Table 1

Example of Indicator Variable Coding Scheme
for Number of Bedrooms

Actual Number of Bedrooms	Value of Intercept (Constant Term)	Value of BEDO	Value of BED2	Value of BED3	Value of BED4
0	1	1	0	0	0
1	1	0	0	0	0
2	1	0	1	0	0
3	1	0	0	1	0
4	1	0	0	0	4
5	1	0	0	0	5

Several variables (such as POOR, DEFECT, and BADHALL) are linear combinations of indicator variables. This constrains the implicit price of each condition included in one of these variables to be the same. The interpretation of individual coefficients is therefore difficult since we can't separate the effects of the several conditions in a variable. This form is used whenever it is likely that there would be insufficient observations of a condition in some cities to permit inclusion of a separate variable.

A few independent variables are treated as continuous, such as age of the structure. It's not feasible to construct a large enough number of indicator variables in such cases, so we add quadratic (squared) and sometimes cubic terms in addition to the linear to permit flexible estimation. For example, if only the linear variable AGE1 were included in the regression, a negative coefficent estimate would imply that rents or value decline at a constant rate with age. Adding a quadratic term (AGE1SQ) allows us to determine whether rents decline faster in earlier or later years.

Dependent Variables

The variables we wish to explain represent expenditures for housing services. CRENTLN is the dependent variable in the renter regression, and is the natural logarithm of monthly contract rent, as reported by the respondent. The dependent variable in the owner regressions is VALUELN. The enumerator asks homeowners the current market value of their dwelling, but instead of writing down the response, checks a box indicating which of fifteen intervals the response falls in. We recode these intervals to their midpoints, except for the highest interval.

Since the top interval is open-ended, and the distribution of house values varies widely from city to city, we estimate a different value for this interval for each city. The average property tax bill of people in the top category is estimated from a preliminary pass through the survey data. The tax rate for these people is estimated after a careful perusal of the tax rate information in the AHS published reports. The estimated average tax bill divided by the estimated average tax rate yields the estimated average value in this category.

Structural Variables

The first group of variables listed in Exhibit 2 includes a relatively straightforward set of dwelling characteristics such as number of bathrooms (B1, B2, B3) and bedrooms (BEDO, BED1, and so on), number of other kinds of rooms (R1, R2, etc.), types of heating and cooling systems (SHEAT, RHEAT, EHEAT, ROOMAC, CENTAC), structure type (SFATT, SFDET, DUPLEX, ELEVP, NGT50), and age of structure (AGE1, AGE1SQ, AGE1CB, DAGE). There are other structural variables which are related to what is loosely termed housing quality, such as the absence of plumbing, presence of holes or cracks in interior surfaces, basement or roof leaks, and the presence of rats, among others.[1] These include NORAD, POOR, NOPRIVCY, NOUT, BADHALL, and DFECT.

Neighborhood Variables

Several variables measure the quality of the respondent's neighborhood. Most of these are based upon the opinion of the household. Such opinion data are not generally used in economic studies, so it is interesting to see if opinions are systematically related to rent and value.

The first three neighborhood variables are constructed from the household's rating of the street upon which the unit is located (EXCELN, GOODN, POORN).[2] A fair rating is the omitted category. The next variable, ABANDON, is constructed from the interviewer's answer to the question: Is there abandoned or dilapidated housing on the street? (Yes = 1, No = 0. The occupant is asked a similar question and the correlation between occupant and interviewer responses is quite high.) Finally, two renter variables are constructed from questions about specific neighborhood conditions (LITTER and NOSHOPS). LITTER takes on the value 1 if there is trash or litter on the respondent's street,

[1] The presence of rats, which is used in determining the value of the variable DFECT, is the one exception to our rule of deleting observations with missing values for any variable. Since recent movers are not asked this question, we assign them the mean response.

[2] The respondent was asked to rate the street in Wave I surveys, the neighborhood in Wave II and Wave III cities.

0 otherwise. NOSHOPS is a similar variable for the absence of convenient shopping.

Locational Variables

Other neighborhood variables represent geographical location. Variables in this set are all indicator variables for the county in which the unit is located or whether the unit is in the central city of the SMSA (CC1).[1] These variables undoubtedly represent many things such as the distance from the center city area and the quality of public services of the county. Forty-two of the fifty-nine surveys identify central city locations; the other seventeen have smaller populations and central cities are not identified because of Census confidentiality requirements. Seventeen SMSAs have at least one additional locational (county) variable. The Allentown SMSA has a county variable but none for central city. New York, with seven variables, has the most locational information.

BLACK and SPAN are indicator variables which equal one if the household head is black or Spanish, respectively. The persistence of residential segregation leads us to interpret these as neighborhood variables, since most minority households live in minority neighborhoods.

[1] Note there are forty-two central cities identified but only 40 SMSAs with the variable CC1. In the Philadelphia regressions, central city is the omitted category. In New York, the central city is represented by five variables. Locational variables are listed in the separate data appendix, available from the authors.

ntract Conditions

OWDS is the ratio of the number of persons in the household to the al number of rooms. CLOT, CLOTSQ, and DLOT are constructed from the ngth of time the tenant has resided in the unit. The first is a line term, the second quadratic, and the third a dummy for those who moved into their dwellings prior to 1950.

These coefficients are interpreted as price differentials faced by households which live in crowded units, or households who have resided in the unit for a long time. It is expected that long-time renters receive discounts. The coefficient of the crowding variable is expected to be positive for renters and negative for owners, reflecting the costs of faster depreciation. The hypothesis is that crowded dwellings depreciate faster because of harder use. Owners of crowded dwellings would find their value decreasing faster; landlords would require higher rents to recoup the additional costs.

It has been hypothesized that live-in landlords charge lower rents to attract desirable tenants, since they have to face them daily (Merrill, 1977). LLBLG is an indicator variable included in renter regressions for landlord living in the building. If this hypothesis is true we expect a negative coefficient for LLBLG.

We want our rental coefficients to reflect the price of housing structure and location, but some renters pay for additional services such as furniture, parking, and utilities. Indicator variables are used to identify differences in contract rent due to these additional services. FURNINC and PARKINC take on, respectively, the value one if furniture or parking are included in contract rent, and are zero otherwise. HEATINC and NHUINC are similar variables for heat and non-heat utilities included in contract rent.

Measuring Inflation

Housing prices do not remain constant over time, and the Annual Housing Survey is given over the course of a year (April to March). The month of interview is recoded into the variable Q. The first month of the survey, April, is zero, May is one, and so on. The semi-log functional form of the regression allows us to interpret the coefficient of Q as the average monthly percentage change in the price of housing.

Renters often pay for utilities as well as for housing structure and location. It is quite possible that housing utility inflation rates differ from inflation rates for other characteristics. The variable QHEAT is another time trend, similar to Q, except that it is zero whenever heat is not included in rent. The coefficient of Q then measures inflation in rents due to changes in the price of structure and location. QHEAT measures the difference in inflation rates between those who pay extra for heat and those who do not.

Locational differences in demand for housing, as well as differences in supply costs, can result in differing rates of inflation in different locations in the same SMSA. The variable FORAY is an interaction term which measures the difference between inflation in the central city and its suburbs. It is entered in the forty-two owner regressions for which we have the necessary locational information.

The Specification Search

Now that we have described the variables in some detail we will discuss the method used to arrive at this specification. Briefly, the current specification is an extension of that used by Follain and Malpezzi (1980a). The criteria used to choose variables are: (1) consistency with the theory of hedonic indexes outlined in Section 2.1, and (2) the variables yield estimates of the correct sign, and statistically different from zero, in preliminary regressions.

Why the Specification Search is Important

The goal of the search is a model which may be applied to fifty-nine different SMSAs. It is desirable to fit the same specification to each SMSA for the following reasons. First, analysis of the individual coefficients is greatly complicated by estimation of different models in different locations. Secondly, it is very costly and time consuming to fit one hundred eighteen different models. Third, the model chosen does incorporate most relevant information available from the Annual Housing Survey. This model performs well in every SMSA except Honolulu (see Chapter III).

How the Experimentation was Carried Out

As noted, the specification we employ is based on that used by Follain and Malpezzi. Their specification search strategy employed the following four steps:

(1) Intensive experimentation and estimation was carried out for the Los Angeles SMSA--one of the SMSAs in Wave I with a sample of fifteen thousand housing units. Wave II and Wave III SMSAs were not yet available when the research began. The products of this stage were several different specifications and a long list of variables.

(2) Several specifications produced by stage one were estimated for six other SMSAs: Boston, Dallas, Detroit, Minneapolis, Phoenix and Pittsburgh. From this stage, a smaller list and two specifications were selected.

(3) These two specifications were estimated for all SMSAs in Wave I.

(4) After one modification based upon stage three, two specifications were estimated in all thirty-nine SMSAs.

The results of this estimation were carefully perused for several months by members of the Housing Division of The Urban Institute, as well as others,[1] and several modifications were suggested. These improvements were tested in the following four steps:

(1) Several new variables were tested in Pittsburgh and Phoenix as part of an evaluation of the original AHS hedonic indexes (Ozanne, Andrews, and Malpezzi, 1979).

(2) More experimentation was carried out in three Wave III SMSAs: Baltimore, Denver and Raleigh. A preliminary specification was chosen for each tenure group.

(3) These specifications were estimated in fifteen SMSAs.

(4) Examination of the stage three results resulted in several changes, and a final owner and renter model were chosen. These were used to estimate the results presented here. These are the models described above.

Summary of Changes in the Hedonic Specification

For those readers familiar with the Follain and Malpezzi specification we summarize the major changes in the model estimated. This

[1] Suggestions for specification changes were also made by Edgar Olsen of HUD, and Sally Merrill and Dan Weinberg of Abt Associates.

list does not include every minor change in the way information is recoded into variables, but briefly outlines key differences. There are seven:

(1) New Dependent Variables. The old specification used by Follain and Malpezzi (F&M) used log of gross rent (contract rent plus utility payments). We use contract rent, relying on HEATINC and NHUINC (variables explained above) to account for utilities included in contract rent. Our coefficients are now interpreted as changes in rent for structure and location only, given a change in independent variables (dwelling characteristics). Also for owners, the open-ended value category now varies by SMSA.

(2) More Flexible Variable Construction. More extensive use of indicator variables and higher order (square and cube) terms results in fewer constraints in estimation. For example, the estimated price of a third bedroom is no longer constrained to be the same as that of a second bedroom.

(3) Recent Movers are now Included. Several service breakdown variables which performed poorly (wrong sign, insignificant) have been dropped. Examples are water and sewer breakdowns, and toilet breakdowns. Since recent (less than 90 days) movers were not asked the questions used to construct these variables, they were dropped from the F&M sample. We retain all recent movers. In particular, this assures a more reliable estimate of the inflation rate.

(4) Census-Allocated Responses are Dropped. For several key variables, including rent and value, the Census Bureau coders allocate

responses to respondents who do not answer the questions. When these observations are dropped the predictive power of the model is noticeably improved.

(5) Several Old Neighborhood Variables are Dropped. F&M included seven neighborhood variables constructed from the opinion questions in the AHS. Several performed perversely, and these are no longer included. Those that remain are the general neighborhood rating, now coded in binary form, and LITTER and NOSHOPS. The latter two are included in the renter regressions only.

(6) Property Tax Rates are Dropped from the Owner Model. The tax rate capitalization hypothesis states that the value of otherwise identical houses will vary by the differences in the present value of the future stream of tax payments (negative), and of services (positive). F&M included the log of the property tax rate in their owner model to account for capitalization. However, inclusion of this variable is likely to result in biased and inconsistent estimation. The tax rate is constructed from property taxes divided by value. That is, the dependent variable is used to construct one of the independent variables, so that regressor is correlated with the error term, violating one of the important assumptions of regression analysis.[1] Test regressions indicated that a tax rate variable probably picked up more of the

[1] When an estimate is unbiased, one expects to estimate the true value of the parameter on average. When an estimate is consistent, adding more observations gives more precise estimates. If a regressor is correlated with the error term, the estimates no longer have these desirable properties. See Wonnacott and Wonnacott (1970), chapter 7.

error term than any capitalization effect, so it was deleted from the final specification.[1]

(7) Several New Variables are Added. ROOMAC indicates the presence of room air-conditioners. SPAN is an indicator variable for Spanish head of household. It measures the premium or discount paid by Spanish households for housing of constant quality (insofar as our other variables account for a unit's quality). Like BLACK, it probably reflects neighborhood characteristics. LLBLG is a variable for the landlord's presence in the building.

City to City Differences in the Model

As noted above, one of the estimation objectives is to use the same model in each SMSA, for two reasons: computational efficiency, and cross-SMSA comparison of coefficients. There are two kinds of exceptions to this rule.

First, if there are no observations of a particular characteristic in the sample for an SMSA, the variable representing that characteristic must, of course, be dropped from the regression. Table 2 presents the modifications made to several SMSA models because of this data problem.

[1] See Thomas King (1977) for more on tax capitalization.

Table 2

Variables Dropped from Individual Regressions

SMSA	Tenure	Deleted Variable	Variable Description
Miami	Owners	SHEAT	Steam or hot water heating
Honolulu	Owners	SHEAT	
Honolulu	Renters	NORAD	Rooms without heat
Birmingham	Renters	PARKINC	Parking included in rent
Memphis	Renters	PARKINC	
Raleigh	Renters	PARKINC	
San Antonio	Renters	PARKINC	

Second, different SMSAs have different locational variables, because some public use files provide more locational information than others. Details on the interpretation of these variables is presented below in Section 3.3. Finally, some SMSA-specific models were estimated for those cities where our model performed less well than in most SMSAs. These results are discussed in Chapter IV.

Omitted Variables, and their Likely Effects

Now that we have discussed the model in some detail, it is useful to consider what is left out. The Annual Housing Survey does not contain information needed to construct several variables commonly used in hedonic estimation of rents and house values. In particular, several studies have emphasized the importance of distance to the central business district (CBD) or other employment centers (e.g., Muth, 1969), the area of the house and lot size (e.g., Noto, 1976), and objective neighborhood information (e.g., Kain and Quigley, 1970).[1]

To assess the effects of this omitted information on hedonic estimates, Ozanne, Andrews and Malpezzi (1979) estimated hedonic indexes using a data source which had some of the omitted information.[2] They concluded that the absence of this information made little difference in the predictive power of the equation. This means that work which relies on predicted rents and values, such as price index construction, may not be seriously affected by omitted variable bias.

Some problems remain regarding the interpretation of individual coefficients. Ozanne, Andrews and Malpezzi find that the estimates of individual coefficients are biased by lack of square footage, locational, and objective neighborhood information. Studies relying on estimates of individual coefficients are more likely to be affected by omitted variable bias than studies using predicted rents or values. Individual coefficients will be biased if omitted variables are correlated with some included variables. Ozanne, Andrews and Malpezzi find, for example, that the correlation between BLACK and omitted neighborhood characteristics imparts a downward bias to the race coefficient. On the other hand, omission of these neighborhood characteristics does not affect estimates of SMSA-wide inflation from

[1] Examples of "objective" neighborhood information used in other studies are median census tract income, school expenditures, and crime rates.

[2] The data were from the Demand Experiment of the Experimental Housing Allowance Program (EHAP) and were available for low income renters in Pittsburgh and Phoenix.

the variable Q, since the monthly samples are independent of location. Of course, cross-SMSA comparisons of biased estimates still yield useful information if the nature of the bias is known.[1]

[1] Examples of such studies include Follain and Malpezzi (1980b, 1980c, 1980d, 1980e).