errors clustered at the tract level reported in brackets. Controls for lot size, square feet, number of rooms, number of bedrooms, number of bathrooms, year built and county fixed effects also included but not reported. [§] Elasticities calculated at medians of pollutants, which are 23.68 for PM10, 0.0367 for SO2, 0.1034 for O3. [‡] Willingness to pay calculated for marginal 1 $\mu\text{g}/\text{m}^3$ change in PM10 and 1 ppb change in other pollutants, annualized at rate of 0.07 for median house price of \$ 417,800.

Table 5: Implicit Price of Pollution: Fixed-Effect Estimates (N=72,059)

	All Pollutants [†]			Single Pollutant [†]		
	Coeff	Elast. [§]	WTP [‡]	Coeff	Elast. [§]	WTP [‡]
PM10 ($\mu\text{g}/\text{m}^3$)	0.0049 [0.0005]	0.1151 [0.0126]	142.19 [15.52]	0.0058 [0.0006]	0.1367 [0.0138]	168.79 [17.07]
SO2 (ppm)	-1.8543 [0.1776]	-0.0680 [0.0065]	-54.23 [5.19]	-2.0447 [0.1773]	-0.0750 [0.0065]	-59.80 [5.19]
O3 (ppm)	-3.0620 [0.0960]	-0.3165 [0.0099]	-89.55 [2.81]	-3.3510 [0.0983]	-0.3464 [0.0102]	-98.00 [2.87]

Standard errors clustered at the tract level reported in brackets. Controls for lot size, square feet, number of rooms, number of bedrooms, number of bathrooms, year built and county fixed effects also included but not reported. [†] The single pollutant regressions are run separately for each pollutant, whereas the other estimates are run with all pollutants in a single regression. [§] Elasticities calculated at medians of pollutants, which are 23.68 for PM10, 0.0367 for SO2, 0.1034 for O3. [‡] Willingness to pay calculated for marginal $1 \mu\text{g}/\text{m}^3$ change in PM10 and 1 ppb change in other pollutants, annualized at rate of 0.07 for median house price of \$ 417,800.

Table 6: Implicit Price of Pollution: Efficient Markets (N=72,059)

	All Pollutants [†]			Single Pollutant [†]		
	Coeff	Elast. [§]	WTP [‡]	Coeff	Elast. [§]	WTP [‡]
PM10 ($\mu\text{g}/\text{m}^3$)	-0.0032 [0.0007]	-0.0761 [0.0172]	-94.01 [21.23]	-0.0036 [0.0007]	-0.0842 [0.0164]	-103.94 [20.25]
SO2 (ppm)	-4.8301 [0.2817]	-0.1772 [0.0103]	-141.26 [8.24]	-6.0748 [0.2606]	-0.2229 [0.0096]	-177.66 [7.62]
O3 (ppm)	-5.8211 [0.1841]	-0.6018 [0.0190]	-170.24 [5.38]	-6.1656 [0.1797]	-0.6374 [0.0186]	-180.32 [5.26]

Standard errors clustered at the tract level reported in brackets. Controls for prior sales price, lot size, square feet, number of rooms, number of bedrooms, number of bathrooms, year built and county fixed effects also included but not reported. [†] The single pollutant regressions are run separately for each pollutant, whereas the other estimates are run with all pollutants in a single regression. [§] Elasticities calculated at medians of pollutants, which are 23.68 for PM10, 0.0367 for SO2, 0.1034 for O3. [‡] Willingness to pay calculated for marginal $1 \mu\text{g}/\text{m}^3$ change in PM10 and 1 ppb change in other pollutants, annualized at rate of 0.07 for median house price of \$ 417,800.

was not adequately controlled for by the house fixed effect. This is the typical sort of bias encountered in the hedonic valuation of air pollution – desirable unobservables may evolve over time in conjunction with worsening air pollution (e.g., the opening of new businesses, or other forms of economic growth). The house fixed effect is unable to control for this sort of evolving omitted attribute.

O3 appears to be an exception, producing comparable results across the cross sectional and fixed-effect models. One explanation for this result has to do with the process in which ground-level ozone is formed. In particular, ozone is the outcome of a photochemical process that can be easily altered by variations in weather patterns – for example, it can be shut-down by thick fog or cloud cover, both of which can be quite common in the Bay Area. This may be a source of exogenous variation in ozone that helps identify its effect separately from those of time-invariant omitted attributes. Recall also that this intuition is supported by the low correlations of O3 with the other pollutants as described in Table 11.

Given concerns about time-varying unobservables, it is at this point in the research process where previous work has turned to some quasi-random source of variation in pollution to accurately identify MWTP. There is no natural source of quasi-random variation in our Bay Area data set, so we instead turn to the model based on our assumption of rational home-buyers. The results of this model for the pollution variables are described in Table 3. We first consider the results for PM10 in detail. Whereas the fixed-effect estimates of the MWTP for PM10 had a counterintuitive sign, estimates from our efficient housing market model imply a statistically significant MWTP to avoid an additional microgram of PM10 per cubic meter ranging between \$94 and \$104.

Of the pollutants that we study, particulate matter has received the most attention in the hedonics literature; Smith and Huang (1995) survey that literature from 1967 to 1988 in the context of a meta-analysis. While those papers focused on the sensitivity of house prices to TSP,¹¹ elasticities are still comparable to our results. Smith and Huang find elasticities that tend to lie between -0.04 and -0.07. Our results are therefore on the high-end of this

¹¹Prior to 1987, the EPA measured the concentration of a wide range of particulate matter of various sizes, denoted by total suspended particulates (TSP). After 1987, the EPA switched its focus to "inhalable coarse particles" with diameters between 2.5 and 10 micrometers, and "fine particles" with diameters less than 2.5 micrometers. PM10 refers to any particle with a diameter smaller than 10 micrometers. These particles, which are the focus of our analysis, are considered to have greater adverse health consequences because of their potential to travel deep into the lungs and even into the bloodstream (<http://www.epa.gov/air/particlepollution/basic.html>). They may not, however, be as visible as larger particles, and may therefore have lower amenity costs.

range (i.e., -0.076 to -0.084).

Similar biases appear to be present for O3 and SO2, although in neither case is the bias as severe as in the case of PM10. In the case of SO2, MWTP rises from \$54 to \$141 in the case of the model with all pollutants when time-varying unobservables are accounted for, and \$60 to \$178 for the single pollutant case. In the case of O3, MWTP rises from \$90 to \$170 in the setting with all pollutants and from \$98 to \$180 for the single pollutant model. In contrast to the comparison between the cross-sectional and fixed effect models (where the addition of fixed effects did little to affect MWTP estimates), it appears that time-varying unobservables do bias downward estimates of the MWTP to avoid O3.

To place our ozone estimates in the context of other estimates in the literature, Tra (2010) and Sieg et al (2004) find a MWTP of \$62 and \$67 respectively for a 1% reduction in ozone.¹² Sieg et al (2004) also survey the literature and find that the MWTP to avoid ozone ranges from \$8 to \$181. Our estimates are at the high end of that range, between \$170 to \$180. However, this is not surprising given that other papers in the literature have not controlled for the correlation of ozone with time-varying unobservables, which appear to bias estimates toward zero based on our comparison with the fixed effects and cross sectional approaches.

In all cases, our estimates of the implicit price of pollution are robust to including all pollutants and one pollutant at a time. Given that other pollutants are like time-varying unobservables in the single-pollutant regression, this could lend useful support that our model effectively accounts for time-varying unobservables that might otherwise bias estimates. However, interestingly the fixed effects estimates appear to be equally robust across the multiple and single-pollutant settings. Importantly, the bias from dropping pollutants is not straightforward as it is a function of the within-house correlation between PM10 and the other two pollutants (conditional on other controls) and the implicit price of the omitted pollutants. To understand why the fixed effects estimator is robust to dropping time-varying pollutants, we run an auxiliary regression of the pollutants indexed k and k' as

$$(x_{j,k,t'} - x_{j,k,t}) = \alpha_{0,t,t'} + (x_{j,k',t'} - x_{j,k',t})\alpha_1 + z'_j\alpha_{2,t,t'} + \zeta_{j,k,t,t'}.$$

The resulting estimates of α_1 for different pollutants are reported in Table 7. Each cell

¹²They use the average of the top 30 1-hr concentrations over the course of the year as their measure of ozone. In contrast, we use the maximum 1-hr concentration over the course of the year, which is likely to be more variable.

corresponds to a separate regression of the row variable on the column variable. For instance, the estimate of α_1 from the regression of the within-house change in PM10 on SO2 is 0.00047 and -0.00058 for O3.

Using the implicit prices of SO2 and O3 estimated in the fixed effect model when all pollutants are included, we can approximate the bias on the estimated implicit price of PM10 from dropping SO2 and O3 as $0.00047 \times (-1.8543) - 0.00058 \times (-3.0620) = 0.0009$. Comparing our fixed effect estimates of the implicit price of PM10 across the two settings, we find a bias in the single pollutant regression of $0.0058 - 0.0049 = 0.0009$. Applying the same formulas, we approximate the bias for SO2 as -0.1890 and for O3 as -0.2902. These approximate well with the bias shown in comparing the single and multiple pollutant regression in Table 5, -0.1904 for SO2 and -0.2890 for O3. These calculations show that the finding that the fixed effects is robust to dropping pollutants is not evidence that time-varying unobservables do not matter. On the contrary, because in some cases the pollutants are positively correlated and others negatively, and because the estimated implicit prices in the fixed effects model are positive for PM10 and negative for the other pollutants, the biases from dropping pollutants cancel each other out. In contrast, the fact that our model predicts a negative effect of PM10 while the fixed effects model predicts a counterintuitive positive effect and larger implicit prices suggests that time-varying unobservables are important.

Table 7: Coefficients from Auxiliary Regressions of Pollutants

	PM10	SO2	O3
PM10 ($\mu\text{g}/\text{m}^3$)	—	36.54	-29.51
SO2 (ppm)	0.00047	—	0.0785
O3 (ppm)	-0.00058	0.1202	—
	PM10 (Drop SO2 & O3)	SO2 (Drop PM10 & O3)	O3 (Drop PM10 & SO2)
Bias	0.0009	-0.1809	-0.2902

Controls for prior sales price, lot size, square feet, number of rooms, number of bedrooms, number of bathrooms, year built and county fixed effects also included but not reported. Each cell corresponds to a separate regression with the dependent variable in the row and the independent variable in the column. The bias is calculated from these coefficients using the estimates from the regression with all pollutant in Table 5.

Table 8 describes the results of all three models for non-pollution housing attributes.

Table 8: Implicit Price of House Attributes

	Efficient Market IV		Fixed Effect		Cross Section	
	Coeff	Elast. [§]	Coeff	Elast. [§]	Coeff	Elast. [§]
	Sold in period 3 and 2				Sold in period 3	
Lot size	5.35E-06	0.0368	-1.62E-07	-0.0011	1.44E-05	0.0991
	[4.69E-07]	[0.0032]	[4.52E-07]	[0.0031]	[1.05E-06]	[0.0072]
Square feet	8.96E-05	0.1523	-5.66E-05	-0.0963	3.31E-04	0.5629
	[7.11E-06]	[0.0121]	[6.18E-06]	[0.0105]	[1.58E-05]	[0.0269]
No. of bedrooms	8.21E-03	0.0173	1.65E-02	0.0347	-1.20E-02	-0.0253
	[4.29E-03]	[0.0090]	[4.03E-03]	[0.0085]	[5.52E-03]	[0.0116]
No. of rooms	2.83E-03	0.0199	-4.56E-03	-0.0321	1.37E-02	0.0962
	[2.19E-03]	[0.0154]	[2.08E-03]	[0.0146]	[3.01E-03]	[0.0212]
No. of bathrooms	1.09E-02	0.0352	-6.63E-03	-0.0214	5.20E-02	0.1682
	[4.27E-03]	[0.0138]	[4.13E-03]	[0.0133]	[6.73E-03]	[0.0218]
Year built	-9.19E-04	-1.8065	-1.47E-03	-2.8839	-6.07E-04	-1.1944
	[1.45E-04]	[0.2858]	[1.27E-04]	[0.2493]	[2.75E-04]	[0.5416]
	Sold in period 3 and 1				Sold in period 2	
Lot size	7.49E-06	0.0515	-5.44E-07	-0.0037	1.44E-05	0.0991
	[9.35E-07]	[0.0064]	[7.79E-07]	[0.0054]	[9.76E-07]	[0.0067]
Square feet	1.74E-04	0.2958	-5.68E-05	-0.0966	4.07E-04	0.6930
	[1.10E-05]	[0.0187]	[1.16E-05]	[0.0197]	[1.30E-05]	[0.0221]
No. of bedrooms	-3.07E-03	-0.0065	3.87E-03	0.0081	-3.33E-02	-0.0699
	[5.46E-03]	[0.0115]	[6.47E-03]	[0.0136]	[5.83E-03]	[0.0123]
No. of rooms	4.22E-03	0.0297	-3.08E-03	-0.0217	2.38E-02	0.1677
	[3.33E-03]	[0.0234]	[3.91E-03]	[0.0275]	[3.79E-03]	[0.0267]
No. of bathrooms	2.51E-02	0.0811	-8.33E-03	-0.0269	4.98E-02	0.1608
	[7.17E-03]	[0.0232]	[8.76E-03]	[0.0283]	[7.07E-03]	[0.0228]
Year built	-1.24E-03	-2.4436	-6.49E-04	-1.2762	2.01E-04	0.3958
	[2.47E-04]	[0.4854]	[2.33E-04]	[0.4588]	[3.21E-04]	[0.6302]
	Sold in period 2 and 1				Sold in period 1	
Lot size	6.05E-06	0.0416	-5.44E-07	-0.0037	1.38E-05	0.0950
	[6.35E-07]	[0.0044]	[7.79E-07]	[0.0054]	[9.67E-07]	[0.0067]
Square feet	1.45E-04	0.2473	8.53E-06	0.0145	4.39E-04	0.7475
	[8.93E-06]	[0.0152]	[6.78E-06]	[0.0115]	[1.18E-05]	[0.0201]
No. of bedrooms	-3.61E-03	-0.0076	1.14E-02	0.0239	-5.07E-02	-0.1067
	[4.24E-03]	[0.0089]	[3.62E-03]	[0.0076]	[6.64E-03]	[0.0140]
No. of rooms	6.82E-03	0.0480	-1.93E-03	-0.0136	2.19E-02	0.1542
	[2.55E-03]	[0.0180]	[2.14E-03]	[0.0151]	[4.08E-03]	[0.0287]
No. of bathrooms	1.52E-02	0.0491	5.31E-03	0.0172	3.86E-02	0.1248
	[5.05E-03]	[0.0163]	[4.69E-03]	[0.0152]	[7.44E-03]	[0.0240]
Year built	-1.22E-03	-2.3979	-1.87E-03	-3.6856	9.84E-04	1.9350
	[2.07E-04]	[0.4078]	[1.36E-04]	[0.2678]	[3.20E-04]	[0.6294]

Standard errors clustered at the tract level reported in brackets. These parameter estimates are taken from the same regressions for which the pollutant coefficients are reported in Table 6 (column 1), Table 5 (column 1), and Table 4 (panel A, columns 1,4,and 7). [§] Elasticities calculated at means of house characteristics as reported in Table 1.

For the efficient housing and fixed-effects models, these estimates describe the difference in the implicit price of each attribute over time (e.g., the difference between the period 3 and period 2 coefficients on lot-size in the efficient housing model is a statistically significant 5.35×10^{-6}). We see that the implicit prices of many attributes associated with larger homes tend to rise over time under the efficient housing model, while they more often fall under the fixed-effects model (although this is by no means uniform across all attributes and many of the estimates are insignificant). The value of newer homes (i.e., year-built) falls over time in both of these specifications. The cross-sectional coefficient estimates for non-pollution housing attributes in each period can be easily interpreted as implicit prices.

5 Conclusion

Our paper demonstrates a new approach to controlling for unobserved product attributes in hedonic models. In particular, we show how an assumption about the rationality of home-buyers with respect to the evolution of the omitted housing attribute can be exploited to identify implicit prices in the context of either fixed or time-varying unobserved product attributes. We then describe an estimator that can be applied to settings where repeat sales data are available and our rationality condition is likely to hold.

We use our estimator to recover a consumer's marginal willingness to pay for clean air in the Bay Area. Particularly appealing features of our identification strategy are that (i) it can be easily applied to data from a single, well-defined housing market, and (ii) our main identification assumption is a necessary condition for something that can be tested (i.e., that available information at time t should not predict economically significant excess returns in periods $t + 1$ and beyond). We find supporting evidence that this assumption is valid for the housing market in the Bay Area.

We estimate the implicit price of three of the EPA's criteria air pollutants (PM10, SO2, and O3). In contrast to fixed-effects methods (which just control for time-invariant omitted attributes) or cross-sectional methods (which ignore correlated omitted attributes altogether), our estimates of the implicit price indicate that consumers value pollution reductions, and that their MWTP to avoid pollution is significantly larger in magnitude than that found by other models. Particularly in the case of PM10, it appears that failing to control for omitted attributes at all or only controlling for time-invariant unobserved attributes leads to

the wrong sign on the estimate of the potential benefits of a pollution reduction policy. Our estimates suggest that SO₂ is also prone to a large bias from ignoring time-varying unobservables. Time-varying unobservables associated with ground-level ozone, while important, do not appear to lead to as large of a bias.

To be clear, while our approach works well in this context, we are not claiming that it will be superior to quasi-random approaches in all applications. The identifying assumptions in our approach and quasi-random approaches are not nested, and the plausibility of either set of assumptions depends on the particular application and data one is using. Our approach may be preferable when a legitimate source of quasi-randomness cannot be found, or when one is available but it generates insufficient exogenous variation in the variable of interest. On the other hand, quasi-randomness will be preferable if there is reason to suspect that our rationality assumption is violated (e.g., if there is strong serial correlation in the stochastic innovation of the omitted housing attribute). In empirical work, we advise applied researchers to test the sensitivity of results to alternative identifying assumptions when they are available.

Appendix

A Nonparametric Hedonic Pricing Equations: Generalization

Consider the generalized system of hedonic pricing equations:

$$\ln(p_{j,t}) = \alpha_t + h_t(\bar{x}_{j,t}) + \xi_{j,t} \quad \text{for } t = 1, \dots, T$$

where we normalize $h_t(0) = 0$ and $h_t(\cdot)$ is nonparametrically specified. Similar to (4), we obtain

$$\begin{aligned} \ln(p_{j,t'}) &= \alpha_{t'} + h_{t'}(\bar{x}_{j,t'}) + \xi_{j,t'} & (9) \\ &= \alpha_{t'} + h_{t'}(\bar{x}_{j,t'}) + \gamma(t, t') [\ln(p_{j,t}) - \alpha_t - h_t(\bar{x}_{j,t})] + \eta(j, t, t') \\ &= (\alpha_{t'} - \gamma(t, t')\alpha_t) + \gamma(t, t') \ln(p_{j,t}) \\ &\quad - \gamma(t, t')h_t(\bar{x}_{j,t}) + h_{t'}(\bar{x}_{j,t'}) + \eta(j, t, t'). \end{aligned}$$

To deal with endogeneity of $\bar{x}_{j,t'}$ (which may be correlated with $\eta(j, t, t')$), we exploit the process that describes the evolution of $\bar{x}_{j,t}$ over time. We assume

$$\bar{x}_{j,t'} = g_{t,t'}(\bar{x}_{j,t}, w_{j,t}) + \bar{v}_{j,t,t'} \quad E[\bar{v}_{j,t,t'} | I_t] = 0 \quad (10)$$

In words, the innovation in the observed attributes is orthogonal to time t information where $w_{j,t}$ denotes other observable variables in I_t . Also we assume that

$$\eta(j, t, t') = \tau(t, t')\bar{v}_{j,t,t'} + \varepsilon_{j,t,t'} \quad (11)$$

where $\tau(t, t')$ is the $1 \times K$ parameter vector. These assumptions imply that $\bar{x}_{j,t'}$ evolves according to the process described in (10), but that the innovation in $\bar{x}_{j,t'}$ may be correlated with the innovation in the omitted attribute. Applying assumptions (10) and (11) to (9),

we obtain

$$\begin{aligned} \ln(p_{j,t'}) &= (\alpha_{t'} - \gamma(t, t')\alpha_t) + \gamma(t, t') \ln(p_{j,t}) \\ &\quad - \gamma(t, t')h_t(\bar{x}_{j,t}) + h_{t'}(\bar{x}_{j,t'}) + \tau(t, t')\bar{v}_{j,t,t'} + \varepsilon_{j,t,t'}. \end{aligned} \tag{12}$$

Our identification and estimation methods are then based on the following moment condition

$$E[\varepsilon_{j,t,t'} | I_t, \bar{v}_{j,t,t'}] = 0.$$

This moment condition states that after controlling for time t information I_t and the innovation in the observed attributes $\bar{v}_{j,t,t'}$, the innovation in the omitted attribute has an expected value of zero. This moment condition motivates the use of a control function approach – in the first step, estimate equation (10) and obtain fitted residuals, $\widehat{v}_{j,t,t'}$; in the second step, include $\widehat{v}_{j,t,t'}$ ($K \times 1$) as additional regressors in (4), so we use one control for each endogenous product attribute $x_{j,k,t'}$, $k = 1, \dots, K$.

A.1 Identification and Estimation

In order to show how the model described in Section 2 generalizes to more than two transactions, suppose for each house j we observe transaction prices on three occasions, denoted by $t_a(j)$, $t_b(j)$, and $t_c(j)$ such that $1 \leq t_a(j) < t_b(j) < t_c(j) \leq T$. The result in this section can also be extended to four or more repeat sales. Here we assume that all the elements in $\bar{x}_{j,t}$ are time-varying. If some components in $\bar{x}_{j,t}$ are time invariant, we need to assume that implicit prices of those time invariant attributes are constant over time by the reason explained in Section 2.3.

We write (12) for several time periods (assuming that we have enough observations of

transaction prices for each time period¹³) such that

$$\begin{aligned}\ln(p_{j,t_b}) &= \alpha_{t_b} - \gamma(t_a, t_b)\alpha_{t_a} + \gamma(t_a, t_b) \ln(p_{j,t_a}) - \gamma(t_a, t_b)h_{t_a}(\bar{x}_{j,t_a}) \\ &\quad + h_{t_b}(\bar{x}_{j,t_b}) + \tau(t_a, t_b)\bar{v}_{j,t_a,t_b} + \varepsilon_{j,t_a,t_b} \\ \ln(p_{j,t_c}) &= \alpha_{t_c} - \gamma(t_b, t_c)\alpha_{t_b} + \gamma(t_b, t_c) \ln(p_{j,t_b}) - \gamma(t_b, t_c)h_{t_b}(\bar{x}_{j,t_b}) \\ &\quad + h_{t_c}(\bar{x}_{j,t_c}) + \tau(t_b, t_c)\bar{v}_{j,t_b,t_c} + \varepsilon_{j,t_b,t_c}.\end{aligned}$$

Estimation proceeds based on the following moment conditions:

$$E[\bar{x}_{j,t_b} - g_{t_a,t_b}(\bar{x}_{j,t_a}, w_{j,t_a}) | 1, \bar{x}_{j,t_a}, w_{j,t_a}] = 0 \quad (13)$$

$$E[\varepsilon_{j,t_a,t_b} | 1, \ln(p_{j,t_a}), \bar{x}_{j,t_a}, \bar{x}_{j,t_b}, \bar{v}_{j,t_a,t_b}, w_{j,t_a}] = 0 \quad (14)$$

$$E[\bar{x}_{j,t_c} - g_{t_b,t_c}(\bar{x}_{j,t_b}, w_{j,t_b}) | 1, \bar{x}_{j,t_b}, w_{j,t_b}] = 0 \quad (15)$$

$$E[\varepsilon_{j,t_b,t_c} | 1, \ln(p_{j,t_b}), \bar{x}_{j,t_b}, \bar{x}_{j,t_c}, \bar{v}_{j,t_b,t_c}, w_{j,t_b}] = 0 \quad (16)$$

where $w_{j,t}$ denotes other observable covariates in I_t . From (13) we identify $g_{t_a,t_b}(\cdot)$ (along with \bar{v}_{j,t_a,t_b}), and from (14) we identify $\alpha_{t_b} - \gamma(t_a, t_b)\alpha_{t_a}$, $\gamma(t_a, t_b)$, $h_{t_a}(\bar{x}_{j,t_a})$, $h_{t_b}(\bar{x}_{j,t_b})$, and $\tau(t_a, t_b)$. Similarly from (15) we identify $g_{t_b,t_c}(\cdot)$ (along with \bar{v}_{j,t_b,t_c}) and from (16) we identify $\alpha_{t_c} - \gamma(t_b, t_c)\alpha_{t_b}$, $h_{t_b}(\bar{x}_{j,t_b})$, $h_{t_c}(\bar{x}_{j,t_c})$, and $\tau(t_b, t_c)$.

These may be run as two separate sets of estimations – i.e., one is based on (13) and (14) and the other based on (15) and (16). We note, however, that $h_{t_b}(\bar{x}_{j,t_b})$ is over-identified from the moment conditions, which suggests combining all the moment conditions and performing a nonlinear nonparametric estimation. Having set-up the moment conditions in (13)-(16), one can cast them into Ai and Chen (2003)'s framework and estimate all the parameters (including nonparametric functions) simultaneously. The consistency of the estimators and the asymptotic normality of the parametric components can be obtained following Ai and Chen (2003).

Once we estimate the hedonic function, another parameter of interest will be the weighted average derivative of the log housing price ($\ln(p_{j,t'})$) with respect to the observed characteristic $x_{j,t}$ for $t \leq t'$, defined by

$$E \left[\lambda(x_{j,t}) \frac{\partial \ln(p_{j,t'})}{\partial x_{j,t}} \right]. \quad (17)$$

¹³To be precise, $\frac{1}{J} \sum_{j=1}^J 1\{t_a(j) = t \text{ or } t_b(j) = t \text{ or } t_c(j) = t\} \xrightarrow{J \rightarrow \infty} C > 0$ for all $t = 1, \dots, T$.

This can be estimated using its sample analogue,

$$\frac{1}{J} \sum_{j=1}^J \lambda(x_{j,t}) \frac{\partial \ln(p_{j,t'})}{\partial x_{j,t}} \approx \frac{1}{J} \sum_{j=1}^J \lambda(x_{j,t}) \frac{\partial \widehat{\ln(p_{j,t'})}}{\partial x_{j,t}}$$

where $\widehat{\ln(p_{j,t'})}$ is the fitted log price function. The weight $\lambda(\cdot)$ satisfies $\lambda(\cdot) \geq 0$ and $\int \lambda(x) dx = 1$. In the literature, estimated implicit prices $\frac{\partial \widehat{\ln(p_{j,t'})}}{\partial x_{j,t}}$ are commonly used to recover valuation for non-market amenities such as clean air or public school quality. Since $\frac{\partial \ln(p_{j,t'})}{\partial x_{j,t}}$ varies across homeowners in the population, the function $\lambda(x_{j,t})$ allows the researcher to aggregate these heterogeneous marginal benefits. The scalar (17) summarizes the average marginal benefit from an increase in the amount of some characteristic $x_{j,t}$ in time period t and is often used to measure welfare from a policy change.

References

- Ackerberg, D., L. Benkard, S. Berry, and A. Pakes (2007), “Econometric Tools for Analyzing Market Outcomes,” *Handbook of Econometrics*, 6A, Edited by J. Heckman and E. Leamer, North Holland.
- Ai, C. and X Chen (2003), “Efficient Estimation of Models With Conditional Moment Restrictions Containing Unknown Functions,” *Econometrica*, 71(6), 1795-1843.
- Arellano, M. and S. Bond (1991), “Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations,” *Review of Economic Studies*, 58, 277-297.
- Bartik, TJ (1987), “The Estimation of Demand Parameters in Hedonic Price Models,” *Journal of Political Economy*, 95(1): 81-88.
- Bajari, P and CL Benkard (2005), “Demand Estimation with Heterogeneous Consumers and Unobserved Product Characteristics: A Hedonic Approach,” *Journal of Political Economy*, 113(6), 1239-1276.
- Bishop, K and C Timmins (2010a), “Hedonic Prices and Implicit Markets: Consistently Estimating the Marginal Willingness to Pay for Differentiated Products Without Instruments,” *Duke University Working Paper*.

- Bishop, K and C Timmins (2010b), "Recovering the Marginal Willingness to Pay to Avoid Ozone: Using Panel Data to Easily Estimate Hedonic Inverse Demand Functions," *Duke University Working Paper*.
- Black, SE (1999), "Do Better Schools Matter? Parental Valuation of Elementary Education," *The Quarterly Journal of Economics*, 114(2): 577-599.
- Blundell, R. and S. Bond (1998), "GMM Estimation with Persistent Panel Data: An Application to Production Functions," Working paper, W99/4, The Institute for Fiscal Studies.
- Brown, JN and HS Rosen (1982), "On the Estimation of Structural Hedonic Price Models," *Econometrica*, 50(3): 765-68.
- Case, KE and R. J Shiller (1989), "The Efficiency of the Market for Single-Family Homes," *The American Economic Review*, 79(1): 125-137.
- Chay, KY and M Greenstone (2005), "Does Air Quality Matter? Evidence from the Housing Market," *Journal of Political Economy*, 113(2), 376-424.
- D'Ippoliti D, F Forastiere, C Ancona, N Agabity, D Fusco, P Michelozzi, and CA Perucci (2003), "Air Pollution and Myocardial Infarction in Rome: A Case-Crossover Analysis," *Epidemiology*, 14: 528-535.
- Dockery DW, CA Pope, X Xu, JD Spengler, JH Ware, ME Fay, BG Ferris, and FE Speizer (1993), "An Association Between Air Pollution and Mortality in Six U.S. Cities," *New England Journal of Medicine*, 329: 1753-1759.
- Dougherty, A. and R Van Order (1982), "Inflation, Housing Costs, and the Consumer Price Index," *The American Economic Review*, 72(1): 154-164.
- Ekeland, I, JJ Heckman, and L Nesheim (2004), "Identification and Estimation of Hedonic Models," *Journal of Political Economy*, 112(S1): S60-S109.
- Epple, D (1987), "Hedonic Prices and implicit Markets: Estimating Demand and Supply Functions for Differentiated Products," *Journal of Political Economy*, 95(1): 59-80.
- Ghio AJ, C Kim, and RB Devlin (2000), "Concentrated Ambient Air Particles Induce Mild Pulmonary Inflammation in Healthy Human Volunteers," *American Journal of Respiratory and Critical Care Medicine*, 162(3 Pt.1): 981-988.

- Greenstone, M. and J. Gallagher (2008), “Does Hazardous Waste Matter? Evidence from the Housing Market and the Superfund Program,” *The Quarterly Journal of Economics*, MIT Press, 123(3): 951-1003.
- Heckman, J, RL Matzkin, and L Nesheim. “Simulation and Estimation of Hedonic Models,” Center for Economic Studies & Ifo Institute for Economic Research (CESifo), Working Paper No. 1014, August 2003.
- Hong YC, JT Lee, H Kim, EH Ha, J Schwartz, and DC Christiani (2002), “Effects of Air Pollutants on Acute Stroke Mortality,” *Environmental Health Perspectives*, 110: 187-191.
- Lin M, Y Chen, RT Burnett, PJ Villeneuve, and D Kerwski (2002), “The Influence of Ambient Coarse Particulate Matter on Asthma Hospitalization in Children: Case-Crossover and Time-Series Analyses,” *Environmental Health Perspectives*, 110: 575-581.
- Mendelsohn, R (1985), “Identifying Structural Equations with Single Market Data,” *Review of Economics and Statistics*, 67(3): 525-29.
- Muller N., D. Tong, and R. Mendelsohn (2009), “Regulating NOx and SO2 Emissions in Atlanta,” *The B.E. Journal of Economic Analysis and Policy*, Berkeley Electronic Press, 9(2).
- Murphy, K and R Topel (1985), “Estimation and Inference in Two-Step Econometric Models,” *The Journal of Business & Economic Statistics*, 3-4, 370-379.
- Norris G, SN YoungPong, JQ Koenig, TV Larson, L Sheppard, and JW Stout (1999), “An Association Between Fine Particles and Asthma Emergency Department Visits for Children in Seattle,” *Environmental Health Perspectives*, 107: 489-493.
- Pope CA, RT Burnett, MJ Thun, EE Calle, D Krewski, K Ito, and GD Thurston (2002), “Lung Cancer, Cardiopulmonary Mortality, and Long-Term Exposure to Fine Particulate Air Pollution,” *JAMA*, 287:1132-1141.
- Rosen, S (1974), “Hedonic Prices and Implicit markets: Product Differentiation in Pure Competition” *Journal of Political Economy*, 82(1):34-55.
- Samet JM, DM DeMarini, and HV Malling (2004), “Do Airborne Particulates Induce Heritable Mutations?” *Science*, 304(5673):971-972.

Sieg H, VK Smith, HS Banzhaf, and R Walsh (2004), "Estimating The General Equilibrium Benefits Of Large Changes In Spatially Delineated Public Goods" *International Economic Review*, 45(4): 1047-1077, November.

Slaughter JC, T Lumley, L Sheppard, JQ Koenig, and GG Shapiro (2003), "Effects of Ambient Air Pollution on Symptom Severity and Medication Use in Children with Asthma," *Annals of Allergy, Asthma, and Immunology*, 91: 346-353.

Smith, VK and JC Huang (1995), "Can Markets Value Air Quality? A Meta-Analysis of Hedonic Property Value Models," *Journal of Political Economy*, 103(1), 209-227.

Tinbergen, J (1956) "On the Theory of Income Distribution," *Weltwirtschaftliches Archiv*, LXXVII: 155-175.

Tolbert PE, JA Mulholland, DD MacIntosh, F Xu, D Daniels, OJ Devine, BP Carlin, M Klein, J Dorley, AJ Butler, DF Nordenberg, H Franklin, PB Ryan, and MC White (2000), "Air Quality and Pediatric Emergency Room Visits for Asthma in Atlanta, Georgia," *American Journal of Epidemiology*, 151: 798-810.

Tra, CI (2010) "A discrete choice equilibrium approach to valuing large environmental changes," *Journal of Public Economics*, 94(1-2): 183-196, February.

Tsai SS, WB Goggins, HF Chiu, and CY Yang (2003), "Evidence for an Association Between Air Pollution and Daily Stroke Admissions in Kaohsiung, Taiwan," *Stroke*, 34(11): 2612-2616.