

Hedonic Price Regressions with Omitted Product Attributes

Patrick Bajari

University of Minnesota and NBER

Jane Cooley

University of Wisconsin

Kyoo il Kim

University of Minnesota

Christopher Timmins

Duke University and NBER

September 21, 2009

Abstract

Hedonic techniques are commonly used to recover implicit prices for the attributes of differentiated products. In the context of housing markets, these include amenities such as clean air, public safety, school quality, proximity to open space, and distance from Superfund sites. Estimated implicit prices are used to value the benefits of policies designed to protect or improve these amenities. However, if unobservable housing attributes are correlated with the amenity of interest, OLS estimates of implicit price estimates will be biased. In this paper, we propose a strategy using assumptions from economic theory to consistently estimate implicit prices. The first assumption is that market prices reflect characteristics of the home, including those that are not directly observed by the econometrician. The second assumption is that housing markets are informationally efficient in at least a limited sense. We assume that home characteristics observable to buyers cannot be used to earn excess returns. Using data describing housing transactions in California's Bay Area between 1990 and 2006, we find evidence supporting our informational efficiency assumption. Applying our estimator, we recover implicit prices for four of the EPA's "criteria" air pollutants – particulate matter (PM10), sulfur dioxide (SO2), nitrogen oxides (NOx), and ground-level ozone (O3). In contrast to simple cross-sectional or fixed effect estimators, marginal willingnesses to pay for a reduction in all four pollutants (considered individually or together) are all statistically significant with the expected sign. The difference in results is particularly large for PM10, a pollutant with serious health consequences. In the case of NOx and SO2, time-varying unobservables also appear to play an important role in biasing simple fixed effect estimates. Our results suggest, however, that controlling for time-varying unobservables may not be an important issue in the case of O3.

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 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 from an estimate of its implicit price.

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 number of square feet, the lot size and 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 to the econometrician. If the unobserved attributes are correlated with the observed attributes, ordinary least squares estimates of the implicit prices will be biased. When confounding unobservables are time-invariant, they can be accounted for with fixed effects if panel data are available. When confounding unobservables vary over time, 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 (2007), and Black (1999) have proposed quasi experimental approaches to this problem, exploiting 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 the product 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 which may confound estimates of implicit prices. Second, the identifying assumptions are transparent and the estimators are simple to implement, frequently only requiring the use of canned regression packages. However, even if the authors recognize that this assumption is incorrect, a large number of papers that use hedonic regressions assume that omitted attributes are exogenous for several reasons. First, a source of quasi-randomness that generates exogenous variation in the product characteristic of interest for a given policy question may not be available in a particular data set. Second, even if a natural experiment is located, the implicit prices may not be precisely estimated or standard diagnostics may suggest the instruments are weak. Third, there may be controversy about the validity of the identifying assumptions, for example, if there is the possibility that the instruments

are themselves endogenous. These limitations suggest that it would be useful to have alternative assumptions to identify implicit prices. At a minimum, this would provide an alternative set of identifying assumptions to check the robustness of results from natural experiments or provide an estimation method when quasi-randomness is not present.

In this paper, we propose a method for consistently estimating implicit prices in housing markets that does not require quasi-randomness. Instead, our identifying assumptions are derived from economic theory. Our method uses an unbalanced panel of repeat sales and home attributes. The first assumption is that home price in a local, geographically separated market at a point in time, can be written as a function of a home's attributes. Importantly, for our application, we assume that this includes attributes that are observed to the home buyers but not the economist. 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 to econometrically measure the "curb appeal" of a home. However, anyone who has purchased a home knows this is very important to many buyers. Our first assumption implies that curb appeal is priced by the market even if the econometrician fails to measure it. As a consequence, the residual from a hedonic regression allows the econometrician to price home attributes that she does not directly observe.

Our second identifying assumption is that housing markets are *informationally efficient* in the following, limited sense: information about a home's characteristics at the time of purchase cannot be used to make excess returns. In other words, housing markets price current and anticipated changes in the characteristics of a home. Thus, innovations in the characteristics of a home cannot be used to earn excess returns given current information. For example, suppose that a home has particularly high curb appeal because it was built by skilled craftsmen in the 1900's. Older homes may be subject to more wear and tear than newer ones. Our assumption implies that buyers rationally anticipate that this curb appeal may depreciate at a fast rate and that market prices adjust accordingly.

In our paper, we demonstrate that these two assumptions allow us to consistently estimate implicit prices using repeat sales data. The intuition behind our estimator is straightforward. Suppose that we observe a home that 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 product attributes

in 1998. If the market price was abnormally positive (negative) controlling for the covariates in the econometrician’s data set, we infer that the home had a large positive (negative) value for characteristics that were not observed by the economist. More formally, the hedonic equation allows us to form a control function to impute the market value of omitted attributes.

Our second condition provides us with a moment equation similar to well known GMM estimators in financial econometrics. We should expect the value of omitted attributes to evolve over time. However, our market efficiency assumption implies that the innovation in the omitted attribute must be orthogonal to the agent’s information in 1998, at the time of purchase. This moment equation permits the identification of the implicit prices in the hedonic, while taking into account potentially confounding unobservables. We show that our strategy can allow for a flexible functional form by casting our problem in the framework of Ai and Chen (2003). Approaches that exploit quasi-randomness may require linearity, as in the case of regression discontinuity, or 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 to the real world functioning of housing markets. For example, our first assumption will not hold perfectly because home prices are often determined by negotiation and therefore cannot be explained perfectly by the characteristics of the home. At the same time, we argue that there are not many opportunities for a free lunch in a housing market with many buyers and sellers. Finding “steals” in the housing market is the exception, not the rule, and only rarely can a buyer find twice the home for half the price.

Our second assumption is also an approximation to real world housing markets. A fraction of buyers may be able to forecast which homes will have a particularly strong appreciation rate. At the same time, competition is likely to limit the returns from speculation using publicly observed information on home characteristics. Our application is a rich data set from the San Francisco Bay Area. This is a market with a large number of buyers and sellers, including many sellers who actively engaged in buying, repairing and then reselling homes in our sample. Our assumption allows for the possibility that buyers anticipated profiting, perhaps handsomely, from buying and selling their homes. What our assumption rules out is that buyers were able to use commonly observed home characteristics to earn abnormal profits. Case and Shiller (1989) find support for this assumption, arguing that within a given metropolitan area, it is difficult to find variables that predict excess returns in housing. In our application, we find support for our second assumption in the sense that the characteristics such as square feet, lot size and air pollution at the time of purchase cannot be used to generate economically significant returns in a local housing market.

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 four of the EPA's "criteria pollutants" (i.e., pollutants used by the EPA in setting emissions regulations) – particulate matter (PM10), sulfur dioxide (SO2), nitrogen oxides (NOx), and ground-level ozone (O3) – all of which are known to have adverse health consequences and inflict 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. Moreover, if changes in pollution are correlated with changes in these omitted variables, a simple fixed effect estimator will not correct the bias.

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 informationally efficient. Using our estimator, we recover implicit prices for four of the EPA's "criteria" air pollutants described above. In contrast to simple cross-sectional or fixed effect estimators, marginal willingnesses to pay for a reduction in all four pollutants (considered individually or together) are all statistically significant and have the expected sign. The elasticity of housing prices with respect to PM10 (considered alone) is -0.15, which is closer to Chay and Greenstone's (2005) instrumental variables estimates (-0.21 to -0.35) than other estimates in the literature. Moreover, unlike Chay and Greenstone (2005), our approach does not require that we treat the US as being comprised of a single, unified housing market. Rather, we are able to derive estimates with data from just a single metropolitan area. Considered together, PM10, SO2, NOx, and O3 exhibit house price elasticities of -0.10, -0.08, -0.04, and -0.16, respectively. While controlling for time-varying unobservables appears to be important for NOx and SO2, it looks to be particularly important for PM10, but not to make much difference for O3.

At a minimum, this application demonstrates that our estimator is able to generate estimates with a priori plausible signs when fixed effects methods can not. Moreover, the biases in the estimates are in the direction that economic and econometric theory suggest.

This paper proceeds as follows. Section 2 describes the estimator based on the efficient housing market hypothesis. Section 2.1 describes identification in the most general terms. Section 2.2 and 2.3 describe a number of modifications and restrictions that we make to the general model before applying it to data. Section 3 describes the data that we use for our application. Section 4 presents results from our model, and compares them to results from a traditional fixed effects and cross-sectional specification. 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 static 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 reasonable. A home $j = 1, \dots, J$ 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, let ξ_j denote 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 will only imperfectly reflect features such as the curb appeal of a home or its state of repair, features that 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 a consumer $i = 1, \dots, I$, 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 preference 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 frequently 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 has been frequently criticized in the literature. For instance, 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. Small (1975) argues that demand estimates in differentiated product markets are typically biased because the economist only imperfectly observes the product attributes that enter into a consumer's utility.¹ Thus, failure to correct for this omitted variable

¹I have entirely avoided...the important question of whether the empirical difficulties, especially correlation

would bias upward estimates of implicit prices of desirable attributes. In empirical applications, the only proposed solutions have relied 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 have the limitations discussed in the introduction.

We propose an alternative approach to estimating implicit prices. We begin by considering a flexible specification, but later impose restrictions on the model to facilitate identification with the data in our application. We consider cases in which there are data on repeat sales so that the price of home j is observed over 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 prices for each house.

To simplify notation, consider the case where there is a single observed, potentially time-varying characteristic, $x_{j,t}$ (all results apply if this were a vector of characteristics). Suppose the system of hedonic pricing equations is:

$$\begin{aligned} \ln(p_{j,1}) &= \alpha_1 + h_1(x_{j,1}) + \xi_{j,1} \\ &\vdots \\ \ln(p_{j,T}) &= \alpha_T + h_T(x_{j,T}) + \xi_{j,T}, \end{aligned} \tag{1}$$

where we normalize $h_t(0) = 0$ and $h_t(\cdot)$ is nonparametrically specified. 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 change over time periods. For example, a noisy neighbor may move in next door to home j or a fungus make it necessary to cut down large existing neighborhood trees. The second is that the price of the omitted attributes could change over time. In the above equations, $\xi_{j,t}$ can be interpreted as the market price of home characteristics that are unobserved by the econometrician. Future home prices will not be known with certainty by the agent. As a result, the implicit price of ‘‘curb appeal’’ could evolve over time. In our model, we will assume that the omitted product

between pollution and unmeasured neighborhood characteristics, are so overwhelming as to render the entire [hedonic] method useless. I hope that...future work can proceed to solve these practical problems...The degree of attention devoted to this [problem] is what will really determine whether the method stands or falls.’’ [Small (1975, p.107) quoted in Chay and Greenstone (2005)]

attribute evolves according to a first order Markov process,

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

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 innovation in the omitted attribute.

Let I_t denote the information available to the buyer at time t . A necessary (but not sufficient) condition for informational efficiency is that the price of any home j at time periods t and t' must satisfy

$$\ln(p_{j,t'}) - \ln(p_{j,t}) = E[\ln(p_{j,t'}) - \ln(p_{j,t})|I_t] + \zeta_{t,t'}. \quad (3)$$

The left hand side variable in the above equation is the log gross return on home j between t and t' , $\ln(p_{j,t'}) - \ln(p_{j,t})$. The right hand side is the expected value of $\ln(p_{j,t'}) - \ln(p_{j,t})$ given current information and the forecast error $\zeta_{t,t'}$. By the above equations, it follows that:

$$\begin{aligned} \zeta_{t,t'} &= (\alpha_{t'} - \alpha_t) - E[\alpha_{t'} - \alpha_t|I_t] + (h_{t'}(x_{j,t'}) - h_t(x_{j,t})) - E[h_{t'}(x_{j,t'}) - h_t(x_{j,t})|I_t] \\ &\quad + (\xi_{j,t'} - \xi_{j,t}) - E[\xi_{j,t'} - \xi_{j,t}|I_t] \\ &= (\alpha_{t'} - \alpha_t) - E[\alpha_{t'} - \alpha_t|I_t] + (h_{t'}(x_{j,t'}) - h_t(x_{j,t})) - E[h_{t'}(x_{j,t'}) - h_t(x_{j,t})|I_t] \\ &\quad + (\gamma(t, t') - 1)\xi_{j,t} + \eta(j, t, t') - (\gamma(t, t') - 1)\xi_{j,t} \\ &= (\alpha_{t'} - \alpha_t) - E[\alpha_{t'} - \alpha_t|I_t] + (h_{t'}(x_{j,t'}) - h_t(x_{j,t})) - E[h_{t'}(x_{j,t'}) - h_t(x_{j,t})|I_t] \\ &\quad + \eta(j, t, t') \end{aligned}$$

Our second major assumption is that $E[\eta(j, t, t')|I_t] = 0$. That is, the innovation in the omitted attribute is orthogonal to the time t information set I_t . As a consequence, a rational agent's forecast of $\xi_{j,t'}$ a time t is $\gamma(t, t')\xi_{j,t}$. Combining this assumption with the above equation implies that:

$$\begin{aligned} E[\zeta_{t,t'}|I_t] &= E[\alpha_{t'} - \alpha_t|I_t] - E[\alpha_{t'} - \alpha_t|I_t] + E[h_{t'}(x_{j,t'}) - h_t(x_{j,t})|I_t] \\ &\quad - E[h_{t'}(x_{j,t'}) - h_t(x_{j,t})|I_t] + E[\eta(j, t, t')|I_t] \\ &= E[\eta(j, t, t')|I_t] \\ &= 0 \end{aligned} \quad (4)$$

The assumption that $E[\eta(j, t, t')|I_t] = 0$ implies a limited form of informational efficiency. In particular, the agents cannot use information about home characteristics at time t to make excess returns in the housing market as measured by $\ln(p_{j,t'}) - \ln(p_{j,t})$. It is important to note that we are not taking a stance on whether excess returns can be made in the housing market using other information. For example, there may be forecastable time trends in home prices due to macroeconomic variables, mean reversion in prices or other information which might be profitably exploited by an investor.

2.1 Lagged Prices and Consistent Estimation of the Hedonic

In this section, 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 $\ln(p_{j,t}) - \alpha_t - h_t(x_{j,t})$ for $\xi_{j,t}$, we get:

$$\begin{aligned} \ln(p_{j,t'}) &= \alpha_{t'} + h_{t'}(x_{j,t'}) + \xi_{j,t'} & (5) \\ &= \alpha_{t'} + h_{t'}(x_{j,t'}) + \gamma(t, t') [\ln(p_{j,t}) - \alpha_t - h_t(x_{j,t})] + \eta(j, t, t') \\ &= (\alpha_{t'} - \gamma(t, t')\alpha_t) + \gamma(t, t') \ln(p_{j,t}) \\ &\quad - \gamma(t, t')h_t(x_{j,t}) + h_{t'}(x_{j,t'}) + \eta(j, t, t'). \end{aligned}$$

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 have been correlated with an observable characteristic such as test scores in local public schools. This means a regression based on (5) will produce inconsistent estimates of hedonic functions. However, by exploiting the process that the evolution of $x_{j,t}$ follows over time, we can still obtain consistent estimates for all the parameters in (5). We assume

$$x_{j,t'} = g_{t,t'}(x_{j,t}) + v_{j,t,t'} \quad E[v_{j,t,t'}|I_t] = 0 \quad (6)$$

that is, the innovation in the observed attributes is orthogonal to time t information. Also we assume that

$$\eta(j, t, t') = \tau(t, t')v_{j,t,t'} + \varepsilon_{j,t,t'}. \quad (7)$$

These assumptions imply that $x_{j,t'}$ evolves according to the process described in (6), but that the innovation in $x_{j,t'}$ may be correlated with the innovation in the omitted attribute. Applying

assumptions (4) and (7) to (5), we obtain

$$\begin{aligned} \ln(p_{j,t'}) &= (\alpha_{t'} - \gamma(t, t')\alpha_t) + \gamma(t, t') \ln(p_{j,t}) \\ &\quad - \gamma(t, t')h_t(x_{j,t}) + h_{t'}(x_{j,t'}) + \tau(t, t')v_{j,t,t'} + \varepsilon_{j,t,t'}. \end{aligned}$$

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

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

This moment condition states that after controlling for time t information I_t and the innovation in the observed attributes $v_{j,t,t'}$, the innovation in the omitted attribute has an expected value of zero. This moment condition motivates us to use a control function approach.² In the first step we estimate equation (6) and obtain fitted residuals, $\hat{v}_{j,t,t'}$. In the second step, we include $\hat{v}_{j,t,t'}$ as an additional regressor in (5).

This procedure differs considerably from standard methods for estimating hedonics where an identifying assumption for consistent estimation is that $E[\xi_j | \bar{x}_j] = 0$, which as we discussed above is likely to be strong in many applications.

Our approach uses the information in lagged prices, $p_{j,t}$ to impute a lagged value of the omitted attribute. For example, if the price for home j was unusually high after controlling for \bar{x}_j , we would infer that $\xi_j = \ln(p_{j,t}) - \alpha_t - h_t(x_{j,t})$ was also large. This is where our first economic assumption has bite. We assume that prices reflect attributes that are observed to consumers, but not the economist. We assume that the innovations in the observed attributes are orthogonal to current information. As in the previous section, if this were not true, it would be possible to earn excess returns in the housing market. We also assume that after controlling for I_t and $v_{j,t,t'}$, the innovation in the omitted attribute $\varepsilon_{j,t,t'}$, is mean zero. As we discussed in the last section, this is necessary to prevent buyers from earning excess returns in the housing market given time t information. Putting it all together, we achieve identification through our two economic assumptions- prices reflect product attributes, including those that are unobserved to the economist and buyers cannot use current information to make excess returns in housing. We shall develop this more formally in the next section.

²Note that, in a linear estimation framework, this will be equivalent to 2SLS.

2.2 Identification and Estimation

To consider a general setting, 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 $1 \leq t_a(j) < t_b(j) < t_c(j) \leq T$ (we later consider the case when we observe only two transactions in detail). For now, we also assume that $x_{j,t}$ is time-varying, but relax this assumption later.

We write (5) for several time periods (assuming that we have enough observations of transaction prices for each time period³) such that

$$\begin{aligned} \ln(p_{j,t_b}) &= \alpha_{t_b} - \gamma(t_a, t_b)\alpha_{t_a} + \gamma(t_a, t_b)\ln(p_{j,t_a}) - \gamma(t_a, t_b)h_{t_a}(x_{j,t_a}) \\ &\quad + h_{t_b}(x_{j,t_b}) + \tau(t_a, t_b)\nu_{j,t_a,t_b} + \varepsilon_{j,t_a,t_b} \end{aligned} \quad (8)$$

$$\begin{aligned} \ln(p_{j,t_c}) &= \alpha_{t_c} - \gamma(t_b, t_c)\alpha_{t_b} + \gamma(t_b, t_c)\ln(p_{j,t_b}) - \gamma(t_b, t_c)h_{t_b}(x_{j,t_b}) \\ &\quad + h_{t_c}(x_{j,t_c}) + \tau(t_b, t_c)\nu_{j,t_b,t_c} + \varepsilon_{j,t_b,t_c}. \end{aligned} \quad (9)$$

Estimation proceeds based on the following moment conditions:

$$E[x_{j,t_b} - g_{t_a,t_b}(x_{j,t_a}) | 1, x_{j,t_a}] = 0 \quad (10)$$

$$E[\varepsilon_{j,t_a,t_b} | 1, \ln(p_{j,t_a}), x_{j,t_a}, x_{j,t_b}, \nu_{j,t_a,t_b}] = 0 \quad (11)$$

$$E[x_{j,t_c} - g_{t_b,t_c}(x_{j,t_b}) | 1, x_{j,t_b}] = 0 \quad (12)$$

$$E[\varepsilon_{j,t_b,t_c} | 1, \ln(p_{j,t_b}), x_{j,t_b}, x_{j,t_c}, \nu_{j,t_b,t_c}] = 0. \quad (13)$$

From (10) we identify $g_{t_a,t_b}(\cdot)$ (along with ν_{j,t_a,t_b}), and from (11) we identify α_{t_a} , α_{t_b} , $\gamma(t_a, t_b)$, $h_{t_a}(x_{j,t_a})$, $h_{t_b}(x_{j,t_b})$, and $\tau(t_a, t_b)$. Similarly from (12) we identify $g_{t_b,t_c}(\cdot)$ (along with ν_{j,t_b,t_c}) and from (13) we identify α_{t_b} , α_{t_c} , $\gamma(t_b, t_c)$, $h_{t_b}(x_{j,t_b})$, $h_{t_c}(x_{j,t_c})$, and $\tau(t_b, t_c)$.

These may be run as two separate sets of estimations – i.e., one is based on (10) and (11) and the other based on (12) and (13). We note, however, that α_{t_b} and $h_{t_b}(x_{j,t_b})$ are over-identified from the moment conditions, which motivates us to combine all the moment conditions and perform a nonlinear nonparametric estimation. Having set-up the moment conditions in (10)-(13), we cast them into Ai and Chen (2003)'s framework where we 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

³To 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 $t = 1, \dots, T$.

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

$$E \left[\tau(x_{j,t}) \frac{\partial \ln(p_{j,t'})}{\partial x_{j,t}} \right]. \quad (14)$$

This can be estimated using its sample analogue,

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

where $\widehat{\ln(p_{j,t'})}$ is the fitted log price function. The weight $\tau(\cdot)$ satisfies $\tau(\cdot) \geq 0$ and $\int \tau(x) dx = 1$. 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 $\tau(x_{j,t})$ would allow the researcher to aggregate these average marginal benefits. The scalar (14) would summarize 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.

2.3 Time-invariant Covariates and Model Restrictions

When $x_{j,t}$ has no time-varying components, some of the parameters in the full model are not identified without further restrictions. With the time invariant covariate z_j , equation (5) becomes

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

We cannot therefore identify $h_{t'}(z_j)$ separately from $h_t(z_j)$ – a multicollinearity problem. Restricting α_t to be fixed over time, the above equation becomes

$$\ln(p_{j,t'}) = \alpha_0 (1 - \gamma(t, t')) + \gamma(t, t') \ln(p_{j,t}) - \gamma(t, t')h_t(z_j) + h_{t'}(z_j) + \eta(j, t, t'). \quad (15)$$

With these restrictions, we can identify $\gamma(t, t')$ from the coefficient on $\ln(p_{j,t})$ and α_0 from the constant term. By further assuming $h_t(z_j) = h(z_j)$, that function is also identified. Alternatively, one can normalize $h_1(z_j) = 1$. Then $h_t(z_j)$, $t > 1$ is identified recursively up to this normalization using the fact that $-\gamma(t, t')h_t(z_j) + h_{t'}(z_j)$ 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' or let $\gamma(t, t')$ take same values for time periods of

interest. Even when some elements of x are time varying, if that variation is not sufficient, one can still impose the above restrictions to obtain more reliable and robust estimates in the estimation.

2.4 Simple Parametric Model with Two Transactions

Consider the case when only two transactions per house are available in the data. Moreover, impose simple parametric functional forms on the process governing the evolution of $x_{j,t}$ and on the way that variable enters the hedonic price function, $h_t(x_{j,t}) = \beta x_{j,t} \forall t$. The model simplifies to

$$\begin{aligned} v_{j,t_a,t_b} &= x_{j,t_b} - \pi_0(t_a, t_b) - \pi_1(t_a, t_b)x_{j,t_a} \\ \ln(p_{j,t_b}) &= \alpha_{t_b} - \gamma(t_a, t_b)\alpha_{t_a} + \gamma(t_a, t_b)\ln(p_{j,t_a}) + \gamma(t_a, t_b)\beta x_{j,t_a} \\ &\quad + \beta x_{j,t_b} + \tau(t_a, t_b)v_{j,t_a,t_b} + \varepsilon_{j,t_a,t_b} \end{aligned} \quad (16)$$

where t_a denotes the time period of the first sale and t_b denotes the time period of the second sale. Clearly all the parameters are identified. More importantly, the estimation procedure is a simple application of two step nonlinear least squares. The coefficients in the second step equation (16) are nonlinear functions of $\theta(t_a, t_b) = (\alpha_{t_a}, \alpha_{t_b}, \gamma(t_a, t_b), \beta, \tau(t_a, t_b))'$ after we estimate $\pi_0(t_a, t_b)$ and $\pi_1(t_a, t_b)$ in the first step. We obtain estimates by solving

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

where $\hat{v}_{j,t_a,t_b} = x_{j,t_b} - \hat{\pi}_0(t_a, t_b) - \hat{\pi}_1(t_a, t_b)x_{j,t_a}$ and

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

Note that the first step estimation contributes to the asymptotic variances of the second step estimators. We can obtain correct standard errors by extending Murphy and Topel (1985) to a

nonlinear least squares. Denote

$$\begin{aligned}
\sqrt{J}(\widehat{\pi}(t_a, t_b) - \pi(t_a, t_b)) &\rightarrow_d N(0, V(t_a, t_b)) \\
G(\cdot; \theta(t_a, t_b)) &= \frac{\partial}{\partial \theta(t_a, t_b)} g(\cdot; \theta(t_a, t_b)) \\
\Omega_0(t_a, t_b) &= E[\widehat{\varepsilon}_{j,t_a,t_b}^2 G(\cdot; \theta(t_a, t_b)) G(\cdot; \theta(t_a, t_b))'] \\
Q_0(t_a, t_b) &= E[G(\cdot; \theta(t_a, t_b)) G(\cdot; \theta(t_a, t_b))'] \\
Q_1(t_a, t_b) &= E[G(\cdot; \theta(t_a, t_b)) \tau(t_a, t_b) x_{j,t_a}]
\end{aligned} \tag{17}$$

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

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

A consistent estimator of the variance matrix $\Sigma(t_a, t_b)$ is obtained using the following sample counterparts of (17):

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

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 four of the EPA's "criteria" air pollutants - particulate matter (PM10), sulfur dioxide (SO2), nitrogen oxides (NOx)

and ground-level ozone (O₃).⁴ 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 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 last time at which that house appears in the sample. 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 cuts to clean the data. First, 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 at a very high or very low rate relative to other houses on the market at the same time. We expect that this will help control for houses whose structural attributes changed (for better or worse) over that pair of years.

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, 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

⁴The list of criteria pollutants also includes 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>)

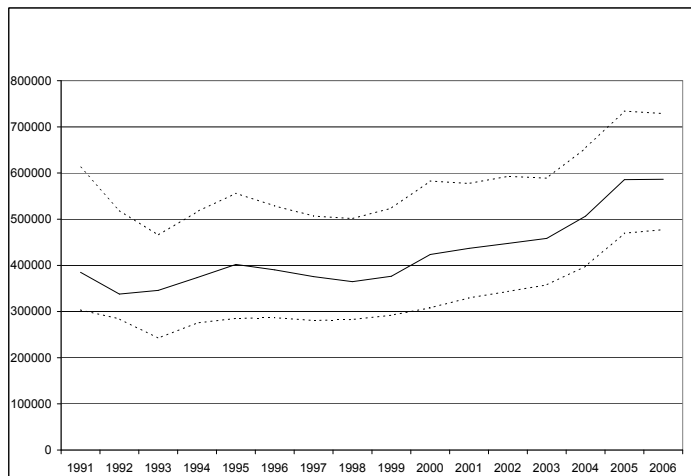
a 70,000 square-foot lot. We finally drop all homes that sell more than two times in the seventeen year period we are considering. This is done primarily for the sake of convenience, as it allows us to implement our estimator using a simple non-linear least squares routine like that found in standard statistical packages. In the end, these cuts leave us with data describing repeat transactions for 74,892 unique housing units. Table 1 summarizes the attributes of these houses.

Table 1: House Attributes (N=74,892)

	Mean	Std Dev	Minimum	Maximum
Lot size	6,983	5,799	1,000	69,900
Square feet	1720	642	500	5,000
No. of bathrooms	1.990	0.6536	1	5
No. of bedrooms	3.236	0.8258	1	5
No. of rooms	6.828	2.279	0	110
Year built	1967	22.16	1873	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 four of the EPA's major criteria air pollutants.⁵ This number 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, or more recent regulations that have allowed for trading in NOx. 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.⁶ 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 more direct 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 also create haze that impairs visibility. Acid rain (or fog), which is produced when SO₂ reacts with water and other chemicals in the air, can also 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 an air district is in compliance with state

⁵Information 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/>).

⁶The 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.

regulations.

Third, we consider nitrogen oxide (NO_x), which is actually a name given to a variety of chemical compounds including nitrogen dioxide, nitric acid, nitrous oxide, nitrates, and nitric oxide. NO_x is particularly damaging because of the ways in which it combines with other pollutants. First, it will combine with SO₂ to create acid rain. It also reacts with ammonia to create fine particles. On its own, NO_x can restrict visibility. Most importantly, NO_x reacts with volatile organic compounds (VOC's) in the presence of sunlight to create photochemical smog, which can damage lung tissue and is extremely dangerous for children and people with asthma.

Finally, we include a measure of 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; repeated exposure to even moderate levels can lead to permanent lung damage. In addition to its health consequences, O₃ also 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 our pollution measures over the sample period. To make the numbers more easily interpretable on the same graph, we express PM10 pollution in $(\mu\text{g}/\text{m}^3) \cdot (1/1000)$.

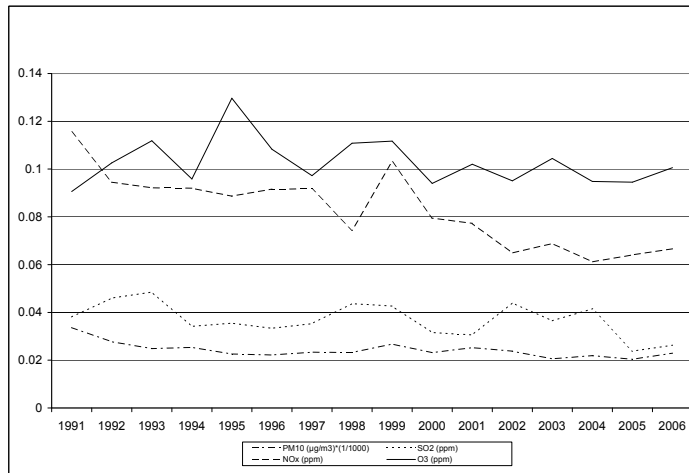
Collinearity problems make it difficult to concurrently measure the MWTP to avoid each of these pollutants. Table 2 describes the correlations across all four pollutants observed at the time of every transaction in our sample. Because some correlations are high (particularly that between PM10 and NO_x), we apply the model to one pollutant at a time, as well as to all the pollutants simultaneously.

Table 2: Correlations of Pollutants

	PM10	SO2	NO _x	O3
PM10	1.0000			
SO2	0.1210	1.0000		
NO _x	0.6766	0.0335	1.0000	
O3	0.2300	0.0948	0.2055	1.0000

A final feature of these pollutants that we do not deal with is the fact that the disutility from each may be a complicated nonlinear function of the concentrations of all the other pollutants.

Figure 2: Median One Hour Maximum Pollution Concentrations



This is a result of the photochemical processes through which they interact. See Muller, Tong, and Mendelsohn (2008) for an example of research that considers these interactions.

4 Results

4.1 Testing the Efficient Housing Market Hypothesis

We provide empirical evidence in support of the efficient housing market hypothesis by running a test that approximates Case and Shiller (1989)'s test using our data set. Let t_a , t_b , and t_c denote the time periods of three housing transactions in the data with $t_a < t_b < t_c$. 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 the county fixed effects. The average returns of previous sales are obtained as the fitted values from the regression of $\ln(p_{j,t_b}) - \ln(p_{j,t_a})$ on the year dummies of the first and the second sales, and the county dummies.

We find coefficients of most attributes are insignificant and even though coefficients of some attributes (square feet and number of bedrooms) are statistically significant, their economic magnitudes in terms of marginal returns are negligible and less than 1%. We also confirm that the

Table 3: Efficient Market Hypothesis Testing (N=16656)

	Avg. return	Lot size	Sqft	Bathroom	Bedroom	PM10	SO2	NOx	O3
Coeff	0.0074	0.0000	-0.0000	-0.0024	0.0077	-0.0022	0.9948	-0.5102	0.2854
t-stat	0.85	1.59	-12.96	-1.50	8.51	-3.36	8.60	-3.57	1.46

t-statistics are calculated from clustered robust standard errors, clustered by county. The dependent variable is the annualized return.

Table 4: Excess returns in dollar amounts

Purchase of home with	Avg. return 10% higher	Lotsize 100 sf larger	Sqft 100 sf larger	Bathroom 1 more	Bedroom 1 more
Home price at purchase	Excess returns per year				
0.4M	297	14	-920	-944	3,086
1M	742	36	-2,300	-2,360	7,715
2M	1,484	71	-4,600	-4,719	15,429

These values are calculated based on estimates in Table 3.

average returns from previous sales do not forecast excess returns. Table 3 summarizes the results.

To give a sense of how much knowing housing attributes could help to generate excess returns, we calculate dollar amounts of excess returns from the estimation results in 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 bathroom is larger by 1, and the number of bedrooms is larger by 1, 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 Table 4 and we see that the amounts are small compared to the home prices. For example, when the home price is 1 million at the time of purchase, one would obtain excess returns of 36 dollars per year by purchasing home with a larger lot size by 100 square feet and would obtain 7,715 dollars of annual excess returns by purchasing home with five bedrooms compared to four bedrooms. One can interpret calculations of excess returns in other cases similarly.

4.2 Measuring 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. These periods correspond to periods of non-attainment, attainment, and non-attainment (respectively) in O3 in the Bay Area, defined according to EPA rules. They also correspond (roughly) to periods of depreciation, appreciation, and very rapid appreciation in this housing market.

We report results for three different econometric models. First, we carry out constrained specification of the model described in equation (5). We restrict $\gamma(1, 3) = \gamma(1, 2)\gamma(2, 3)$, and $h_t(x_{j,t}) = x'_{j,t}\beta$. $\psi_{t,t'}$ replaces the intercept in equation (5), $\psi_{t',t} = \alpha_{t'} - \gamma(t, t')\alpha_t$, and $Z'\delta_{t,t'}$ controls flexibly for any attributes that do not vary over time.⁷ This implies the following specification:

Efficient Housing Market Model

$$\begin{aligned}\ln(p_{j,3}) &= \psi_{2,3} + \gamma(2, 3) \ln(p_{j,2}) - x'_{j,2}\gamma(2, 3)\beta + x'_{j,3}\beta + z'_j\delta_{2,3} + \eta_{j,2,3}, \\ \ln(p_{j,3}) &= \psi_{1,3} + \gamma(2, 3)\gamma(1, 2) \ln(p_{j,1}) - x'_{j,1}\gamma(2, 3)\gamma(1, 2)\beta + x'_{j,3}\beta + z'_j\delta_{1,3} + \eta_{j,1,3}, \\ \ln(p_{j,2}) &= \psi_{1,2} + \gamma(1, 2) \ln(p_{j,1}) - x'_{j,1}\gamma(1, 2)\beta + x'_{j,2}\beta + z'_j\delta_{1,2} + \eta_{j,1,2},\end{aligned}\tag{18}$$

where the subscripts $\{1, 2, 3\}$ correspond to each of the three time periods, $x'_t \equiv \{PM10, SO2, NOx, O3\}$ and z includes the housing attributes described in Table 1 and a vector of county fixed effects. Depending upon in which two of these 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 when $t = 1$ and $t' = 3$, and the third equation applies when $t = 1$ and $t' = 2$.

Given the linearity of this specification, we implement a 2SLS approach to deal with the endogeneity of $x_{t'}$. In particular, we first estimate the following regression equations, which describe the evolution of the pollution variables

$$\begin{aligned}x_{j,3} &= \pi_0 + \pi_1 x_{j,2} + County'_j\vartheta + \zeta_{j,2,3}, \\ x_{j,3} &= \pi_0 + \pi_1(\pi_0 + \pi_1 x_{j,1} + County'_j\vartheta) + County'_j\vartheta + (\pi_1\zeta_{j,1,2} + \zeta_{j,2,3}), \\ x_{j,2} &= \pi_0 + \pi_1 x_{j,1} + County'_j\vartheta + \zeta_{j,1,2}.\end{aligned}$$

⁷In 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 $\delta_{t,t'} = \phi_{t'} - \gamma(t, t')\phi_t$.

where *County* denotes a vector of county dummies. 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 $t' = 2$. Note that we restrict the coefficients on the county dummies, along with the other coefficients governing the evolution of $x_{j,t}$, to be the same over time periods. We then use the predicted values of $x_{j,t'}$ pollutants based on information in period t when estimating equations (18).

In addition, we also estimate a house fixed effect model, that incorporates the same constraint that the derivative of $\ln(P)$ with respect to each pollutant is constant over time. The house fixed effect model is the traditional approach taken with panel data. Whereas our model is intended to control for time-varying correlated unobservables, the house fixed effect model can only control for time-invariant correlated unobservables. To the extent that the answers from the two approaches differ, the relative importance of time-varying versus time-invariant unobservables can be inferred.

House Fixed Effect Model

$$\begin{aligned}\ln(p_{j,3}) - \ln(p_{j,2}) &= \rho_{2,3} + (x_{j,3} - x_{j,2})'\beta + z_j'\chi_{2,3} + u_{j,2,3}, \\ \ln(p_{j,3}) - \ln(p_{j,1}) &= \rho_{1,3} + (x_{j,3} - x_{j,1})'\beta + z_j'\chi_{1,3} + u_{j,1,3}, \\ \ln(p_{j,2}) - \ln(p_{j,1}) &= \rho_{1,2} + (x_{j,2} - x_{j,1})'\beta + z_j'\chi_{1,2} + u_{j,1,2}.\end{aligned}\tag{19}$$

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

Finally, we estimate a simple cross-sectional model, maintaining the assumption used in the previous two models that the slope of the hedonic price function is constant over time. Unlike the previous two models, however, this approach does nothing to control unobservables (time-varying or time-invariant) that are correlated with pollution:

Cross-Sectional Model

$$\begin{aligned}\ln(p_{j,3}) &= \alpha_3 + x'_{j,3}\beta + z'_j\phi_3 + \xi_{j,3}, \\ \ln(p_{j,2}) &= \alpha_2 + x'_{j,2}\beta + z'_j\phi_2 + \xi_{j,2}, \\ \ln(p_{j,1}) &= \alpha_1 + x'_{j,1}\beta + z'_j\phi_1 + \xi_{j,1}.\end{aligned}\tag{20}$$

Of the pollutants that we study, particulate matter has received the most attention in the hedonics literature. Chay and Greenstone (2005) provide a good set of comparison results for this pollutant. They measure the value of total suspended particulate (TSP) reductions by looking at

Table 5: Implicit Price of Pollution (N=74,892)

	A. Efficient Housing Markets Hypothesis								
	Without Instruments*			With Instruments*			Single Pollutant†		
	Coeff	Elast.§	WTP‡	Coeff	Elast.§	WTP‡	Coeff	Elast.§	WTP‡
PM10 ($\mu g/m^3$)	-0.0082 [0.0003]	-0.1844	-311.2	-0.0057 [0.0003]	-0.1290	-217.8	-0.0068 [0.0003]	-0.1544	-260.6
S02 (ppm)	-2.3851 0.0790	-0.0838	-90.8	-2.2216 [0.0770]	-0.0780	-84.6	-2.4969 [0.0779]	-0.0877	-95.1
O3 (ppm)	-2.1489 [0.0687]	-0.2135	-81.8	-1.5984 [0.0598]	-0.1588	-60.9	-1.5791 [0.0588]	-0.1569	-60.1
B. Fixed Effects Model									
PM10 ($\mu g/m^3$)	0.0035 [0.0003]	0.0790	133.3				0.0031 [0.0003]	0.0695	117.3
S02 (ppm)	-1.0467 [0.0787]	-0.0368	-39.9				-0.9374 [0.0762]	-0.0329	-35.7
O3 (ppm)	-1.6736 [0.0630]	-0.1663	-63.7				-1.7587 [0.0627]	-0.1747	-67.0
C. Constrained Cross Sectional Model‡									
PM10 ($\mu g/m^3$)	-0.0054 [0.0006]	-0.1216	-205.3				-0.0064 [0.0006]	-0.1444	-243.8
S02 (ppm)	-1.4237 [0.1484]	-0.0500	-54.2				-1.7234 [0.1453]	-0.0605	-65.6
O3 (ppm)	-1.5646 [0.0931]	-0.1555	-59.6				-1.5837 [0.0918]	-0.1574	-60.3

Heteroskedasticity robust standard errors in brackets. Controls for prior sales price (in the efficient housing market model), lot size, square feet, number of rooms, number of bedrooms, number of bathrooms, year built and county fixed effects also included but not reported. * with instruments indicates that instrument for future pollution with measures at time of purchase. † 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 means of pollutants, which are 22.55 for PM10, 0.0351 for SO2, 0.0994 for O3. ‡ Willingness to pay calculated for marginal 1 $\mu g/m^3$ change in PM10 and 1 ppb change in other pollutants, annualized at rate of 0.07 for average house price of \$ 543,896. ‡ constrained cross sectional estimates restrict the implicit price for the pollutant to be the same at different periods, but do not control for unobserved product attributes.

Table 6: Implicit Price of House Attributes (N=74,892)

	Efficient Market IV		Fixed Effect		Const. Cross Section	
	Coeff	Elast. [§]	Coeff	Elast. [§]	Coeff	Elast. [§]
	Sold in period 3 and 2				Sold in period 3	
Lot size	2.34E-06*	5.7995	1.04E-06*	2.9117	1.10E-05*	10.5781
	[4.04E-07]		[3.56E-07]		[1.04E-06]	
Square feet	6.48E-05*	0.1115	7.45E-06	0.0128	4.73E-04*	0.8130
	[5.91E-06]		[4.68E-06]		[1.27E-05]	
No. of bedrooms	2.91E-04	0.0009	5.20E-03	0.0168	-1.87E-02*	-0.0606
	[3.12E-03]		[2.91E-03]		[6.98E-03]	
No. of rooms	5.80E-04	0.0001	-2.27E-04	0.0000	3.64E-03	0.0005
	[8.09E-04]		[7.27E-04]		[4.03E-03]	
No. of bathrooms	8.19E-03	0.0163	7.89E-03	0.0157	3.44E-02*	0.0684
	[4.46E-03]		[4.21E-03]		[9.75E-03]	
Year built	-1.16E-03*	-2.2793	-1.26E-03*	-2.4744	-1.46E-03*	-2.8680
	[1.12E-04]		[1.07E-04]		[2.43E-04]	
	Sold in period 3 and 1				Sold in period 2	
Lot size	6.06E-06*	0.0423	9.65E-07*	0.0067	1.10E-05*	0.0766
	[4.41E-07]		[4.09E-07]		[6.93E-07]	
Square feet	1.66E-04*	0.2850	-3.24E-05*	-0.0558	4.64E-04*	0.7989
	[5.94E-06]		[6.70E-06]		[8.42E-06]	
No. of bedrooms	6.34E-03*	0.0205	1.48E-02*	0.0480	-3.24E-02*	-0.1047
	[3.19E-03]		[3.88E-03]		[4.63E-03]	
No. of rooms	2.61E-03*	0.0004	-2.43E-03	-0.0004	1.52E-02*	0.0022
	[1.19E-03]		[2.48E-03]		[2.77E-03]	
No. of bathrooms	2.96E-02*	0.0589	1.10E-02*	0.0219	2.89E-02*	0.0576
	[4.58E-03]		[4.87E-03]		[6.33E-03]	
Year built	-2.67E-03*	-5.2474	-2.61E-03*	-5.1313	-1.03E-03*	-2.0183
	[1.18E-04]		[1.28E-04]		[1.60E-04]	
	Sold in period 2 and 1				Sold in period 1	
Lot size	2.34E-06*	0.0164	1.04E-06*	0.0072	1.10E-05*	0.0765
	[4.04E-07]		[3.56E-07]		[1.04E-06]	
Square feet	6.48E-05*	0.1115	7.45E-06	0.0128	4.73E-04*	0.8130
	[5.91E-06]		[4.68E-06]		[1.27E-05]	
No. of bedrooms	2.91E-04	0.0009	5.20E-03	0.0168	-1.87E-02*	-0.0606
	[3.12E-03]		[2.91E-03]		[6.98E-03]	
No. of rooms	5.80E-04	0.0001	-2.27E-04	0.0000	3.64E-03	0.0005
	[8.09E-04]		[7.27E-04]		[4.03E-03]	
No. of bathrooms	8.19E-03	0.0163	7.89E-03	0.0157	3.44E-02*	0.0684
	[4.46E-03]		[4.21E-03]		[9.75E-03]	
Year built	-1.16E-03*	-2.2793	-1.26E-03*	-2.4744	-1.46E-03*	-2.8680
	[1.12E-04]		[1.07E-04]		[2.43E-04]	

Heteroskedasticity robust standard errors in brackets. * indicate statistically significantly different from 0 at 95% level. These parameter estimates are taken from the same regressions for which the pollutant coefficients are reported in Table 5: panel A, column 4; panel B, column 1; panel C, column 1. [§] Elasticities calculated at means of house characteristics as reported in Table 1.

Table 7: Implicit Price of Pollution: Unconstrained Cross Sectional Estimates (N=74,892)

	Period 1			Period 2			Period 3		
	Coeff	Elast. [§]	WTP [‡]	Coeff	Elast. [§]	WTP [‡]	Coeff	Elast. [§]	WTP [‡]
A. Regressions Controlling for All Pollutants									
PM10 ($\mu g/m^3$)	0.0144 [0.0006]	0.3252	548.9	0.0180 [0.0009]	0.4055	684.6	-0.0337 [0.0008]	-0.7605	-1283.8
S02 (ppm)	2.4958 [0.2462]	0.0877	95.0	8.7280 [0.3023]	0.3066	332.3	-3.6673 [0.1232]	-0.1288	-139.6
O3 (ppm)	-2.2252 [0.1769]	-0.2211	-84.7	-0.2005 [0.1103]	-0.0199	-7.6	-2.2920 [0.1325]	-0.2277	-87.3
B. Separate Regression for Each Pollutant									
PM10 ($\mu g/m^3$)	0.0170 [0.0006]	0.3836	647.6	0.0267 [0.0008]	0.6032	1018.3	-0.0372 [0.0008]	-0.8384	-1415.4
S02 (ppm)	2.7123 [0.2372]	0.0953	103.3	10.7170 [0.2861]	0.3765	408.0	-4.7565 [0.1256]	-0.1671	-181.1
O3 (ppm)	-2.3586 [0.1703]	-0.2343	-89.8	-0.4110 [0.1082]	-0.0408	-15.6	-2.5400 [0.1157]	-0.2524	-96.7

Heteroskedasticity robust standard errors 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 means of pollutants, which are 22.55 for PM10, 0.0351 for SO2, 0.0994 for O3. [‡] Willingness to pay calculated for marginal 1 $\mu g/m^3$ change in PM10 and 1 ppb change in other pollutants, annualized at rate of 0.07 for average house price of \$ 543,896.

how the median house price in a county varies with TSP concentration, assuming a national housing market and ignoring the wage gradient described in Roback (1982).⁸ Looking at cross-sectional evidence from 1970 and 1980, they find statistically weak correlations between TSP and county median house prices, and the sign of the estimated effect varies with the specification. The same is true even when fixed effects are used to control for permanent unobserved differences between counties. This suggests a need for controls to deal with time-varying unobservables. Because Chay and Greenstone treat the US as a single housing market, they are able to observe variation in EPA attainment status for TSP across locations within a year. When the EPA declares an area to be out of attainment, local officials are required to implement strict regulations to bring pollution levels down. Using a variety of tests, Chay and Greenstone find that this creates an exogenous source of variation in TSP (i.e., one that only affects housing values through its effect of reducing TSP); TSP attainment status can therefore be used as an instrument. Because we restrict our attention to what we believe to be a clearly defined housing market (i.e., the Bay Area), and since that area is treated as a single entity for the purposes of air pollution regulation, within-year variation in attainment status is not available to us to use as an instrument. We instead rely on the efficient housing market hypothesis. Using their instrumental variables strategy, Chay and Greenstone find a housing price elasticity with respect to TSP between -0.2 and -0.35. Those results are statistically significant and substantially larger than previous results in the literature.⁹ Our results (in particular, when we consider PM10 in isolation, as is done for TSP by Chay and Greenstone) are more similar than the rest of the literature; we find an elasticity of -0.15. This number falls to -0.10 when we consider all four pollutants together. Like Chay and Greenstone (2005), our PM10 results are highly variable across the cross-sectional and fixed effect specifications.

Looking at the remaining pollutants, we can draw some conclusions about the nature of correlated unobservables. For example, willingness to pay for a marginal reduction in O₃ is virtually identical in the efficient housing market, fixed effects, and cross-sectional specifications, regardless of whether it is considered alone or together with the other pollutants. This suggests that unobservables (time-varying or time-invariant) are not a serious concern. In the case of SO₂, however, the efficient housing market marginal willingness to pay is significantly larger than both

⁸Prior 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>)

⁹Previous papers that did not control for time-varying unobservables (many of which ignored time-invariant unobservables as well) found smaller (often counter-intuitively signed) marginal willingnesses to pay. Smith and Huang (1995) survey the literature from 1967 to 1988 that values marginal reductions in particulate matter in the context of a meta-analysis. They find elasticities that tend to lie between -0.04 and -0.07.

the fixed effects and cross-sectional estimates, regardless of whether the pollutants are considered individually or together. This suggests that time-varying unobservables are an important source of potential bias for SO₂.

Finally, in the case of NO_x, marginal willingness to pay is similar for the efficient housing markets and fixed effects estimators when all pollutants are considered together. However, when NO_x is considered by itself, the efficient housing market elasticity rises (from -0.04 to -0.06) while that associated with the fixed effects estimates falls (from -0.05 to 0.00). This suggests important correlations between NO_x and the other time-varying pollutants.

5 Conclusion

Our paper demonstrates a new approach to controlling for unobserved product attributes in hedonic models. In particular, we show how the assumption that a market is informationally efficient 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 informational efficiency 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 it can be easily applied to data from a single housing market, and that it relies on a set of testable assumptions in that the available information should not predict excess returns. We find evidence that the housing market in the Bay Area is in fact informationally efficient.

We estimate the implicit price of four of the EPA's criteria air pollutants (PM₁₀, NO_x, SO₂, and O₃). In contrast to fixed effects methods (which just control for fixed unobservables) or cross sectional methods (which ignore unobservable attributes), our estimates of the implicit price are generally larger in magnitude and indicate that consumers value lower levels of pollution. Particularly in the case of PM₁₀, it appears that failing to control for omitted product attributes or only controlling for fixed unobserved attributes significantly understates the potential benefits of policies aimed at reducing pollution. Our estimates suggest that other pollutants (e.g., SO₂ and NO_x) are also prone to bias from ignoring time-varying unobservables. Unobservables (time-varying or time-invariant) associated with ground-level ozone, on the other hand, do not appear to be a source of bias.

To be clear, while our approach seems to work well in this context, we are not claiming

that it is 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 the assumptions depends on the application and data at hand. Our approach may be preferable when a source of quasi-randomness cannot be found, generates insignificant estimates or is not plausibly exogenous. Quasi-randomness may be preferable if there is reason to suspect that the housing market was failing to function efficiently. In empirical work, we advise applied researchers to test the sensitivity of results to alternative identifying assumptions when they are available.

References

- Ai, C. and X Chen (2003), "Efficient Estimation of Models With Conditional Moment Restrictions Containing Unknown Functions," *Econometrica*, 71(6), 1795-1843.
- 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.
- Black, SE (1999), "Do Better Schools Matter? Parental Valuation of Elementary Education," *The Quarterly Journal of Economics*, 114(2): 577-599.
- 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.
- 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.
- 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.
- 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.
- 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.
- 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.