APPENDIX A
Model Documentation

In this Appendix we present a detailed discussion of the development of the GSD public housing operating cost model. Following a general overview of the modeling process provided for non-technical readers, this appendix presents detailed information about:

- Data exclusions
- Variables that were tested but were not retained in the model
- Alternative specifications of variables that were retained in the model
- Alternative model specifications that were tested
- Statistical tests that were conducted to verify the robustness of our final specification.

1. GENERAL MODELING APPROACH

The following discussion is intended for non-technical readers who want to understand the details of GSD's approach to modeling operating costs. In this discussion we explain how to interpret the results of this important part of the research. In order to communicate GSD's plans accurately and precisely, the discussion uses (and defines) some technical statistical terms.

GSD developed a cost model through the use of multiple regression analysis. This is a statistical technique that allows an outcome measure (called the dependent variable) to be expressed as the result of the combination of characteristics that affect it (called the explanatory variables or independent variables) multiplied by their respective regression coefficients. In this case, GSD modeled operating costs per unit month, by utilizing data on the factors that drive operating costs (such as the age of a property, the market it is located in, etc.).

The final regression coefficients are the key product of the modeling exercise. Each explanatory variable (cost factor) has a regression coefficient that expresses in quantitative terms the extent to which the explanatory variable is found by the model to determine the dependent variable. Generally, an explanatory variable that has a larger coefficient is more strongly correlated with the dependent variable than an explanatory variable that has a smaller coefficient.[1] If the sign of the coefficient is negative, then the explanatory variable is negatively related to the outcome variable, meaning that an *increase* in the value of the explanatory variable is associated with a *decrease* in costs; if the sign of the coefficient is positive, then the explanatory variable is positively correlated with the dependent variable. The regression model holds other factors constant, while estimating each coefficient. Thus, the coefficient can be interpreted as the independent effect of each explanatory variable on the outcome, holding constant the values of all other variables in the model. The sections below discuss the details of the dependent variable (the operating cost that GSD is predicting) followed by a discussion of the explanatory variables tested by GSD.

GSD used the most common form of regression, Ordinary Least Squares (OLS).[2] OLS is a mathematical

[1] This statement assumes the two explanatory variables being compared have the same units of analysis. For example, both may be measured in dollars. The statement does not hold if the two explanatory variables have different units of analysis. Also, this statement assumes that both coefficients are statistically significant. In addition to coefficients, the model also generates a measure of the statistical significance of each coefficient, called the standard error of the coefficient. If a coefficient is not statistically significant, it means the relationship between the explanatory variable and the outcome variable cannot be precisely estimated, and GSD cannot with confidence interpret the coefficient as being different from zero, even though the coefficient might be very large.

[2] GSD considered the use of alternative estimators, including least average deviation (LAD) and iterative re-weighted least squares or robust regression, but found that the model results were not materially different, suggesting that the OLS results were not influenced by outliers.

optimization algorithm that produces the regression coefficients for a given model and a set of observations. We performed the estimation using the statistical software packages SAS and Stata.

2. WHY WE CHOSE FHA OVER OTHER DATA

The Federal Housing Administration (FHA) provides mortgage insurance for approximately 1.5 million multifamily units. The owners of these properties are required to submit audited financial statements to HUD on an annual basis. Since 1998, these financial statements have been submitted electronically. For the purposes of the Cost Study, GSD has assembled data on these properties from multiple sources or files, as described below.[3] The FHA inventory can be divided into two sub-sets: FHA unassisted and FHA assisted.

- **FHA Unassisted.** There are approximately 500,000 FHA unassisted units, representing at least 3,000 properties. These are properties with no underlying mortgage interest rate reduction or rental subsidy program.[4] FHA mortgage insurance is for non-luxury housing, and the average income of households living in FHA unassisted housing is estimated to be comparable to the income of households living in non-FHA-financed, non-luxury market rate apartment housing.[5]

- **FHA Assisted.** There are approximately 1,000,000 FHA assisted units, representing at least 11,000 properties. These are properties that are assisted with either a mortgage interest reduction program or a rental assistance program and are also insured with FHA. These properties are commonly divided into "older assisted" and "newer assisted" properties. The older assisted properties are those properties developed under the Below Market Rate Interest (BMIR) and Section 236 programs. These were not rental assistance programs, although most of these units either originally had Rent Supplement or Rental Assistance Program subsidies (later converted to project-based Section 8) or have since received project-based Section 8 assistance. The newer assisted properties are those Section 8 New Construction, Substantial Rehabilitation, and Moderate Rehabilitation properties that were developed with FHA mortgage insurance.[6]

Based on its review of potential databases, GSD decided to use the FHA database as the primary source of data for developing the operating Cost Model. An analysis of several large databases reviewed by GSD, as compared with FHA, is found in the Draft Research Design.[7] The basis for recommending FHA as the primary source for data is as follows:

---

[3] Although we makes reference throughout this document to the "FHA database", GSD actually had to construct this information on FHA housing from several different data sources at HUD. The process of constructing this database is described below, and is also explained in more detail in the Draft Research Design (July 9, 2001), located at: http://www.gsd.harvard.edu/research/research_centers/phocs/documents.html

[4] As with any multifamily housing, however, these properties may house families with Section 8 Housing Choice Vouchers.

[5] Because there is no underlying subsidy to these properties, HUD does not require any reporting of tenant incomes/demographics. However, 36 percent of FHA unassisted properties are in the central cities of metropolitan areas; 33 percent are in neighborhoods in which the median income is below 50 percent of area median income; and 24 percent had gross rents less than $500 per month in 1995. Meryl Finkel, et al., *Status of HUD-Insured (or Held) Multifamily Rental Housing in 1995,* Prepared for U.S. Department of Housing and Urban Development by Abt Associates, Inc., May 1999.

[6] HUD's database on the financial characteristics of FHA-insured multifamily properties also includes properties with direct loans made under the Section 202 direct loan program. For simplicity, this document refers to the entire database as an FHA database.

[7] The Draft Research Design is located at: http://www.gsd.harvard.edu/research/research_centers/phocs/documents.html. It should be noted that one of the findings from the analysis of different databases is the relative uniformity of accounting "Charts of Accounts." This was an area where GSD had previously anticipated a problem. Throughout the multifamily industry, there is growing uniformity in the accounting of operating revenue and expenses. This is especially true as real estate attracts institutional investors, who, to reduce transaction costs, prefer a standard way of comparing one property's

- **The FHA database is large, with wide geographic coverage.** The database includes financial data on 14,260 properties and almost 1.5 million units. All regions of the country and all types of locations — central city, balance of metropolitan area (suburbs), and non-metropolitan — are well represented. Table A.1 shows the number and percentage of FHA units and developments in each type of location and each region. The maps show the geographic distribution of FHA developments and of FHA units. (These two maps are similar but not identical because of variations in property size.) Figure A.1 shows the distribution of FHA developments, and Figure A.2 shows the distribution of FHA units.

Table A.1: Numbers of Units and Properties in FHA Database by Census Region and Type of Location
Data from 1996, 1997, and 1998 Annual Financial Statements[1]

	Central City		Balance of Metro Area		Non-Metro		Unknown		Total	
	Units	% of Total Units	Units	% of Total Units	Units	% of Total Units	Units	% of Total Units	Units	% of Total Units
Northeast	171,933	12%	81,423	6%	11,998	1%	19,865	1%	285,219	20%
Midwest	215,980	15%	131,346	9%	43,144	3%	21,524	1%	411,724	28%
South	298,369	20%	144,990	10%	69,465	5%	29,979	2%	542,803	37%
West	143,809	10%	70,024	5%	18,195	1%	5,657	<1%	237,685	17%
Total	830,091	56%*	427,783	29%	142,802	10%	77,025	5%	1,477,431	100%

	Central City		Balance of Metro Area		Non-Metro		Unknown		Total	
	Props.	% of Total Props.	Props.	% of Total Props.	Props.	% of Total Props.	Props.	% of Total Props.	Props.	% of Total Props.
Northeast	1,200	9%	701	5%	149	1%	141	1%	2,191	17%
Midwest	1,779	14%	1,069	8%	769	6%	207	2%	3,824	29%
South	2,286	17%	1,098	8%	1,028	8%	273	2%	4,685	36%
West	1,383	11%	693	5%	315	2%	76	1%	2,467	19%
Total	6,648	50%	3,561	27%	2,261	17%	697	5%	13,167	100%

* Because of rounding, percentages may not add to 100 percent.
[1] The final version of the Cost Model uses data from 1998, 1999, and 2000 Annual Financial Statements.

financial records with another. Among the different public sponsors of assisted housing – HUD/FHA, state and local finance agencies, tax credit issuers, etc.– the formats for operating budgets and year-end financial statements are also becoming more similar.

Figure A.1

Distribution of FHA-Insured Multifamily Units by Census Region and Type of Location

Data from 1996, 1997, and 1998 Annual Financial Statements

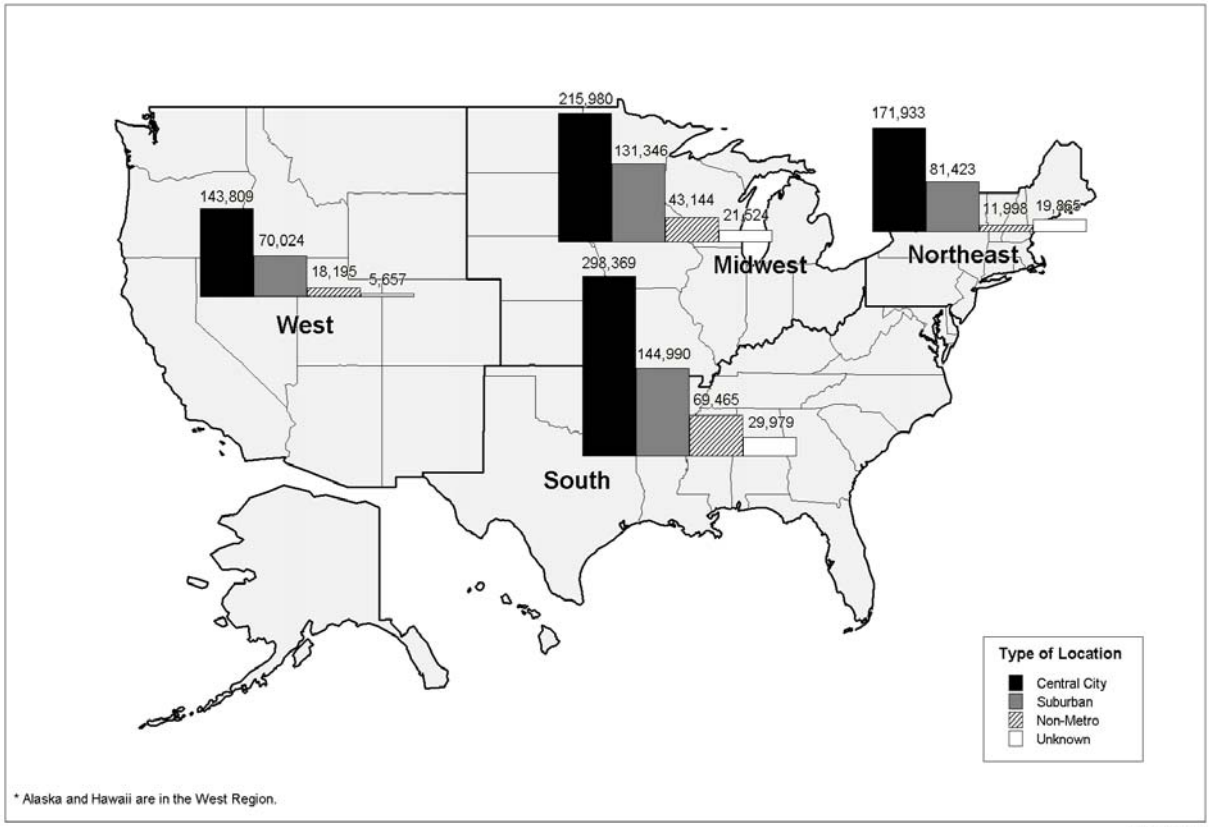
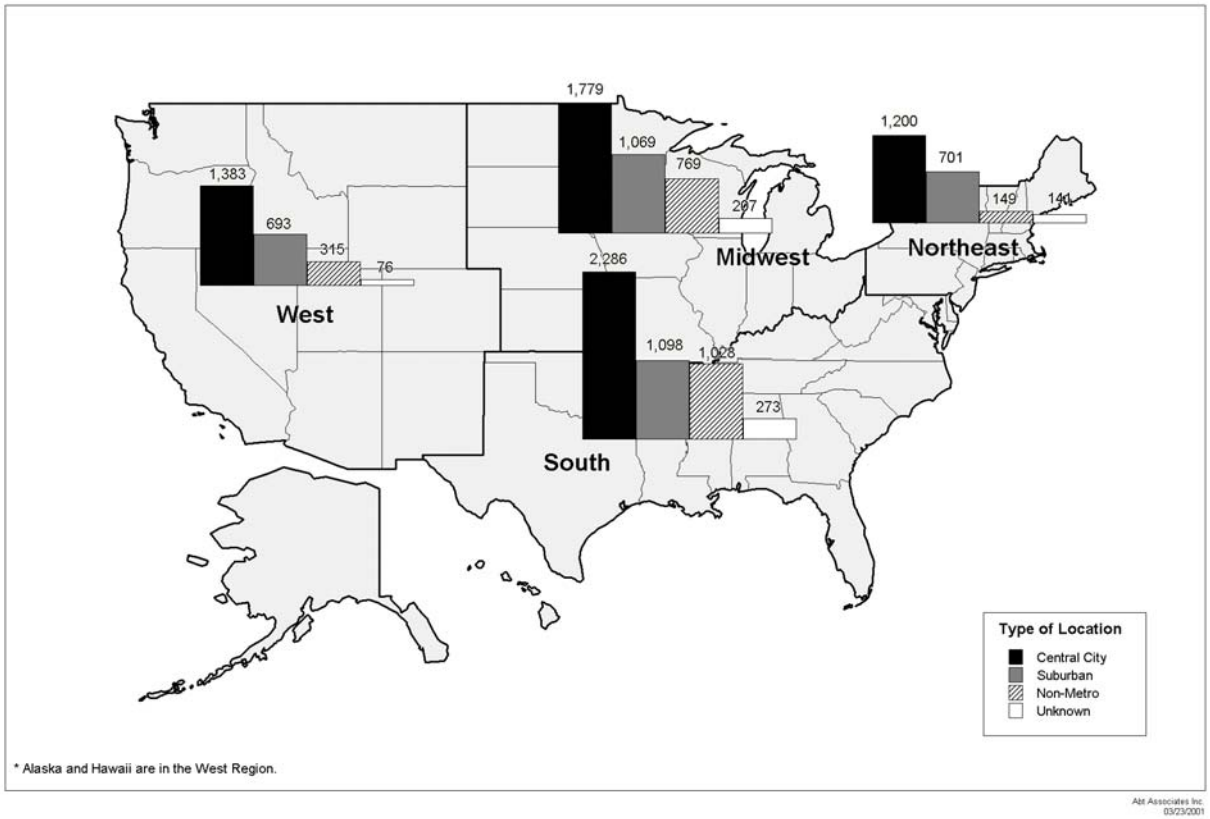


Figure A.2

Distribution of FHA-Insured Multifamily Developments by Census Region and Type of Location
Data from 1996, 1997, and 1998 Annual Financial Statements



- **The properties in the FHA database are the most comparable of the larger databases to public housing.** Tables A.2 – A.7, located at the end of this section, show the degree of overlap between the FHA multifamily stock and the public housing stock for a number of variables that may drive the costs of operating multifamily housing.[8] To be useful for creating benchmarks for the operating costs of public housing properties, a database for private multifamily housing need not have the same average characteristics of the public housing stock. However, it does need to include enough properties with each of the characteristics that are found in a number of public housing properties and that are likely to influence operating costs.

  In terms of geographic coverage, Table A.2 indicates that the units of the FHA stock are distributed quite similarly to those in the public housing stock across the nine Census divisions. In terms of property characteristics, as can be seen from Table A.3, while properties larger than 300 units or with three or more bedrooms are not typical within the FHA stock, there are a substantial number of properties with these characteristics. Public housing has a larger percentage of units with four or more bedrooms than the FHA multifamily stock. Nevertheless, FHA assisted properties include more than 20,000 units with four or more bedrooms and close to 147,000 three-bedroom units (Table A.4). Table A.5 shows that townhouses are a much more common type of building in public housing than in the FHA stock. Nonetheless, between FHA assisted and unassisted properties, there are more than 100,000 such units.

  There are not nearly as many FHA properties that are more than 30 years old as there are public housing properties (Table A.6). However, there are almost 34,000 units in FHA assisted properties and almost 24,000 units in FHA unassisted properties that have mortgages that are at least 30 years old.

  Table A.7 presents data (for FHA assisted properties) for resident characteristics that may affect the operating costs of multifamily housing. For example, in 34 percent of FHA assisted properties, half or more of the residents are single-parent families with children, which some focus group participants have suggested as a potential cost driver. Because GSD wants to test whether location in a distressed neighborhood increases operating costs, GSD has used a common indicator of neighborhood distress, whether more than 40 percent of the population of a census tract has income below the federal poverty level. GSD has identified more than 1,200 FHA assisted properties in such neighborhoods.

- **The FHA database has the most extensive property characteristics.** The FHA database also contains the most property variables of any other large existing database. Data elements that are available in the database GSD has created for FHA-insured multifamily properties include:

  - property size;
  - unit mix (numbers of bedrooms);
  - building type;
  - building age (both age of mortgage and date of first occupancy are available);
  - occupancy type (family or elderly/disabled);
  - location type (central city, balance of metro area, or non-metro);
  - state;
  - MSA;
  - census tract;
  - sponsor type;
  - an indicator of financial soundness;
  - an indicator of physical distress;

[8] Total units or property counts may not be the same across tables because of missing values in the stratifying variables. For example, Table A.5, showing the distribution of building type, does not include any properties or units whose building type information is missing in the database.

- information on the income, source of income, and demographic characteristics of occupants, and;
- an indicator of the level of distress of the census tract in which the property is located.

Many of these variables are also available for the public housing stock, so that a model based on them could be applied to public housing to derive benchmark operating costs for public housing properties with different sizes, building types, bedroom size distributions, locations, and tenant characteristics. Other variables are not relevant to public housing (e.g., sponsor type) or obtainable for public housing at the property level (e.g., financial soundness), but may be useful for ensuring that a model of operating costs reflects the costs of well-managed multifamily properties.

- **The FHA data are accessible.** In addition to being one of the largest, the FHA database is the most readily accessible of any of the major databases on assisted housing. Importantly, the data that make up the FHA database are available on an on-going basis, should they be needed to re-estimate a model to reflect changes that may occur over time to the factors that drive the costs of operating multifamily housing.

    - The Annual Financial Statements are available each year, with a time lag of three to six months following the end of the year.

    - Mortgage and property characteristics data are updated periodically as properties join or leave the FHA-insured inventory or when property characteristics change.

    - Information on financial and physical inspection scores are likely to be updated periodically.

    - Tenant characteristics data are reported annually to HUD (for assisted properties).

    - US Census data for 1990 are used to provide tract-level neighborhood characteristics in the operating cost model. If plans for fielding a rolling American Community Survey as an alternative to the Census long form go ahead, Census tract characteristics will be available in the future much more frequently than every decade.

- **Data Integrity.** Annual Financial Statements are submitted by the owner of each property or his/her agent at the end of the property's fiscal year. They are subsequently audited, and any values changed by the independent audit are changed during the next year's submission. Owners submit this information knowing that it is subject to audit.

  Since 1998, financial statements have been submitted electronically. Data are then subjected to extensive quality control checks by HUD's Real Estate Assessment Center (REAC), as they are used to determine whether a property is financially troubled or at risk of becoming financially troubled. Users of the operating cost elements of these financial data believe the information to be accurate.

  Property characteristics data are entered into HUD's Real Estate Management System (REMS) by the HUD field office staff person responsible for the property, the Asset Manager, and are used directly in program operations. Some of the REMS data fields are used directly in the process for submitting Annual Financial Statements, governing which screens are available to an owner or agent who logs onto the system for submitting those statements.

  GSD has made extensive efforts to investigate the quality of individual variables on property characteristics that are believed likely to be associated with property-to-property differences in operating costs.

Inspection scores developed by REAC are subject to variations among observers in judging scores appropriate for individual elements on the inspection form. However, for the purposes for which GSD used the REAC scores, these variations were not important. GSD used a REAC score cutoff to remove from the analysis properties that may be physically troubled; see the discussion under REAC scores for more information.

Data on tenant characteristics may be subject to integrity issues, including data entry errors and lack of accurate submission of data by tenants or management staff; however, these issues are likely to be the same across regions or program or housing types, or from year to year. Tenant characteristics were only available for the FHA assisted housing stock. However, GSD verified that there are very strong correlations between tenant characteristics and, respectively, neighborhood characteristics and unit size. The strong correlations found provided justification for including only tract characteristics and property characteristics (including unit size) in our Cost Model, which allowed us to use both the assisted and unassisted FHA housing stock.

- **Accounting Processes.** Operating costs are reported in Annual Financial Statements in categories that correspond well to line items in public housing's Chart of Accounts.

- **The FHA database has multi-year data for most properties.** There are two reasons why it us desirable to have data on operating costs for more than one year.

  First, because of year to year fluctuations in operating costs at the property level, it is desirable to consider the operating costs that serve as a benchmark for public housing operating costs to be costs for particular properties that have been averaged over two or three years.

  Second, if a model based on FHA costs is used to benchmark the costs of public housing properties, it may be desirable to use an inflation index derived from FHA operating cost data to update those benchmarks to future years. (This is not the only possible source of such an inflation index, however. For example, a wage index derived from Bureau of Labor Statistics data could also be used for this purpose, as could the Operating Cost Adjustment Factors published by HUD.)

- **Cost patterns track other databases.** Patterns of costs for FHA are similar to patterns of costs found in the other databases that GSD reviewed, including real estate industry databases and the Rural Housing Service (RHS). The RHS of the Department of Agriculture, formerly the Farmers Home Administration, provides direct loans to approximately 360,000 multifamily housing units (13,127 properties) in rural areas. About 12,969 (99 percent) of these properties have rental subsidies provided through RHS and an additional 893 properties (7 percent) receive Section 8 project-based assistance (a number of properties have both types of subsidy). The real estate industry databases that GSD reviewed included data from the Institute for Real Estate Management (IREM), and data from the National Apartment Association (NAA). IREM publishes two sets of apartment income and expense data: Conventional apartments and Federally assisted apartments. These surveys/reports include approximately 600,000 unassisted units and 100,000 federally assisted units. Like IREM, NAA publishes an annual survey for conventional and federally assisted properties, which represent approximately 620,000 and 95,000 units, respectively.

Table A.2: Distribution of Units by Census Divisions

	FHA				PUBLIC HOUSING	
	Unassisted		ASSISTED		Number of Units	Percent of Units
	Number of Units	Percent of Units	Number of Units	Percent of Units		
New England	11,694	2%	77,329	7%	72,446	6%
Middle Atlantic	62,883	12%	151,109	14%	318,890	25%
East North Central	110,718	20%	209,637	20%	190,291	15%
West North Central	49,754	9%	76,048	7%	66,305	5%
South Atlantic	129,236	24%	194,162	18%	222,792	17%
East South Central	43,115	8%	80,817	8%	125,080	10%
West South Central	48,852	9%	92,392	9%	184,010	14%
Mountain	41,224	8%	43,683	4%	31,261	2%
Pacific	48,699	9%	124,648	12%	75,056	6%
TOTAL	546,175	100%	1,049,825	100%	1,286,131	100%

Table A.3: Property Size (Total Number of Units in the Property)

	FHA				PUBLIC HOUSING	
	Assisted		Unassisted			
	Number of Properties	Percent of Properties	Number of Properties	Percent of Properties	Number of Properties	Percent of Properties
Less than 50 units	3,001	27%	333	10%	6,571	47%
50 to 99 Units	3,678	33%	687	21%	3,543	25%
100-199 Units	3,281	30%	1,328	41%	2,409	17%
200-249 Units	596	5%	377	12%	495	4%
250-299 Units	193	2%	211	6%	228	2%
300 or more Units	242	2%	332	10%	673	5%
TOTAL	**10,991**	**100%**	**3,268**	**100%**	**13,919**	**100%**

Table A.4: Distribution of Units of Different Bedroom Sizes (# of Bedrooms in Unit)

	FHA				PUBLIC HOUSING	
	Assisted		Unassisted			
	Number of Units	Percent of Units	Number of Units	Percent of Units	Number of Units	Percent of Units
0 Bedrooms	58,714	6%	24,259	5%	94,950	7%
1 Bedrooms	429,415	43%	169,416	37%	405,488	32%
2 Bedrooms	337,344	34%	222,104	49%	396,502	31%
3 Bedrooms	146,675	15%	36,736	8%	299,729	23%
4 or more Bedrooms	20,134	2%	2,327	1%	89,463	7%
TOTAL	**992,282**	**100%**	**454,842**	**100%**	**1,286,132**	**100%**

Table A.5: Building Type Distribution

	FHA				PUBLIC HOUSING	
	Assisted		Unassisted			
	Number of Units	Percent of Units	Number of Units	Percent of Units	Number of Units	Percent of Units
Detached	11,172	1%	3,548	1%	35,257	3%
Row-type/Townhouse	74,176	7%	29,001	5%	297,370	23%
Semi-Detached	24,631	2%	5,611	1%	120,592	9%
Walkup	466,034	44%	357,037	67%	146,963	11%
Hi-rise/ Elevator	318,060	30%	94,072	18%	389,731	30%
Mixed	164,894	16%	42,438	8%	296,201	23%
TOTAL	1,058,967	100%	531,707	100%	1,286,114	100%

Table A.6: Property Age*

	FHA				PUBLIC HOUSING	
	Assisted		Unassisted			
	Number of Units	Percent of Units	Number of Units	Percent of Units	Number of Units	Percent of Units
< 15 years	153,756	15%	386,541	70%	63,901	5%
15-30 years	863,456	82%	138,833	25%	482,972	38%
30+ years	34,816	3%	23,573	4%	739,258	57%
TOTAL	**1,052,028**	**100%**	**548,947**	**100%**	**1,286,131**	**100%**

* In the case of FHA properties, this is the age of the mortgage rather than the age of the property.

Table A.7: Tenant and Tract Characteristics: Public Housing and FHA Assisted Properties*

Variable Means	Public Housing N**	Public Housing Mean	All FHA Properties, Assisted N	All FHA Properties, Assisted Mean
% HHs w/income < $5,000	10,639	20.51	8,659	16.03
% single parent families (with children)	10,679	38.74	8,671	33.80
Average household size	10,709	2.36	8,671	1.96
% age 62 +	10,675	31.62	8,671	41.27
% of households w/majority of income from work	10,408	23.80	8,545	28.61
% households w/majority of income from AFDC/TANF/GA	10,408	18.43	8,545	11.82
Average HH income as % of local median	10,537	24.17	8,311	26.15
% w/disability, as % of HHs < 62 years	9,057	30.56	6,401	26.94
% poor in tract	8,399	37.05	8,676	23.33
Variable Distributions	**N**	**Distribution**	**N**	**Distribution**
Percentage of property households with majority of income from AFDC/TANF/GA				
0 - 9 %	5,066	36%	4,922	52%
10 - 19%	2,083	15%	1,433	15%
20 - 39%	2,457	17%	1,724	18%
40 % or more	802	6%	466	5%
data missing	3,637	26%	978	10%
Percentage of property households that are single parent families with children				
0 - 24 %	3,693	26%	3,813	40%
25 - 49%	2,360	17%	1,661	17%
50 – 74%	3,187	23%	2,270	24%
75 % or more	1,439	10%	927	10%
data missing	3,366	24%	852	9%
Distribution of properties by census tract poverty rate				
0 - 9% poor in tract	1,045	7%	1,837	19%
10 – 19% poor in tract	2,128	15%	2,740	29%
20 – 29% poor in tract	2,035	14%	1,708	18%
30 – 39% poor in tract	1,385	10%	1,169	12%
40% or more poor in tract	1,806	13%	1,222	13%
data missing	5,646	40%	847	9%

*Data Source: Picture of Subsidized Housing and FHA Database.
**N = number of properties for which GSD currently has data. Means are weighted by the number of units in the property; frequency distributions are not weighted.

3. HOW WE ASSEMBLED THE FHA DATA

The dataset used for analysis in this study combined information from several sources. These sources are described below.

- **HUD's Office of Housing Real Estate Management System (REMS).** This administrative database contains a wealth of information at the development level. For example, it includes variables on the number of units in each property, the distribution of units by bedroom size (i.e., one-bedroom, two-bedroom, etc.), building type (high rise, garden, townhouse, etc.), mortgage sponsor type (for-profit, non-profit, limited dividend), occupancy type (family, elderly/disabled), HUD program (section of the authorizing legislation) and the location of the property.

- **HUD's Office of Policy Development and Research (PD&R*) A Picture of Subsidized Households* 1998 database.** From this, GSD has added to the FHA database variables describing the characteristics of the tenants occupying each assisted property, including income, source of income, and size and structure of the household. Data on tenant characteristics are aggregated to the property level from the Tenant Rental Assistance Characteristics (TRACS) system. TRACS is a household-level data system to which assisted housing property managers report each month's information from the certification of income and characteristics of household members that is required at program admission and annually thereafter.

- **1990 Census of Population and Housing.** From the US Census, GSD has added variables that can serve as proxies for the level of distress of the neighborhood represented by the census tract in which the property is located.

- **Real Estate Assessment Center (REAC) Physical Inspection Scores.** HUD's Real Estate Assessment Center has inspected each property in the FHA database at least once. GSD has obtained from REAC both the most recent overall physical inspection score for each property and a set of sub-scores associated with capital needs.

- **Office of Housing's Field Office Multifamily National System (FOMNS).** This database includes additional property characteristics, such as original age of construction, square footage, heating and cooling system, building materials, and management type.

- **Office of Housing's F-47 database.** This database contains the mortgage endorsement date information, which is useful for determining the age of a property's mortgage.

- **Fair Market Rents.** This database contains HUD-estimated fair market rents by unit size for the entire country. These data are used in conjunction with actual average property rents to form a proxy for housing quality.

Table A.8: Data Elements Tested in the Cost Model*, by Source

Variable	Notes	Data Source
Operating Costs	See definition below	REAC Annual Financial Statements
Property Size (number of units)		REMS
Property Age	Age of first mortgage	F-47
Property Age	Age of most recent mortgage	F-47
Property Age	Age dated from initial occupancy	REMS
Number of bedrooms per unit	(Percentage of units of each size)	REMS
Building Type	(detached, semi-detached, garden, walk-up, high-rise, row-house, mixed)	REMS
Clientele	(Family or Elderly)	REMS
Central City		REMS
Census Tract Characteristics	(poverty rate; percentage of single parent households; etc.)	1990 Census of Population and Housing
Tenant Demographics	(percent employed; average family size; etc.)	HUD's A Picture of Subsidized Households database
Number and percentage of capital defects		REAC
Physical inspection score		REAC
Mortgage subsidy		REMS
Assistance Type	(Unassisted, Older Assisted, Newer Assisted, Section 202)	REMS
Ownership Type	(For profit, non-profit, limited dividend)	REMS
Percentage of units with project-based Section 8 assistance		REMS
Average contract rent		REMS
Fair Market Rent		REMS
Average square feet per unit		FOMNS
"Troubled" Indicator	(whether property ever received a "troubled" designation)	REMS
Property Owner	Used to identify multiple properties held by a single owner	REMS

* Note: each of these data elements is described in more detail below.

4. DATA CLEANING

4.1 Outliers in the outcome variable

Prior to conducting any analysis of the FHA data we eliminated observations with extreme values for the outcome variable, operating costs per unit per month. In all regression models the inclusion of invalid outliers produces less precise estimates, and in models based on means, such as ordinary least squares, the inclusion of outliers will produce inaccurate results if outliers are more extreme at one tail of the distribution than another. For outcome variables that are truncated at one end (such as operating costs, which are truncated at zero) there is the risk that outliers in the upper tail could bias estimates upwards.

Our first step in handling outliers was to eliminate extreme values of the outcome variable. We were fairly conservative in our initial data cleaning, and we only dropped values below $50 or greater than $800. After Field Testing, however, we narrowed our analysis to a more restrictive range, reflecting the range that our testers felt were plausible: only observations with operating costs between $135 and $650 were permitted in the final model.

As a second step towards reducing spurious variability in the data, we used a three-year average value of operating costs as our outcome variable. Not every observation had all three years of data; however, we restricted our sample to observations with at least two years of operating cost data.

Finally, in order to test whether outliers might be influencing our model results, we re-estimated our model using two estimation techniques that are more robust to the presence of outliers than Ordinary Least Squares. We estimated the model using Least Absolute Deviation (LAD) regression and Iteratively Reweighted Least Squares (IRLS)[9]. LAD is a simple version of quantile regression in which estimates are fit to the median, rather than the mean, of the outcome data. Because median values – unlike means – do not weight outliers more heavily than other observations, LAD produces parameter estimates that are relatively insensitive to the presence of outliers. Similarly, IRLS functions to automatically reject extreme outliers and give little weight to large outliers. First, extreme outliers (observations where Cook's D statistic > 1) are dropped from the model. Secondly, IRLS performs an iterative series of regressions, with each regression providing smaller weights to observations that have the largest residuals from the previous regression. Thus, IRLS produces parameter estimates in which the influence of outliers has been minimized.

Results from the IRLS and LAD models were broadly consistent with the estimates obtained in OLS models. Specifically, none of the parameter estimates that were significant in the OLS model changed signs in the IRLS or LAD models**.** The findings indicate that outliers were not significantly influencing our model results. Therefore, given the familiarity of OLS to the statistical community and its advantageous asymptotic properties, we decided to use OLS for our final model.

4.2 Identifying properties with falsely inflated or deflated costs

This study was mandated to estimate the operating costs of well-managed public housing. We therefore needed to exclude from our analysis database properties that had exceptionally low costs simply because they were being allowed to deteriorate. Including such properties would generate cost estimates that would not reflect the costs of operating well-managed housing.

Similarly, we needed to avoid including properties that might have falsely inflated costs. Specifically, we were concerned that two types of ownership structure – Limited Dividend and Non-Profit – provided incentives for property owners to spend more than necessary on property maintenance because of their inability to remove cash flow from the property in the form of profit.

[9] IRLS was implemented using the RREG command in STATA. LAD was implemented using the QREG command in STATA.

4.3 Properties with potentially deflated costs

- **"Older Assisted" Properties**
  Several members of the study consulting team were concerned that Older Assisted properties may have operating costs that did not fully reflect costs, because their operating costs are constrained by budget-based rents negotiated between owners and HUD. We found, however, that Older Assisted properties have among the highest levels of reported operating costs. (See Table A.31 in the supplemental tables at the end of this section.) Therefore, we decided that including these properties did not threaten to bias downward our estimate of operating costs.

- **Troubled Properties**
  Another concern arose around properties that received the designation of "Troubled" by an FHA field office. We considered excluding properties that are or were declared "Troubled" out of concern that the status may have resulted from cumulative neglect of property. If "Troubled" properties had low operating costs because they were poorly maintained, their inclusion in our analysis sample could bias downward our estimate of the costs of running well-managed housing. However, we found that these properties in fact had among the highest operating costs in the sample (See Table A.32). Furthermore, excluding these properties did not change any cost relationships in the model. Therefore, we decided to leave "troubled" properties in the model.

- **Properties with poor REAC inspection scores**
  We considered excluding properties with low REAC physical inspection scores, based on the hypothesis that low scoring properties may have lower operating costs because they were constrained by restricted rents, or because owner decisions to spend little on the property may have lead to physical distress.

  However, we found that lower-scoring properties had among the highest costs in our sample. Furthermore, excluding properties with REAC physical inspections scores below 60 did not change any cost relationships in the model. (We also tested below 30, with the same results.)

  We decided to exclude properties in the lowest 5% of REAC physical inspection scores (scores < 56) because our model is designed to capture cost relationships among well-managed properties.

4.4 Years of data used

We used three years of FHA data (1998, 1999, and 2000). Operating costs were inflated to year 2000 dollars and averaged over the three-year period to smooth out year-to-year fluctuations using the housing component of the Bureau of Labor Statistics' Consumer Price Index. 1998 data was multiplied by 1.0574 and 1999 data was multiplied by 1.03478. Only observations with at least two years of data were included in the analysis file.

4.5 Properties that were dropped from the model

Although we began with a sample size of 17,493 observations with at least one year of data from the 1998, 1999, or 2000 Annual Financial Statements, our final sample size for the estimation model consisted of 10,554 observations. Observations were dropped from the analysis sample for several reasons. A large number were dropped when we restricted the sample to observations with at least two years of data. Others were dropped because they did not meet one of our few selection criteria. Finally, some were dropped because they had missing values for one of the key model variables.

The following list presents the number of observations with missing values for each of the model variables (variables with no missing values are not listed):

- age (1559 missing)
- distribution of units by bedroom size (177 missing)

- building type (433 missing)
- family/senior (6 missing)
- census tract poverty rate (1501 missing)
- rent to FMR ratio (992 missing)

1,969 observations do not meet one or more of our sample restriction criteria, and hence are not included in the model. Properties that did not meet the restrictions are:

- Bottom 5% of the physical inspection score (n = 841)
- Unassisted Senior properties (219)
- Senior properties with large units (545). Definition: property is designated as senior, and also has either average bedroom size $\geq$ 1.5, or average bedroom size $\geq$ 1.2 and 100 or more units that are 2 bedroom or larger)

Among the 15,524 observations that are not excluded by our sample restriction criteria, 11,893 have no missing values for the analysis variables. Of these, 10,554 have at least 2 years of operating cost data.

Table A.9: Number of Included Observations

Number of Years of Operating Cost data	Any Missing Values for Model Variables?		
Frequency	No	Yes	Total
0 years	16	12	28
1 year	1323	562	1885
2 years	4446	1787	6233
3 years	6108	1270	7378
Total	11893	3631	15524

5. MODEL DEVELOPMENT

5.1 Operating Costs (Dependent Variable)

The value GSD sought to predict is operating cost per unit month. In this section we discuss the construction of the outcome variable.

5.1.1 PUM definition

The unit of measurement for the outcome variable is costs *per unit per month*, excluding utilities and real-estate taxes. Costs per housing unit (as opposed to operating costs for entire properties, or, at the other extreme, operating costs per bedroom) were chosen as the natural outcome variable.[10]

[10] Note, however, that the unit of analysis is the property. This model specification gives equal weight to all properties regardless of the number of units. An alternative specification – with observations weighted by size – did not yield results that were inconsistent with our unweighted specification, although it did produce slightly smaller impacts for the number of large units and for property age.

5.12 Chart of accounts

The specific components of the dependant variable are all line items reported in the Statement of Profit and Loss portion (formerly HUD Form-92410) of the Annual Financial Statement (AFS). They include:

- Total administrative expenses (Line 6200/6300)
- Total operating and maintenance (Line 6500)
- Total taxes and insurance (Line 6700) minus real estate taxes (Line 6710)

To account for non-recurring capital expenses, we subtracted the Replacement or Painting Reserve releases that are included as part of the expenses reported on the Statement of Profit and Loss.[11]

5.13 Log Transformation

GSD changed the dependent variable from PUM (dollars) to the logarithm of PUM. The effect of transforming the dependent variable is that the regression coefficients express the *percentage* change created in the dependent variable by a unit change in the explanatory variable (rather than a dollar change). The regression coefficients produced by the model will be percentage changes. To use the example of building type: it may be more accurate to conclude that an elevator building costs 5 percent more to operate than to conclude that it costs $10 more to operate per unit month regardless of the base cost to which that $10 is added.

5.2 Identification of Cost Factors (Explanatory Variables)

In this section of the appendix we discuss in detail each of the explanatory variables used in the model. We discuss the theoretical reasons why each explanatory variable is expected to be associated with operating costs; we discuss the alternative variable specifications that were tested, including any interactions with other explanatory variables; and we comment on the empirically observed relationship between the explanatory variable and operating costs.

5.2.1 Property Size

Property size (the number of units in a property) is a variable that we expected to be highly correlated with operating costs. We constructed our initial models to allow us to test two basic hypotheses from industry experts about the relationship of size to operating costs:

- Standard economic theory leads us to expect increasing returns to scale in operating costs, as fixed costs are spread over more units. Thus, lower per unit costs are expected as the total number of units increases.

- There may be diseconomies of scale above a certain size. This could result from overcrowding and corresponding social distress, leading to vandalism, crime, and an increase in wear and tear.

We initially tested size entered linearly and in quadratic form. In a simple model with size entered linearly and the dependent variable (per unit monthly operating costs) entered in log form, the coefficient on size is negative and highly significant (every 100 unit increase in size is associated with a 1.2 percent decrease in costs). When entered with size squared, the coefficient on size is negative and highly significant (b = -.000019), and the coefficient on size squared is positive but very small and of borderline significant (b = 7.1E-8), implying that returns to scale are not constant, but level off. (The inflection point – the point where increased size actually leads to increased costs – does not occur until size exceeds 2,635 units

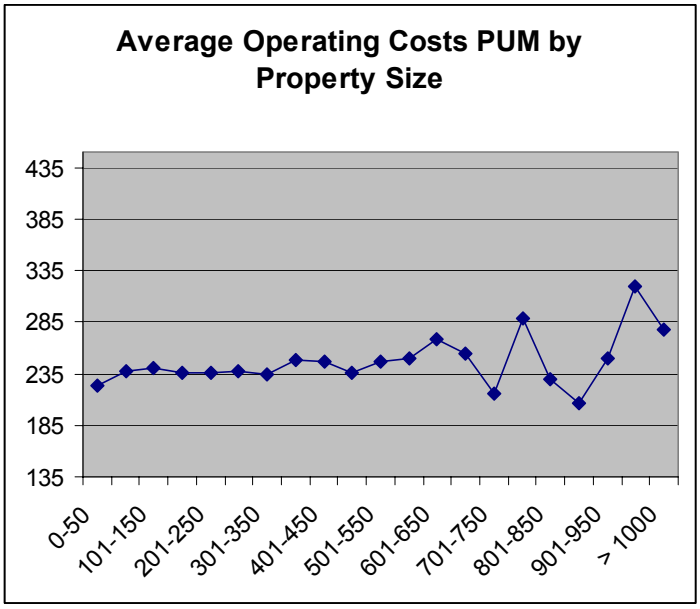
[11] Specifically, Line 3 of Part II of the Statement of Profit and Loss, *Replacement or Painting Reserve releases that are included as expense items on this Profit and Loss statement*. The mean value was $8.50 PUM and the median was $0.60 PUM.

according to the quadratic specification.) A cubic specification did not yield significant coefficients on size squared and size cubed.

Because the quadratic specification suggested – as expected – a non-linear relationship between size and costs we decided to enter the size variable in categories. Industry experts suggested that economies of scale begin to occur between 150 – 200 units. We tested many different breakpoints at 50 – 100 unit intervals. We found a significant breakpoint at 150 units. We did not find additional breakpoints at higher intervals despite exhaustive testing. We tested for breakpoints at 300, 500, 750, and 1000 units, as well as finer intervals. In our final specification, we include a single dummy variable for size greater than or equal to 150 units. The coefficient on this dummy ranged from about – 1.5 to – 3.5 percent, depending on other details of the model specification. In our final specification, size above 150 units is associated with a 1.5 percent reduction in estimated per-unit costs.

There was considerable concern among some participants in the model development process that we did not identify any additional costs associated with very large size properties (Hypothesis 2.) Indeed, when one looks at simple average costs per unit by property size, costs do appear to increase once property size exceeds approximately 350 units (see chart below). However, it appears that those costs typically associated with very large properties are, in fact, well captured by other model variables. High costs do not appear to be associated with very large properties *per se*, once characteristics of the neighborhood (measured by the tract-level poverty rate), the metropolitan area, and typical household size (proxied by average number of bedrooms per unit) are taken into account. To make sure that the relationship between property size and cost was not obscured by confounding variables, we tested the effect of property size separately for two subgroups of properties: family buildings and high-rises. These are the subgroups in which very large and very expensive properties have most commonly been noticed. However, even for these subgroups, we did not observe a point in the data at which size becomes positively associated with costs. We also tested the effect of property size for the subgroup of family high-rises (the interaction of the two subgroups); for this subgroup, too, we found no point at which size becomes positively associated with costs in the multivariate model.

Figure A.3



5.2.2 Property Age

Property age was expected to be one of the strongest predictors of property operating costs. Older buildings were expected to be more expensive to operate, as aging infrastructure such as heating and cooling systems require more expensive routine maintenance. Age did indeed prove to be a key model variable; however, the relationship between age and operating costs was not as straightforward as initially expected.

It should be noted that one limitation of our measure of age is that we had no ability to capture the number of years since major renovations took place, either for FHA or public housing. Substantial renovations can have significant impacts on operating costs, but we did not have access to data on modernization.

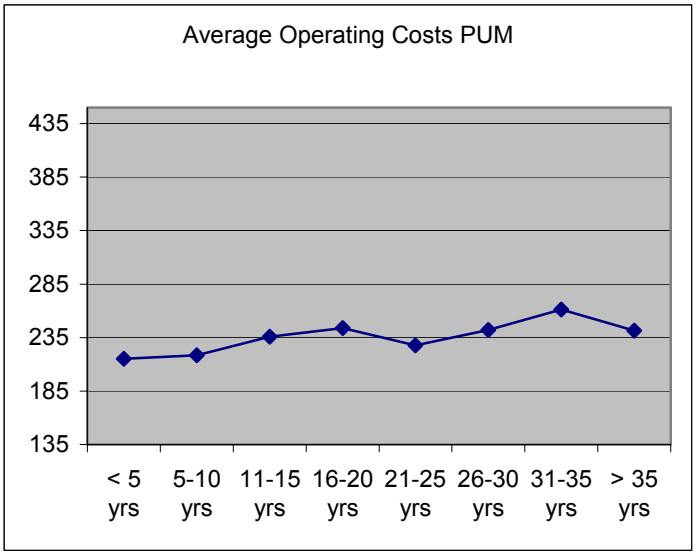
- **Measures of Age.**

There were several alternative measures of property age available. We considered three options: the age of the property's current mortgage; the age of the property's first mortgage; and the date of property occupation. The variables came from different sources: current mortgage age and first mortgage age come from the F-47 data, while the date of first occupation comes from the REMS data. Each of these age measures had different advantages and drawbacks. For the purpose of estimating operating costs the most relevant age may be the number of years since a property was last modernized. Because many properties that are modernized are also refinanced in the same year, current mortgage age may come closest to capturing the number of years since a property was last modernized. On the other hand, several participants in the operating cost project argued that the number of years since initial construction was the appropriate measure of property age, because different construction eras had different building standards which in turn effect maintenance costs. Among our three measures of age, age since initial occupation comes closest to measuring the number of years since initial construction. Finally, age based on final endorsement date of the first mortgage produces values very similar to age based on the year of first occupation.

In practice, all three variables are extremely highly correlated with one another, and they produced very similar parameter estimates when used in the model. We attempted to enter both age of current mortgage and age since first occupation in a single model, and found the two variables to be highly collinear. We ultimately decided to use age based on the final endorsement date of the first mortgage, because it is conceptually and empirically very close to the building age variable we constructed in the public housing stock, which is based on the Date of Full Availability (DOFA).

Figure A.4 shows the average operating cost of the FHA properties by age, measured by the age of the first mortgage, not controlling for any other property characteristics. It can be seen in the exhibit that there appears to be an overall upward trend in costs as age increases, but the trend is not monotonically increasing.

Figure A.4



- **Controlling for Quality.**
  Property age is highly correlated with several other property characteristics that are themselves determinants of operating costs. If these correlates of age are omitted, the relationship of age to operating costs can be masked. Consequently, in models that are not fully specified, property age appears to have little relationship to operating costs.

  A variable that masks the effect of another variable when omitted from the model is described as a confounding variable. The most important confounding correlate of age is property quality or market segment. Properties that are maintained to a poor standard of quality, and are consequently positioned in the lower end of the rental market, are cheaper to operate than properties that are maintained to a high standard of quality. Consider a hypothetical city in which older properties are positioned in the lower end of the rental market, due to their undesirable location, renter preference for new buildings, and years of neglect. In this city, newer properties are used to serve the high end of the rental market, and these new properties are maintained at a high standard of quality, with more frequent physical maintenance as well as numerous amenities and services. If we were to examine the relationship between property age and operating costs in this city, we would observe that newer buildings were more expensive to operate than older buildings. However, this simple analysis is incorrect, because we are not comparing buildings of comparable quality. If we were to compare buildings that were all maintained to the same level of quality, but which varied in age, we might instead find that older buildings are in fact more expensive to operate and maintain than newer buildings. Thus, it is essential to include a model variable that controls for property quality when we are attempting to estimate the relationship between property age and operating costs. We measured each property's market segment – a proxy for property quality – using the ratio of the property's average rent to the area Fair Market Rent (FMR). For a detailed description of this measure see the section on that variable below. In all of the models discussed in the current section, controls for property quality have been included.

- **Specification of the Age Variable.**
  When entered in a linear specification that includes the rent-to-FMR ratio, we found that age has a positive, statistically significant association with operating costs: every 10 years of age is associated with approximately a 5 percent increase in costs. When age is entered in a

quadratic specification, the quadratic term was statistically significant. The quadratic specification implies that age is negatively associated with costs for approximately the first 5 years, and then becomes positively associated with cost.

Because there appeared to be a non-linear relationship, we entered age in categories. After much testing we found that the FHA properties fell into four cost categories by age, with older properties being the more expensive. The four categories identified were: under 15 years; 15 – 20 years; 21 – 25 years; and over 25 years. In application to public housing, the relationship between age and cost is smoothed so that there are no discontinuous jumps between age categories.

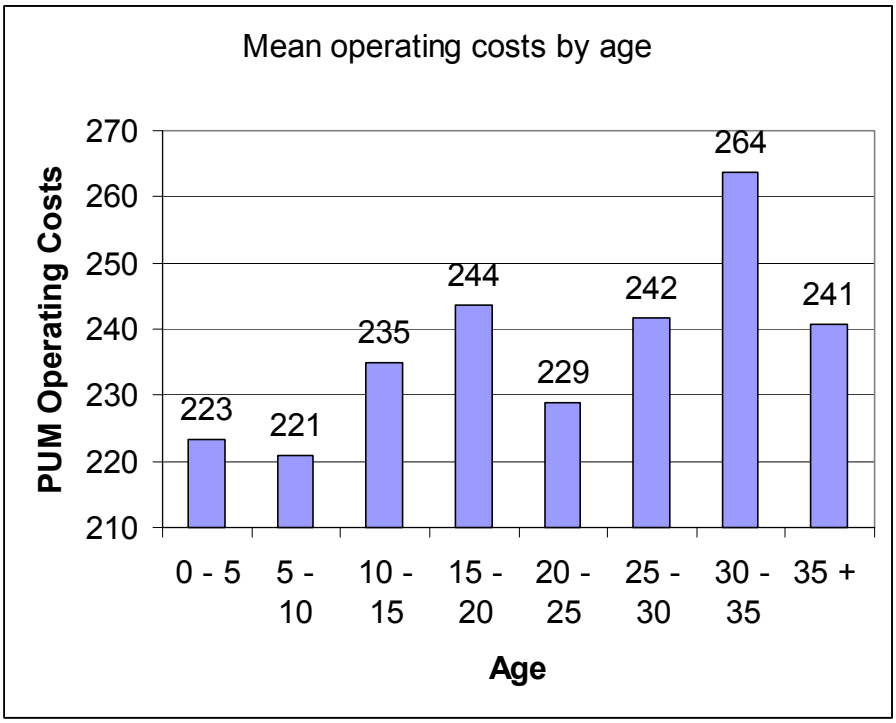
In research working group meetings with industry representatives, considerable concern was expressed about whether the FHA cost model captures increases in costs above 30 years of age. Approximately 4.6 percent of the FHA housing stock is 30 years or older. In comparison, 57 percent of the public housing stock is older than 30 years, and approximately 25 percent are older than 35 years. It is true that the FHA housing stock cannot support analyses of the relationship between age and operating costs outside of the range of ages represented in the FHA housing stock. Property age in the FHA housing stock ranges from 0 years (buildings with the final endorsement date of the first mortgage in the year of analysis) to 52 years old; the mean and median property age in the FHA stock are both 18 years old, and the 75th percentile property is 25 years old. The 90th percentile FHA property is 28 years old; and the 99th percentile property is 33 years old.

There are over 700 properties in the FHA stock that are 30 years or older. Of these older FHA properties, the mean property is 32 years old and the median is 31 years old; the 75th percentile property is 32 years old; and the 90th percentile property is 35 years old. There are 96 properties that are 35 years or older. Model work showed that properties over 30 years had about the same or slightly lower operating expenses than properties 21-29 years old, indicating possibly that operating cost increases level off after climbing for the first 25 or so years, instead of continuing to increase with age (although capital needs will continue to increase as buildings age).

We feel that our cost model estimated using the FHA stock provides a good estimate of the relationship between cost and age for properties that are 35 years old or less in the public housing stock (approximately 75 percent of the public housing stock). Based on our field testing (see Chapter 5), we do not believe that the model underestimates properties more than 35 years old.

Figure A.5 presents mean operating costs by age in the FHA housing stock, not controlling for any other factors.

Figure A.5



Age	N	Mean	Standard Deviation	Minimum	Maximum
0 - 5	1966	223	80	53	682
5 - 10	1874	221	88	53	747
10 - 15	2127	235	97	56	699
15 - 20	4986	244	91	53	762
20 - 25	1822	229	84	53	650
25 - 30	2870	242	87	53	745
30 - 35	250	264	107	71	718
35 +	61	241	95	93	460

- **Age, Type of Subsidy, and Property Ownership Structure.**
  Two additional potential confounding variable in the relationship of age to operating costs are the type of property subsidy and the property ownership type. In particular, owners of limited dividend and non-profit properties face incentives to reinvest all available revenue back into a property, whether or not such expenditures are necessary, because owners of such properties cannot extract any surplus revenue as profit. Thus, we might expect limited-dividend and non-profit properties to report higher operating expenses than for-profit properties at any given level of housing quality; these expenses, however, may be falsely inflated and inaccurately reflect true costs. Because non-profit and (especially) limited-dividend properties tend to be older than for-profit FHA-insured properties, falsely inflated costs in these two groups will generate an upward bias in the estimated relationship of age to operating costs. (See Table A.10.)

Table A.10

Mortgage Sponsor Type:	Average Age of Mortgage	Std. Dev. Age	Average PUM Operating Costs	Std. Dev. Operating Costs	N
For-Profit	14	8	226	78	6703
Non-Profit	17	7	229	94	5670
Limited Dividend	22	7	260	96	3362
Other	23	10	243	102	221

In order to address this problem, we ran the cost models interacting age with ownership type, and tested whether the relationship between age and operating costs varied significantly by ownership. We found that, although the coefficients on the main effects for limited-dividend ownership and non-profit ownership are significant, the interactions of age and limited-dividend ownership and non-profit ownership are small and not statistically significant[12].

Property subsidy type does not bear a direct relationship to property owner incentives structures. However, subsidy type is strongly correlated with property age, as different subsidies dominated housing production programs in different eras. (See Table A.34.) Therefore, in order to isolate the effect of age, the cost model includes controls for whether or not a property is assisted[13] and for the percentage of units that receive Section 8 tenant-based assistance.

Table A.11

Subsidy Type:	Average Age of Mortgage	Std. Dev. Age	Average PUM Operating Costs	Std. Dev. Operating Costs	N
Unassisted	10	9	220	82	3258
Older Assisted	23	8	247	89	4664
Newer Assisted	18	3	243	90	4036
Section 202	14	4	223	91	3983

- **Age and Capital Deficiencies**.
  Capital deficiencies were considered another possible confounding variable that could mask or alter the relationship of age to operating costs. As discussed below, the percentage of REAC-identified capital deficiencies in a property provide a measure of physical deterioration expected to be highly correlated with both age and operating costs.

  However, inclusion of the capital deficiencies measure in multivariate models did not significantly alter the estimated relationship of age to operating costs. Indeed, the inclusion of capital deficiencies slightly strengthens the estimated relationship between age and operating costs. We also tested whether there was a significant interaction between capital deficiencies and age in the cost model, and found that there was none. It should be borne in mind, when interpreting the impact of capital deficiencies, that these measures provided by REAC were a more accurate measure of habitability than of capital needs. See Table A.34 at the end of this appendix for a list of the capital defects examined.

[12] The interaction between age and "Other" is significant. The coefficient on the interaction term is almost the same size as the age coefficient, but negative, yielding a net age effect of zero for the "other ownership type" group.

[13] Assistance status is entered in the model interacted with geographic location.

5.2.3. Unit Size

Unit size refers to the number of bedrooms per unit in a property. This is an extremely important variable in the cost model, as it proxies a key measure of tenant demographic composition: average household size. Efficiencies and one-bedroom units tend to house senior citizens or disabled persons, while families live in two bedroom and larger units. Thus, the presence of families with children is reflected by the number of units with two or more bedrooms, while the presence of large families is reflected in the number of larger units. Public housing industry group experts early on informed us that families housed in large (3 or more bedroom) units pose particular challenges to property maintenance, as such families have a higher ratio of children to parents, and increase the likelihood that unsupervised children and teens will be living in the units.

We tested a range of specifications for this variable, with models that included average unit size, the percentage of large (3 or more bedroom) units, and combinations of the two. All specifications of this variable are highly significant in the model. We selected the specification that we felt was the most comprehensive and straightforward: we included the percentage of units of each size in the model, excluding one category (the percentage of efficiency and one-bedroom units) as our reference category.

Note that we have a separate variable that indicates whether a building primarily houses senior citizens. Thus, the unit size measure is able to distinguish the cost impact of having more or less units of different sizes, within either a primarily family or primarily senior building.

5.2.4 Clientele: Senior/Family

We used two criteria to determine whether a property should be classified as a family property or a senior property. The REMS database contains information on clientele served. All properties except those specifically designated for "families" are classified as non-family buildings. (This includes the small number of properties specifically designed to serve particular medical populations.) In addition, we used a formula provided by HUD's office of Policy Development and Research to identify properties that should be considered "senior" buildings. This formula, based on unit size and the total number of units in the property, identified buildings that contained primarily non-family units. Any properties which did not meet the REMS-based definition of a "family" property but also did not meet HUD's definition of a "senior" property were dropped from the sample.

The indicator for Senior buildings was consistently associated with lower costs in the model, regardless of specification. This indicator is entered as a simple dummy variable.

We did test whether there were significant interactions between property clientele, building type, and size (see the discussion under Property Size above). However, such interactions did not prove significant.

5.2.5 Building Type

The FHA dataset allows us to identify six different building types: detached, semi-detached, row/townhouses, garden-style/walk-ups, highrises/elevator buildings, and mixed types. We collapsed detached and semi-detached properties into one building type, as each of these categories contained only about 4.5 percent of the sample, and industry experts reported cost structures to be comparable across the two types.

Table A.12

Operating Costs by Building Type[14]	N	25th Percentile	Median	Mean	75th Percentile
Detached	783	$129	$183	$205	$253
Row/Townhouse	2370	$173	$212	$224	$259
Semi-detached	718	$168	$202	$216	$250
Walk-up	8315	$177	$215	$231	$268
High-rise / Elevator	4493	$187	$231	$254	$303

There was a strong expectation among all participants in the model development process that high-rise properties would evidence significantly higher operating costs than other types of properties. In raw means, as seen in Table A.12, high-rises in the FHA housing stock do indeed have notably higher operating costs than any other group of properties.

Surprisingly, however, in our multivariate models building type proved to be a largely insignificant explanatory variable. This finding, and in particular the finding that high-rise buildings were no more expensive than any other types, strongly went against the expectations of industry experts. It was suggested that the GSD test whether other cost drivers, such as property size and property clientele, were confounding the relationship of building type and costs. In order to address this concern, GSD tested several interaction terms that would allow the relationship of building type to cost to vary across other dimensions. We tested two-way interactions between high-rise and building size, property clientele (family vs. senior), and a three-way interaction between high-rise, property clientele, and building size. None of these interactions yielded the expected result that high-rises are significantly more expensive to operate, controlling for other cost drivers in the model.

5.2.6 Central City/Suburb/Rural location

One of the key drivers of labor and land costs is property location. Our primary measures of location are the geographic indicator variables that identify the metropolitan statistical area (for metropolitan properties) or the state or region (for non-metropolitan properties). In addition to these geographic indicators, we also include a central city/suburb designation, which provides further differentiation within metropolitan areas. Properties that are not located in metropolitan areas do not receive a central city/suburb designation.

The central city designation indicates whether or not a metropolitan property is located in the major residential and employment center within its MSA. Note that the central city designation does not indicate that a property is located in the "inner city", or is located in a high poverty community. Neighborhood characteristics are measured by the census-tract poverty rate.

The model estimate for Central City indicates that properties operating in central city locations cost 2.6 percent more to operate per-unit per-month than suburban properties with similar characteristics located in otherwise similar neighborhoods.

Below is the definition of "central cities" of Metropolitan Statistical Areas (MSAs), excerpted from the White House Office of Management and Budget, Notice "Revised Standards for Defining Metropolitan Areas in the 1990s," published at 55 FR 12154, March 30, 1990.

The central city/cities of the MSA are:

[14] There were also 342 properties that received a designation of "mixed" building type. These properties contained multiple buildings of different types. We do not know enough about this small group of properties to apply their building type coefficient to public housing.

- The city with the largest population in the MSA;
- Each additional city with a population of at least 250,000 or with at least 100,000 persons working within its limits;
- Each additional city with a population of at least 25,000, an employment/residence ratio of at least 0.75, and at least 40 percent of its employed residents working in the city;
- Each city of 15,000 to 24,999 population that is at least one-third as large as the largest central city, has an employment/residence ratio of at least 0.75, and has at least 40 percent of its employed residents working in the city;
- The largest city in a secondary noncontiguous urbanized area, provided it has at least 15,000 population, an employment/residence ratio of at least 0.75, and has at least 40 percent of its employed residents working in the city;

Each additional city in a secondary noncontiguous urbanized area that is at least one-third as large as the largest central city of that urbanized area, that has at least 15,000 population and an employment/residence ratio of at least 0.75, and that has at least 40 percent of its employed residents working in the city.

5.2.7 Demographic Measures

We considered two types of demographic measures: 1. Measures of the characteristics of property residents (property-level measures), and 2. Measures of the characteristics of the neighborhood in which a property is located.

Theoretically, both property level and neighborhood level tenant characteristics can influence costs, through related but different mechanisms. Neighborhood characteristics influence operating costs primarily through wear and tear on property (litter, vandalism) and through security costs.

Resident characteristics influence operating costs through wear and tear associated primarily with the number of children in the unit, especially the number of unsupervised children in the unit. It is possible that adults who are not working generate more wear and tear on a property than employed adults, because they spend more time in their units. (This may be offset by the fact that adults who are not at work may provide more supervision for children.) In addition, adults engaged in certain types of behavior such as drug use may cause damage to their units or fail to maintain them; however, none of our data sources provide information about such behavioral factors.

We do not have direct measures of local crime and vandalism rates or the average number of unsupervised children per unit for any of our properties. However, we have several measures that provide excellent proxies for these and other neighborhood and tenant characteristics.

We tested the following census tract level variables, which were obtained from the 1990 Census of Population and Housing:

- percentage of persons in poverty
- percentage of persons employed
- percentage of families on welfare
- percentage of households that are headed by a single parent
- percentage of persons who are African-American
- percentage of persons who are elderly
- percentage of persons who are non-elderly and disabled

We tested the following property-level demographic measures, obtained from the 1998 Picture of Subsidized Households:

- percentage of families with the majority of income from work
- average family size
- percentage of families on welfare
- percentage of families that are headed by a single parent
- percentage of persons who are African-American
- percentage of persons who are elderly
- percentage of persons who are non-elderly and disabled

We also tested the following property-level measures, obtained from the FHA's REMS database, which are discussed in more detail above:

- percentage of large units (3 or more bedrooms per unit)
- average number of bedrooms per unit

Each of these sources of data has advantages and disadvantages. The tract-level measures have the advantage of being very well-defined. Census tract measures have been subjected to quality control by the Census Bureau, and missing values exist only for properties where we have incomplete geographic identifiers. There are two disadvantages using the Census measures. First, the 1990 Census data is now 10 years old. Many neighborhoods will have changed in character over the decade, introducing measurement error into this variable. Second, the Census data measures neighborhood characteristics rather than characteristics of the residents of particular properties. As discussed above, there are theoretical reasons to expect both of these types of characteristics to influence costs.

The property-level demographic characteristics obtained from HUD's "A Picture of Subsidized Households" database have the advantage that they directly measure characteristics of property residents. Conceptually, tenant characteristics are very important cost drivers. However, there are several disadvantages with using data from the Picture of Subsidized Households. Most importantly, the data is only available for FHA properties that receive Section 8 subsidies. Thus, any analysis using these measures cannot include unassisted properties or properties that receive a mortgage subsidy but no Section 8 subsidy. Second, even among assisted properties, the data from Picture of Subsidized Households is not as reliable as Census data. The Picture data was not subject to the extensive editing and quality control measures received by the Census data, and there are several variables from Picture of Subsidized Households with a high percentage of missing values. Because of these limitations, we primarily used the demographic variables from Picture of Subsidized Households to determine which tract-level variables were most highly correlated with property-level demographics.

Finally, the number of bedrooms per unit (a proxy for household size) has the advantage of being a well-defined, precisely measured variable that is available for nearly all properties. The disadvantage of this measure is that it is not a direct measure of tenant demographic characteristics: in properties where units are either over-crowded or under-occupied, the number of bedrooms per unit will be an imprecise measure of household size. Empirically, however, we found that average number of bedrooms per unit proves to be a very powerful proxy for household demographic characteristics. The correlation between the average number of bedrooms per unit and average family size exceeds 90%, and the correlation between the average number of bedrooms per unit and the percentage of single parent households in the property exceeds 80%.

Both neighborhood and property-level demographics were found to be significantly correlated with operating costs: for example, the simple correlation between average household size in a property and operating costs is .29, while the simple correlation between the percentage of single parent households in the neighborhood and operating costs is .28.

We decided to use just one tract-level variable, because these measures are so highly correlated with each other. When we entered more than one tract-level variable in the operating cost model simultaneously, coefficient estimates on the tract-level variables became unstable, suggesting a significant degree of colinearity among the regressors. We decided to use the tract poverty rate as our measure of neighborhood quality, primarily because it is a widely accepted measure of socioeconomic distress.

Because the variables available from the "A Picture of Subsidized Households" data set were only available for a subset of properties, we decided not to use any of the property-level demographic variables. However, because of the extremely high correlation between the number of bedrooms per unit and the demographic variables average family size, we felt that bedrooms per unit adequately captures family size, which is one of the most important demographic measure. Another key demographic measure is the percentage of elderly residents: however, this variable too is well captured by other model variables. Our measure of whether a property houses primarily families or elderly persons is discussed below.

To summarize, tenant and neighborhood characteristics are measured in our model through the following variables:

- the census tract poverty rate (measuring neighborhood quality)
- the number of bedrooms per unit (measuring family size)
- family/elderly property designation (measuring whether residents are primarily elderly)

It should be noted that we were unable to obtain a measure that effectively captures the costs associated with units that are occupied by non-elderly disabled tenants. The variable that should measure this in the "A Picture of Subsidized Households" data is very poorly populated; therefore, we were unable to use tenant-level demographic information to verify whether the tract-level data would make a good proxy. The tract-level measure of the percentage of non-elderly disabled persons was strongly (and significantly) correlated with the tract poverty rate (rho = .62, p-value < .0001) but only weakly correlated with operating costs (rho = .06, p-value < .0001) (See Table A.13). Regressions that included the percentage of non-elderly disabled persons as the only tract-level variable showed a positive association with costs, as did, of course, regressions that included the percentage of poor persons as the only tract-level variable. However, when both tract-level measures are included in the model, the coefficient on the percentage of non-elderly disabled persons becomes negative, suggesting that the positive relationship between operating costs and the percentage of non-elderly disabled persons exists only because of the positive relationship between the percentage of poor persons and the percentage of non-elderly disabled persons. Therefore, it did not seem useful to include the tract-level measure of the percentage of non-elderly disabled persons in the cost model. Cost adjustments reflecting the additional costs of a non-elderly disabled population were separately reviewed in the public housing case studies.

Finally, the inclusion of census tract measures also implied that we should test the robustness of our variance-covariance estimates to within-tract correlation of errors[15]. We ran our model using the Huber-White correction for clustering within census tracts. In fact, the correction led to little change in our estimated standard errors. No doubt this is because the properties in the FHA database are spread out over a very large number of census tracts.

[15] Because multiple observations may exist in the same census tract, inaccurately small standard errors could be estimated if similar properties resided in the same tracts and within-tract clustering was not taken into account.

Table A.13: Pearson Correlation Coefficients for Selected Tract and Property Level Demographic Characteristics

	Rho Prob > |r| under H0: Rho=0 Number of Observations								
	opcost	pctpoor	pctsph	pctwelf	avbed	pic_size	pic_sp1	pic_welf	pic_wage
pumavgrfrfree	1.000	0.182	0.281	0.287	0.185	0.285	0.225	0.320	0.161
PUM Operating Costs		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001 10285	<.0001	<.0001
	17452	15962	15797	15797	17305	10285		10127	10127
pctpoor	0.182	1.000	0.667	0.832	0.106	0.197	0.167	0.276	0.074
Census Tract: % Below Poverty	<.0001		<.0001	<.0001	<.0001	<.0001	<.0001 9536	<.0001	<.0001
	15962	15992	15827	15827	15863	9536		9387	9387
pctsph	0.281	0.667	1.000	0.748	0.243	0.338	0.351	0.400	0.226
Census Tract: % Single Parent Households	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001 9420	<.0001	<.0001
	15797	15827	15827	15827	15698	9420		9273	9273
pctwelf	0.287	0.832	0.748	1.000	0.086	0.181	0.158	0.322	0.042
Census Tract: % with majority income from welfare	<.0001	<.0001	<.0001		<.0001	<.0001	<.0001 9420	<.0001	<.0001
	15797	15827	15827	15827	15698	9420		9273	9273
Avbed	0.185	0.106	0.243	0.086	1.000	0.919	0.863	0.578	0.779
Average Number of Bedrooms	<.0001	<.0001	<.0001	<.0001		<.0001	<.0001 10248	<.0001	<.0001
	17305	15863	15698	15698	17344	10248		10090	10090
Pic_size	0.285	0.197	0.338	0.181	0.919	1.000	0.903	0.667	0.805
Tenant Characteristics: Avg. Household Size	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001 10293	<.0001	<.0001
	10285	9536	9420	9420	10248	10293		10135	10135
Pic_sp1	0.225	0.167	0.351	0.158	0.863	0.903	1.000	0.720	0.786
Tenant Characteristics: % Single Parent Households	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	10293	<.0001	<.0001
	10285	9536	9420	9420	10248	10293		10135	10135
Pic_welf	0.320	0.276	0.400	0.322	0.578	0.667	0.720	1.000	0.388
Tenant Characteristics: % w/majority of incomc from welfare	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001 10135		<.0001
	10127	9387	9273	9273	10090	10135		10135	10135
Pic_wage	0.161	0.074	0.226	0.042	0.779	0.805	0.786	0.388	1.000
Tenant Characteristics: % w/majority of incomc from wages	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001 10135	<.0001	
	10127	9387	9273	9273	10090	10135		10135	10135

5.2.8 REAC variables

We obtained two measures of property characteristics from the Research and Development team of the Real Estate Assessment Center (REAC) at HUD: the annual physical inspection score, and an index of capital deficiencies.

REAC conducts annual physical property inspections of the nearly 33,000 rental properties that are either owned, insured, or subsidized by HUD. In addition to constructing several physical inspection scores – five area scores and an aggregate score – REAC inspectors also collect information on the count of observed capital defects at each property. The total number of possible capital defects is also recorded, permitting the construction of the percentage of capital deficiencies.

REAC provided GSD with both of these measures of property and management quality. The use of physical inspection scores in the modeling process was somewhat controversial, as several members of public housing industry groups argued that physical inspection scores were too variable (across inspectors and across years) to provide valid information. In order to address this concern, GSD did not use the physical inspection scores as a continuous measure of quality. Instead, GSD constructed an indicator variable to identify those properties that had scored particularly low – in the bottom 5 percent of all properties, with a score of less than 56. GSD considered *extremely* poor physical inspection scores to be an indicator of poor property management and, as mentioned above (under Data Cleaning), we dropped these properties from the analysis sample.

The second REAC measure, the percentage of capital deficiencies, was intended to proxy for capital needs in the model. (See list of capital defects in Table A.34 at the end of this section.) As mentioned above, it should be borne in mind that the measures obtained from REAC may in fact provide a more accurate measure of habitability than of capital needs. As expected, the percentage of capital deficiencies is significantly correlated with operating costs in raw correlations, and is also, as expected, positively correlated with age, central city location, and tract poverty rate. (See Table A.14.)

Table A.14: Correlations Between Percent Capital Defects and Other Model Variables

	Operating Costs	Tract: Percent Poor	Age	Central City
rho	0.19837	0.2141	0.14164	0.12776
p-value	<.0001	<.0001	<.0001	<.0001
N	15979	14675	14692	16013

Somewhat surprisingly, coefficient for the capital deficiencies measure was very small and rarely statistically significant in the multivariate model, regardless of the details of model specification. In the final Cost Model, therefore, we did not include this variable.

5.2.9 Ownership Type, Assistance Type, and Project-Based Section 8 Assistance

Properties that receive FHA mortgage insurance can be divided into two categories: unassisted properties, and those that receive some form of assistance. Unassisted properties are those that receive no underlying mortgage interest rate reduction or rental subsidy program.

FHA assisted properties are those properties that are assisted with either a mortgage interest reduction program or a rental assistance program and are also insured with FHA.

FHA assisted properties have three main ownership types: unlimited dividend, limited dividend, and non-profit. In a limited dividend property, the owner is restricted in the cash flow that can be distributed each

year, in accordance with the regulatory agreement. All else being equal, a rational owner will want to keep operating costs down and increase cash flow. However, if an owner cannot profit from excess cash (or the profit is limited), and if the property has the resources, the owner is likely to plow those funds back into the property, either to refurbish the property or to provide a higher level of service than would otherwise be required.

GSD constructed several measures to capture the relationship of assistance type and operating costs. First, we include a dummy variable for mortgage subsidy. This variable has a small and marginally significant negative coefficient. Second, we included a categorical measure of the percentage of units for which a property receives project-based Section 8 assistance. Third, we include dummy variables for limited-dividend properties. Fourth, we create a dummy variable for all unassisted and assisted for-profit properties. This variable, however, is not simply entered as a main effect in the model; instead it is interacted with the geographic dummy variable. (See discussion below under Local Cost Adjusters section.)

These variables are included primarily to accurately specify cost relationships in the FHA dataset. In application to public housing, the mortgage subsidy variable is not applied. However, two of the variables are applied to public housing.

First, in application, all public housing properties are treated as if they have the costs associated with 100 percent Section 8 project-based units. This is equivalent to a 6.4 percent increase over the estimated costs for a property with no Section 8 assisted units, all other characteristics held constant.

Secondly, in application, all public housing properties are treated as if they have costs equivalent to those faced by non-profit owners. As is described below in the section on Local Cost Adjustors, GSD has estimated the average cost differential between non-profit and for-profit properties, holding all other model variables constant. This non-profit differential is assumed to reflect the additional costs that are associated with operating outside of the private sector. It is assumed that the costs faced by non-profits are equivalent to the costs faced by public housing agencies, in that both types of organizations face similar external oversight and share similar corporate structures. GSD found that non-profits had costs that were 12% higher than for-profits; however, as described in Chapter 1, and resulting from both public policy concerns and results of field testing, GSD applied a 10% coefficient for non-profit ownership to public housing.

5.2.10 Local Cost Adjustors

One of the most important cost drivers in the model is geographic location. Because the FHA housing stock is a national database, cost differences across metropolitan areas and between rural and urban areas can be extremely large, reflecting the great diversity in wages and costs of living found throughout the United States. As mentioned in Chapter 1, GSD took extreme care to construct the most detailed possible measure of geographic location, while still ensuring sufficient sample size in each area to produce statistically significant cost estimates. It was decided that a geographic area must have at least 25 observations; geographic areas with fewer than 25 observations were grouped into a higher level of aggregation.

Geographic indicator variables were constructed as follows. Properties in metropolitan areas were grouped according to their PMSA. Properties in non-metropolitan areas were grouped at the state level, so that all properties in the non-metro portion of each state were grouped together.

If there were fewer than 25 observations in a PMSA, then observations within that PMSA were grouped together with other properties from small metropolitan areas within the state. Thus, an additional category of state-level metropolitan area was created. If the number of properties in the state-level metro area was still fewer than 25, then these metropolitan properties were grouped with other metropolitan properties within the census division.

Similarly, if there were fewer than 25 properties within the non-metropolitan portion of a state, those rural properties were grouped with other rural properties within the census division.

Initially, GSD constructed a total of 213 local areas, representing both metropolitan and rural areas across all regions of the country. In order to identify market-driven differences in costs across geographic areas, however, GSD decided after some discussion that the best measure of geographic cost differences would be based on the cost differentials found among unassisted and for-profit property owners. These owners would be most responsive to market pressures, and therefore cost differentials within this group are thought to most accurately reflect differences in wages and costs of living.

For this reason, we recreated our local area measures, creating one measure for unassisted and for-profit properties, and one measure for assisted properties that are not for-profit, in each geographic area. Because we intended to base our cost differentials on the coefficients based on the unassisted geographic areas, GSD set the minimum sample size criteria to apply to the unassisted and for-profit housing stock. Thus, local areas were constructed according to the rules described above, except that the minimum sample size of 25 had to apply to the unassisted and for-profit portion of the housing stock. In the cases of Toledo, Ohio, Orange County, California and Pacific Census Division non-metro, there were only 24 for-profit properties, but GSD decided to still retain those values and not group with other areas. With this restriction, GSD was able to construct a total of 129 local areas.

Finally, after construction of our local areas according to the general rules described above, some changes were made to reflect special circumstances. For example, we have divided the New England census division into two parts (north and south) since the data indicate that housing costs are significantly different between the two. Similarly, properties in Washington and Oregon statewide metro areas are grouped into a single category, rather than lumping them into the California statewide metro areas. In the end, our final specification of the cost model contains a total of 78 local areas.

Cities or towns that are part of a Consolidated Metropolitan Statistical Area (CMSA) but that did not have enough properties to get an individual local estimate were given an adjustment to reflect that they were part of a larger housing market and not quite the same as smaller metropolitan areas in the state or census division. These areas received an average of the statewide or divisionwide estimate and the estimate for the Primary Metropolitan Statistical Area in the CMSA. For the very large CMSA of New York, there were sufficient properties to create an estimate for the surrounding communities in the CMSA. For the CMSAs of San Francisco and Miami/Fort Lauderdale, GSD generated one coefficient for the entire CMSA. In the case of Boston, there were sufficient properties within the PMSA to create a unique coefficient and the remaining communities in the CMSA received the statewide metro coefficient.

The final set of local areas defined in the model include:

- Alaska statewide metro areas
- Chicago, IL PMSA
- Indianapolis, IN MSA
- Detroit, MI PMSA
- Cleveland-Lorain-Elyria, OH PMSA
- Columbus, OH MSA
- Dayton-Springfield, OH MSA
- Toledo, OH MSA
- Milwaukee-Waukesha, WI PMSA
- East North Central census division wide metro areas
- Birmingham, AL MSA
- Mobile, AL MSA
- Lexington, KY MSA
- Knoxville, TN MSA
- Nashville, TN MSA
- East South Central census division wide metro areas
- Hawaii statewide metro areas
- New York, NY PMSA

- Balance of New York CMSA (excluding NY PMSA)
- Pittsburgh, PA MSA
- Philadelphia, PA-NJ PMSA
- Mid Atlantic census division wide metro areas
- Phoenix-Mesa, AZ MSA
- Tucson, AZ MSA
- Denver, CO PMSA
- Colorado statewide metro areas
- Salt Lake City-Ogden, UT MSA
- Las Vegas, NV-AZ MSA
- Mountain census division wide metro areas
- Boston, MA-NH PMSA
- New England (North) census division wide metro areas
- New England (South) census division wide metro areas
- Los Angeles-Long Beach, CA PMSA
- Orange County, CA PMSA
- Sacramento, CA PMSA
- San Francisco-Oakland-San Jose, CA CMSA
- California statewide metro areas
- Oregon and Washington statewide metro areas
- Seattle-Bellevue-Everett, WA PMSA
- Portland-Vancouver, OR-WA PMSA
- Puerto Rico statewide metro areas
- Miami-Fort Lauderdale, FL CMSA
- Florida statewide metro areas
- Atlanta, GA MSA
- Georgia statewide metro areas
- Baltimore, MD PMSA
- Greensboro-Winston-Salem-High Point, NC MSA
- Raleigh-Durham-Chapel Hill, NC MSA
- North Carolina statewide metro areas
- South Carolina statewide metro areas
- Norfolk-Virginia Beach-Newport News, VA-NC MSA
- Richmond-Petersburg, VA MSA
- Charlotte-Gastonia-Rock Hill, NC-SC MSA
- Washington, DC-MD-VA-WV PMSA
- South Atlantic census division wide metro areas
- Kansas City, MO-KS MSA
- West North Central census division wide metro areas
- Little Rock-North Little Rock, AR MSA
- Dallas, TX PMSA
- Houston, TX PMSA
- West South Central census division wide metro areas
- Cincinnati, OH-KY-IN PMSA
- Louisville, KY-IN MSA
- Minneapolis-St Paul, MN-WI MSA
- St Louis, MO-IL MSA
- Alaska statewide non-metro areas
- East North Central census division wide non-metro areas
- East South Central census division wide non-metro areas
- Hawaii statewide non-metro areas
- Mid Atlantic census division wide non-metro areas
- Mountain census division wide non-metro areas

- New England census division wide non-metro areas
- Pacific census division wide non-metro areas
- South Atlantic (north) census division wide non-metro areas
- South Atlantic (south) census division wide non-metro areas
- West Virginia statewide non-metro areas
- West North Central census division wide non-metro areas
- West South Central census division wide non-metro areas

For-profit properties located in the Cleveland-Lorain-Elyria, OH PMSA serve as the reference category in the cost regression model. This PMSA was chosen as the reference category because it had costs that were close to average across all MSAs. The choice of reference category is arbitrary (it does not change the model results), but it makes the model coefficients on the other metro areas easier to interpret when they are presented relative to a baseline value that is close to a "typical" metropolitan area.

As mentioned above, GSD needed to identify the overall difference in costs between for-profit and non-profit properties. However, as discussed in this section, for-profit properties were also used to define the geographic cost adjustments. Therefore, to identify the overall differential between non-profit and for-profit, it was necessary, in effect, to take the average difference between non-profit and for-profit properties over all of the metropolitan areas. An overall non-profit adjustment factor, calculated essentially by averaging the difference between for-profit and non-profit costs over all geographic areas, is estimated simultaneously with the interaction model. The overall difference in estimated costs between non-profit properties and for-profit properties, holding all other factors constant, is 12 percent (reduced, as discussed earlier, to 10 percent). In application to public housing, all public housing properties receive this average non-profit adjustment factor of 10 percent above the estimate costs for for-profit properties.

Note that GSD had to make a trade-off between examining smaller subsets of the data, and providing unique cost estimates the maximum possible number of geographic areas. It would have been reasonable to have subdivided each metropolitan area into central City and suburb, on the grounds that different metropolitan areas have different *cost differentials* between central city and suburb. Our model specification does not do this: instead, we have one aggregated Central City dummy that captures the average, national difference in costs between central cities and suburban areas. Because GSD felt it was crucial to interact each geographic area with ownership status, we were limited in our ability to conduct additional interactions while still maintaining a large number of geographic areas with sufficient sample size.

Finally, there were several geographic areas for which GSD could not directly apply the rules described above. For Puerto Rico, Hawaii, Alaska, and a portion of the New England Census Division, there were insufficient numbers of for-profit properties to construct either the metropolitan or non-metropolitan geographic area dummies as described.

- For the Puerto Rico metro area, the Puerto Rico non-metro area, the New England (North) Census Division metro area, and the New England Census Division non-metro area, the following rule was applied: all properties in the geographic area were combined to make one geographic dummy, and the estimate based on those geographic dummies was considered to be an estimate of non-profit operating costs. In other words, when the model is applied, PHAs located in one of the geographic areas discussed are not assigned an additional non-profit add-on, because their geographic cost adjustor is already based on non-profit housing stock.

- There are no PHAs located in the Hawaii non-metro area, so no adjustment was made. For metropolitan Hawaii, and for both metropolitan and non-metropolitan Alaska, where there are not enough combined FHA properties, GSD recommends further study in those markets. Still, GSD kept the model estimates as "placeholders" since they appear to be within a reasonable range of what further field research in those markets might find. GSD held discussions with the Alaska Housing Finance Corporations and reviewed the range of operating costs for assisted properties that they finance. Those costs generally ranged from $200-$300 PUM. Consequently, the model

estimate for Alaska, $308 PUM, seemed reasonable. Similarly, the model estimate for Hawaii, $353 PUM, seemed reasonable in relationship to operating costs in California markets, although, as mentioned, supplemental research in those markets is recommended.

Finally, there are no FHA properties in Guam or the Virgin Islands and, hence, GSD was not able to construct any model estimates. For these two markets, like Hawaii and Alaska, GSD recommends further field work that would examine the operating costs of some suitably comparable properties and, relying on professional expertise, result in a recommended expense level. Resources under this study did not allow GSD to conduct such an analysis.

5.2.11 Rent-to-FMR ratio

As mentioned above in the discussion of Property Age, housing quality is an extremely important variable in the Cost Model. Omitting housing quality obscures the relationship between operating costs and other cost drivers that may also be associated with quality – this is extremely evident when we consider the relationship between Age, Housing Quality, and Operating Costs.

The ratio of a property's rent to the FMR is included in the model as a measure of housing quality. Housing quality is a cost driver because it costs more to provide a higher level of housing service.[16] Just as additional bedrooms provide more housing service, and increase operating expenses, so do a variety of other housing traits that we cannot measure directly with the variables in the FHA database.

Including additional measures of housing quality is expected to increase the overall explanatory power of the model and, by reducing omitted variable bias, give us coefficients on other variables that better match our expectations.

- **Rent/FMR as a Measure of Housing Quality.**

  For properties that charge market rents, the ratio of the property's average rent to the local market's fair market rent provides a summary indicator of the amount of housing service provided by the typical unit in that property. Rent serves as a summary measure of the quantity and quality attributes of the typical unit in a property. The FMR is used to deflate the rent variable for area differences in housing costs. The FMR is not a true price index for rental housing, but it provides a good approximation.

- **Interacting Rent/FMR with Assistance Type.**

  Profit-maximizing property owners will attempt to minimize their operating costs, whatever the level of quality they are providing (stated differently, whatever the segment of the market they are serving). So, for market-rate, profit-maximizing housing operators, there are well defined and market determined links from housing quality to rent/FMR and from rent/FMR to operating costs.

  Assisted properties do not necessarily follow this model. First, because assisted properties do not charge market rents, rent/FMR cannot be taken as a "clean" indicator of housing quality. There may be some correlation between rent/FMR and housing quality in assisted housing, but the calibration is uncertain and varies from property to property. A second issue arises if assisted property owners, on average, do not maximize the service provided for a given level of operating cost, because they are not under market pressures to do so. In this instance, the relationship of rent/FMR to operating costs will differ between market rate and assisted properties.

[16] Economists often think of housing consumption in terms of "units of housing services provided," where a unit is a synthetic quality/quantity amalgamation.

Because we expect the relationship of rent/fmr to operating costs to differ for assisted and unassisted properties, rent/fmr is included in the model interacted with assistance type. In application of the Cost Model to public housing, we assume that well-managed public housing should provide housing with a quality level roughly comparable to the median unassisted property in the FHA housing stock. The median unassisted property in the FHA housing stock had a ratio of rent/fmr of approximately 1, meaning that rents charged were very close to their fair market rent (conditional on unit size). PHAs, of course, do not charge market rents. In the application of the model, therefore, GSD assigned costs to public housing under the assumption that public housing should have costs equivalent to those faced by an unassisted property charging rents that are close to the fair market rent in their area[17].

- **Variable Construction.**

The numerator is per-unit-monthly rent revenue. The denominator is the FMR for the MSA (or county, for rural properties), which varies by unit size. For each property, the ratio of rent/fmr was constructed as the average per-unit-monthly rent revenue, divided by the weighted average of FMRs for units of different sizes, where the weights used correspond to the distribution of units within the property.

After the size-weighted rent/fmr ratio was constructed, we grouped the variable into five categories. As can be seen in the table below, the median value for rent/fmr was close to 1 in the unassisted stock, implying that the median unassisted property in the FHA housing stock was of a quality level close to average for its market area.

Table A.15: Distribution of Rent / FMR Ratio for Unassisted and Assisted Properties

Quantile	All	Unassisted	Assisted
Sample Size	16501	2947	13511
100% Max	9.27	6.13	4.14
99%	2.17	2.96	2.11
95%	1.76	1.78	1.76
90%	1.58	1.44	1.59
75% Q3	1.30	1.21	1.32
50% Median	1.04	1.02	1.05
25% Q1	0.83	0.88	0.81
10%	0.65	0.75	0.63
5%	0.55	0.67	0.54
1%	0.39	0.44	0.39
0% Min	0.00	0.00	0.00

The five categories into which we grouped the rent-to-FMR ratio were:

- Rent / FMR less than 0.7
- Rent / FMR 0.7 - 0.9
- Rent / FMR 0.9 - 1.1

[17] In practice this means that when the model is applied to public housing, all public housing units are treated as if their rent/fmr ratio is within the range of 0.9 – 1.1.

- Rent / FMR 1.1 - 1.6
- Rent / FMR greater than 1.6

Each of these categories was interacted with assistance type (assisted/unassisted).

- **Alternative Specifications Tested.**

  We tested including the rent/fmr ratio as a continuous variable; it was highly significant. However, we decided to enter the variable in categorical ranges for ease of interpretation and application. Break-points were set so that the sample was distributed fairly evenly across categories.

  Coefficients on each category of rent/FMR are very similar for assisted and unassisted properties; t-tests showed that the coefficients on most categories were not significantly different for the assisted and unassisted stock. However, we retained the interaction term in the model because rent/FMR captures a different theoretical concept in the unassisted stock.

- **Simple Correlations between Rent/FMR and Other Cost Drivers.**

  In simple correlations, rent/FMR is, generally, positively correlated with factors that are indicators of higher quality multifamily properties. This can be seen in the table below. Rent/FMR is positively correlated with the REAC physical inspection score; and it is negatively correlated with the percentage of REAC capital deficiencies; property age; average unit size; and the percentage of large (3+ bedroom) units. The one exception to this pattern is that the rent/FMR is positively correlated with the poverty rate in the census tract; however, this correlation is not very strong.

Table A.16: Simple Correlations Between Rent/FMR and Other Cost Drivers

Pearson Correlation Coefficients
Prob>|r| under H0: Rho=0
Number of Observations

% Capital Deficiencies	Property Age	Average number of bedrooms per unit	Percentage of large (3+ bedroom) units	REAC physical inspection score	% Poor families in census tract	PUM Operating Costs
-0.095	-0.301	-0.270	-0.238	0.140	0.040	0.240
<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
15143	15107	16501	16501	16155	15799	16468

- **Simple Correlation with Operating Costs.**

  Properties can have high costs for three quite different reasons:

  A) They require a high level of service because of aging infrastructure and/or a difficult tenant population and neighborhood environment.

  B) They are being maintained at a high level of quality.

  C) They are being poorly managed.

  On average, property owners facing market pressures will not be able to charge rents in excess of the rents charged by other comparable properties. Because GSD uses rent/FMR based on the unassisted housing stock only in the application of the Cost Model, we expect

that high rent/FMR ratios in our model do not reflect poor management, but on average reflect higher costs due to reasons 1 or 2.

As seen above, rent/FMR is generally correlated with other property characteristics that are predictive of lower costs. For example, rent/FMR is negatively correlated with property age and the percentage of capital defects, both predictors of higher costs.

Despite these correlations, rent/FMR is positively correlated with operating costs. Together, this set of correlations strongly supports the argument that rent/FMR is capturing housing quality.

5.2.12 Property Owner

A subset of the properties in the FHA dataset (less than 3 percent) are owned by a person who owns at least one other FHA property. Thus, it is possible that model residuals are correlated among properties owned by the same person. To ensure that our results were robust to correlated errors, we estimated the model using Huber-White estimated standard errors that controlled for clustering at the level of the property owner. In fact we found that running these robust standard errors produced very little change from the OLS standard errors, probably because such a small percentage of the dataset is held by owners who own more than one FHA-insured property.

5.2.13 Average number of square feet per unit (by unit size)

We tested whether average square feet per unit was a cost driver. For each property we had 4 square foot variables: average square foot for efficiencies, average square foot for one-bedroom units, average square foot for two-bedroom units, and average square foot for three-bedroom and larger units. Square feet per unit was hypothesized to be a cost driver for the following reasons:

1. Larger units mean more physical space to clean and maintain, potentially increasing costs.

2. Larger units are easier to overcrowd, possibly increasing the number of residents per unit, thereby possibly increasing wear and tear and hence costs.

3. Larger units mean more room for residents, possibly decreasing social tensions and therefore possibly decreasing building wear and tear due to social problems.

4. Larger units might be a proxy for higher quality properties. Quality is a critical determinant of operating costs (maintaining buildings at high levels of quality requires larger expenditures) but can only be measured indirectly. Our primary measure of quality is the rent/FMR ratio discussed above. We tested whether including the square foot variables changed the coefficient on the rent/FMR variables, and whether the square foot variable had a similar effect on other model coefficients (especially age) as did the rent/FMR variable.

We found that:

1. In simple correlations, square feet per 0, 1, 2, and 4 bedroom units were positively correlated with operating costs.

2. In multivariate models, only average square feet per 1-bedroom unit is significantly correlated with operating costs. We found a small, positive correlation: a 100 square foot increase is associated with a 0.4 percent increase in operating costs.

3. Including square foot per unit does not change the coefficients on the RENT/FMR variables. Also, it does not substitute for the rent/FMR variables. When we drop rent/FMR from the models but include square feet, we no longer see the expected positive relationship between property age

and property costs. Thus square foot does not appear to be controlling for quality the way that Rent/ FMR does.

4. Including square foot does not increase model R^2 or adjusted R^2.

For all of these reasons, we decided not to retain square foot per unit in the final model. In addition, square footage is not a variable that is currently collected for public housing units, so this variable would have been difficult to apply had it proven important.

5.3. Summary of Variable Definitions

This section provides a concise summary of how each variable used in the final version of the model was constructed.

- Operating costs (the dependent variable): Operating costs, measured per unit per month, are constructed by taking the sum of three line items reported in the Statement of Profit and Loss portion of the Annual Financial Statements, minus reserve releases. We entered the natural logarithm of this value as the dependent variable in the model.

  The three line items are: total administrative expenses (line 6200/6300); total operating and maintenance expenses (line 6500); and total taxes and insurance (line 6700) minus real estate taxes (line 6710). From this sum, we subtracted the Replacement or Painting Reserve Releases that are included as part of the expenses reported on the Statement of Profit and Loss.

  Costs from three years of Annual Financial Statements were included: 1998, 1999, and 2000. Costs from 1998 and 1999 were inflated to year 2000 dollars using the Bureau of Labor Statistics' Consumer Price Index housing component. Thus, 1998 values were multiplied by 1.0574, and 1999 values were multiplied by 1.0348.

- Property Size: Property size was entered as a single dummy variable that equaled one for properties with 150 or more units, and equaled zero for a property with less than 150 units.

- Property Age: The underlying continuous variable from which we constructed our model variables is the age of property in year 2000 as measured from the final endorsement date of first mortgage. We used four age dummies in the model. The dummy variables are defined as: age less than 15 years, age 15 – 20 years, age 21 – 25 years, and age greater than 25 years. The reference category is age less than 15 years.

- Unit Size: We measured the distribution of large and small units by entering four continuous variables in the model: the percentage of two bedroom units; the percentage of three bedroom units; the percentage of four bedroom units; and the percentage of five or more bedroom units. The reference category is the percentage of efficiencies and one bedroom units.

- Clientele (Senior / Family): We entered a single dummy variable for "senior" properties in the model. (The reference category was family properties.) The variable is from the REMS database.

- Building Type: We entered four building type dummies in the model. These are: detached and semi-detached; row / townhouses; high-rise / elevator; and mixed. The reference category is garden-style / walk-up.

- Central City / Suburb: We entered a single dummy variable indicating central city status in the model. The reference category is suburb.

- Poverty Rate of Census Tract: We entered the 1990 Census tract poverty rate in four categories: poverty rate is 21 – 30 percent; poverty rate is 31 – 40 percent; and poverty rate is greater than 40 percent. The reference category is a poverty rate of 0 – 20 percent.

- Mortgage Subsidy: We entered a dummy variable that indicated whether the property was receiving a mortgage subsidy.

- Percentage of Assisted Units: We entered the percentage of units in the property for which the property is receiving project-based Section 8 assistance in four categories: 1 – 20 percent; 21 – 80 percent; 81 – 99 percent; and 100 percent. The reference category is 0 percent of units receiving project-based Section 8 assistance.

- Ownership Type: (For-Profit, Non-Profit, Limited Dividend). We grouped properties into three categories based on ownership type of the mortgage sponsor: limited dividend ownership; non-profit ownership; and for-profit ownership. Note that the for-profit ownership type category includes all properties that were unassisted (all properties that did not receive HUD rental assistance) and a small number of assisted properties with unlimited dividend ownership structure.

  We did not enter the ownership type dummies into the model directly. Instead, we interacted each of the three ownership type dummies with each of the geographic area dummies. We then entered these interacted terms in the model. Thus, for every geographic area X, there are three model terms: Area X*For-Profit (a dummy variable for all for-profit properties in Area X); Area X*Non-Profit (a dummy variable for all non-profit properties in Area X); and Area X*Limited Dividend (a dummy variable for all limited-dividend properties in Area X).

  When applying the cost model to the public housing stock, we only used the for-profit area coefficients (and thus ignored the limited dividend and non-profit area coefficients). To account for the non-profit operating environment of public housing, an overall add-on, equal to 12%, was added to the predicted cost estimate of each public housing development.

  The non-profit add-on was obtained by running the model in SAS PROC GLM procedure. PROC GLM creates an aggregate difference between the for-profit and non-profit properties, over all of the metropolitan and non-metro areas, at the same time that it estimates each area * ownership interaction term. Conceptually, this is like taking a weighted average of the difference between each for-profit area coefficient and the non-profit area coefficient, weighted by the number of properties in the area, and averaged over all areas. The reason we had to do this was because 1) we wanted to know the overall difference in costs between for-profit and non-profit, but, 2) we also wanted to accurately identify the cost differential between for-profit and non-profit in each individual area separately.

- Geographic Areas: We identified a total of 78 geographic areas in the model. Each geographic area is interacted with the three ownership dummies. For example, consider Pittsburgh, PA. There are three dummy variables entered for Pittsburgh: Pittsburgh * For-Profit (a dummy variable equal to one for all for-profit properties in Pittsburgh and zero for all other properties), Pittsburgh * Non-Profit (a dummy variable equal to one for all non-profit properties in Pittsburgh and zero for all other properties), and Pittsburgh * Limited Dividend (a dummy variable equal to one for all limited dividend-profit properties in Pittsburgh and zero for all other properties).

- Housing Quality (Rent – to – FMR Ratio): The underlying variable is defined as the average per-unit rent charged in the property, divided by the weighted average of the FMRs for units of different sizes in the area where the property is located. The FMR is weighted according to the mix of unit sizes in the property. After constructing the rent:FMR ratio, We then constructed 5 dummy variables based on the rent:FMR ratio: 1. Rent/FMR < .07; 2. Rent/FMR = .07 - .09; 3. Rent/FMR = .091 – 1.1; 4. Rent/FMR = 1.11 – 1.6; 5. Rent / FMR > 1.6. We then interacted each of these dummies with a dummy variable for Assistance Type, producing a total of ten dummies:

Rent/FMR < .07 * Assisted, Rent/FMR < .07 * Unassisted, Rent/FMR=.07 - .09*Assisted, etc. We entered eight of these ten dummies in the model. The omitted (reference) dummies are: Rent / FMR= .091 – 1.1 * Assisted, and Rent / FMR= .091 – 1.1 * Unassisted. We had two omit two categories (rather than just one) to avoid multi-colinearity, because the model already controls for Assistance through the combination of the subsidy variables and the percent Section 8 assisted variables.

6. STATISTICAL PRECISION OF THE MODEL

This section discusses several issues related to the statistical precision of the model. We first present a general discussion of the overall fit of the model. We then turn to a discussion of the predictive accuracy of the model, overall and for various subsets of properties. Finally, we present a brief discussion of our handling of outliers in the data.

6.1 OVERALL MODEL FIT AND SOURCES OF UNEXPLAINED VARIATION

The overall R^2 statistic for the cost model is .53. This means that the variables in the model explain just over half of the overall variation in operating costs in the FHA dataset. For example, if the model variables (age, size, location, clientele, etc) perfectly predicted costs, then the R^2 would = 1. If the model variables were unrelated to costs, then the R^2 would = 0. R^2= .53 is quite good for a cross-sectional model (in other words, for a model that cannot use property fixed-effects to control for unobserved property-specific characteristics.) However, there is of course a significant amount of variation in operating costs that the model is unable to explain. As discussed in Chapter 1, the 47 percent of the variation in costs that is not explained could result from two types of factors.

First, the unexplained variation could, of course, reflect the omission of relevant variables that are not captured in the database. For example, some commenters have raised the issue of local crime rates. To the extent that our measures of neighborhood quality do not fully capture levels of crime, any variation in costs that reflect variation in local crime rates will contribute to the unexplained variation in the model. Instead of producing predicted costs that are higher in high-crime areas and lower in low-crime areas, our model will average out these costs, providing one number for properties that may face quite different crime rates. Of course, to the extent that crime rates vary by central city versus suburb, by metropolitan area, and by region of the country, we will have controlled for them in our geographic area variables. We also found in the field testing conducted as part of the overall study that the model did not produce operating cost estimates that were too low to provide needed security for properties in central city locations. So, we discuss crime rates here only as an explanatory example.

Second, another source of unexplained variation (the 47 percent of variation in costs that is not explained by the mode) is, simply, differences in the choices made by owners and managers. Because management practices differ, two different owners may manage the same property with the same tenants and in the same location for quite different costs. Our model does not have a variable that indicates “efficient management”, and even among owners facing the pressures of market competition, levels of efficiency will vary substantially. Our model will average out these costs, providing one number for otherwise identical properties that are actually managed quite differently. Our model also does not have a variable that measures management quality or the level of management services provided, except indirectly and through the rent-to-FMR variable. Among properties with similar positions in the rental market, managers may provide very different levels of staffing for certain management functions or higher or lower levels of routine maintenance.

The existence of important omitted variables (the first source of unexplained variation) would be, indeed, a valid critique of the model. If it is found, for instance, that costs are significantly higher in high-crime areas, it would be reasonable that an out-of-model adjustment should be made to predicted costs for properties in high-crime areas. It should be recognized, however, that a downward adjustment for properties in low-crime areas should also be made, if such a process is implemented.

The second source of variation discussed, however, does not call into question the validity of the model predictions. The fact that the model predicts costs with a large fraction of unexplained variation does not mean that the model predictions are inaccurate: it means that the model is predicting costs for typical properties, averaged over a diverse range of management styles. The model predicts the costs that are sufficient for a "typical" manager of FHA housing to operate a property with a given set of characteristics. and in a given location. These predicted costs will be lower than the actual costs of managers who operate at atypically high budgets, and will be higher than the actual costs of managers who operate at atypically low budgets. There is no way to eliminate from the model this inevitable variation in management choices and, we believe, no need to do so.

6.2 PREDICTIVE ACCURACY OF THE MODEL

While the R^2 statistic provides a summary measure[18] of how much observed variation in the data is explained by the model, a different approach to evaluating model reliability is to examine the model's forecasting accuracy. Because the purpose of the cost model is to accurately predict costs[19], measures of forecast accuracy provide a better measure of the functionality of the model than does the R^2 and similar statistics.

Standard errors measure the amount of variation, or noise, around a predicted value. They are thus an estimate of the level of uncertainty around our prediction. Once standard errors have been estimated, *confidence intervals* can be created which indicate the range around our prediction within which a certain percentage of our data is expected to fall. For example, with normally distributed errors, once we have an estimated standard error ($\hat{SE}$), we know that 95 percent of our sample observations will fall within their predicted value plus or minus 1.96 * $\hat{SE}$.

Two measures are used to assess the level of uncertainty around the model predictions: these are the standard error of the prediction, and the standard error of the forecast[20].

The standard error of the linear prediction ("STDP") measures the uncertainty of the prediction that originates from the uncertainty of the estimated model coefficients. This is model-generated uncertainty, and it is the measure that we use to assess the precision of our model-generated estimates. The confidence intervals based on the STDP are intervals around our prediction for the mean; these are commonly referred to simply as the model confidence intervals.

The standard error of the forecast ("STDF") measures the uncertainty of the prediction that originates from both the prediction's standard error (the STDP) and our inherent uncertainty due to unmeasured characteristics of the individual properties, or the residual standard error. These two sources of error combined are called the forecast standard error. This is the measure that we use to test whether the model predicts accurately for individual observations, accounting for unmeasured individual variation (for example, unmeasured differences in management style.) The confidence intervals based on the STDF are commonly called prediction intervals – the intervals around our predictions for individual observations.

When we assess the accuracy of our model's ability to predict costs for a typical property with a given set of characteristics, we use the STDP and its associated confidence interval. This is because we are not attempting to predict unmeasured individual variations; rather, we want to know how accurately our model

[18] There are several other summary statistics available that serve similar purposes, such as the model root mean squared error or RMSE, which is discussed below.

[19] If the primary purpose of the Cost Model were to identify the precise functional form of the causal relationships between cost drivers and the dependent variable, then we would care less about forecast accuracy and more about classical regression statistics. For our purposes, however, we care principally about having a model that produces accurate forecasts.

[20] This explanation of the standard error of the prediction and the standard error of the forecast is based on a discussion in "The STATA Reference Manual Release 6, Volume 1" TX: Stata Press, 1999. p. 8.

predicts costs for a typical property. The confidence interval around the model prediction tells us the range of uncertainty around our prediction for a typical property. When applying our model to public housing properties, the confidence intervals tell us how sure we are about the precision of each prediction.

When we are testing the accuracy of our model for forecasting individual observations, we use the STDF and its associated prediction interval. We use the prediction interval for purposes of testing how well our model predicts individual values in the hold-out sample. Because we are attempting to predict values for individual properties – not to predict values for typical properties, averaging over individual variation in unmeasured characteristics – we must use the STDF, which accounts for unmeasured individual variation.

6.3 SIZE OF THE CONFIDENCE INTERVALS AROUND THE MODEL PREDICTIONS

As discussed above, the confidence intervals based on the standard error of the prediction measures the accuracy of our model predictions. Overall, the model predictions had a mean confidence interval of $28, and a median confidence interval of $26. The average confidence interval was plus or minus 12 percent of the point estimate. As can be seen below, the confidence intervals for over 90 percent of properties in the estimation sample were between 7 percent and 19 percent of the point estimate, clustered around a median of 11 percent.

Table A.17: Distribution of Confidence Intervals

Percentiles	CI / Predicted Cost
1%	0.062
5%	0.071
10%	0.075
25%	0.085
50%	0.107
75%	0.136
90%	0.166
95%	0.190
99%	0.288

6.4 TESTS ON THE HOLDOUT SAMPLE

One advantage of the FHA dataset was that it was large enough to permit us to construct both a holdout sample and a development sample. A holdout sample is a random sub-sample of the original dataset that is set aside and not used in the development of the analytical model. Because this sample was not used in model development, it can be used to test whether the regression model has been over-fit to the development sample. If extensive testing of alternative variable specifications is conducted, a final model may appear to fit the data very well – but much of this good fit may result from the selection of a specification that happens to fit well the idiosyncrasies of that particular sample, rather than from selecting a specification that accurately reflects the underlying causal relationships between regressors and the dependent variable. If the model coefficients generated through analysis of the model development sample

are good predictors of the dependent variable in the holdout sample, we can be confident that over-fitting was not a significant problem.

The majority of our tests for predictive accuracy apply the model coefficients, developed on our analysis sample, to a "holdout sample" (a sub-sample of our dataset which was not used for model development), to test how accurately our model predicts costs on a "fresh" set of data. We examine the predictive accuracy of the model in the holdout sample overall and for meaningful subsets of the data.

In addition to the tests on the holdout sample, presented here, we also present below a decomposition of the explained variance in the sample, to show which types of characteristics are explaining the majority of the sample variance[21].

Our holdout sample was a 25 percent stratified random sample of the original dataset. The dataset was stratified by property age and property size to ensure that the holdout sample and the model development sample were evenly distributed along these two dimensions. The holdout sample had 2,828 observations, and the development sample (the remaining 75 percent of the original sample) had 8,715 observations.

6.4.1 Comparison of R^2 Statistics

The R^2 statistic in a regression model measures the percentage of variation in the dependent variable, which is explained by the model, as a fraction of the total variation in the dependent variable.

$$R^2 = \frac{[\sum_i (y_i - \bar{y})(\hat{y}_i - \hat{\bar{y}})]^2}{\sum_i (y_i - \bar{y})^2 (\hat{y}_i - \hat{\bar{y}})^2}$$

Where y_i = the observed value of the dependent variable for observation i, $\hat{y}_i$ = the predicted value of the dependent variable for observation i, $\bar{y}$ = mean (y_i), and $\hat{\bar{y}}$ = mean($\hat{y}_i$). Thus, R^2 is the squared correlation between the observed values of y and the predicted values of y generated by the regression equation.

In the model run on the development sample, the model R^2 = .5430. A comparable statistic can be constructed for the holdout sample. When calculated for the holdout sample (using predicted values based on the development sample regression coefficients), the holdout sample R^2 = .5248.

This comparison indicates that the model's explanatory power for predicting cost is nearly as large in the holdout sample as it is in the development sample, indicating that explanatory power is not due to over-fitting the model to idiosyncrasies of the development sample.[22]

6.4.2 Comparison of Actual Costs and Model-Predicted Costs

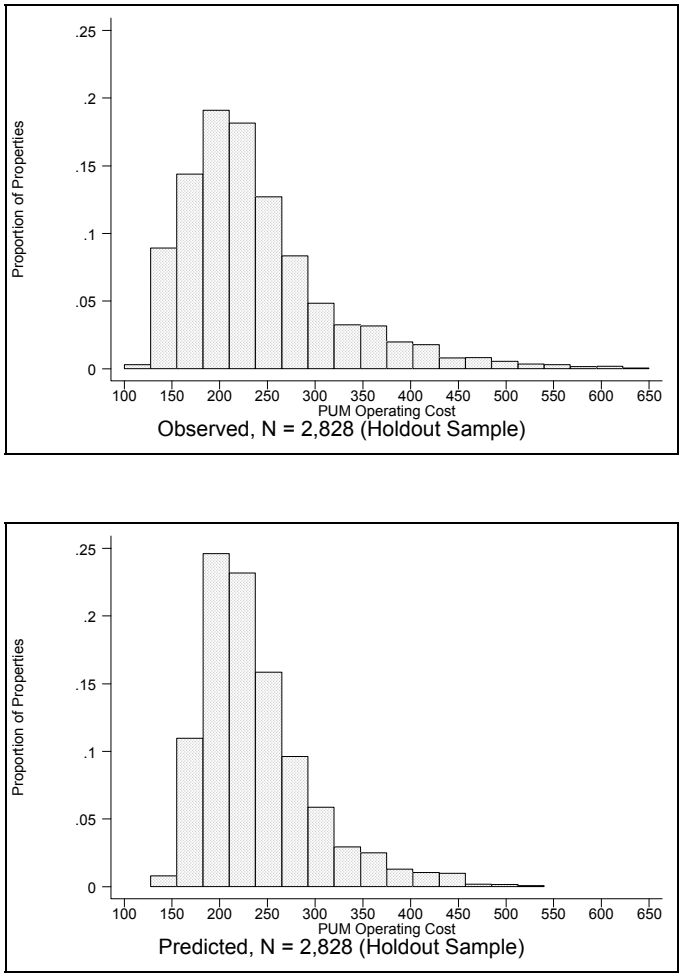
A comparison of the distribution of actual and predicted operating costs in the hold-sample provides an illustration of how well the model predicted over the entire distribution. As is expected in an OLS regression model, which attempts to fit to the mean of the data, the predicted distribution is somewhat more compressed than the actual cost distribution[23]. Nevertheless, the overall distribution of predicted costs in the holdout sample is quite similar to the overall distribution of actual costs.

[21] The Cost Model explains about 53 percent of the overall variation in the dependent variable. "Decomposing" the variance means determining how much of this explained variation is explained by each of the independent variables.

[22] Adjusted R^2 in the 75 percent sample is .5282, and adjusted R^2 in the holdout sample is .4738. Adjusted R^2 provides some adjustment for the number of independent variables in the model; with each additional variable, adjusted R^2 will increase only if the t-statistic on the variable is > 1; if the t-ratio is < 1 the adjusted R^2 will fall.

[23] As expected, Chi-square tests of the two distributions indicate that there are statistically significant differences between them.

Figure A.6: Distribution of Actual and Predicted Costs in the Hold-Out Sample



Distribution of observed and predicted costs in the holdout sample

	N	MEAN	5TH PCTILE	10TH PCTILE	25TH PCTILE	50TH PCTILE	75TH PCTILE	90TH PCTILE	95TH PCTILE
Observed	2,828	$238	$145	$157	$185	$220	$269	$348	$401
Predicted	2,828	$239	$171	$179	$198	$225	$264	$315	$358

6.4.3 Prediction accuracy for key subsets of properties in the holdout sample.

In this section we discuss the predictive accuracy of the model for key subsets of properties in the holdout sample. The purpose of this section is to explore whether the model predicts costs with lower accuracy for particular types of properties. We examine properties along the following characteristics: geographic

location (Census division, state, and metropolitan area[24]); building type; property age; property size; clientele (senior or family); and property ownership structure.

We employed several different measures of predictive accuracy. The measure that we believe to be most useful is the Mean Absolute Percentage Error (MAPE). This measure is equal to the mean, over all observations in the subset, of the absolute value of the difference between each observation's observed and predicted cost, divided by that observation's observed cost.

$$\text{MAPE} = \Sigma_i \left(|(\mathbf{y} - y_i)| / y_i\right) / N$$

where N = the number of observations in the subset of interest.

There are two particularly useful features of the MAPE:

1. Over-predictions and under-predictions do not cancel each other out. If a model is predicting very inaccurately, but some predictions are far too high and other predictions are far too low, it is important to use a measure of predictive accuracy that is based on the absolute value of the prediction error.
2. Percentage differences give larger weight to errors that are large relative to the size of the observed cost. A $5 error in predicting a cost of $400 is less significant than a $5 error in predicting a cost of $150.

The other measures of predictive accuracy we used to test the model are:

Mean Absolute Error (MAE)	$\Sigma_i \left(|(\mathbf{y} - y_i)|\right) / N$
Mean Error (ME)	$\Sigma_i \left(\mathbf{y} - y_i\right) / N$
Mean Percentage Error (MPE)	$\Sigma_i \left((\mathbf{y} - y_i) / y_i\right) / N$

These measures also have some utility. The MAE is useful to the extent that we care equally about all deviations of predictions from observed costs (in the same units as the dependent variable), regardless of the size of the observed cost. The ME and MPE both tend to drastically understate overall error, because positive and negative errors cancel each other out. However, they are useful statistics to check for bias, i.e., that no systematic errors (systematic over-prediction or under-prediction) are occurring in particular subsets of the data. Thus, in most of our tables we will focus on the MAPE, we will also discuss the MAE and MPE statistics when they provide useful additional information.

6.4.5 Overall Distribution of the Absolute Percent Error

The MAPE in the holdout sample overall is 17, meaning that the difference between predicted and actual costs for the typical property in the holdout sample is equal to 17 percent of actual cost. In addition to the Mean Absolute Percent Errors, it is of interest to observe the entire distribution of the Absolute Percent Errors (APEs). The distribution below shows that, for the median property in the holdout sample, the APE was 14 percent; 25 percent of properties had APEs of 6 percent or less; and 25 percent of properties has APEs greater than 24 percent.

[24] Actually, for this finer level of geography, we examine metropolitan areas for metro properties, and states for non-metropolitan properties. Metropolitan areas were defined to have at least 25 unassisted properties in the 75 percent sample. Metropolitan areas with fewer observations were aggregated into higher levels of geography. For further details, see the description under "Model Development." In this section we are examining predictive accuracy for all types of properties (assisted and unassisted) combined within each geographic area. Note that sample size may be quite small for any given geographic area in the holdout sample.

Table A.18: Distribution of the MAPE in the hold-out sample

	Percentiles
1%	0
5%	1
10%	3
25%	6
50%	14
75%	24
90%	35
95%	45
99%	66

This distribution does indicate that there was a significant fraction of properties in the holdout sample for which the model predictions deviated from observed costs by 25 percent or higher.

The Mean Error in the holdout sample was just $1, indicating that there was no systematic over-prediction or under-prediction in the sample overall. The Mean Absolute Error was $40, indicating that for the average property in the holdout sample, the difference between predicted cost and absolute cost was $40.

6.4.6 Predictive Accuracy of the Model Across Census Divisions

Examining the distribution of errors by Census division in the holdout sample, it is clear that in most census divisions the MAPE was close to, or slightly lower than, the overall average of 17 percent. One region stands out as having a somewhat higher MAPE, however: New England, where the MAPE is 22 percent.

Table A.19: Mean and Median Absolute Percentage Error by Census Division

Census Division	N	Mean APE	Median APE
East North Central	484	15	12
East South Central	249	16	13
Mid Atlantic	322	20	15
Mountain	170	17	14
New England	143	22	18
Pacific	398	19	16
Puerto Rico & VI	25	20	16
South Atlantic	537	17	13
West North Central	254	15	13
West South Central	246	15	12

The MPE in New England is 10 percent – also the largest MPE of any region – indicating that there is some evidence for systematic over-prediction in this region, as there is in the Mid Atlantic, where the MPE is 7 percent.

Table A.20: Mean Percentage Error by Census Division

Census Division	N	Mean Percentage Error	Median Percentage Error
East North Central	484	5	3
East South Central	249	-1	-2
Mid Atlantic	322	7	4
Mountain	170	4	2
New England	143	10	5
Pacific	398	3	2
South Atlantic	537	6	4
West North Central	254	5	5
West South Central	246	4	2

Examining the MAPEs at a state level (see listing at the end of this document), it is apparent that the high MAPE in New England is driven by an exceptionally high level of error in Maine, where the MAPE = 36 percent. Examining MPEs at the state-level, we see that Maine also has a very high Mean Percent Error of 33, indicating that there is significant systematic overprediction in this state. The reason for this level of error in Maine appears to be the imprecision of the local geographic dummies in this state. In fact, metropolitan New England as a whole was grouped into just three geographic regions: Northern New England, Southern New England, and the Boston metropolitan area. This relatively crude geographic grouping led to three of the five New England states having MAPEs within the top eight nationwide, with Connecticut (MAPE=22, MPE=8) and Massachusetts (MAPE=22, MPE=12), both having large and positive aggregate errors. The aggregation led to a significant overestimation of costs in Maine, where costs are among the lowest in New England. Vermont and Rhode Island, in contrast, both have quite low MAPEs, and smaller, negatively-signed MPEs. (Vermont: MAPE=18, MPE= -3; RI: MAPE=13, MPE=-5.)

While there are some unusually high MAPEs among the New England states, which drive the aggregate New England MAPE up, most of the other Census Divisions have MAPEs that seem comparable to the national MAPE. The Division with the second highest MAPE is the Mid Atlantic, containing New York, New Jersey, and Pennsylvania. The Mid Atlantic MAPE is influenced by the relatively high degree of prediction error (and some degree of over-prediction) in New Jersey (MAPE=27, MPE=9). All other Census Divisions have MAPEs close to or lower than the national average.

6.4.7 Predictive Accuracy of the Model Across Building Types

There were no notable differences in the predictive accuracy of the model in the holdout sample across building types. Perhaps unsurprisingly, given the small sample size for this building type, "mixed" properties had a slightly higher MAPE (19 percent) than the national average. However, the difference is not analytically important. The range of MPEs is similarly comparable across building types, with all types having an MPE of 4 or 5 percent, except for detached / semi-detached, with an MPE of 2 percent.

Table A.21: Mean Absolute Percentage Error by Building Type

Building type	Freq.	Mean APE	Median APE
Detached/Semi-detached	238	17	14
High-rise/Elevator	762	18	14
Mixed	32	19	16
Row/Townhouse	401	17	13
Walk-up/Garden	1,395	17	13

6.4.8 Predictive Accuracy of the Model Across Age Categories

There were no notable differences in the predictive accuracy of the model in the holdout sample across age categories. MAPES across age categories ranged from 16 to 18 percent. MPEs for all age categories were 4 percent, except for the oldest age category, where there was a slightly higher level of overprediction in the holdout sample (MPE=6 percent).

Table A.22: Mean Absolute Percentage Error by Age Category

Agegroup	N	Mean APE (MAPE)	median APE
0-15	912	18	15
16-20	932	17	13
21-25	361	16	13
26+	623	17	13

6.4.9 Predictive Accuracy of the Model Across Property Size Categories

There were no notable differences in the predictive accuracy of the model in the holdout sample across property size categories. MAPEs across size categories were either 17 or 18 percent, while MPEs were either 4 or 5 percent.

Table A.23: Mean Absolute Percentage Error by Property Size

Sizegroup	Freq.	Mean APE (MAPE)	Median APE
size 1-80	1,411	18	14
size 81-150	835	17	13
size 150+	582	17	13

6.4.10 Predictive Accuracy of the Model by Property Age and Size Combined

Examining the intersection of property age and property size, again there are few notable differences in the predictive accuracy of the model. Throughout the model development process, concern has been expressed about the ability of the model to accurately capture costs in large, older developments. Within the 25 percent holdout sample, costs were just as well predicted among the largest and oldest properties as they were among other types of properties.

Table A.24: Mean Absolute Percentage Error by Property Age and Size

Agegroup		size 1-80	size 81-150	size 150+
age 0-15	N	496	194	222
	MAPE	20	15	17
	Med APE	16	12	15
age 16-20	N	521	288	123
	MAPE	17	17	16
	Med APE	14	14	12
age 21-25	N	142	130	89
	MAPE	15	17	18
	Med APE	13	15	12
age 26+	N	252	223	148
	MAPE	16	18	17
	Med APE	13	13	12

6.4.11 Predictive Accuracy of the Model by Type of Clientele

There were no notable differences in the predictive accuracy of the model in the holdout sample by type of clientele (family vs. elderly). The MAPE for family properties was 17 percent (MPE = 5 percent), and the MAPE for senior properties was 18 percent (MPE = 4 percent.)

Table A.25: Mean Absolute Percentage Error by Property Clientele

occtype	N	MAPE	Med. APE
Elderly	1,176	18	15
Family	1,652	17	13

6.4.12 Predictive Accuracy of the Model by Ownership Status

There were no notable differences in the predictive accuracy of the model in the holdout sample by ownership status. The MAPE ranged from 16 to 18 percent, and MPEs ranged from 4 to 5 percent.

Table A.26: Mean Absolute Percentage Error by Ownership Status

Ownership	N	MAPE	Med. APE
For Profit	1,201	17	13
Limited Dividend	579	16	13
Non Profit	1,048	18	15

6.5 Do Actual Costs in the Holdout Sample Fall Within the Forecast Interval?

In this section we examine the extent to which actual costs in the holdout sample fall within the forecast intervals generated by the model around each predicted value. The forecast interval is constructed such that, when the model is used for prediction, the actual outcome will be expected to fall within the forecast

interval in 95 out of 100 cases. The forecast interval for the predicted cost of each individual property is constructed so that it captures both the uncertainty (i.e., standard error) associated with the estimated model coefficients and the uncertainty associated with the residual of the regression. The mathematical formula for calculating these forecast intervals is standard and can be found in most of the graduate level statistics and econometrics texts, such as William H. Greene's *Econometric Analysis* (1993, 2nd edition: page 165).[25]

In the holdout sample overall, the observed costs fall within the forecast interval for 95 percent of all observations. The fact that the forecast intervals include observed costs in the holdout sample to such a large degree provides further evidence that the fit of the model reflects causal relationships between the independent and dependent variables.

It should be noted, however, that the 95 percent forecast intervals were quite large. The mean value of the forecast interval in the holdout sample was $101 on either side of the prediction. The mean predicted cost in the holdout sample was $239, so that the mean forecast interval was +/- 42% of the predicted value of the dependent variable. In order to test the percentage of observations that fell within narrower intervals, we constructed a range of forecast intervals between +/- 5 percent of predicted costs, up to +/- 25 percent of predicted costs. The results, shown below, show that over half of the holdout sample observations fell within +/- 15 percent of the model predicted costs.

Table A.27: Accuracy of Cost Model Estimates Using User-defined Forecast Intervals

Definition of Forecast Interval	Proportion of Holdout Sample with Observed Costs fall within the Forecast Intervals
Predicted Cost ± 5%*Predicted Cost	19.5%
Predicted Cost ± 10%*Predicted Cost	37.4%
Predicted Cost ± 15%*Predicted Cost	55.0%
Predicted Cost ± 20%*Predicted Cost	68.4%
Predicted Cost ± 25%*Predicted Cost	80.0%

Stratifying the sample by Census Division, the percentage of properties with actual costs falling within the forecast interval ranges from 91 percent (in New England) to 98 percent (in West North Central.) These percentages exactly parallel the distribution of MAPEs across the divisions, of course, and reflect the same underlying causes. The relatively low percentage in New England reflects the fact that the state with the lowest percentage of properties with actual costs falling within the forecast interval range is Maine, at 73 percent. (See table at end.)

Table A.28: Percent of observations for which observed costs fall within the forecast interval

Census Division	N	Mean
East North Central	484	0.97
East South Central	249	0.95
Mid Atlantic	322	0.92
Mountain	170	0.96
New England	143	0.91
Pacific	398	0.94
South Atlantic	537	0.95
West North Central	254	0.98
West South Central	246	0.97

[25] Calculations were performed using the PREDICT command and STDF option in the STATA software package.

As in the comparison of MAPEs, we see extremely little difference in the percentage of properties with actual costs falling within the forecast interval range, when properties are grouped by building type, age, size, clientele (family / senior), or ownership type. Because these results exactly parallel the MAPE results, we have not included the statistics here.

6.6 Decomposition of Variance

The cost model explains about 52 percent of the observed variation in the dependent variable in the development sample. In this section, we break down the explained variation (the 52 percent) by groups of independent variables. We have categorized our independent variables into three groups that are conceptually quite distinct from one another: property characteristics, environmental variables, and geographic variables.

Property characteristics include building age, number of units, the distribution of large and small units in the property, building type, and the clientele served by the development. These variables are the most straightforward to interpret and apply. Property characteristics account for 23 percent of the total explained variation in the model.

Environmental variables include variables that reflect either the policy environment or the market environment. This group includes the HUD mortgage subsidy variable, the percentage of units that receive project-based Section 8 subsidy, and the ratio of Rent to Fair Market Rent. The first two of these variables reflect the policy environment faced by the property. The last of the three reflects the position of the property in the rental housing market. Note that these variables are included to improve the accuracy of the model. These are not variables that can be applied directly to the public housing stock. In application, it has been decided that public housing developments will be treated as if they had 100 percent of units receiving project-based Section 8 subsidy; as if they did not receive a HUD mortgage subsidy; and as if they were providing housing of a quality level approximately at the median for their FMR area. (For further discussion of these decision rules, please see the main body of the chapter on Model Documentation.) Environmental characteristics accounted for 8 percent of the total explained variation.

Geographic variables include the geographic dummy variables described in the main body of this chapter, as well as the neighborhood poverty rate and the central city dummy variable. The geographic dummies explain by far the largest portion of the variance, accounting for 69 percent of the explained variation. It should be noted that as the model is currently constructed, the geographic dummies are interacted with an environmental characteristic – the ownership type indicator. Geographic dummies are interacted with an indicator for whether the property is unassisted/for-profit or has some form of assisted or non-profit or limited dividend ownership. In application, the geographic dummies based on the unassisted/for-profit sample are applied.

Table A.29

EXPLANATORY VARIABLE	PROPORTION OF COST VARIATION EXPLAINED
Property Characteristics	12.3%
Development size	0.2%
Age	1.3%
Unit size	5.7%
Building type	2.3%
Occupancy type	2.8%
Environmental Variables	4.0%
HUD assistance status	1.0%
Rent-to-FMR ratio	3.0%
Location	36.6%
Geographic Dummies	30.1%
Central City	2.8%
Neighborhood Poverty Rate	3.7%
All Variables	52.9%

Table A.30: Measures of Deviation between Predicted and Observed Costs in the 25 percent holdout sample, by State

state	N	Mean Absolute Percentage Error	Median Absolute Percentage Error	Mean Percentage Error	Median Percentage Error	Percentage with observed costs falling within forecast interval
AK	7	22	21	2	8	86%
AL	89	15	13	-1	-3	98%
AR	40	15	11	10	5	100%
AZ	40	16	12	-1	-5	95%
CA	277	19	16	3	2	94%
CO	51	15	12	4	2	98%
CT	48	22	18	8	7	94%
DC	25	14	14	0	-6	100%
DE	6	23	25	-23	-25	100%
FL	106	19	14	7	6	92%
GA	72	17	12	3	4	94%
HI	10	20	19	-15	-14	80%
IA	38	16	14	8	9	100%
ID	8	16	4	0	-2	88%
IL	71	19	15	8	7	93%
IN	88	14	9	5	0	99%
KS	26	14	14	3	-1	100%
KY	64	15	12	4	3	97%
LA	47	14	12	-2	-1	98%
MA	52	22	17	12	8	90%
MD	59	15	12	1	0	95%
ME	22	36	35	33	35	73%
MI	76	15	12	5	4	95%
MN	87	14	12	4	5	100%
MO	62	17	13	7	3	97%
MS	47	15	11	-4	-6	94%
MT	12	22	20	22	20	92%
NC	113	16	14	8	8	100%
ND	4	24	20	5	2	100%
NE	26	17	13	3	4	92%
NH	14	17	15	1	4	93%
NJ	45	27	24	9	5	82%
NM	15	16	12	8	6	100%
NV	15	21	21	-4	-14	100%
NY	163	18	13	4	1	95%
OH	182	14	12	3	2	98%
OK	33	15	11	6	2	94%
OR	33	19	16	-4	-11	91%
PA	92	19	14	10	7	92%
PR	25	20	16	-7	-12	96%

state	N	Mean Absolute Percentage Error	Median Absolute Percentage Error	Mean Percentage Error	Median Percentage Error	Percentage with observed costs falling within forecast interval
RI	22	13	10	-5	-5	100%
SC	50	17	14	8	6	96%
SD	10	11	12	6	7	100%
TN	67	18	13	1	1	91%
TX	126	16	13	4	0	97%
UT	17	15	14	3	1	94%
VA	80	18	13	7	3	90%
VT	7	16	17	-3	-7	100%
WA	71	19	15	10	11	97%
WI	50	18	12	10	7	94%
WV	26	17	12	7	1	92%
WY	12	19	18	11	14	100%

7. ASSESSING THE INFLUENCE OF OUTLIERS

We did not do a search for outliers using statistical methods that tested for observations with unusually high influence or leverage. Instead, we tested our model for robustness to outliers in two ways. First, we ran a Median Regression (also known as a Quantile Regression or a Least Absolute Deviation (LAD)), and compared the coefficients to those derived from our OLS regression. Median Regression differs from OLS in that it produces coefficient estimates that minimize the distance between predicted values and the median value of the dependent variable, rather than the mean value. The effect of the median regression is to give less weight to outliers, because medians are much less influenced by outliers than means are. Comparing the LAD results and the OLS results, we noted that no coefficients which were significant in the OLS model changed sign in the LAD model, and we took this as confirmation that our model was not highly sensitive to outliers. We also compared the OLS results to results from a procedure called IRLS (iteratively re-weighted least squares) which does, in fact, detect and drop observations with high leverage. Again results indicated that our model estimates were not heavily influenced by outliers, so we decided not to do a test for and elimination of outliers.

We did drop from the model observations where operating costs fell below $135 or above $650. We made this decision not as a result of data exploration, but rather on basis of the field-testing results, which suggested that observations with operating costs outside this range were implausible and likely represented data errors.

Supplemental Tables

Table A.31

Average Operating Costs, by Assistance Type and Mortgage Ownership
Analysis Variable: oc pumavgrfrfree

Grp	N Obs	N	Mean	Std Dev	Median	Minimum	Maximum
Unassisted	2555	2521	224	90	208	53	800
Older Assisted	3540	3521	251	95	234	56	830
Newer For Profit	2172	2170	228	76	213	53	733
New/202 LD & NP	3950	3923	243	107	220	53	794

Table A.32

Ever Troubled	N Obs	Label	Mean	Minimum	Maximum
0	11264	Percentage of Capital Deficiencies	0.029	0	0.37
		Physical Inspec. Score	85.77	9.0	100
		PUM Operating Costs	236.35	52.67	830.32
1	874	Percentage of Capital Deficiencies	0.059	0	0.54
		Physical Inspec. Score	72.60	4.0	100.0
		PUM Operating Costs	267.46	56.49	797.48

Table A.33

Operating Costs, Surplus Cash > 0 Analysis Variable : oc pumavgrfrfree							
Grp	N Obs	N	Mean	Std Dev	Median	Minimum	Maximum
Unassisted	1625	1610	219	83	206	53	752
Older Assisted	1753	1749	243	85	229	68	830
Newer For Profit	1610	1609	224	73	211	53	601
New/202 LD & NP	1345	1335	242	107	216	53	790

Operating Costs, Surplus Cash ≤ 0 Analysis Variable: oc pumavgrfrfree							
Grp	N Obs	N	Mean	Std Dev	Median	Minimum	Maximum
Unassisted	930	911	234	99	213	56	800
Older Assisted	1787	1772	258	104	239	56	797
Newer For Profit	562	561	240	83	222	57	733
New/202 LD & NP	2605	2588	243	107	222	55	794

Operating Costs, Net Profit > 0 Analysis Variable: oc pumavgrfrfree							
Grp	N Obs	N	Mean	Std Dev	Median	Minimum	Maximum
Unassisted	1251	1237	219	85	206	53	752
Older Assisted	2449	2441	245	89	229	58	830
Newer For Profit	1317	1316	222	75	210	53	733
New/202 LD & NP	1087	1075	249	108	224	53	794

Operating Costs, Net Profit <= 0 Analysis Variable: oc pumavgrfrfree							
grp	N Obs	N	Mean	Std Dev	Median	Minimum	Maximum
Unassisted	1304	1284	229	93	210	56	800
Older Assisted	1091	1080	265	106	245	56	797
Newer For Profit	855	854	237	77	221	57	670
New/202 LD & NP	2863	2848	240	106	218	58	790

Table A.34:
REAC Capital Defect Measures

Description	Defect Name
Roads	Settlement/Heaving
Parking Lots/Driveways/Roads	Settlement/Heaving
Fencing and Retaining Walls	Missing Sections
Fencing and Gates	Missing Sections
Retaining Walls	Damaged/Falling/Leaning
Storm Drainage	Damaged/Broken/Cracked
Storm Drainage	Damaged/Obstructed
Refuse Disposal	Inadequate Outside Storage Space
Refuse Disposal	Broken/Damaged Enclosure-Inadequate Outside Storage Space
Foundations	Cracks/Gaps
Foundations	Spalling/Exposed Rebar
Foundations	Leaking
Walls	Cracks/Gaps
Walls	Damaged Chimneys
Doors	Damaged Frames/Threshold/Lintels/Trim
Windows	Damaged Sills/Frames/Lintels/Trim
Roofs	Damaged/Torn Membrane/Missing Ballast
Roofs	Missing/Damaged Shingles
Roofs	Ponding
Roofs	Leaks
Fire Escapes	Visibly Missing Components
HVAC	Fuel Supply Leaks
HVAC	Boiler/Pump Leaks
Fire Protection	Missing Water Diffusers (Sprinkler Head)
Electrical System	Evidence of Leaks/Corrosion
Elevators	Not Operable
Domestic Water	Central Hot Water Supply Inoperable
Kitchen	Refrigerator-Missing/Damaged/Inoperable
Kitchen	Refrigerator-Missing/Damaged/Inoperable
Kitchen	Range/Stove – Missing/Damaged/Inoperable
Kitchen	Range/Stove – Missing/Damaged/Inoperable
Trash Collection Areas	Chutes - Damaged/Missing Components
Trash Collection Areas	Trash Room Door - Damaged/Inoperable
Bathroom	Water Closet/Toilet - Damaged/Clogged/Missing
Bathroom	Lavatory Sink - Damaged/Missing
Bathroom	Shower/Tub - Damaged/Missing
Kitchen	Countertops/Cabinets - Missing/Damaged
Kitchen	Cabinets - Missing/Damaged
Kitchen	Countertops – Missing/Damaged
Kitchen	Range/Stove – Missing/Damaged/Inoperable
Kitchen	Refrigerator-Missing/Damaged/Inoperable
Walls	Bulging/Buckling
Ceiling	Bulging/Buckling
HVAC System	Inoperable
Floors	Bulging/Buckling
Electrical System	Evidence of Leaks/Corrosion

Table A.35: Sample Size by Geographic Area

Geographic Area	For-profit	Non-profit	Limited Dividend	Total
Alaska statewide metro areas	3	0	5	8
Chicago, IL PMSA	107	73	30	210
Indianapolis, IN MSA	84	25	7	116
Detroit, MI PMSA	78	51	23	152
Columbus, OH MSA	98	35	28	161
Dayton-Springfield, OH MSA	42	23	12	77
Toledo, OH MSA	24	19	9	52
Milwaukee-Waukesha, WI PMSA	54	23	15	92
East North Central census division wide metro areas	286	178	100	564
Birmingham, AL MSA	34	19	6	59
Mobile, AL MSA	29	9	8	46
Lexington, KY MSA	27	13	10	50
Knoxville, TN MSA	25	14	10	49
Nashville, TN MSA	34	15	21	70
East South Central census division wide metro areas	157	94	84	335
Hawaii statewide metro areas	0	28	0	28
New York, NY PMSA	86	109	146	341
Balance of New York CMSA (excluding NY PMSA)	63	106	92	261
Pittsburgh, PA MSA	56	68	27	151
Philadelphia, PA-NJ PMSA	28	63	27	118
Mid Atlantic census division wide metro areas	99	175	111	385
Phoenix-Mesa, AZ MSA	46	35	10	91
Tucson, AZ MSA	27	11	5	43
Denver, CO PMSA	39	50	25	114
Colorado statewide metro areas	37	13	11	61
Salt Lake City-Ogden, UT MSA	29	22	15	66
Las Vegas, NV-AZ MSA	30	6	4	40
Mountain census division wide metro areas	62	35	18	115
Boston, MA-NH PMSA	30	60	55	145
New England (North) census division wide metro areas	0	64	0	64
New England (South) census division wide metro areas	52	92	112	256
Los Angeles-Long Beach, CA PMSA	133	120	75	328
Orange County, CA PMSA	24	20	5	49
Sacramento, CA PMSA	46	15	21	82
San Francisco-Oakland-San Jose, CA CMSA	75	139	27	241
California statewide metro areas	155	118	38	311
Oregon and Washington statewide metro areas	62	49	31	142
Seattle-Bellevue-Everett, WA PMSA	46	52	11	109
Portland-Vancouver, OR-WA PMSA	25	36	12	73
Puerto Rico statewide metro areas	0	121	0	121
Miami-Fort Lauderdale, FL CMSA	26	41	8	75

Geographic Area	For-profit	Non-profit	Limited Dividend	Total
Florida statewide metro areas	105	120	52	277
Atlanta, GA MSA	44	32	33	109
Georgia statewide metro areas	48	10	21	79
Baltimore, MD PMSA	93	40	46	179
Greensboro-Winston-Salem-High Point, NC MSA	43	31	10	84
Raleigh-Durham-Chapel Hill, NC MSA	36	32	7	75
North Carolina statewide metro areas	42	26	10	78
South Carolina statewide metro areas	89	28	20	137
Norfolk-Virginia Beach-Newport News, VA-NC MSA	45	19	22	86
Richmond-Petersburg, VA MSA	34	7	10	51
Charlotte-Gastonia-Rock Hill, NC-SC MSA	43	33	10	86
Washington, DC-MD-VA-WV PMSA	158	49	60	267
South Atlantic census division wide metro areas	57	38	37	132
Kansas City, MO-KS MSA	44	42	25	111
West North Central census division wide metro areas	130	106	71	307
Little Rock-North Little Rock, AR MSA	27	16	11	54
Dallas, TX PMSA	32	19	14	65
Houston, TX PMSA	43	27	4	74
West South Central census division wide metro areas	219	208	132	559
Cincinnati, OH-KY-IN PMSA	58	36	55	149
Louisville, KY-IN MSA	32	34	7	73
Minneapolis-St Paul, MN-WI MSA	125	54	47	226
St Louis, MO-IL MSA	64	34	23	121
Alaska statewide nonmetro areas	6	1	4	11
East North Central census division wide nonmetro areas	131	154	60	345
East South Central census division wide nonmetro areas	145	128	86	359
Hawaii statewide nonmetro areas	0	16	2	18
Mid Atlantic census division wide nonmetro areas	29	45	17	91
Mountain census division wide nonmetro areas	66	61	39	166
New England census division wide nonmetro areas	0	91	0	91
Pacific census division wide nonmetro areas	24	49	39	112
South Atlantic (north) census division wide nonmetro areas	84	96	33	213
South Atlantic (south) census division wide nonmetro areas	71	56	41	168
West Virginia statewide nonmetro areas	27	12	8	47
West North Central census division wide nonmetro areas	108	135	95	338
West South Central census division wide nonmetro areas	58	130	61	249
Cleveland-Lorain-Elyria, OH PMSA	45	28	32	105
Total	4,763	4,282	2,498	11,543