Hedonic Methods and Housing Markets

Chapter 2: A Brief Survey and Interpretation of Hedonic Parameters

Edward Coulson
Department of Economics
Penn State University
fyj@psu.edu

In this chapter we discuss regression-based methods for estimating hedonic prices. Hedonic regressions are a way of statistically estimating the relationship between a property's characteristics and its market value, and thus a way of determining the value of the property itself. The result of this regression estimation is a hedonic function like equation (1.1) that can solve the problem of appraisal, as discussed in the previous chapter. By way of introducing hedonic regression, we briefly take a look at the primary method of solving the appraisal problem for single family homes, comparable sales[1].

A caricature of comparable sales would include the following steps. A property is in need of appraisal. A few salient characteristics of the property are registered– this may be as few as square feet, age, and number of bathrooms, but will likely include more. Recent sales records are searched for properties that are (a) in the vicinity of the appraisal, and with (b) a "comparable" set of salient characteristics. By comparable, we can not mean that the two properties are exactly the same; if properties were exactly the same, there would be much less need for hedonic analysis. The prices of these comparable sales are used as benchmarks for providing a valuation of the unit under investigation. If the comparisons were exact, the benchmarking would be easy, but this is not going to be the case. . One appraisal handbook suggests finding comparable sales that match "neighborhood, zoning, location, date of sale, size of dwelling, lot size, room count, quality, condition, amenities, and price" in that order (Hill and Hill, 1990). Interestingly, to these authors it seems far more important to make the properties most comparable in their spatial characteristics rather than their physical ones. This seems to be the case because no comparison is going to be

---

[1]Appraisal is sometimes done using the *income method*. The income method determines the present discounted value of the income that the property could generate. This method is, on that account, limited to rental properties (usually multi-unit properties). Note that for rental properties there is (typically) a transaction each month, and so the appraisal problem is solved.

perfect, and "subjective adjustments" need to be made to the comparable properties' sale prices. It seems evident from their subsequent discussion that "subjective adjustments" of physical characteristics is far easier than adjustments due to spatial differences. Six pages are devoted to comparing unlike neighborhoods; perhaps one -tenth of that amount to physical adjustments.

2.1 The basics of hedonic regression

A differentiated, or heterogeneous commodity is one in which the characteristics of the product are fundamental to its value in the marketplace. All commodities are heterogenous to a certain extent but the heterogeneity is particularly apparent in the real estate market. We define a hedonic function to be a mathematical form that links the characteristics, collectively defined as X to the price of the real estate product, P. Thus:

$$P = H(X)$$

The easiest way to think about the hedonic function is to follow the lead of Haas or Andrew Court, and for the time being assume that the way you combine the characteristics is by making H(X) a *linear* function:

$$P = a_0 + a_1 X_1 + a_2 X_2 + \ldots + a_k X_k \qquad (2.1)$$

where X_1 through X_k are the attribute levels for k selected attributes, and a_1 through a_k are the weights assigned to the particular attribute. Suppose that X_1, the first characteristic is yard area, or lot size. The linear function (2.1) implies that if X_1 goes up by one square foot, the price of the property rises by a_1 dollars. In the language of calculus, we are saying this:

$$\partial P / \partial X_i = a_i ,$$

i.e. that the change in P due to a change in X is constant and equal to a_1 and it is on this basis that we say that the price of a square foot of land is a_1 dollars.

The question we address is how to determine the hedonic prices for a particular housing market. Hedonic studies generally use a statistical tool called multiple regression analysis. About regression there is much that can be said, so the present work will leave much unsaid. The following is merely an outline of the process[2].

Databases

We first need a database. A database is a collection of observations on a set of real estate properties, which includes the required characteristic and price information. There are several such sources[3]:

1. Transactions data: This would be a database of transactions in a particular real estate market, often the outgrowth of the local real estate market's Multiple Listing Service, or from the local taxing authority's records. Any listing of actual transactions is, as such, highly desirable for hedonic function estimation, because the goal of the analysis is to predict transaction prices.

2. Surveys: Because they often come from local authorities or realtors, transactions are local in nature. This will certainly suffice when interest is focused on some aspect of the local real estate market. For some questions it will be of interest to create hedonic functions for more broadly based markets. Since massing a number of transactions databases into a unified whole is quite costly, in time and/or money, one may have recourse to surveys taken by agencies of the federal government or other institutions. Prominent among these is the *American Housing Survey*, conducted biannually by

---

[2]There are dozens of excellent textbooks which will introduce the reader to the mechanics of regression analysis. Particular favorites are Wooldridge (2000) and Stock and Watson (2006)

[3]Green and Malpezzi (2003) contains an extensive discussion of real estate data sources.

the Bureau of the Census, which contains information on a national sample of tens of thousands of housing units and their characteristics. The major drawback of the AHS for hedonic analysis is that the housing values are estimates by the residents, and are not necessarily derived from actual transactions. (Kiel and Zabel (1999) find that residents over-estimate the value of the property, but that this over-estimate is not related to any particular characteristic, so that estimation of hedonic price functions is not affected.)

3. Appraisal data: The most complete source of housing price data in local markets is often through property tax assessor data. Because of the necessity of taxing every piece of property, the assessor has complete data on the characteristics of each property in his or her jurisdiction as well as the assessed value. Like the AHS or other surveys, the value of the property is an estimate– literally, an assessment– and this can be problematic: where did the assessment come from? It might have come from the assessor's own hedonic estimate, in which case any use of it for a fresh hedonic study may merely replicate what the assessor has found. Or it may merely replicate idiosyncratic assessment practices in a particular jurisdiction. However there are occasions when the use of appraiser's data may be appropriate, particularly when evaluating the hedonic price of an attribute with little representation in smaller transactions or survey databases.

However the data is gathered, the researcher needs to put into a database[4]. (S)he must then decide which characteristics will be part of the function, and do the necessary checking of the data to ensure consistency and accuracy to the extent possible. With transactions data, for example, one usually would like to rid the database of any transaction that is not arms-length; with survey data one might want to inspect the data for coding errors, such as houses with 50 square feet of living space;

[4]There are several data management/statistical software packages that provide regression analysis. Those that are geared particularly toward econometric regression and also provide more advanced extensions include Stata, LIMDEP, RATS and Eviews

and the like.

Regression

The paramters of the function must be estimated: what are the best guesses of a_1 through a_k in equation (2.1)? One might think that this is like an algebra problem, in the sense that the parameters are unknowns to be solved for. But as long as the number of observations in the sample exceeds the number X's (which is a truism) there is no combination of a's that makes the hedonic function true for all observations. That is, let N be the number of observations in the sample and index the observations by i=1...N. Then we might wish that

$$P_i = a_0 + a_1 X_{1i} + a_2 X_{2i} + \ldots + a_k X_{ki}$$

were true for every i, but it can't be so. The number of equations (N) is greater than the number of unknowns (k+1). What can be true instead is that each of the N equations misses by some amount e_i:

$$P_i = a_0 + a_1 X_{1i} + a_2 X_{2i} + \ldots + a_k X_{ki} + e_i$$

and then the estimation problem is one where the goal is to make these *errors* as small as possible through appropriate choice of the a_i's. This provides the basis for claiming that the hedonic prices are estimated in a manner that best represents how those characteristics relate to the sale prices of the units. Regression analysis does this in a particular way, by choosing the a_j's to minimize *the sum of the squared errors*[5].

---

[5]More formally, the following is typically proposed. Again, retaining the assumption of linearity, one assumes that the true hedonic relationship is given as

For any house with attributes X*, in or out of the data set, we can estimate the price as

$$P^* = \hat{a}_0 + \hat{a}_1 X_1^* + \ldots + \hat{a}_k X_k^*$$

where P* is now becomes a forecast or appraisal for a property with attributes X*. Note that the error term has been set at zero. This is appropriate because this is the error term's expected value; the least squares technique ensures that this will happen in the course of calculating the parameters.

To see how this works in a very simple database, see Spreadsheet 1, which gives a example database of 10 units. The columns are labeled with abbreviations for each attribute: Number of Bathrooms, Number of Rooms, and Exterior Square Feet. The last column is the market price of the housing unit (P). If the hedonic function were to be estimated on the basis of just this data (and if it were linear) it would have the form

---

$$P = Xa + e$$

where P is an nx1 vector of observations on housing prices, X is a matrix with n rows and k columns, representing for each of the n observations, values for each of k different housing attributes, and e is the nx1 vector of error terms. The usual, or "classical" regression assumptions are that (a) X is of full rank; (b) zero correlation between the X variables and the error term; and (c) the error vector is normally distributed with constant mean of zero, constant variance σ^2 , and zero covariances for each pair of errors. Below, we examine the consequences of these assumptions being violated, but for now assume that the regression retains its classical form.

Multiple regression analysis chooses the a_i's such that

$$S = \Sigma_i e_i^2$$

is minimized. In matrix language this is done through the formula

$$\hat{a} = (X'X)^{-1} X'P$$

where $\hat{a}$ is a column vector containing numerical estimates of each of the hedonic coefficients. .

$$P = a_0 + a_1 * LivingArea + a_2 * ExteriorSquareFeet + a_3 * NumberofBaths$$

Just by looking at the data displayed in the spreadsheet we can observe that there does seem to be a relationship between the price and the quality of the units, as measured by these attributes. Generally speaking, one can for instance observe that the larger the Living Area the higher the price. Demonstrating this pictorially, Figure 1 plots the prices versus the Living Area for the 10 units in the database. The numbers in the figure reference the observation number from the Spreadsheet; there seems to be a clear positive relationship between these two variables. One might calculate the hedonic function simply on the basis of this single characteristic:

$$P = a_0 + a_1 * LivingArea$$

For this data, the line (as given by the formula in footnote 4) is

*P = 10448.54+43.61*Living Area*

and this is the straight line plotted in Figure Note that the intercept is indeed $10,448 and the slope of the line – the price of a square foot of Living Area – is $43.61. The square markers (with accompanying observation numbers) in Figure 2 represent the predicted prices of houses in the database. Those that are on the estimated line above or below the actual data (with the same accompanying number) are the actual property prices. Thus the vertical distance between the actual and predicted price is the error for that transaction. Some property prices are very well-predicted. Indeed, for observations 5 and 6 the error is so small that the actual and predicted values are indistinguishable on the graph. Note further that observations 2 and 3 have nearly identical living area sizes; because size is the only attribute in this example, they have nearly identical predictions. Since observation 2's actual price is somewhat lower than that of observation 3, it seems clear that this model is incomplete. This is even more evident when observations like number 10 are examined. This home has the highest price in the sample even though it is not nearly the largest, and because of that the error for this observation is large in absolute value; its price is seriously

underpredicted. An examination of the Spreadsheet immediately discloses the reason for this, that it has by far the largest lot of any property, and this clearly contributes to its high price. We can see early on that the omission of important attributes can play a large role, and we return to this in more detail later on.

If we therefore admit the other available variables from the spreadsheet into our regression model the least squares line will be of the form:

*P = -50,875+44.34*Living Space + 0.89*Exterior Square Feet+18211*Number of Baths*

Thus, according to this expanded model, a square foot of living space costs $44.34, a square foot of land costs 89 cents, and $18,211 will get you a house with an extra bathroom. The negative value for the intercept term should not concern us. In essence, it is the appraisal for a house with zero square feet on the inside and the outside and zero baths. This is a nonsensical calculation. The last column (labeled Predicted Price 2) of Spreadsheet 1 gives the prediction made by this model for each of the twelve units. Note that the size of the errors is generally much reduced– indeed this an immediate consequence of using more information to predict the prices[6]. In particular, observation number 10's error has been considerably reduced by taking into account its large lot. There are still problems, though. Compare observations 2 and 3 once more. With identical interior sizes, the predictions from the first equation were quite similar. In taking into account lot size, the predicted price for observation 2 is somewhat higher (noting that they have the same number of bathrooms). And yet the *actual* price for observation 2 is lower than for number 3. There must yet other property attributes that are missing from this model that cause this to be the case.

Serious attempts at hedonic modeling typically include a wealth of information on the

---

[6]To be more precise, the average squared error must fall with the addition of additional components to the model. The previously included variables can explain whatever they explained before, and the new information can only improve upon this, on average. This is not to say that a prediction of an individual property price might not deteriorate.

characteristics of the houses in the database, much more so than the primitive attempts above. Attention is turned to an example of this.

A better example

Table 1 reproduces a typical hedonic study, albeit an unusually complete one, especially for its time. The source is Palmquist (1983), who created a data set using FHA mortgage insurance applications, in this case from the Atlanta, GA metropolitan area. The census tract for each unit was given in the FHA record, allowing Palmquist to link each observation with tract characteristics and other spatial data. The first column describes the characteristic and the number in the second column are the hedonic prices that are estimated.

Take a few simple examples: the coefficient for "Lot Size (sq. ft.) is given as 0.0813, meaning that each square foot of lot adds another 8.13 cents to the price of the house. Similarly, every additional bathroom adds an additional $1,821.32 to the price and every garage space adds $1451.09 (rounded off).

One thing to note about *linear* hedonic functions (as all of our examples so far have been) is that the hedonic price of characteristics is constant. This may not accurately reflect what happens in the marketplace. However, with a bit of additional specification one can overcome this limitation. An example is contained in the third and fourth rows of Table 1, which contain entries for "Improved Living Area" and "Improved Living Area^2". This is exactly what it seems; not only is Living Area as before inserted into the function, but also for each entry in the database, the number of square feet is squared, and included as a separate "characteristic". The purpose of this is to allow this single attribute to have a nonlinear impact on price. The idea is that there are "diminishing returns" to interior area, so that the price of the house increases at a decreasing rate. Palmquist's Atlanta hedonic price function appears in part as

$$P_i = \ldots 15.0765 * InteriorArea - 0.22 * InteriorArea^2 \ldots$$

therefore the "price" of Interior Area is

$$\partial P / \partial InteriorArea = \$(15.0576 - .0044 * InteriorArea)$$

That is, increases in Interior Area will continue to add to the price of the house, but the incremental addition to the house price diminishes (by 0.44 cents) with each additional square foot. This quadratic representation is fairly common in hedonic studies because the phenomenon it captures is thought to be applicable to numerous attributes. It does however indicate that at some point houses can become so big that the bigger they are, the lower their price, holding other things constant. That point occurs when the price/derivative above reaches zero; in the present case this is 3422 square feet. A 3600 square foot house would be cheaper than a 3500 square foot house with the same attributes. This is an unavoidable consequence of using squares in the hedonic function, however it is often the case that these maximum prices occur outside the range of the data under consideration, and therefore perhaps becomes an irrelevant consideration, which is presumably the case here. The number of houses in Atlanta in which applied for FHA mortgage insurance and also were larger than 3422 square feet must be fairly small.

A number of the rows in Table 1 begin with the entry "=1 if". These are attributes which are either present or absent in the housing unit and cannot otherwise be enumerated in the same way that space or the number of bathrooms can. Called "dummy", "binary" or "indicator" variables, the rows are to be read as indicating the price increase or decrease which is predicted due

to the presence of the indicated variable. Thus the presence of a dishwasher adds about $1710, a fireplace adds about $1114, and a swimming pool roughly $3274.

The condition of the property was also indicated. Properties were categorized into one of four grades: Excellent, Good, Fair and Poor. In the Atlanta sample, none of the houses were classified as Poor, so this indicator can be discarded. The hedonic model had indicator variables that equalled one if the house was in Excellent Condition or Fair Condition. It is important to note that for any set of variables that groups all of the observations into one of those categories, one of the categories must be omitted from the hedonic model. The excluded category is implicitly represented in the intercept term, and the coefficients listed for Excellent and Fair Condition are to be understood as the change in price as the condition changes from Good to the indicated category. Thus a unit in Excellent condition has an expected price $1007 higher than a similar unit in Good condition, and a unit in the Fair category will see a price drop of $2227.

Similarly, note that there is an indicator variable for "brick or stone externior" and this row of Table 1 indicates that a house with such a facade will increase in price by $1852. Compared to what, though? In fact, because there are no other indicators for other types of exteriors, this dollar amount is to be compared to every other type of exterior, all of which are included in the "other" category, the value of which is folded into the intercept term. All exteriors that are neither brick nor stone are thus assumed to have the same hedonic value.

Finally, Palmquist's hedonic estimates indicate that neighborhood and environmental characteristics are important to the price of a housing unit. The racial, income, age, education and job type distributions are all used as characteristics in the hedonic index[7]. Also the distribution of the physical attributes of structures is importance: the number of older buildings the number of job sites and the level of crowding are all included. Finally, the level of air pollution in the tract is

[7]Not all of these are significant determinants in the usual statistical sense.

included the model to reflect environmental quality of the location

Instant Appraisal

The estimates herein provide a ready-made appraisal calculator for 1977 Atlanta. One can simply enter in the given implicit prices into the spreadsheet and if it is desired to appraise a house with a specific set of characteristics, X* one need only set up the spreadsheet accordingly. For example the third column of Table 1 proposes a set of "typical" characteristics-- an X*-- and the fourth column is the product of the second and third, and gives the contribution of each of the specified characteristics-- that is, $a_j X_j$*-- to the appraisal. Thus we have:

$$\hat{P}_i = \hat{a}_0 + \hat{a}_1 X_{1i} + \hat{a}_2 X_{2i} + \ldots + \hat{a}_k X_{ki}$$

where $\hat{P}$ is the *appraisal* of the property, as distinguished from the actual value; the difference between the two is the error term. This is simply a more sophisticated version of the predictions made in the previous section. The estimated weights are designated as $\hat{a}_j$ to distinguish them from the putatively "true" weights a_j. The sum of these contributions, given at the bottom of column 4 is the predicted house price for the house with X*, in this case $35,695, Computer-assisted mass appraisal will do this sort of thing with a few mouse clicks or lines of code for any number of observations in a data base. Note that in the appraisal process the error term is set to zero, because for any given regression estimate, that is its average, or expected value. Of course, for any property that is actually in the database we know what the error term is because we know the real sale price.

The important issue-- and the whole point of appraisal-- is that we wish to apply the above formula to properties not in the data set-- for which P is unknown. In such a case one does the obvious thing and apply the formula to the new unit and assume that the value of the error is zero.

That is the appraised value. The virtue of regression analysis is that it leads to predictions that have several appealing traits, *given the information in the database.*

1. On average the appraisal-predictions will be correct. This is because the average value of the error is zero.
2. The predictions have a "minimum variance" property. Under certain statistical conditions, they will make errors that are smaller than any other prediction method.

So predictions that arise from regression models are best, in this sense.

Statistical adequacy

How accurate will these appraisals be? Within the sample, this can be easily measured. Regression analysts generally have two measures of how well the regression coefficients capture the variation of prices in the sample. The first is the coefficient of multiple determination-- R^2. This is a measure of the percent of the variation in housing prices is "explained" by the attributes. In the case of Palmquist's Atlanta model this turns out to be about 80% . The second measure is the "standard error of the regression". This is the square root of the average squared error made by the regression. It is useful as it provides a benchmark for the model's accuracy; if the errors are normally distributed one can make the statement that the true price is within approximately two standard errors, plus or minus, of the estimated value[8].

The precision of the individual coefficients is measured by their own standard errors. Again,

[8]More formally, with a large enough sample, the probability that the true value is within 1.96 standard errors of the estimated value is 95%. This is a typical standard by which to measure accuracy, but it is neither a universal, nor even a desirable one. Please see the econometrics text from footnote 1 for more on this complex subject. Palmquist does not provide this datum for his estimates.

under the usual assumption of normally distributed errors, the true value of the coefficient is roughly within two standard errors of this estimate. If this plus/minus interval does not contain zero, then the usual language is that the particular attribute is *statistically significant* in the determination of housing prices.

Why regression-based appraisal will go wrong

Why should there be errors? Why doesn't the hedonic function give back the exact sale price for every property. Roughly speaking, the mistakes that regression functions make can be divided into two categories: those which affect the variance and those that affect the bias. The first type are those that occur because of the sheer randomness in the world. Recording errors are perhaps the most important of these, but also our lack of knowledge about the financing arrangements, the impatience of the seller, and the like, can cause the price of two otherwise identical houses to be quite different. And so errors will occur precisely because these things are unknown. The more important these things are within the context of a particular housing database, the larger the errors will be. As long as the errors are not unduly associated with particular housing attributes, the hedonic prices are unaffected, though one's ability to appraise will be reduced, because such omissions can increase the standard error of the regression.

A second, perhaps more important reason for mistakes in hedonic analysis is that some of the attributes that go into the formation of the house's price may not be observed. Imagine that the database is derived from a Multiple Listing Service book with actual sale prices. There are any number of characteristics which might therein be recorded, including the improve square footage, the unimproved square footage, number of bathrooms, type of exterior, etc. But there are certain things which might not be included, particularly those involving the attributes of the neighborhood, such as the amount of traffic, the presence of multiple-unit buildings, and the distance to shopping

and work locations. (To be sure, sometimes these data are available as in the Palmquist example above. Addresses are given in MLS listings, and geographical information systems (GIS) can aid the translation of such information into spatial characteristics of the unit, but this is an uncertain process. Spatial issues are discussed further below.) Even careful recording of the physical characteristics may not be enough to completely capture things which can be very important in the marketplace, such as the condition or floorplan of the unit.

Property information is necessarily incomplete and inaccurate. If it were otherwise, prediction would be easy, and appraisers would be redundant. The problem of *omitted characteristics* is one of the most troubling to the hedonic analyst. No matter how complete the listing of characteristics in the database there will be some attributes of the home that remain hidden from the investigator.

A simple example will help to clarify the consequences. Imagine that, in a given housing market, that a swimming pool adds $10,000 to the price of a house, and the presence of central air conditioning adds $5000. These are the a_i's for these two characteristics. For purposes of the example suppose that every house *in the database* has both or neither of these items, and that only the presence of air conditioning is recorded-- air conditioning is one of the X variables, and swimming pools are the omitted characteristic. The regression calculations will then "observe" that every house with air conditioning has a price of $15000 greater than similar houses without the a/c, and give an a_i associated with a/c of $15000. The omission of swimming pools from the list of X's causes an *upward bias* in the estimated price of air conditioning.

In the context of the database this has mild consequences: the errors in the database itself will still average out to zero. Appraisal within the sample will still be as accurate as if you did have knowledge about the presence or absence of swimming pools. The problem arises the first time the hedonic index gets used to predict or appraise a house that has air conditioning but no pool. The

appraiser will, in invoking the formula (2.1) assign a value of $15,000 to air-conditioning, and overappraise the property by $10000.

We can extrapolate from this example to a more general (though still only approximate) rule. When a characteristic is omitted from the hedonic regression, the weights assigned to the attributes that *are* included are biased upwards to the extent that they are positively correlated with the excluded attributes, and biased downwards to the extent that they are negatively correlated with the excluded attributes[9]. This *omitted variable bias*, as it's called, is most damaging when there is a high degree of such correlation. In the example, swimming pools and air conditioning were very highly correlated, so the omission of one from the hedonic function cause the value of the other to be grossly overstated. If the omitted characteristic is totally uncorrelated with anything in the X's then the weights will not have any bias, although the predictions themselves will be less accurate.

Indeed, one of the ways that researchers think their way out of these issues is to say that there are hundreds of such omissions, and that the law of large numbers applied to the hundreds of biases means that they all cancel each other out, with little net impact on the final quality of the appraisal. But this escape hatch is closed when important attributes are omitted from the analysis.

Yet another way out is if the true weight on the omitted characteristic is zero (that is, the omitted characteristic has no value). In that case there is no bias resulting from the omission (and indeed the errors don't get any larger, either). In the contrived example above, if the price of the swimming pool were in fact zero, air conditioning would get its correct price of $5000. The bottom line of course is that the safest thing to do is to make sure that the database includes all of the important characteristics and worry about whether they belong in the hedonic function afterwards.

A different set of problems can arise when the *included* characteristics are correlated with one

---

[9]It's not this simple, actually. The direction of the bias for any of the a_i's depends in a complicated fashion on the correlation of the particular X_i on all of the omitted characteristics and all of the included characteristics.

another. In the above example, suppose that the database and hedonic function did include both swimming pools and air conditioning among the X variables. In this situation there is no way the regression could separate the distinct influences of the two characteristics on the housing price. All that it could determined was that their joint price was $15000. Indeed, a human assessor would have precisely the same problem. Again, within the database this isn't really an issue-- in fact only one characteristic need be included in the model– call it "pool and air-conditioning" – and it will have a weight of $15000. But when a new unit is appraised, and it has only one of these characteristics, the prediction method is going to break down. The unfortunate aspect of this is that if two variables are nearly perfectly correlated throughout the housing market, there is nothing to be done about this problem

The example above basically assumed that pools and air conditioning were perfectly correlated. Such extremely perfect correlations do not occur in real life. There will be some houses with a/c and without pools, and vice-versa, and so breakdown need not occur. The two characteristics can be separate X's in the hedonic function, and separate a_i's can be estimated via regression analysis. Suppose that houses with pools *nearly* always have air-conditioning. Such near- perfect correlations can cause the regression model to provide misleading, and even bizarre results . To see this, imaging that *all but one* of the houses with a/c have pools. That one house is going to determine the price of pools, because it is the only avenue through which the data can provide information on the separate value of air conditioning and swimming pools. If this one particular house is more or less average in every other way this will not be a problem, and the hedonic price of pools and air conditioning will be sensible. However if this particular house is unusual in any other way the hedonic price of pools will reflect not just the pool price but also any other idiosyncratic characteristics of that unit. If, for example, the unit in question was sold at $20,000 below the "market" price because the seller was particularly impatient, then the price of

pools will be estimated to be (roughly speaking) $20,000 less than the true value-- i.e. instead of a coefficient close to the true value of $10,000, the estimated value of pools will be -$10,000. Now, this is a silly estimate, and the analyst will know it's silly, not only from the sign but also from the fact that the standard error for this coefficient will be very high-- and it will be high precisely because it is based only on the information from one unit. So while the analyst might know to remove swimming pools from the regression model, he is left without a reliable hedonic model because his model doesn't have swimming pools in it.

Here is a another example, more realistic than the 10 member database discussed above. Using a data set on 102,000 homes in Fort Worth, Texas, suppose we wish to estimate the hedonic price of bathrooms[10]. If bathrooms are the only member of X, the estimated hedonic relationship is

$$P = -55{,}347 + 72{,}674 * Bathrooms$$

thus the prediction for every home with one bathroom is $17,327, and each additional bathroom adds $72,674 to the estimated sales price of the house. This is stupid in so many different ways it is hard to know where to begin. First, it would be ridiculous to think that every house with one bathroom would have the same price. Furthermore, the hedonic price of bathrooms seems rather high-- certainly it is well above construction cost, though in later chapters we will observe that this is not the most important criterion. Even so, the R^2 for this regression is 0.43; bathrooms alone can explain 43% of the variation in value. However, the standard error of the regression is $47,038, which seems a bit high, high enough to signal a problem with this simple model.

Can we do better in determining the hedonic price of bathrooms? Of course we can.

---

[10]For more information on this data set, see Leichenko, Coulson and Listokin (2002)

Bathrooms are surely highly correlated with the size of the house. Merely adding interior square feet to the regression gives us dramatically different results:

$$P = -68{,}192 + 16{,}211 * Bathrooms + 65.52 * IntSqFt$$

The hedonic price of bathrooms falls dramatically to $16,211. Why such a fall in the coefficient? When interior square feet is not in the regression, and bathrooms alone must explain price variation, the coefficient must not only account for the increase in price due to the extra bathroom, but also the increased square footage. In this data the average one-bathroom house has 1096 square feet, and the average two-bathroom house has 1839 square feet. Even more dramatically, the average three-bathroom house has 3429 square feet. So in the first equation, the coefficient on bathrooms had to account not only for the extra bathroom but the extra 800 -1400 square feet that houses with more bathrooms have. Clearly the extra square feet are not entirely due to the bathrooms alone. Since bathrooms and square footage are so strongly correlated, a hedonic price of bathrooms that is estimated in the absence of square footage can not be credible. The weight attached to bathrooms is covering for both characteristics. Once square footage is available as a factor in the pricing model the amount that bathrooms can and should explain is reduced.

The addition of other variables can alter the bathroom coefficient, although not so dramatically. Adding exterior square footage yields the hedonic function

$$P = -70{,}270 + 16527 * Bathrooms + 63.26 * IntSqFT + 0.52 * ExtSqFt$$

There are only small changes to the Bathrooms and Interior Square Feet coefficients, indicating that exterior square feet are only mildly correlated with the first two coefficients. (This is the case; the relevant correlation coefficient is only 0.16). However, adding the vintage of the house does produce a significant change:

$$P = -479873 + 12{,}721 * Bathrooms + 63.13 * IntSqFt$$
$$0.53 * ExtSqFt + 212.67 * YrBlt$$

The coefficient on bathrooms has dropped to $12,721, when the vintage of the house is included. This indicates, roughly speaking, that vintage and the number of bathrooms are positively correlated Evidently newer houses contain more bathrooms. Indeed, in the database the average one-bathroom house was built in 1943, while the average two-bathroom house was built in 1964, and the average three-bathroom house in 1972. This correlation might be due to the simple fact that newer houses are also bigger on average (as discussed above). But we can infer that the correlation is due to more than mere size precisely because Interior Square Feet has already been included in the regression. Thus we know that age and bathrooms are correlated even after accounting for the difference in size. That is to say, a 1900 square foot home built in 1990 will have, on average, more bathrooms than a 1900 square foot house built in 1960.

There are three lessons to take away from the above examples. The first is that all of the important characteristics need to be included in the regression model in order to obtain accurate hedonic prices and accurate appraisals. The second is that if such comprehensive inclusion involves high degrees of correlation, then an individual characteristic's contributions to the pricing process are going to be hard to measure. And the third is that leaving characteristics out of the regression

model on that account is a procedure fraught with danger.

2.3 Functional Form

The functional form of a hedonic price index matters. It matters because the functional form determines the way that attribute prices vary with attribute levels. Therefore a functional form that is too restrictive-- that is, imposes restrictions that are incorrect-- will have poor predictive power. Unforturnately, neither theory nor experience provide foolproof advice on this matter. To the contrary, absent a full knowledge of the distributions of tastes, incomes and other consumer characteristics on the one hand, or the distribution of physical and other housing characteristics on the other, the pricing function which matches households to units can be of any form whatsoever (See Chapter 5). Therefore any functional form for the hedonic price index is possible when confronting actual data.

There are, nevertheless, some theoretical considerations. As Rosen(1974) pointed out in his seminal paper, there are circumstances under which a linear functional form (as exemplified in the above examples will be expected. To quote Rosen:

> "A buyer can force [the hedonic function] to be linear if certain types of arbitrage activities are allowed. Arbitrage is assumed impossible in what follows... on the assumption of indivisibility. This amounts to an assumption that packages cannot be untied....Similarly, assume sellers cannot repackage existing products in this manner or do not find it economical to do so, as might be the case with perfect rental markets and zero transaction costs". (38-39)

In other words, linear pricing is the norm when competitive pressures can force the untying of tied bundles. To take a prosaic example, a 10 pound box of detergent costs $10 and a 20 lb. box costs $15; this is an example of nonlinear pricing because the first 10 pounds of detergent has a

different price ($10) than the second 10 pounds ($5). Such nonlinear pricing in principle creates a profit opportunity. It might motivate some arbitrageur to purchase a number of 20-pound boxes and break down the contents to smaller packages. The arbitrageur pays, in effect, $.75 a pound; notional purchasers of the smaller box are paying $1.00 a pound, and so the arbitrageur and the customer could presumably agree on some price in between. The detergent seller would hate this of course because of the lost profits on the small boxes that are no longer being purchased. The only way of getting around it is to linearize the price– that is, sell the 20-pound box for exactly twice the ten-pound box. The point, however, is that nonlinear pricing– in detergent and many other products does exist. The fact that arbitrageurs do not, in fact, stand around the entrances to supermarkets enticing customers with this bargain indicates that it is not, to quote Rosen, "economical to do so", presumably because transaction costs are not zero.

If this does not happen with the easily repackaged commodity of laundry detergent, the case for linear pricing in housing markets is even weaker. Physical housing characteristics are for the most part, tied together in an inseparable bundle. One cannot detach a bathroom from a house and sell it on Ebay. Because of this, linearity should not be assumed in a housing hedonic function as a matter of course.

There are degrees of closeness to linearity, however, that might be exploitable. Goodman (1988) found that rental price indices are "more linear" than indices for owner-occupiers. Goodman writes that this indicates that "landlords are more willing to combine, alter and divide housing units". Coulson (1989) found that multi-unit rental properties have a more linear index than either rental or owner-occupied detached units, but that owner-occupied units in multi-unit structures were the least linear of all indexes. Evidently, transaction costs matter a lot. Landlords who are owners of a number of contiguous units can easily combine, alter and divide, but when the contiguous units are owned by different owner-occupiers it becomes much more difficult to combine housing

attributes.

Individual attributes might have differing degrees of flexibility. Coulson (1989) speculates that to the extent that property lines are easily drawn and redrawn, land might enter the hedonic function "more linearly" than other attributes. Using a flexible functional form (see below) he finds that that is indeed the case. On the other hand Colwell and Munneke (1999) speculate that land assembly and reassembly in densely-built markets might be difficult, and that as a result the price of land has strong nonlinearity. The lack of ease in redrawing boundaries in densely-built areas would seem to naturally result in nonlinearities due to the zoning, transactions, and demolition costs.

Anyway, the usual view is that the functional form is in fact nonlinear, and as a result the *semilog* functional form has become perhaps the most widely used functional form in hedonic studies. It takes the form

$$\log P_i = a_0 + \Sigma_j a_j X_{ij} + e_i$$

and the coefficients are known as semi-elasticities. Roughly speaking, the coefficients give the *percentage* increase in price due to a unit increase in X. Recall that the hedonic price of a characteristic is the increase in the price of the housing unit due to an increase in X– the derivative, and in the linear case this is a_i. When we use the semilog form above the characteristic price/derivative is:

$$\partial P / \partial X_j = a_j e^{a_0 + a_1 X_1 + \ldots + a_n X_n + e} = a_j P$$

Note the nonlinear pricing. Assuming a_i is positive, P rises with X, and so the price of X at the

margin is continually increasing.

For characteristics with binary measures the situation is a bit different. As Halvorsen and Palmquist (1980) point out, if X is a dummy variable (or other discretely measured attribute) care must be taken. For small values of a_j the "percentage" interpretation is valid (as this is the same as assuming that log(1+a)=a, a well-known approximation for small a), but if a_j is large then the actual discrete difference should be calculated. That is, if X_1 is a characteristic (like a swimming pool) that only takes on a value of 1 if the house has a pool, and zero otherwise, then you can't really take the derivative of price with respect to X_1. Instead you calculate the hedonic price of the swimming pool as the difference between the appraisal price with a swimming pool and that without:

$$changeinhouseprice = e^{a_0 + a_1 + \ldots + a_k X_k + e} - e^{a_0 + \ldots + a_k X_k + e}$$

Note that the first term $a_1X_1=a_1$, because $X_1=1$, and in the second term $a_1X_1=0$, because $X_1=0$. Note further an important point, that applies in both the continuous and discrete attribute cases, that the calculation of the hedonic price in the semilog functional form depends on the value of all of the other X variables. The greater are (any or all of) X_2 through X_k the greater the hedonic price of swimming pools.

An adjustment is also need when using the semilog form in the appraisal/prediction process: What Granger and Newbold (1986) would call the *naive* prediction would be to insert values of X* and calculate

$$(\log P)^* = \Sigma_j a_j X_{ij}^*$$

and then provide a forecast of P by using the inverse logarithmic function:

$$\hat{P}^* = e^{\log P^*}$$

However, it should be remembered that as the prediction is the expected value, and the expected value of a nonlinear function is not equal to the nonlinear function of the expected value, then the naive forecast of P* is in fact biased. A standard adjustment to the appraisal/forecast is:

$$\hat{P}^* = e^{(\log P^* + 0.5\sigma^2)}$$

where σ is the standard error of the regression.

Another popular functional form is the *double-log* or *log-linear* functional form where if X is positive for all housing units, we write

$$\log P = a_0 + a_1 \log X_1 + a_2 \log X_2 + \ldots + a_k \log X_k$$

The coefficients are elasticities, giving the percentage increase in P due to a percentage increase in X. Obviously if X takes on negative or zero values this functional form is infeasible, and the variable can be entered linearly. In the case of dummy variables this is obviously the course of action, and the above warning remains in force for the calculation of hedonic prices..

For continuous variables,

$$\partial P / \partial X_j = a_j P / X_j$$

the price can be either increasing in X or decreasing, depending on whether the coefficient is greater or less than one. In that respect the double-log formulation is more flexible than the semi-log. The adjustments for dummy variables and predictions are similar to those stated above for the semilog form.

While all three of these functional forms have remained in common use by hedonic analysts over the years, there has been increasing concern that while all have their strong points, they all have weaknesses in the sense that all restrict the ways in which the hedonic function can be nonlinear. Concerns over the use of specific functional forms and the restrictions that they impose led many researchers since the1970s to consider so-called flexible functional forms, particularly the Box-Cox transformation:

$$P_i^{(\lambda_1)} = \Sigma \beta_j X_{ij}^{(\lambda_2)} + e_i$$

where:

$$X^{(\lambda)} = \frac{X^{\lambda} - 1}{\lambda} \qquad \text{if } \lambda \neq 0$$

$$X^{(\lambda)} = \log X \qquad \text{if } \lambda = 0$$

and similar transformations take place for P. As can easily be seen, when λ (on either side) is set at unity the transformed variable is entered in a linear fashion (with some adjustment to the intercept term). Thus the Box-Cox nests all three of the functional forms described above. When λ_1=1 and λ_2=1, we have a linear functional form; when λ_1=0 and λ_2=1, the semilog results, and when λ_1=0 and λ_2=0 the double log form is being estimated. One also has the rarely-used inverse semi log (λ_1=1, λ_2=0) as well as square roots and quadratics (λ=0.5 , λ=2). The point is that, through

nonlinear estimation techniques, one can simultaneously estimate the a_j's and the λ_1 and λ_2 and let the datat decide which functional form is best.

Note that marginal hedonic prices for the attributes X can be calculated as

$$\partial P / \partial X = \beta_j P^{1-\lambda_1} X^{\lambda_2 - 1}$$

which offers a wide variety of potential responses of the housing price to changes in X. Early applications of this transformation include Goodman (1978) and Linneman (1979). One possible use of this general functional form is to allow for testing the simpler forms against the more highly parameterized alternatives, and several simple tests have been proposed in the econometrics literature. The tests very often reject the simpler forms, although the data usually pick values for the transformation parameter that are moderately close to the semilog. That is, the value of λ_1 is often close to zero and the value of λ_2 is usually larger and closer to one[11].

A comparison is in order, which will highlight why functional forms can make a difference. The sample from Fort Worth, TX is used. As before, the hedonic regressions were run using, only interior floor space, lot size, year built, and number of bathrooms as regressors. Four functional forms were used: linear, semi log, double-log and Box-Cox with transformation of the dependent variable only (i.e. λ_2 is set to one). Table 3 provides the calculation of the price of interior square footage for various - sized houses (at the 10th, 25th, 50th, 75th and 90th percentiles). The other three attributes were set at their sample medians (1 bathroom, 7500 square feet of lot, built in 1953) and

[11]Cropper, Deck and McConnell (1988) simulate hedonic price estimation in the possible presence of unobserved attributes. They find, somewhat surprisingly, that the linear functional form outperforms the semilog and double-log, and might under some decision criteria, do better than the Box-Cox. The Box-Cox ends up being the authors' preferred choice.

derivatives with respect to floor space were calculated. The most important result would seem to be that the linear functional form provides a (constant) price that is far higher than the prices provided by the other three functional forms.

It is possible to expand on this. Coulson (1989) explores the possibility of letting the transformation parameter λ be different for each of the different right-hand side variables. Halvorsen and Pollakowski (1981) suggested the even more highly parameterized quadratic Box-Cox functional form

$$P_i^{(\lambda_1)} = \Sigma_j \beta_j X_{ij}^{(\lambda_2)} + \frac{1}{2} \Sigma_j \Sigma_k \gamma_{ij} X_{ij}^{(\lambda_2)} X_{ik}^{(\lambda_2)} + e_i$$

adding the interaction terms. Interaction terms are terms where two attribute levels are multiplied together, with the resultant product having its own parameter. Doing this for every pair of characteristics certainly increases the flexibility of the hedonic functional form, since now there is an explicit connection between the calculated hedonic prices of one characteristic and the sizes of other characteristics. The above model allows each of them to be first transformed by a Box-Cox function. Thus a large number of simpler forms are nested within this very general framework, and if the transformation parameter for the linear terms is allowed to be different from that for the interactive terms, an even greater number of forms are possible including the quadratic, translog, Leontief, the "ordinary" Box-Cox and all of the simpler forms discussed above. Again, statistical procedures are available to test whether the data will admit simpler functional forms. Halvorsen and Pollakowski find in their example that every restricted functional form is rejected.

Cassel and Mendohlsson (1985) object to the Box-Cox functional form, and by extension other flexible functional forms, on two sensible grounds. One is that in increasing the number of parameters, such as through the addition of the interaction terms above, will naturally decrease the precision in the estimation of each of the individual parameters. The more variables, the greater the correlation possibilities; when those correlations involve unimportant variables, or (more to the point) the interactions of unimportant variables, the increased flexibility comes at the potential cost of wildly inaccurate appraisals of individual properties. Their second objection is that the Box-Cox transformation of the dependent variable is nonlinear. When forecasting from a Box-Cox model the first step is to forecast $P_i^{(\lambda)}$ from $\beta X^{(\lambda)}$; then one must apply the inverse of the Box-Cox transformation to the forecasted dependent variable in order to obtain a forecast of P_i. Since the expectation of a function is not the same as the function of the expectation, a forecast of P obtained must be biased. But as noted, this is true of all functional forms that invoke a nonlinear transformation of the dependent variable, such as the logarithmic transformations.

Indeed, the use of flexible parametric functional forms such as the Box-Cox has abated in recent years. When researchers have found it undesirable to impose the restrictions of simple forms such as linear and semilog, they have eschewed functional form considerations all together in favor of *semiparametric* or *nonparametric* estimation of the hedonic price function. The first use of such methods in a hedonic context is apparently by Stock (1989) although interest in such methods seems to have accelerated upon the appearance of Meese and Wallace (1991). Other early contributions include Coulson (1992), Pace (1993, 1995), Anglin and Gencay (1996), Thorsnes and McMillen (1998), and Yatchew (1998). This last paper is a survey of nonparametric techniques and includes an hedonic example where the influence of location on price is investigated nonparametrically. As an aside, the use of nonparametric techniques in this particular application would seem to be particularly important as also noted in Redfearn (2009).

There are several types of nonparametric or semiparametric methods, and a full survey need not be attempted here. All of them have in common the form

$$P_i = \beta X + m(Z_i) + e_i$$

where now for convenience β and X are considered multi-valued (i.e. vectors of) housing attributes entering into the hedonic function linearly (for convenience; they could certainly be log-transformed), and Z is an attribute (or attributes) subject to the nonparametric transformation. It is possible for the dependent variable, P_i , to be entered in logarithms, and it also possible for X to be empty.

Anyway, the idea is that m(.) is a function that is not subject to any of the restrictions embedded in functional forms, even those as flexible as the Box-Cox, thus allowing for any sort of curvature in the response of price to the level of the attribute, possibly even changing from convex to concave and back again. Thus the data is allowed to almost completely choose the nature of the response of P to Z. There is a danger of allowing too much flexibility, of course. A *really* flexible function would simply provide a binary response for each value of Z. Such a model would lack the property of *smoothness;* a graph of the response of price to Z would be extremely jagged. This would provide no guidance at all for prediction, or anything else; certainly no extrapolation or interpolation of the model is possible, and would make the model irrelevant for appraisal.

Therefore every nonparametric or semiparametric technique recognizes the tradeoff between fitting the data and the smoothness of the hedonic response. At one extreme is the "dummy variable" approach just mentioned, which would have a fantastic fit but would not be particularly smooth; at the other extreme is the linear model, which is as smooth as one can have, but might not fit particularly well.

Speaking *very roughly* many semi- and non-parametric estimators work as follows. Let the data

be ordered according to the values of z. Consider a database consisting of observation z_i and the set of observations "close to " it $[\underline{z}_i...z_i....\hat{z}_i]$. A regression of P on x and z in this limited data would yield a kind of a "local" coefficient. Doing this for each observation in the data would create a serires of coefficientsat each level of z.. Nonparametric techniques basically find different ways of putting these local slopes together to create a hedonic function.

As an example, take the method of locally weighted regressions. In the standard application (see Cleveland and Devlin, 1988) each observation z_i takes its turn as the kernel or center of a subgroup of observations, and the fraction g of the observations closest to z_i comprise a data set to estimate this local slope around this z_i , the fraction being chosen by the investigator. An extra bit of apparatus is added, in that the surrounding observations get less weight in the regression sample, the farther away they are from the central observation. The weight on observation k for the local slope around observation i is calculated using

$$W_{ik}=V(D(x_i,x_k)/D(x_g,x_k))$$

where D is a measure of distance (usually Euclidean distance), x_g is the observation in the limited sample which is farthest from x_i, and V is a weighting function which has a negative derivative, so that nearer observations obtain greater weight in the estimation. The use of

$$V(s)=(1-s^3)^3$$

where $0<s<1$ is chosen by the investigator, is popular for this purpose. This is done for each observation in turn. The use of overlapping samples insures some degree of smoothness in the resulting assembly of local slopes, but clearly the choice of g is critical. When g is small the samples have less overlap; the function will be rougher-- it is as if the number of parameters has increased.

When g is large the function will be smoother, and in the limit when g=1 the samples are the same for each of the local regressions and there is only one slope; the function is linear.

There is little in the way of guidance for the researcher attempting to choose the value of g . The problem is quite similar to the problem of choosing the set of regressors in a parametric regression; adding regressors increases the variance but decreases the bias in a model. Regression models often resort to "model selection criteria" which attempt to weight the importance of bias and variance by trading off the fit of the model . Examples of selection criteria include the adjusted R^2, and the Akaike and Schwarz Information Criteria. Some progress on applying such a selection criterion to nonparametric models has been made by e.g. Engle, Granger, Rice and Weiss (1986) and an application to hedonic regressions is contained in Coulson (1992) but this remains a relatively unexplored area.

2.4 Heteroskedasticity

Heteroskedasticity occurs when the variation in the errors from a hedonic regression around their mean of zero depends in some systematic way on the housing attributes. If, for example, errors get bigger when the interior square footage of a house is larger, this is an example of heteroskedasticity. The existence of heteroskedasticity can have marked consequences for regressions. Since ordinary least squares tries to minimize the variation of the predictions around the actual values, heteroskedastic data will give more weight to those observations where the variability is bigger. It turns out that this doesn't bias the appraisals but it does make them less precise. Appraisal error will be larger on average. It also biases the hypothesis tests one might perform on the coefficient estimates.

A typical thought is that heteroskedasticity is related to the size of the building. Expensive houses are more likely to have very high prices or very low prices *given their size or other signals of quality*

than are small houses. Consider Figure 3, which plots interior square footage against price, from the Fort Worth data set described above. Note how the data "fans out" as it proceeds away from the origin, indicating that the variation of price around its mean is much more extensive for big houses than it is for small ones. A regression of price on interior area would create errors which would fan out in precisely this way. Thus large houses would more influential in creating the weights than they should[12]. It is often the case that functional form decisions have an impact on the extent to which heteroskedasticity can be an issue. Consider Figure 4, which plots the natural log of price against floorspace. Heteroskedasticity would seem to be much less of an issue once the log transform has been used. In fact, the logarithmic transformation actually seems to overcompensate; the data seem to "fan in" as the size increases. The spread of prices for large houses actually seems to be getting narrower for large units. The optimal Box-Cox transformation, using λ=0.254 as above actually seems to equalize the spread across large and small houses, as Figure 5 demonstrates, thus this middle ground seems to be appropriate for this particular instance.

While the above data seem to indicate that heteroskedasticity can be related to the size of the house, in fact heteroskedasticity can arise from any correlation between the variance of the error term and X. A prominent example is discussed by Goodman and Thibodeau (1995, 1999) who hypothesize that heteroskedasticity is related to the building date of the structure. They reason that as housing ages, some owners decide to remodel and some do not, some decide to provide maintenance and some do not. All of these houses will be entered into a multilist data base as having been built in the same year, which is immutable even though the "true" age might be a lot lower due to remodeling, or even higher if there has been poor maintenance. They suggest a weighting scheme

---

[12]Ideally each house should get the same weight when estimating the hedonic coefficients. This is because the information from each house is equally valuable, unless we have some reason for thinking otherwise. When there is heteroskedasticity, some types of houses get more influence than others.

for the observations based on a polynomial of age found by regressing the squared error term from the ordinary least squares hedonic regression on the first four powers of the units age and taking the fitted values as weights.

These corrections notwithstanding, many researchers have opted to focus on the correction to the estimation of the standard errrors, since the estimates themselves remain unbiased. Thus rather than reweighting the observations in order to obtain efficient parameter estimates, the OLS estimates are used to estimate correct standard errors for hypothesis testing purposes. A common and certainly convenient method is to estimate standard errors that are correct regardless of the sort of heteroskedasticity that appears (see White 1980), and this option is now available on most software packages.

2.4 Spatial Dependence

The practice of incorporating spatial characteristics into regression analysis is as old as hedonic regression itself. Interest in this aspect of hedonic regression analysis has been invigorated over the past decade or so with the introduction of new methods and new interest in old methods, along with new data that allows their implementation.

There are a large number of ways in which spatial data might be incorporated. In the following, let *j* index locations 1 through J and *i* index individual housing units 1 through N where it is understood that N differs across locations. (N might, in what follows, be 1 for all databases where a unique address is available for each observation. For now, assume that this is not the case, or that one chooses not to make use of the address.) We thus classify all of the properties in a data base as belonging to one of J different neighborhoods or property developments or census tracts or blocks, depending on the source of the data. Then one might incorporate spatial characteristics in the same way as the original hedonic modelers Ridker and Henning did, by specifying the model

$$P_{ij} = \sum_i \sum_i \sum_k a_k X_{ijk} + \sum_j \sum_k Z_{jk} + e_{ij}$$

where the X_{ijk} are the characteristics of the individual house and the Z_{jk} are the characteristics of the jth location. One will recall that the focus of Ridker and Henning was the air pollution of the jth district, but any number of other characteristics could be so included, and most modern hedonic studies include such characteristics. A further discussion of such spatial characteristics is contained in the next chapter.

The next step in spatial analysis is to recognize that there may be other characteristics of the location that are not captured by these observable and quantifiable data. Other strategies that capture location differences can be pursued.

One particularly straightforward strategy is to treat the overall conditions in a region (that contribute to housing price) as embodied in a single region-specific effect. That is, treat the error term as being comprised of two components:

$$e_{ij} = v_j + w_{ij}$$

where v_j is a component common to all property from region or location j and w_{ij} is an error term idiosyncratic to property i. Much attention in the econometrics literature is focussed on whether this *spatial heterogeneity* is to be treated as part of the model or as part of the error term. In the former case , the v_i are, in effect, the coefficients of J distinct binary variables, one for each region; indeed, this so-called *fixed effects* model is implemented by including the appropriate dummy variables in X[13]. One aspect of fixed models is that the Z_{ij}s must be removed from the regression, since there would

[13]One of the J dummy variables must be omitted, as it will be represented by the intercept term, as in the discussion of exterior material dummies above. This can be automatically implemented as an option in many software packages.

be perfect correlation between them and the set of dummy variables. In some contexts, this is of no consequence, and is in fact a convenience, since the J dummy variables will collectively account for *everything* that makes one location different from another. However if interest centers on the influence of locational characteristics *per se,* then the use of fixed effects will be a drawback.

When the ν_j are treated as part of the (random) error term, the model is known as *random effects*. Again, one is accounting for all of the locational characteristics, but in a slightly different way, by observing the correlation of the error terms across different observations in the same location, and assuming (presumably correctly) that the reason for this positive correlation is because of otherwise unobserved locational characteristics. Because this is done through the error term rather than the regression model, the other locational attributes, Z_{ij} need not be eliminated from the model, and this is a major advantage for the random effects treatment. As is commonly realized, however (see Wooldridge (2002)) there is the possibility of correlation between the random effects component of the error term and the hedonic variables. As an obvious example, the v_i might index neighborhoods in different parts of the city, and therefore access to job locations might be one of the things that contributes to the neighborhood impact on price. However, the monocentric model suggests that larger houses might be built on the less accessible properties, and since square feet will be one of the property characteristics there will be correlation between this Z and the error term. See Wooldridge for extensive discussion of these issues. The implementation of random effects is equivalent to assuming that such correlations do not exist.

While fixed or random effects might be an adequate representation of the (unobserved) local characteristics, they have the distinct disadvantage of not being particularly *spatial*. That is, while such a model groups housing units on a spatial basis, the spatial relationship between different houses within the same group, or the spatial relationship between different groups is not accounted for.

A more sophisticated approach to the spatial characterization of property values would

recognize, to the extent possible, the spatial relationship between various individual housing units. One method of doing this is through the use of *spatial lags*. Basically, spatial lags allow the property price to be a function not only of the Z of its of its own region, but also of the characteristics of neighboring regions or units. Alternatively or additionally, one could allow for *spatial autocorrelation* which allows the error term to depend on error terms of other locations.

There has been a recent blossoming of research and discussion of these possibilities; prominent and more complete sources include Anselin (1988), and Lesage (2000). In this survey we can no more than scratch the surface of this booming field. Take for example, the possibility of spatial lags in the Z's. One might imagine, for example, that home prices depend on local crime rates. In that event it might be the case that not only the crime rate in one's own neighborhood is important, but also crime rates in surrounding neighborhoods. If Zj is the crime rate in neighborhood *j* then one might then include a regressor Z^*_j , the average crime rate across the neighborhoods which border *j*[14]. Or one could let Z^*_j be a weighted average of every other neighborhood in the city, where the weight attached to neighborhood j′ depends in some fashion on the distance between j and j′ with lower weights attached to longer distances.

Such a model could be written like this in matrix form (assuming for simplicity one spatial characteristic):

$$P = X\beta + WZ\gamma + e$$

where P is now the vector of observations on price, X is the nxk matrix of unit characteristics, with the associated kx1 coefficient vector β, and Z is the spatial characteristic. (It is common to normalize W so that the elements of each row add to one.) W is an nxn matrix which maps the observations on

[14]This is an example of "first-order" contiguity, wherein the spatial correlation is assumed to carry over only for neighboring locations. This is a common assumption in this research.

the entire set of Z's (from all locations) into the spatial variables that presumably affect price. One common method of defining W is by use of the principal of *first order coniguity*. In the ith row and jth column of W, insert a 1 if the two observations are contiguous-- that is, if they border each other, and zero otherwise. Then normalize by dividing each observation by the number of ones in that row. Thus WZ is a weighted average of the Z values that surround each house this average will be included as a characteristic. Obviously, manipulation of the W matrix to suit the particular application and data set allows a wide variety of possibilities for the specification of spatial lags.

But in point of fact the use of spatial lags in Z (and X) is relatively rare. Another, more widely used, method is to allow the price of housing to depend on neighboring house prices. In this case the hedonic price model will be written (in matrix form) as

$$P = X\beta + \varphi WP + e$$

where W as before is a spatial weighting matrix that provides the means by which the surrounding prices are aggregated. In this case the idea is that the prices of surrounding properties, in and of themselves, add or subtract to the price of the property being appraised. In one of the models proposed in Can and Megbolugbe (1997) the elements of W to be such that for any given property, WP is a weighted average of the three most recent sales within a two-mile distance; the weights are based on the distance from each of the three sales to the appraised property (and so that the weights sum to one). But clearly the way in which the W matrix is constructed will depend on the particular application.

Finally, one might wish to specify the dependence in the form of spatial autocorrelation in the error term. A model with *spatially autocorrelated errors* would be

$$P = XB + e$$
$$e = \rho We + v$$

where W is again a weighting matrix. The residuals, ν, are assumed to be uncorrelated with each other; all the dependence is accounted for in the W matrix. In this case the correction is in the manner of the usual GLS correction; the second equation is rewritten as

$$(I - \rho W)e = \nu$$

There are two slightly different interpretation of this model. The first is that whatever is causing the spatial correlation, it is unobserved to the analyst (or at least not included among the X variables). Thus it is by construction assumed to not be correlated with the included X's.

A second interpretation is revealed by premultiplying both sides of the hedonic equation by $(I-\rho W)^{-1}$ from which it can be seen that the first equation equivalent to:

$$(I - \rho W)^{-1} P = (I - \rho W)^{-1} \beta X + \nu$$

Thus spatial autocorrelation is seen to be equivalent to spatial lags of both the dependent and independent variable– with similar lag matrices as well. This "common factor" representation of autocorrelation in the time domain was first pointed out by Hendry and Mizon (1978) and applied to spatial econometrics by Anselin (198?).

In the case of spatial autocorrelation, t the covariance matrix of e, which we write as Ω, can be written as

$$\Omega = \sigma^2 (I - \rho W)(I - \rho W)'$$

where σ^2 is the variance of ν (assumed to be homoskedastic). In a pioneering application of these procedures, Dubin (1988) specified Ω as having ones down the diagonal, and each off-diagonal element – that is, the covariance of any two observations was expressed as

$$\mathrm{cov}(e_i, e_j) = \exp(-d_{ij} / \lambda)$$

where d_{ij} is the distance between the two observations, and λ is a parameter to be estimated[15]. This specification captures one of the main ideas behind spatial correlation of any sort, which is that observations which are spatially farther apart have proportionally smaller correlations. This is intuitively reasonable and corresponds with similar requirements for data in the time domain (such as mixing) that are used to confirm the consistency of estimators in the presence of observational dependence.

There is, finally, the spatial error components model (Kelejian and Robinson 1993, Anselin and Moreno (2003)) where (coming somewhat full circle) spatially related locations are assumed to have a common component:

$$P = XB + e$$
$$e = \rho W\zeta + v$$

where the ς are random error terms. The W matrix is interpreted as before. This has much in common with the random effects model described above, in that there are two distinct components to the regression error term, one related to location and one idiosyncratic. The spatial error components model is more flexible in the sense that the relationship between different locations need not merely about whether they are in the same general area or not, but can be defined on the basis of relative contiguity or distance.

2.5 Summary

This chapter has summarized the use of regression analysis to estimate hedonic price functions, and thus both calculate hedonic prices for property characteristics and appraise individual properties. The estimation of hedonic models will depend on the nature of the database and the information on

[15]The parameter ρ would be redundant – that is, it would not be identified in this specification.

characteristics contained therein, particularly with regard to spatial and locational information. The reader should be aware that regression analysis is fraught with difficulty and some background learning on the subject would be appropriate. Fortunately several excellent textbooks and software packages will provide one with the necessary detail to undertake this sort of analysis.

Spreadsheet 1

Observation Number	External Area	Living Area	Number of Bathrooms	Price	Predicted Price	Predicted Price 2
1	15000	1265	2	76000	65,624	55,006
2	15000	1926	2	77500	94,455	84,315
3	9500	1921	2	80500	94,237	79,192
4	24444	2002	2	83500	97,770	96,102
5	43560	1688	2.5	86000	84,074	108,322
6	10960	1768	2.5	87000	87,563	82,814
7	26250	2207	2	91700	106,711	106,802
8	11250	2728	1.5	114000	129,436	107,429
9	8024	2665	3	130000	126,688	129,077
10	57935	2297	2.5	171000	110,637	148,137

Table 1:Hedonic Price Function for Atlanta 1979. (Source Palmquist 1983)

ATTRIBUTE	COEFFICIENT	Value for X*	a_jX_j*
Intercept	-9337.32	1	-9337.32
Lot Area (square feet)	0.0813	40000	3252
Improved Area (square feet)	15.0576	1400	21080.64
Improved Area^2	-0.0022	1960000	-4312
Number of Baths	1821.32	2	3642.64
Year Built	134.4473	70	9411.311
Number of stalls in Garage	1451.094	2	2902.188
Number of stalls in Carport	1198.081	0	0
=1 if garaged is detached	-1006.91	0	0
=1 if wiring is underground	710.0944	1	710.0944
=1 if dishwasher	1710.118	1	1710.118
=1 if garbage disposal	292.5529	1	292.5529
=1 if central air conditioning	1937.391	1	1937.391
=1 if wall air conditioning	604.6657	0	0
=1 if ceiling fan	344.714	0	0

=1 if sold in 1976	-1114.5	0	0
=1 if““excellent condition”	1007.502	1	1007.502
=1 if ““fair condition””	-2227.37	0	0
=1 if brick or stone exterior	622.333	0	0
=1 if full basement	1852.194	1	1852.194
=1 if partial basement	1108.292	0	0
=1 if fireplace	1114.569	1	1114.569
=1 if swimming pool	3274.725	1	3274.725
level of air pollution	-45.47	0	0
median age in census tract	-58.1812	36	-2094.52
median family income in census tract	0.0788	25,000	1970
% of workers in tract with blue-collar jobs	-52.1812	45	-2348.15
% of houses in tract with new occupants (< 5 yrs)	-32.4515	14	-454.321
% of tract population that is non-white	-1516.5	0	0
% of tract population over 24 that is HS graduate	1.2341	68	83.9188

% of structures with >1 person per room	35.9097	0	0
number of work destinations			0
per square mile in tract	16.6915	0	0
			Sum=$35,695.53

Table 2

Hedonic Prices by percentile and Functional Form

Floor Space (Percentile)	Linear	Semilog	Double Log	Box-Cox (λ=.254)
813 (10)	63.13	13.95	9.65	15.62
1017 (25)	63.13	15.90	14.62	18.01
1327 (50)	63.13	19.38	15.43	21.92
1788 (75)	63.13	26.03	16.34	28.70
2371 (90)	63.13	37.80	17.31	39.10

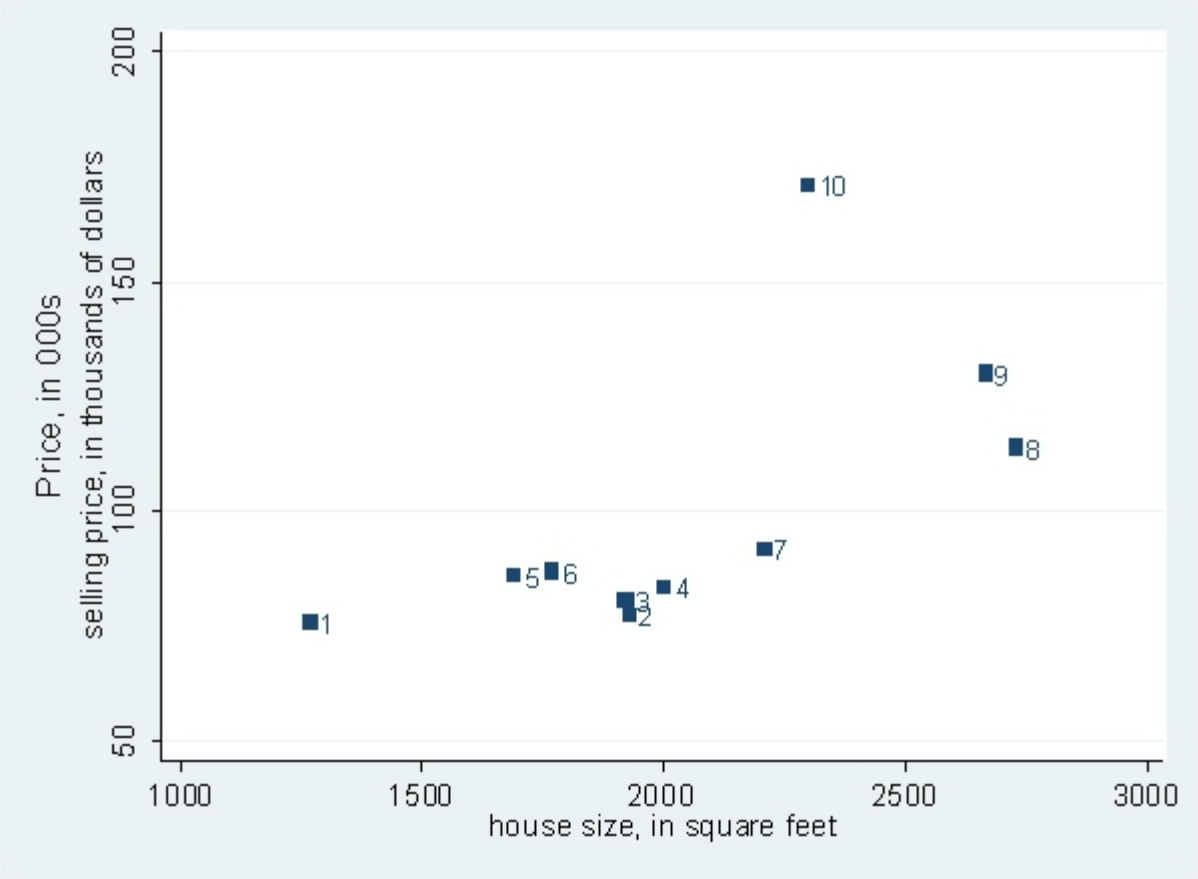
Price, in 000s
selling price, in thousands of dollars
200
150
100
50
1000
1500
2000
2500
3000
house size, in square feet
1
2
3
4
5
6
7
8
9
10

Figure 1

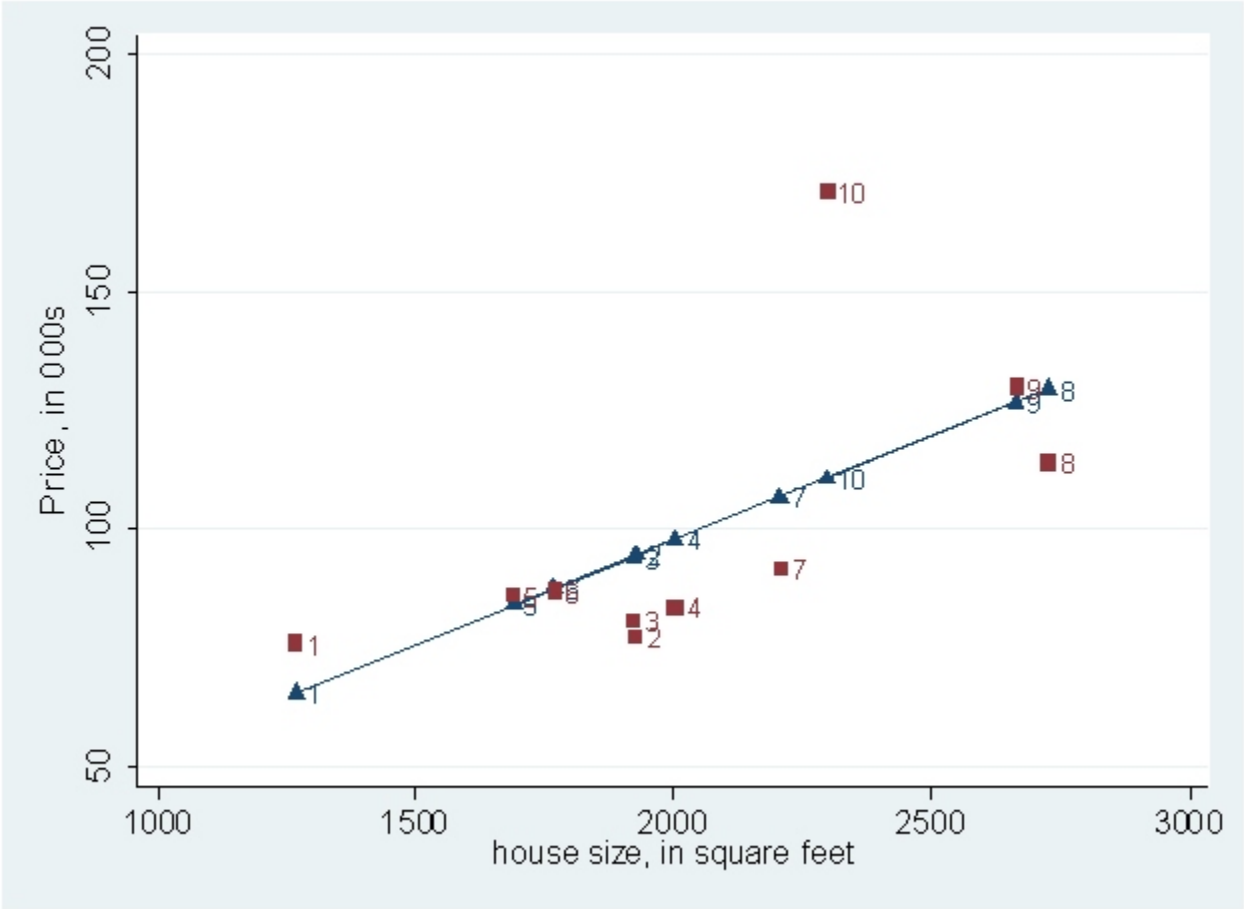
Price, in 000s
50
100
150
200
1000
1500
2000
2500
3000
house size, in square feet

Figure 2

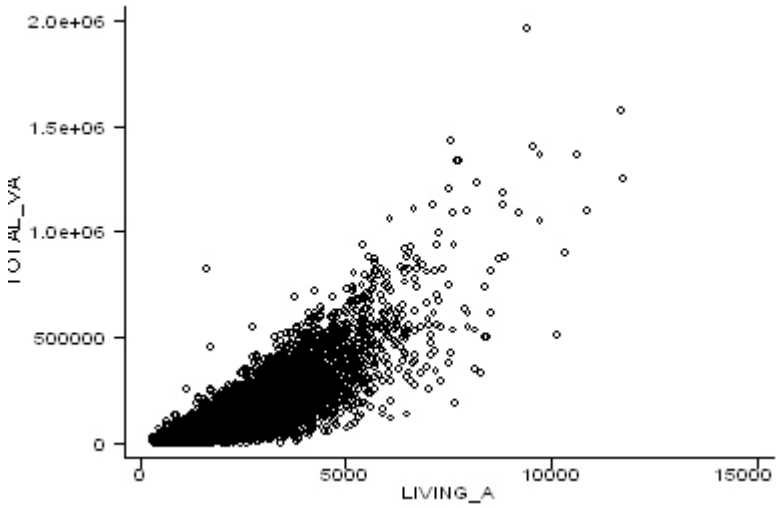
2.0e+06
1.5e+06
1.0e+06
500000
0
TOTAL_VA
0
5000
10000
15000
LIVING_A

FIGURE 3

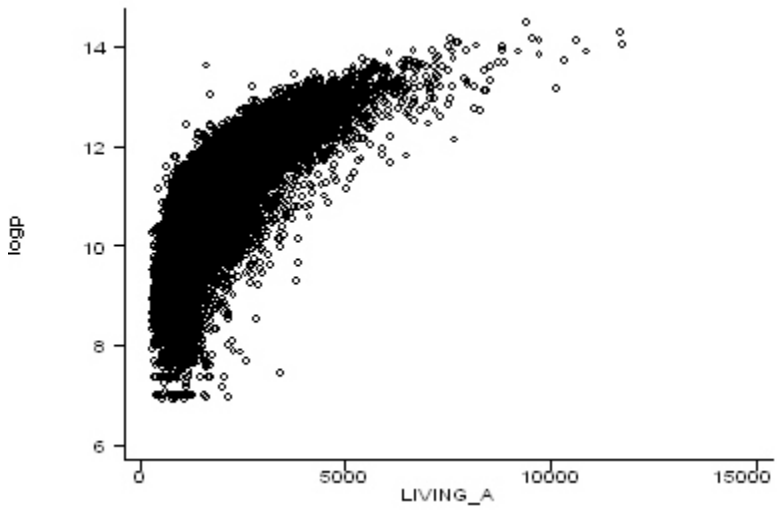

FIGURE 4

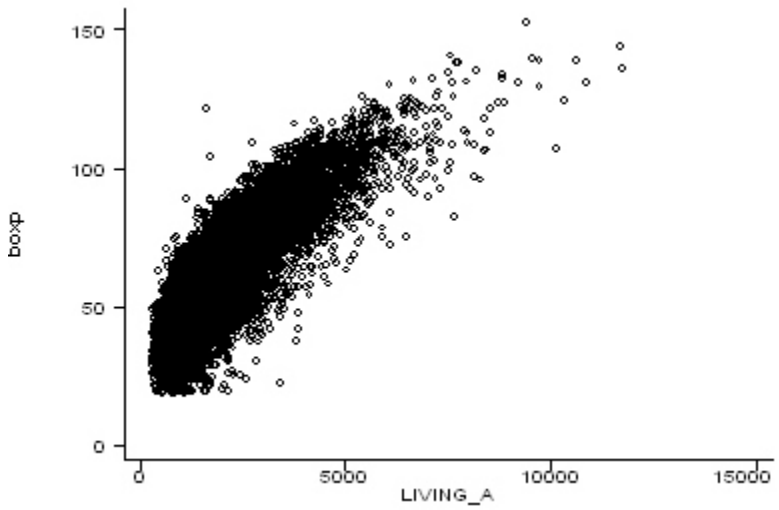

FIGURE 5