CHAPTER II: SPECIFICATION OF THE HEDONIC INDEX

This chapter has two purposes. The first section is designed to make clear what a hedonic regression is. We discuss both the intuitive and theoretical arguments which have been put forth to support the use of hedonic regressions. The second section examines practical considerations in estimating the hedonic relation, such as how a market is defined and the choice of functional form for the equation.

SECTION 2.1: THEORETICAL BASIS

In the first part of section 2.1, the goal is to make clear how the hedonic technique works and why the method is valid. The assumptions upon which the method is based are explained with examples. Given these assumptions, the general hedonic relation is discussed.

Intuition Underlying the Hedonic Index

The first, and least controversial, assumption is that a house is a bundle of size, quality, and locational characteristics. An analogy can be made to a bundle of groceries. Some grocery bundles are bigger and better than others, depending upon the number and type of food items in the bundle. So too with housing. A house embodies many features: bedrooms, baths, a heating system, location, etc. The number and type of features embodied in a particular house distinguish it from other houses.

How can housing bundles be compared? It is simple to compare houses which are identical except for one characteristic. For example, a four-bedroom house contains more housing than an otherwise identical three-bedroom unit. Problems occur when units differ in more than one

characteristic at a time. Does a three-bedroom unit with two baths represent more housing than a four-bedroom house with one bath? It depends, of course, on the value of a bathroom relative to a bedroom. The problem is easily solved in the grocery bundle example because all individual items have clearly marked prices. The more expensive bundle represents more groceries. This follows because the money used to buy the expensive bundle could be used to buy the less expensive bundle and there would still be money left over to buy more groceries.

Unlike groceries, prices of the individual features which comprise a housing bundle are not directly observable. This is where the second assumption comes in. The second assumption is that the rent or value of a housing unit stems from the quantity and type of characteristics it contains, and that the "prices" of the characteristics can be estimated from the rents or values of many units via multivariate regression analysis. A simple example which demonstrates the reasonableness of this assumption concerns the difference in values between two units which differ only with respect to the type of heating system. If one unit has a central heating system and the other has a fireplace, then the difference in the market value of the two units will equal the market valuation of a central heating system relative to a fireplace. Not all examples are so simple, but by pooling together many dwellings it is possible for multivariate regression to determine the relationship between rents, house values, and dwelling characteristics. The estimated regression coefficients are implicit prices which measure the value of each dwelling and neighborhood characteristic. For example, the regressions might determine that a central heating system adds 10 percent to the value of a house.

More formally, the assumptions suggest the following general hedonic relation:

$$R = f(S, N), \text{ where:}$$

R = rent or value,
S = structural charcteristics, and
N = neighborhood characteristics, including location.

If this relationship is true, then a properly specified regression equation applied to appropriate data can provide precise estimates of the relationship. If the relationship is a linear one, then the estimates are interpreted as implicit prices of each of the characteristics which determine rent or value.

Is it likely that the relationship between housing characteristics and rent or value is the same for all types of households in all types of situations? Probably not. Four exceptions seem likely: (1) long-time tenants often receive discounts; (2) large families often pay more than smaller ones for the same unit; and (3) black households pay different amounts for the same unit due to prejudice (the result of prejudice could be higher or lower prices, as we will discuss in Section 3.2). In addition, (4) some renters receive utilities, furniture, and other services in addition to structure and neighborhood. The basic relationship should be modified to reflect these cases. Now,

$$R = f(S, N, C), \text{ where}$$

C = contract conditions, implicit and explicit.

Several of these contract condition variables are tenant characteristics. We emphasize that only those tenant characteristics which affect the prices faced for housing, or the supply prices, are included in the hedonic equation. For example, people with higher incomes can afford

One widely cited basis is the household production model of the consumer proposed by Lancaster (1966) and Muth (1969). In this theory, utility is not a function of commodities (units bought in the market, such as a house), but rather of the characteristics embodied in a commodity, such as number of rooms. A specific functional relationship is assumed to exist between the characteristics and the commodities. Maximization of utility, subject to this relation and a normal budget constraint, gives rise to a hedonic function relating the price of a commodity to the characteristics embodied in it. Muellbauer (1974) has shown, however, that the conditions on both the utility and production functions which must hold in order to derive the hedonic relation in this way are quite restrictive.

In Muellbauer's critical review of the theoretical bases for the hedonic approach, he cites two other general models which give rise to the hedonic regression. The first he calls the Houthakker approach. In its simplest form, this model assumes utility is a direct function of the characteristics of a dwelling, and there exists a price schedule for characteristics, which consumers take as given. Maximization of utility, subject to a standard budget constraint and the price schedule for characteristics, implies a hedonic relation with the properties needed for price index construction and comparative analysis of housing quality. Muellbauer argues that the Rosen (1974) approach is basically an extension of the Houthakker approach. According to Muellbauer, all these models require the assumption that consumers face a fixed and known price schedule for characteristics which is based upon production costs. That is, it reflects supply conditions, not demand conditions.

One widely cited basis is the household production model of the consumer proposed by Lancaster (1966) and Muth (1969). In this theory, utility is not a function of commodities (units bought in the market, such as a house), but rather of the characteristics embodied in a commodity, such as number of rooms. A specific functional relationship is assumed to exist between the characteristics and the commodities. Maximization of utility, subject to this relation and a normal budget constraint, gives rise to a hedonic function relating the price of a commodity to the characteristics embodied in it. Muellbauer (1974) has shown, however, that the conditions on both the utility and production functions which must hold in order to derive the hedonic relation in this way are quite restrictive.

In Muellbauer's critical review of the theoretical bases for the hedonic approach, he cites two other general models which give rise to the hedonic regression. The first he calls the Houthakker approach. In its simplest form, this model assumes utility is a direct function of the characteristics of a dwelling, and there exists a price schedule for characteristics, which consumers take as given. Maximization of utility, subject to a standard budget constraint and the price schedule for characteristics, implies a hedonic relation with the properties needed for price index construction and comparative analysis of housing quality. Muellbauer argues that the Rosen (1974) approach is basically an extension of the Houthakker approach. According to Muellbauer, all these models require the assumption that consumers face a fixed and known price schedule for characteristics which is based upon production costs. That is, it reflects supply conditions, not demand conditions.

The second family of models discussed by Muellbauer is developed from Fisher and Shell's (1971) "simple repackaging hypothesis." This model asserts that each "market good has a quality index which is a function of a set of physical characteristics" and which is "independent of market variables" (Muellbauer, 1974, p. 988). In other words, the quantity of a particular commodity can be expressed as a function of the characteristics and the relation is not influenced by supply or demand.

The thread which unifies these models is that they all give rise to the desired hedonic relation only under conditions which can be viewed as quite restrictive. In other words, the search for a rigorously derived basis for the hedonic relation is incomplete.

Does it make sense to estimate a model which does not yet have firm basis in theory? Yes. The reason is that time and time again attempts to estimate hedonic relations have produced results highly consistent with simple intuition. This kind of empirical support is hard to ignore. The housing prices we observe in the marketplace are related to structural characteristics of dwellings, location, neighborhood and tenant. In this paper we attempt to obtain precise estimates of the relationship between housing characteristics and housing prices.

The situation is analogous to the aggregation problem which haunts the study of macroeconomics. The frequently estimated macroeconomic relations are only derivable from a microeconomic theory of the consumer under restrictive conditions. Yet such relations continue to be studied and estimated. Why? Because the relationships estimated make sense and appear to be quite strong.

In summary, the search for a widely accepted and rigorously derived basis for the hedonic relation is not yet finished. Ideally, the model should not only demonstrate that a hedonic relation does exist, but should also make clear the conditions under which the relation can be used to construct constant quality price and quantity indexes in different markets. Until such a model is developed, the principal basis for hedonic analysis is relatively simple and intuitive. We shall see that most of our results are consistent with such a basis.

SECTION 2.2: EMPIRICAL IMPLEMENTATION

Overview

This section describes in some detail the actual specification used to estimate the hedonic relation. First, we describe the data, which is gleaned from the metropolitan Annual Housing Surveys. This is followed by our justification for using the SMSA as our definition of a housing market. Then we describe the functional form estimated and the construction of variables. The search which led to this specification is also described. In general, our present specification results from improvements to the model described in Follain and Malpezzi (1980a), so we highlight these changes and discuss their importance. Finally, we present evidence on the effect of some omitted variables. Many previous hedonic studies have included variables such as distance from the central business district, lot and house size, and certain kinds of neighborhood information, which are omitted from our specification because the AHS data do not include the information necessary for their construction. We show that the hedonic estimates we obtain are useful despite the lack of some desirable information.

Data

The data used in this study are from the 1974, 1975 and 1976 metropolitan Annual Housing Survey (AHS). The survey is designed and sponsored by the Department of Housing and Urban Development, and is conducted by the Bureau of the Census. Its purpose is to collect data on certain indicators of housing and neighborhood quality. The survey is designed

to be compatible with the decennial Census of Population and Housing, but the AHS includes data on characteristics not included in the Census.

Separate surveys are carried out for the U.S. as a whole and for separate SMSAs. The data used in this study are from 59 SMSA surveys carried out from April 1974 through March 1975 (1974 SMSA surveys), from April 1975 through March 1976 (1975 SMSA surveys), and from April 1976 through March 1977 (1976 SMSA surveys). Although these SMSAs are widely distributed geographically, the SMSA data are not necessarily representative of the country as a whole. An SMSA comprises a central city of at least 50,000 population and one or more contiguous counties. Minneapolis-St. Paul, San Francisco-Oakland and Los Angeles-Long Beach contain two central cities each. Although SMSAs include the rural portions of the contiguous counties rural America is underrepresented. Also, the smaller SMSAs are underrepresented. There are 159 SMSAs with populations of 200,000 or more, but the AHS samples 59 of the largest.

The survey data are from personal interviews with a dwelling's occupants. The enumerators read the questions directly from a copy of the survey, which is reprinted in the printed report available for each SMSA.[1] The survey includes questions about household characteristics such as family size, race, and income; dwelling characteristics such as number of rooms and the presence of various defects; opinions of neighborhood characteristics, and some information on location. There are hundreds of questions in each survey, so no attempt will be made to summarize them here.

---

1. U.S. Bureau of the Census (1976, 1977, 1978).

The survey contains much of the information needed to estimate hedonic equations for the prices of housing characteristics. Structural characteristics are well represented. Lot size and the floor area of the dwelling are not included, although these often make significant contributions to other hedonic studies. There is little objective information on neighborhood characteristics, although there are many questions relating to the occupant's opinion of his surroundings. Finally, there is limited information on the dwelling's location. The effects of the lack of this information will be discussed at the end of this section.

There are two sample sizes. Twelve SMSAs have samples of approximately 15,000 units. In these, about half of those surveyed reside in the central city, and about half outside it. The other 47 SMSA samples are about 5,000 units each, and the number of central city respondents is proportional to the number that actually live there. The sample is selected from three populations: (1) housing units from the 1970 Census of Housing and Population; (2) new construction, sampled from building permits issued since 1970; and (3) new units located in areas not covered by a permit issuing office.

The Census sample is stratified to insure adequate representation of various races, income classes, tenure groups, and family types. The public use copy of the AHS includes a set of weights to make the sample representative of the population as a whole, but the regressions are unweighted. Therefore the means of variables reported in the appendix are not the best estimates of the actual distribution of those characteristics in each SMSA. However, checking these means against weighted

population estimates reveals that they are often similar. We report unweighted means because they best characterize the data underlying the estimated regressions.

Several kinds of dwelling units and households are excluded from the estimation sample, either because they lack information needed for a hedonic regression or because respondents do not pay market prices for housing. Most obviously, vacant units and those households not at their usual residence are excluded. No cooperative or condominium owners are included because there is no survey information on the market value of their dwellings. Those who live in public or subsidized housing, or who do not pay cash rent, are excluded because their rent is not determined by the market. Other excluded categories include hotels, rooming houses, trailers, homes on more than ten acres, and owner-occupied dwellings which are part of commercial establishments or medical offices. Of course, any observations with missing data, or with missing responses allocated by Census, are dropped from the regression.[1] Although this seems like a formidable list of excluded categories, the great majority of non-vacant units remain in the sample. For example, in the Pittsburgh file there are about 4,700 occupied housing units, of which we use 947 in the renter estimation and 2,384 in the owner regression. That is, over 70 percent of the total non-vacant units are still included after sample selection.

---

1. There is one minor exception. See the discussion of the variable DFECT, below.

Market Definition[1]

Just as each market for apples produces a market clearing price for apples, each housing market produces a set of hedonic prices. This means that each set of hedonic prices we estimate must be derived from a set of observations from the same housing market. To use too broad a geographical definition of a housing market would produce biased estimates from an improperly aggregated sample. To use too narrow a definition would produce inefficient estimates because the estimates would not be based on all available information.

Much debate has centered on the precise definition of a housing market (Schnare and Struyk, 1976; Murray, 1978). Although most agree it is no larger than an SMSA, finer breakdowns are possible. The principal geographical possibilities are to divide the SMSA into central city and suburban markets, or even further into census tracts or neighborhoods. It is also possible to think of an SMSA market segmented by the kinds of households they serve. For example, separate markets may exist for blacks and whites, due to racial prejudice. Markets could also be defined in terms of housing quality (De Leeuw and Struyk, 1975).

If one believes in the existence of submarkets within an SMSA, there are basically two ways of dealing with them in the estimation of hedonic equations. First, separate regressions could be estimated for each submarket. This implies rather extreme separation because it assumes all the hedonic prices are different in each submarket. The second alternative is to introduce dummy (or indicator) variables for each submarket. This is

---

1. This discussion is from Follain and Malpezzi (1980a).

is more restrictive than the first alternative in the sense that it forces the coefficients to be equal in each submarket. Only the constant term, or the base price, is allowed to differ across submarkets.

This paper adopts the second alternative. The SMSA is defined as the basic housing market, although the rental market is separated from the owner-occupied market. Owner and renter markets may be closely related, but it is not clear how to compare rents and values, which would be necessary if owners and renters were pooled. Two submarket divisions are hypothesized within each SMSA market--central city versus suburbs and black versus white households. The submarkets are assumed to affect the base price or rent of a unit (constant term), with indicator variables for the race and location of a household. The coefficients of these variables represent the base price differential between the submarkets.

There are several reasons for this treatment of submarkets. The first is the strong a priori belief that the metropolitan area is the appropriate definition of the housing market. This is based upon the traditional urban economic analysis of Alonso (1964), Mills (1967), and Muth (1969). Second, such a definition is most appropriate for the long run purposes of the research effort, of which this paper is one part. In that larger research effort, the hedonic estimates will be used to analyze variation in Fair Market Rents (ceilings for HUD Section 8 rent subsidies) among metropolitan areas.[1] The impact of intermetropolitan differences can be best highlighted by using the SMSA as the subject of

---

1. See Ozanne and Thibodeau (1980), and Follain (1979), for examples of the work to be undertaken.

analysis. Third, the data are neither precise enough nor numerous enough to permit finer breakdowns. Census tracts are not identifiable, and breakdowns by race produce very small samples in many SMSAs. Finally, available statistical tests which can be used to study the existence of submarkets are probably not precise enough to warrant a purely empirical approach to defining submarkets (Schnare and Struyk, 1976).

Earlier work described in Follain and Malpezzi (1980a) included tests for segmentation based upon house quality as measured by household incomes. Analysis of metropolitan housing markets using The Urban Institute Housing Model (De Leeuw and Struyk, 1975) has concluded that markets are separated by housing quality. However, until the hedonic model is estimated, no measure of housing quality is available upon which to divide the sample. Rent or value is a possible measure upon which to divide the sample, but it is rejected because estimates obtained using a sample truncated upon the dependent variable are subject to serious bias. Another possibility is to divide the sample based upon household income (adjusted for household size). The idea is that income is positively correlated with housing quality, so splitting the sample upon income effectively splits the sample upon quality.

This was done for the eight SMSAs with which Follain and Malpezzi carried out their specification search. Separate regressions were run to high-income and low-income households for each tenure type. F-tests were computed, and in only four regressions (of a possible sixteen) was the null hypothesis of equality of the coefficients rejected. This suggests that splitting the sample using income as a proxy for quality is not appropriate.

There is one problem of empirical implementation which is unique to this study: estimating the two equations in 59 SMSAs. Most studies are for a particular market, and the final set of estimates which a particular paper reports is often the end product of much experimentation in that particular market. We do not experiment in each SMSA because such a process would be extremely expensive, and because we want to compare the final set of results across markets.

Choice of Functional Form

There is no strong a priori notion of the correct functional form. In earlier work, Follain and Malpezzi (1980a) estimated a linear functional form as well as a log-linear (semi-log) specification. In our present work we choose the log-linear for five reasons.

First, the semi-log form allows for the joint determination of expenditures in the regression. That is, the semi-log model allows for variation in the dollar value of a characteristic so that the price of one component depends in part on what else is in the house. For example, with the linear model, the value added by central airconditioning to a six room house is the same as to a ten room home. This seems unlikely. The semi-log model allows the value added to vary proportionally with the size and quality of the home. This fact, all else equal, favors the semi-log model.

Second, the coefficients of a semi-log model have simple and appealing interpretation. That is, the coefficient can be interpreted as the percentage change in the dependent variable given a unit change in the independent variable. For example, if the coefficient of a variable representing central airconditioning is .13, then adding it

to a structure adds 13 percent to its value or the rent it commands. (Of course this is really the same as the joint-determination-of-prices advantage described above.)

Third, the semi-log form alleviates the common statistical problem known as heteroskedasticity, or changing variance of the error term. The presence of heteroskedasticity suggests you have not yet derived the best possible estimates. A total lack of heteroskedasticity is rare in applied work, but any simple transformation which reduces it and is otherwise acceptable is useful. Preliminary owner regressions showed that the semi-log model usually exhibited less heteroskedasticity than a linear model. Exhibit 1 presents the results of linear and semi-log regressions for Raleigh owners and renters, including residual plots. (Exhibits are found in numerical order at the end of the text. Tables are displayed within the text.) Heteroskedasticity is indicated if the plots exhibit a widening or narrowing pattern as you move from left to right. Notice that the problem seems greater for owners than for renters, and that there appears to be some improvement in the owner regression when the semi-log form is employed.[1]

Fourth, lacking a strong theoretical reason to prefer one form over the other, another useful criterion is explanatory power. However, the usual measure of explanatory power, the R^2 statistic, cannot meaningfully be directly compared between two regressions with different (linear versus logarithmic) dependent variables. Follain and Malpezzi (1980a, p. 27) reported that, using an appropriate statistical test of

1. For more on heteroskedasticity and the examination of residual plots, see Draper and Smith (1966), Chapter 3.

the hypothesis that the explanatory power of the two regressions is the same, they were only able to reject the semi-log model in favor of the linear model in only a few of their 39 SMSAs.[1] We conclude that explanatory power strongly favors neither model.

Fifth, the model must be computationally feasible. Alternatives to the linear and semi-log forms exist, but they are expensive to implement and difficult to interpret for large data bases such as the AHS.[2] That is why this discussion has revolved around the linear versus semi-log model.

Finally, we note that our independent variables are mostly dummy (or indicator) variables. This allows us maximum flexibility in estimation. Earlier estimates such as those described in Follain and Malpezzi (1980a) often constrained coefficients. For example, if the number of bathrooms is entered as one variable, then the percentage change in rent from adding a second bath is forced to be the same as the change from adding a third. The percentage change may well differ as more are added, so we try to use indicator variables wherever possible. When continuous variables are needed (such as in the age variable) because of the large number of possible values, we try to use higher-order terms (squares and cubes) to allow maximum flexibility. The actual construction of these variables will next be discussed in detail, followed by a description of the specification search that yielded these variables.

---

1. A discussion of the method of comparing explanatory power, known as the Box and Cox test, is in Rao and Miller (1971), pp. 107-11.

2. An example is the general transformation suggested by Box and Cox (1964). Such an estimation would be prohibitively expensive to undertake with our data, and the results are unwieldly for price index construction.

Variable Definitions

The variables which were included in our final specification are defined in Exhibit 2. The means of the variables, the ranges of the means (by SMSA), and some other summary statistics can be found in Exhibit 3. The next few paragraphs set forth our principles of variable construction. Then the variables are discussed group by group.

Variable Construction

Whenever possible, variables are coded as indicator variables.[1] For example, for renters the number of bedrooms is coded into the variables BED0, BED2, BED3, BEDG4. The first 3 take on the value 1 if there are 0, 2, or 3 bedrooms, respectively; otherwise they are 0. BEDG4 takes on the value of the number of rooms if that number is greater than 3, and is 0 otherwise. BEDG4 is not constructed as an indicator variable so we can discern, say, 5-bedroom dwellings from 4-bedroom dwellings without separate indicator variables. This is necessary because until we run the regression we don't know if there are any 5- or 6- or 7- or 8-bedroom dwellings in the sample, so we can't be sure an indicator variable will work. Note that a 1-bedroom dwelling takes on the value 0 for all 4 variables. This is the base case, or the number of bedrooms represented by the constant term. Table 1 gives some examples of how different numbers of bedrooms would be coded using indicator variables.[2]

---

1. Indicator variables are also commonly known as dummy, binary, dichotomous, or 0-1 variables.

2. See Maddala (1977), pp. 132-41 for a good introduction to indicator variables.